ARTIFICIAL INTELLIGENCE

original reports abstract

Risk Prediction Using Bayesian Networks: An Immunotherapy Case Study in Patients With Metastatic Renal Cell Carcinoma

Alind Gupta, PhD[1]; Paul Arora, PhD[1,2]; Darren Brenner, PhD[3]; Jacqueline Vanderpuye-Orgle, PhD[4]; Devon J. Boyne, PhD[3]; Mark Edmondson-Jones, MSc[5]; Elena Parkhomenko, PhD[5]; Warren Stevens, PhD[5]; Shaan Dudani, MBChB[3]; Daniel Y. C. Heng, MD, MPH[3]; Samuel Wagner, MPharm, HBA, PhD[6]; John Borrill, MSc[7]; and Elise Wu, PhD[6]

PURPOSE To address the need for more accurate risk stratification models for cancer immuno-oncology, this study aimed to develop a machine-learned Bayesian network model (BNM) for predicting outcomes in patients with metastatic renal cell carcinoma (mRCC) being treated with immunotherapy.

METHODS Patient-level data from the randomized, phase III CheckMate 025 clinical trial comparing nivolumab with everolimus for second-line treatment in patients with mRCC were used to develop the BNM. Outcomes of interest were overall survival (OS), all-cause adverse events, and treatment-related adverse events (TRAE) over 36 months after treatment initiation. External validation of the model's predictions for OS was conducted using data from select centers from the International Metastatic Renal Cell Carcinoma Database Consortium (IMDC).

RESULTS Areas under the receiver operating characteristic curve (AUCs) for BNM-based classification of OS using baseline data were 0.74, 0.71, and 0.68 over months 12, 24, and 36, respectively. AUC for OS at 12 months increased to 0.86 when treatment response and progression status in year 1 were included as predictors; progression and response at 12 months were highly prognostic of all outcomes over the 36-month period. AUCs for adverse events and treatment-related adverse events were approximately 0.6 at 12 months but increased to approximately 0.7 by 36 months. Sensitivity analysis comparing the BNM with machine learning classifiers showed comparable performance. Test AUC on IMDC data for 12-month OS was 0.71 despite several variable imbalances. Notably, the BNM outperformed the IMDC risk score alone.

CONCLUSION The validated BNM performed well at prediction using baseline data, particularly with the inclusion of response and progression at 12 months. Additionally, the results suggest that 12 months of follow-up data alone may be sufficient to inform long-term survival projections in patients with mRCC.

JCO Clin Cancer Inform 5:326-337. © 2021 by American Society of Clinical Oncology

INTRODUCTION

Immunotherapy with immune checkpoint inhibitors (ICIs) has shifted the paradigm of cancer therapy. By activating the body's immune response against cancer cells, ICIs have been shown to be particularly effective in advanced malignancies at producing durable responses in patients.[1,2] ICIs are also associated with fewer side effects because of their selectivity for tumor cells rather than indiscriminate cytotoxicity.[3-5] Despite their therapeutic promise across multiple cancer indications, studies show significant variability in individual-level response to immunotherapy drugs, driving a need to accurately identify treatment responders.[1,6] By identifying prognostic variables, markers of survival and adverse events (AE), and/or risk stratification criteria, therapies could be individually tailored to patients to maximize therapeutic benefit and improve cost-effectiveness of innovative therapies.

Machine learning methods are increasingly important for prognostic modeling and knowledge discovery in clinical research, offering an opportunity to deliver on the promise of precision medicine.[7] Clinical trials and real-world data sets routinely capture many variables on patient demographics, treatment regimens, and tumor molecular markers among others that could be used to assess optimal targets for treatments. However, traditional methods of analysis typically use only a small set of predefined variables to inform risk and survival predictions. On the other hand, machine learning methods leverage the full suite of available patient data for individual-level risk estimation.[8] Potential applications of this type of analysis include informing adaptive clinical trial design, identifying predictors of surrogate end points, and supporting extrapolation of trial outcomes and clinical decision making. Yet, very few studies have used this more

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on January 7, 2021 and published at ascopubs.org/journal/cci on March 25, 2021: DOI https://doi.org/10.1200/CCI.20.00107

CONTEXT

Key Objective

To develop an interpretable model for predicting survival and adverse events in patients with metastatic renal cell carcinoma (mRCC) treated with immunotherapy.

