Evaluating disease-free survival as a surrogate endpoint for overall survival in muscle-invasive urothelial carcinoma: An analysis of Surveillance, Epidemiology, and End Results-Medicare data

CO205

Leung L,[1] Kanters S,[1] Pourrahmat MM,[1] Friedman HS,[2] Navaratnam P,[2] Reardon G,[3] Patel M,[4] Teitsson S,[5] Kurt M,[4] Lotan Y[6]

[1]Evidinno Outcomes Research Inc., Vancouver, BC, Canada; [2]DataMed Solutions LLC, New York, NY, US; [3]Keck Graduate Institute, Claremont, CA, US; [4]Bristol Myers Squibb, Princeton, NJ, US; [5]Bristol Myers Squibb, Uxbridge, UK; [6]University of Texas Southwestern Medical Center, Dallas, TX, US

Introduction

- Urothelial carcinoma (UC) develops with the growth of abnormal tissues in the urothelial cells lining the mucosal surfaces of the lower urinary tract (including the urethra and bladder) and upper urinary tract (including the renal pelvis and ureters)
- UC is the 10th most common cancer globally.[1] In the United States (US), the age-adjusted incidence rate of bladder cancer using 2015-2019 data from Surveillance, Epidemiology, and End Results (SEER) was 18.7 per 100,000 people, with a death rate of 4.2 per 100,000 people.[2] Five-year relative survival for urinary bladder cancer in the US was 77.1% from 2012-2018[2]
- Overall survival (OS) is the gold standard efficacy measure in oncology; however, obtaining mature OS data in localized cancers poses challenges such as requiring longer follow-up times, larger trial populations, and potential influence by subsequent therapies in the context of a rapidly changing treatment landscape[3,4]
- A common way to address these challenges is to use statistically appropriate and clinically relevant surrogate endpoints (SEs), such as disease-free survival (DFS), that can alleviate the uncertainty in long-term survival and the impact of subsequent therapies on the OS benefit gained by treatment[5]
- Establishing statistically valid SEs can increase the statistical power in measuring treatment effect and enable an earlier assessment of randomized controlled trials (RCTs) by regulatory agencies. Ideally, surrogacy assessments should be performed on individual patient-level data from randomized settings. However, limited availability of patient-level data from RCTs poses a challenge for the assessment of individual-level association between the SEs and OS. Access to such data is often costly and hindered by logistical constraints as well as privacy agreements protecting the data
- Analysis of real-world data can complement clinical trial evidence by increasing generalizability to larger and more inclusive patient populations, health care providers, and health care systems, as well as to settings that reflect patients' and physicians' treatment preferences in day-to-day practice more realistically.[6] Hence, we evaluated the appropriateness of real-world DFS as an SE for OS using data from the SEER-Medicare linked database

Objectives

- To evaluate the appropriateness of DFS as an SE for OS in adults with high-risk muscle invasive UC (MIUC) who have undergone radical surgical resection
- To investigate the impact of prior neoadjuvant treatment on the strength of association via subgroup analysis

Methods

Patient selection

- **Inclusion criteria**
  - Diagnosed from 2009-2017 with pathologically confirmed MIUC originating in the bladder, ureter, or renal pelvis
  - Received radical surgical resection and lymphadenectomy from 2009-2017, with high risk of recurrence
  - ≥66 years old on the index date
  - American Joint Committee on Cancer (AJCC) staging M0 and either N+ or T2-T4a (neoadjuvant therapy, bladder), T2-T4 (neoadjuvant therapy, ureter or renal pelvis), T3-T4a (no neoadjuvant therapy, bladder), T3-T4 (no neoadjuvant therapy, ureter or renal pelvis)
  - Had continuous enrollment in both Medicare plans A and B with no health maintenance organization enrollment one year prior to the index date for baseline assessment of previous treatment or cancers (SEER-Medicare)
- **Exclusion criteria**
  - Partial cystectomy or nephroureterectomy within 90 days prior to or after the index date
  - Urethrectomy or exclusively urethral disease
  - Concurrent radiation therapy for UC or prostatic carcinoma within 120 days post-index
  - Had active, known, or suspected autoimmune disease
  - Had pN× (regional lymph nodes not available) status at diagnosis
  - Diagnosed with hepatitis within 365 days prior to the index date
  - Cancer identified solely at autopsy or by death certificates, or of unknown origin