Knowledge Generated

Bayesian network–based multivariate individual-level prediction model evaluated on face validity, performance against machine learning classifiers, and real-world generalizability. We identify prognostic variables for survival and safety outcomes and show that tumor response within the first year of initiation of nivolumab for second-line therapy is highly prognostic of long-term outcomes in mRCC.

Relevance

Because of individual-level heterogeneity in response to treatment, many novel cancer therapies are efficacious in only a subset of patients receiving treatment. Our findings indicate that individual-level prediction of survival and adverse events in patients with mRCC treated with immunotherapy may be tractable at the time of treatment initiation or shortly thereafter.

promising approach as a result of concerns about interpretability, limited clinician input, missing observations, and challenges with concurrently modeling several correlated outcomes (ie, multivariate predictions).[10-12]

This proof-of-concept study aimed to address these limitations of traditional modeling approaches by using machine learning to develop a multivariate Bayesian network model (BNM) for predicting survival and safety outcomes in patients with metastatic renal cell carcinoma (mRCC) being treated with ICIs. Currently, two validated prognostic models for targeted therapy in RCC are in widespread use—the Memorial Sloan Kettering Cancer Center (MSKCC) risk score (originally published in 1999[9]) and the International Metastatic Renal Cell Carcinoma Database Consortium (IMDC)/Heng risk score.[10] These models use Karnofsky performance status, time from diagnosis to systemic treatment, and three or four serum markers to stratify patients into favorable, intermediate, or poor prognostic risk groups. Although both models are continually updated to improve their prediction accuracies,[11] they were developed prior to the advent of ICIs for cancer therapy. Thus, they may exclude variables that predict safety, response, and/or survival while on immunotherapy.[16]

This study addresses these gaps by building a transparent risk prediction BNM that also offers insight into prognostically significant variables in cancer immunotherapy. The objectives of this study are as follows:

1. To develop an interpretable BNM for jointly predicting overall survival (OS), AE, and treatment-related adverse events (TRAE) within 3 years after initiating immunotherapy,
2. To assess the predictive performance of the BNM,
3. To validate the BNM to assess real-world performance and generalizability, and
4. To identify variables that predict OS, AE, and TRAE within 3 years after initiating immunotherapy (ie, prognostically significant variables).

METHODS

This article is concordant with Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis guidelines.[12] Specific study components are detailed below. Additional details can be found in the Appendix.

CheckMate 025 Data Set

CheckMate 025 (CM-025) is an open-label, randomized, phase III trial that compares the safety and efficacy of nivolumab (a PD1 ICI) versus everolimus in patients with mRCC who had received prior antiangiogenic therapy. Details about CM-025 have been previously described.[13] Participants were at least 18 years old with advanced or metastatic clear cell renal cell carcinoma and had received prior antiangiogenic therapy. Key exclusion criteria were Karnofsky performance status of < 70, past or current CNS metastases, and previous treatment with mTOR (mammalian target of rapamycin) inhibitors or glucocorticoids.[3]

For purposes of this study, baseline data (eg, sociodemographic characteristics, tumor characteristics, and prior treatment types) and 3 years of follow-up data for patients who received treatment (eg, treatment response, disease progression, and subsequent therapies) were used to train the BNM (n = 803 with 75 variables). Outcomes of interest were OS at months 12, 24, and 36 (alive, dead, or missing or censored if lost to follow-up); AE in year 1, 2, and 3; and TRAE in year 1, 2, and 3. AE and TRAE were encoded as number of events of CTCAE grade 3 or above binned as one of 0, 1-2, 3-5, 6+. In all cases, we defined AE and TRAE as the number of AE and TRAE observed from the start of a year until death, loss to follow-up, or the end of the year, whichever happened first. A 3-year database lock that included a maximum follow-up of 5 years was used for the analysis.