Statistical methods

- OS was defined as time from surgery to all-cause mortality, and DFS was defined as time from surgery to disease recurrence or all-cause mortality with a 7-month post-index surveillance period. For sensitivity analyses, a secondary definition of DFS using a 4-month post-index surveillance period was also evaluated
- Validation of SEs in oncological settings require strong association with OS at the patient-level and at the trial-level
- Patient-level (i.e., endpoint) association seeks to address the prognostic or predictive role of the surrogate for OS. Trial-level (i.e., treatment-effect) association seeks to address if the treatment effect on OS can be predicted from the treatment effect on the SE for a prospective dataset or population under proportionality assumptions
- Surrogacy was evaluated using the two-level meta-analytic approach[7] based on the refined set of patient-level data from SEER to evaluate the potential role of DFS as a surrogate for OS:
  - **Individual-level correlation:** Assessed by Spearman's rank correlation in a non-parametric fashion. Copula models were also employed to estimate Kendall's τ as a supportive measure
  - **Treatment-effect correlation:** Assessed by Pearson's correlation through sample size-weighted linear regression (WLR) relating log-transformed treatment effects on DFS and OS (i.e., HR_{DFS} and HR_{OS})

Methods (continued)

- For the treatment-effect correlation, clusters with synthetic experimental and controls arms were created by pairing each patient to another patient on a different treatment using propensity score matching based on age, sex, disease stage, and race/ethnicity. Median sample size was 46 (range: 8-50) in the overall population and 32 (range: 6-38) in the population receiving prior neoadjuvant treatment
- The performance of the WLR model was tested via leave-one-out-cross-validation (LOOCV) as described by the National Institute for Health and Care Excellence (NICE).[8] For each cluster, a WLR model was re-fitted to the data omitting that cluster. HR_{OS} of the omitted cluster along with a 95% prediction interval (PI) were obtained from its HR_{DFS} using the surrogacy equation generated from the WLR, and the actual HR_{OS} in the omitted cluster was then compared to the 95% PI of the predicted HR_{OS}
- Strengths of individual-level and treatment-effect correlations were evaluated according to modified Institute for Quality and Efficiency in Health Care (IQWiG) criteria: weak when the upper limit of the 95% confidence interval (CI) of the correlation measure was ≤0.70, strong when the lower limit of the 95% CI of the correlation measure was ≥0.85, and moderate otherwise
- To evaluate the impact of clustering on the results, the WLR model was fitted ten more times using alternative assignments of patients to clusters
- The utility of the WLR model was evaluated with the surrogate threshold effect (STE), which is the minimum DFS benefit that would translate into statistically significant OS benefit at a default 95% confidence level. Because the calculation of STE requires the sample size of a prospective cluster or population, STE estimations in this study were based on the average sample size of all clusters in each population of interest[9]
- The external validity of the WLR model developed for the overall population was evaluated by predicting HR_{OS} from HR_{DFS} for 9 published RCTs which formed the evidence base for the surrogacy assessment between DFS and OS in Sternberg et al. (2022)[9]

Results

Patient characteristics

- After the application of inclusion and exclusion criteria, there were 1038 patients remaining for surrogacy assessments. Most of these patients were elderly males of non-Hispanic and White ethnicity, diagnosed with stage 3 cancer (**Table 1**). Disease site was mainly in the bladder (90.4%), followed by renal pelvis (6.6%) and ureter (3.0%)

Treatment characteristics

- Most patients did not receive adjuvant therapy (84.2%), but those who did typically received 5+ cycles (64.0%) of a multi-agent chemotherapy regimen (90.2%) (**Table 2**)
- Of the 1038 patients, 433 (41.7%) received neoadjuvant therapy