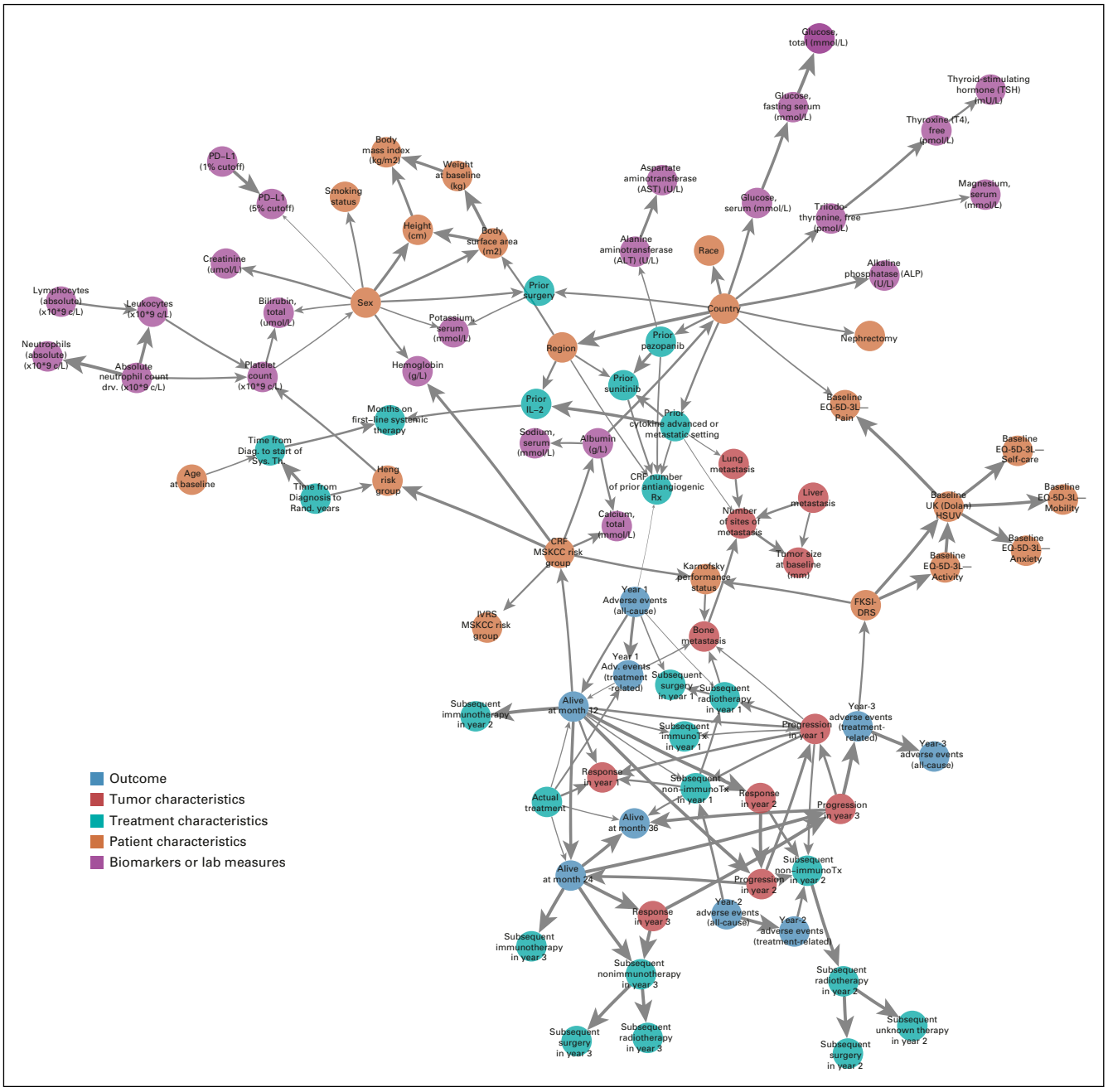

FIG 1. Bayesian network model structure learned from CheckMate 025. Edge linewidth represents log-transformed mutual information between variables (thicker linewidth implies stronger statistical dependence). Disconnected nodes are not shown. Adv. events, Adverse events; CRF, case report form; Diag, Diagnosis; Rand. years, randomization (years); Rx, prescription; Sys. Th, Systemic therapy; Tx, therapy.

IMDC Data Set

The BNM was validated with the latest cut of the IMDC data set containing observational data for patients with metastatic RCC who were receiving nivolumab or everolimus (n = 2,152).[10] Only IMDC centers with nivolumab access and sufficiently reliable data were included. Notably, key exclusion criteria from CM-025 did not apply to patients in IMDC data, and several variables showed moderate-to-high imbalance (see Results on external validation), which makes this a sufficiently distinct cohort for assessing broader model generalizability. However, only 26 (35%) of the baseline variables used for the development of the BNM were available in the IMDC data set, and data on AE were not

available in IMDC data. Thus, only OS predictions (at 12 months) were used to validate the BNM to assess real-world generalizability.

RESULTS

Model Structure

The structure of the estimated BNM, showing nodes (ie, variables) and directed edges (ie, associations), is presented in Figure 1. To evaluate model interpretability, the face validity of the BNM structure was assessed by comparing the model against expected relationships. Specifically, the level of mutual information within clusters of biologically related variables in the model was examined and was determined to be high. These included clusters for glucose measures (serum glucose, fasting serum glucose, and total glucose), thyroid hormones (T3, T4, and thyroid-stimulating hormone), liver enzymes (AST and ALT), immune cells (lymphocytes, leukocytes, neutrophils, and platelets), and PD-1 and its ligand PD-L1. Similarly, summary metrics such as body mass index, health state utility values, and risk scores (ie, MSKCC and IMDC) were clustered with baseline variables used for their calculation (eg, height, weight, health-related quality-of-life scores, Karnofsky performance status, and hemoglobin, respectively) with a high level of correlation or mutual information. Thus, the model structure learned from CM-025 was consistent with expected relationships from literature. Markov blankets[14] of year-1 outcomes are shown in Figure A1. Variables in the Markov blanket, when nonmissing, are the only variables used for prediction as a result of conditional independence assumptions in BNs.

In addition to the graphical structure, we also examined the conditional probability tables underlying the graph. Conditional survival at 24 months, ie, the probability of survival conditional on having survived to month 12, was 65%, which is considerably higher than the marginal probability of survival to 24 months at baseline (46%) (Table 1). Similarly, probability of surviving to 36 months was 33% at baseline, increasing to 46% and 71% conditional on surviving to months 12 and 24, respectively (Table 1). The probability of AE decreased over time, going from 54% to 41% then to 15% conditional on surviving to months 12, 24, and 36, respectively. This indicates a substantial incremental improvement in survival and safety events over time for patients who were able to survive longer.

TABLE 1. Conditional Probability of Survival at Months 12, 24, and 36

		Conditional on	
Probability of	**Baseline**	**Surviving to Month 12**	**Surviving to Month 24**
Survival at month 12	0.70	—	—
Survival at month 24	0.46	0.65 (+ 0.19)	—
Survival at month 36	0.33	0.46 (+0.13)	0.71 (+0.38)

NOTE. Marginal probabilities indicate the probability of survival at baseline. Numbers in parentheses indicate the improvement in conditional survival probabilities compared with the baseline probability.

Classification Performance

Observed mean areas under the receiver operating characteristic curve (AUCs) for the BNM were generally either acceptable ($0.7 < AUC < 0.8$) or fair or poor ($0.5 < AUC < 0.7$). Specifically, the AUCs were 0.74, 0.71, and 0.68 for years 1, 2, and 3, respectively, for OS with event rates (% death) of 30%, 51%, and 62%; 0.60, 0.68, and 0.71 in years 1, 2, and 3, respectively, for AE; and 0.71, 0.64, and 0.72 in years 1, 2, and 3, respectively, for TRAE. However, the inclusion of response and progression as predictors of 12-month OS increased mean AUC to 0.85—ie, in the excellent category ($0.8 < AUC < 0.9$). Note that response and progression during years 1, 2, and 3 are a part of the model; the BNM allows flexibility in choosing whether we include these as predictors, outcomes, or neither depending on the analysis of interest.

Notably, similar mean AUCs were observed when alternative machine learning approaches were used to assess the relative performance of the estimated BNM (Table A1). The study found no significant difference *on average* in model performance across the nine outcomes when comparing the BNM against three machine learning classifiers: Lasso regularized logistic regression (Lasso), support vector machine, and random forest (RF). Comparing the BNM with gold-standard RF classifiers, RF did significantly better at classifying 12-month OS (P value .0029), whereas the BNM performed significantly better at classifying year-1 TRAE (P value $<$ .001).

External Validation

Notable differences were observed between patients in the CM-025 and IMDC data sets. For instance, patients in the IMDC data set had worse Karnofsky performance status, poorer MSKCC risk scores, higher prevalence of bone metastases, lower weight or BMI, and lower survival rates and were more likely to be receiving everolimus. There were also notable imbalances (ie, standardized differences [d] > 0.25) in similar variables across both data sets (Fig A2). Overall, the IMDC data were found to be appropriate for assessing broader generalizability of the BNM.