Outcomes

- Median follow-up times were 29.5 and 35.2 months in the overall population and in the subgroup with prior neoadjuvant treatment, respectively
- In the overall population, median OS and DFS were 46.3 months (95% CI: 37.8, 56.0) and 20.7 months (95% CI: 16.0, 27.8), respectively
- In the subgroup with prior neoadjuvant treatment, median OS was 100.4 months (95% CI: 76.3, not reached) whereas median DFS was not reached (95% CI: 87.8, not reached)
- **Individual-level correlation analyses**
  - In the overall population, Spearman's ρ was 0.85 (95% CI: 0.79, 0.90), and supplementary Kendall's τ estimates ranged from 0.79 to 0.81 across Copula models
  - In the subgroup with prior neoadjuvant treatment, Spearman's ρ was 0.90 (95% CI: 0.83, 0.96), and supplementary Kendall's τ estimates ranged from 0.82 to 0.89 across Copula models
- **Treatment-effect correlation analyses**
  - In the overall population (**Figure 1**), Pearson's correlation coefficient was 0.91 (95% CI: 0.75, 0.97), and the corresponding surrogacy equation was $\log(HR_{OS}) = -0.04 + 1.08 \times \log(HR_{DFS})$, with an STE of 0.73 ($HR_{DFS} < 0.73$ predicts $HR_{OS} < 1$ with 95% probability)
  - In the subgroup with prior neoadjuvant treatment (**Figure 2**), Pearson's correlation was 0.97 (95% CI: 0.83, 1.00), and the corresponding surrogacy equation was $\log(HR_{OS}) = -0.03 + 0.98 \times \log(HR_{DFS})$, with an STE of 0.74 ($HR_{DFS} < 0.74$ predicts $HR_{OS} < 1$ with 95% probability)
- **Internal validation**
  - In LOOCV, the 95% PIs on HR_{OS} covered the actual HR_{OS} for 89% of clusters in both the overall population (16 of 18 clusters; **Figure 3**) and in the subgroup with prior neoadjuvant treatment (8 of 9 clusters; **Figure 4**)
  - Across alternative assignments of patients to clusters, Pearson's correlation estimates ranged from 0.85 to 0.97 in the overall population and from 0.88 to 1.00 in the subgroup with prior neoadjuvant treatment, indicating consistency with the primary analysis
- **External validation**
  - When the surrogacy model for the overall population was applied to 9 RCTs used for surrogacy assessments between DFS and OS, the 95% PI on HR_{OS} covered the observed HR_{OS} for 6 of 9 RCTs (66.7%)
- **Sensitivity analysis (alternate definition of DFS)**
  - In the overall population, the individual-level correlation was 0.69 (95% CI: 0.68, 0.84), and the treatment-level correlation was 0.68 (95% CI: 0.25, 0.89) with an STE of 0.54 and an LOOCV accuracy of 89%
  - In the subgroup with prior neoadjuvant therapy, the individual-level correlation was 0.85 (95% CI: 0.75, 0.92), and the treatment-effect correlation was 0.96 (95% CI: 0.86, 1.00) with an STE of 0.73 and an LOOCV accuracy of 100%

Table 1. Patient characteristics in the data refined by inclusion and exclusion criteria

Characteristic	Categories/units	Overall	Subgroup with prior neoadjuvant treatment
Sample size	Patients	1038	433
Age, median (Q_1, Q_3)	Years	74.1 (70, 79)	72.2 (68.9, 76.3)
Sex, n (%)	Male	749 (72.2)	331 (76.4)
Surgery index year, median (Q_1, Q_3)	Years	2014 (2011, 2016)	2015 (2012, 2016)
Clinical stage, n (%)	2	304 (29.3)	241 (55.7)
	3	453 (43.6)	110 (25.4)
	4	281 (27.1)	82 (18.9)
Race/ethnicity, n (%)	Non-Hispanic White	887 (85.5)	379 (87.5)
	Non-Hispanic Black	39 (3.8)	13 (3.0)
	Hispanic	67 (6.5)	30 (6.9)
	Asian or Pacific Islander	1 (0.1)	1 (0.2)
	American Indian/ Alaska Native	42 (4.0)	8 (1.8)
	Unknown	2 (0.2)	0 (0)

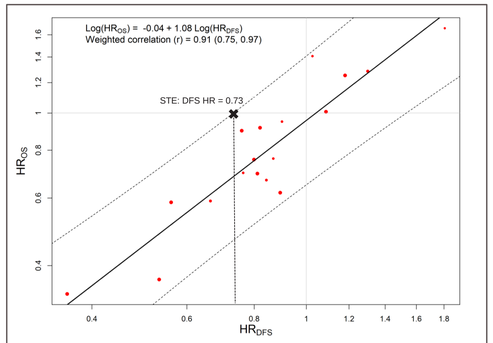
Q_1, first quartile in interquartile range; Q_3, third quartile in interquartile range.

Table 2. Treatment characteristics in the data refined by inclusion and exclusion criteria

Characteristic	Categories/units	Overall	Subgroup with prior neoadjuvant treatment
Sample size	Patients	1038	433
Adjuvant treatment, n (%)	Single-agent chemotherapy	16 (1.5)	7 (1.6)
	Multi-agent chemotherapy	148 (14.3)	20 (4.6)
	None	874 (84.2)	406 (93.8)
Number of cycles of adjuvant treatment, n (%)	1	15 (9.1)	5 (18.5)
	2	7 (4.3)	3 (11.1)
	3	17 (10.4)	5 (18.5)
	4	20 (12.2)	5 (18.5)
	5+	105 (64.0)	9 (33.3)
Duration of adjuvant treatment, median (Q_1, Q_3)	Months	76 (49, 112)	53 (25.5, 100)
Neoadjuvant treatment, n (%)	Single-agent chemotherapy	12 (1.2)	12 (2.8)
	Multi-agent chemotherapy	421 (40.6)	421 (97.2)
	None	605 (58.3)	0 (0)
Duration of neoadjuvant treatment (median; Q1, Q3)	Months	65 (42, 77)	65 (42, 77)

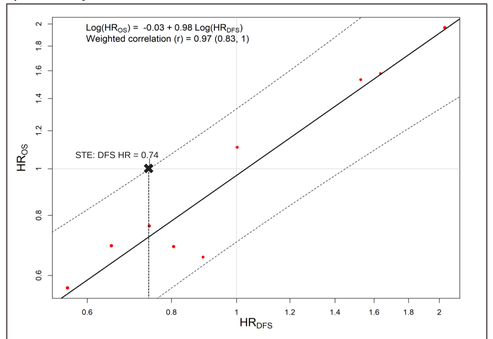
Q_1, first quartile in interquartile range; Q_3, third quartile in interquartile range.

Figure 1. WLR model for treatment-effect correlation in the overall population



The weighted linear regression is graphed as a solid straight line with its corresponding 95% prediction interval boundaries as dotted curves. The red dots are plotted using measures from the reported $\log(HR_{OS})$ on the x-axis against measures of $\log(HR_{DFS})$ on the y-axis for each treatment comparison per treatment cluster. Sizes of the dots are proportional to the corresponding weights of clusters used in the surrogacy equation.
DFS, disease-free survival; HR, hazard ratio; OS, overall survival; STE, surrogate threshold effect.

Figure 2: WLR model for treatment-effect correlation in the subgroup with prior neoadjuvant treatment

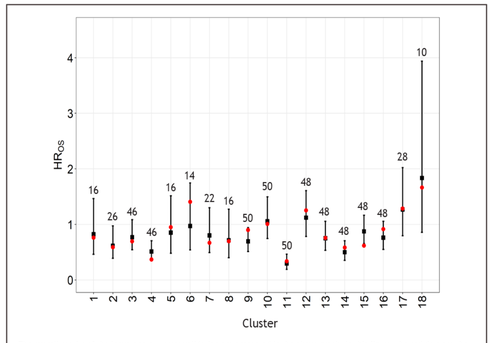


The weighted linear regression is graphed as a solid straight line with its corresponding 95% prediction interval boundaries as dotted curves. The red dots are plotted using measures from the reported $\log(HR_{OS})$ on the x-axis against measures of $\log(HR_{DFS})$ on the y-axis for each treatment comparison per treatment cluster. Sizes of the dots are proportional to the corresponding weights of clusters used in the surrogacy equation.
DFS, disease-free survival; HR, hazard ratio; OS, overall survival; STE, surrogate threshold effect.

Conclusions

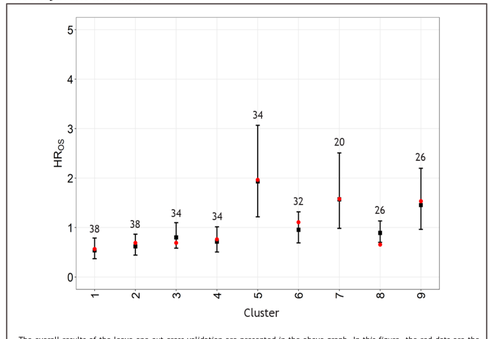
- In both the overall population and the subgroup with prior neoadjuvant treatment, individual-level and treatment-effect correlations were moderate
- Surrogacy equations in both populations had similar intercepts and slopes, with high utility based on STEs (>0.7), but moderate predictive accuracy (<90%) based on LOOCV and external validation
- The strength of individual-level and treatment-effect correlations in the current study are similar to findings previously reported from other investigators despite the differences in data sources[10,11,12,13]
- Estimated surrogacy equations may assist earlier assessments of OS benefit from DFS benefit in the adjuvant treatment of MIUC in the real-world setting
- Compared to data from relatively older randomized settings, utilization of up-to-date RWD in surrogacy assessment can provide more accurate and quicker insights on the impact of subsequent treatments on the strength of the surrogacy relationship

Figure 3. LOOCV results for the WLR model in the overall population



The overall results of the leave-one-out cross validation are presented in the above graph. In this figure, the red dots are the reported HR_{OS} measures, and the black squares are the predicted HR_{OS} measures. Corresponding to the predicted HR_{OS} measures, the 95% prediction intervals are presented as black solid lines per each cluster. Numbers above the vertical whiskers are the corresponding cluster sizes. For complete model validity, reported HR_{OS} measure should fall within the plotted interval for each cluster.
HR, hazard ratio; OS, overall survival.

Figure 4: LOOCV results for the WLR model in the subgroup with prior neoadjuvant treatment



The overall results of the leave-one-out cross validation are presented in the above graph. In this figure, the red dots are the reported HR_{OS} measures, and the black squares are the predicted HR_{OS} measures. Corresponding to the predicted HR_{OS} measures, the 95% prediction intervals are presented as black solid lines per each cluster. Numbers above the vertical whiskers are the corresponding cluster sizes. For complete model validity, reported HR_{OS} measure should fall within the plotted interval for each cluster.
HR, hazard ratio; OS, overall survival.

References

1. Saginala K, et al. Medical Sciences. 2020;8(1):15.
2. Siegel RL, et al. CA Cancer J Clin. 2023; 73(1):17-48.
3. Lux MP, et al. Cancer management and research. 2021;13:8457-8471.
4. Pazdur R. Oncologist. 2008;13 Suppl 2:19-21.
5. FDA. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics Guidance for Industry. December 2018. Available from: https://www.fda.gov/media/71195/download
6. Sherman RE, et al. New England Journal of Medicine. 2016;375(23):2293-2297.
7. Burzykowski T, et al. New York, Springer, 2005.
8. Bujkiewicz S, et al. NICE DSU technical support document. 2019. Available from: https:www.sheffield.ac.uk/nice-dsu/tsds/full-list
9. Kelly RJ, et al. New England Journal of Medicine. 2021;384(13):1191-1203.
10. Sternberg C, et al. Annals of Oncology. 2022; 33(7, Suppl.): S1335-S1336.
11. Sonpavde G, et al. J Urol. 2011;185(2):456-61.
12. Kim HS, et al. Frontiers in Oncology. 2015;5:249.
13. Nuhn P, et al. European Journal of Surgical Oncology, 2012;38(7): 637-642

Acknowledgments

- All authors contributed to and approved the presentation; writing assistance was provided by Evidinno Outcomes Research Inc., funded by Bristol Myers Squibb

Email: lleung@evidinno.com