Results from evaluating real-world performance of the BNM are presented in Figure 2 and Table 2. The estimated AUC for OS at 12 months was 0.71 (compared with 0.74 in CM-025). The study also compared the BNM against the IMDC/Heng risk score commonly used for risk stratification in patients with RCC undergoing systemic therapy. Mean AUCs for OS at 12 months using the BNM and IMDC/Heng risk score (available for 58% of patients in the IMDC data set) were 0.76 and 0.69, respectively. Conversely, in the subset of patients with missing IMDC risk score because of

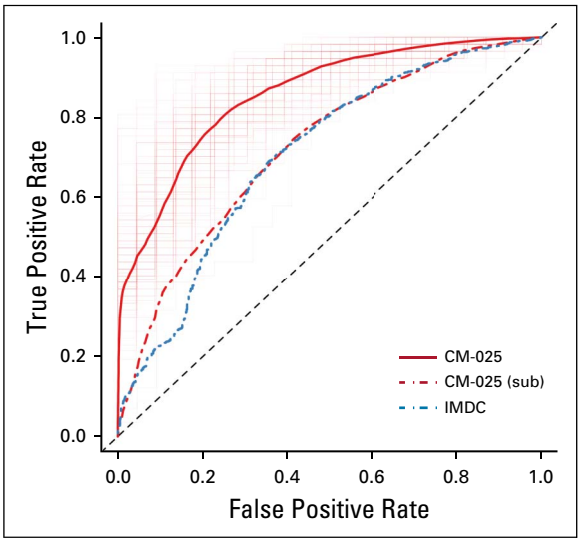

FIG 2. ROC curves for Bayesian network model for 12-month overall survival. CM-025, all available variables in the model; CM-025 (sub), common variables between CM-025 and IMDC from CM-025; IMDC, common variables between CM-025 and IMDC from IMDC. Red curves represent internal cross-validation, and blue curves represent external validation. Faint red lines represent ROC curves from individual runs of 10-fold cross-validation; the bold red line represents their mean. CM-025, CheckMate 025; IMDC, International Metastatic Renal Cell Carcinoma Database Consortium; ROC, receiver operating characteristic.

one or more missing risk criteria, externally validated AUC for OS was 0.64 for BNM. Therefore, the BNM outperforms the IMDC risk score for year-1 OS and can handle missing data for risk prediction. It is worth noting that probabilities were recalibrated to account for differences in 12-month mortality between CM-025 and IMDC data, and probability output showed < 10% deviance from ideal calibration in all cases (Fig 3).

Prognostic Variables

Prognostic variables were identified using their AUCs and are presented in Table 3. For OS, MSKCC risk score, IMDC/Heng risk score, and Karnofsky performance score were prognostic from months 12, 24, and 36 (univariable AUCs between 0.65 and 0.69; see also Tables A2 and A3). Functional Assessment of Cancer Therapy–Kidney Symptom Index–Disease-Related Symptoms[15] and health-related quality-of-life scores (EuroQol EQ-5D) at baseline were more highly prognostic of OS at months 24 and 36 (univariable AUC > 0.6) than month 12 (AUC < 0.6). Serum hemoglobin was highly prognostic, as expected. Interestingly, only 12-month follow-up related to progression, response, and subsequent radiotherapy and nonimmunotherapy was highly prognostic of OS over 3 years; this is possibly due to the majority of these occurring within the first 12 months after treatment initiation. For AE in year 1, sodium; hemoglobin; calcium; platelet count; and MSKCC, Karnofsky, and IMDC/Heng scores were prognostic. For AE in years 2 and 3, subsequent interventions (nonimmunotherapy, radiotherapy, and surgery) were prognostic. Similar to OS, progression and response in year 1 were prognostic of AE in years 1-3. Prognostic variables for TRAE included actual treatment type (everolimus or nivolumab), tumor size, bone and liver metastasis, and subsequent radiotherapy and immunotherapy.

TABLE 2. Areas Under the Receiver Operating Characteristic Curve for External Validation of BNM Trained Using CM-025 on IMDC Data

		Data		
Variable	**Subgroup**	**CM-025 (n = 803)**	**CM-025 (Sub) (n = 803)**	**IMDC (n = 2,153)**
Heng/IMDC risk group	Favorable	0.83 ± 0.17	0.84 ± 0.14	0.74 ± 0.06
	Intermediate	0.80 ± 0.07	0.64 ± 0.08	0.65 ± 0.02
	Poor	0.84 ± 0.10	0.66 ± 0.11	0.69 ± 0.04
MSKCC risk group	Favorable	0.80 ± 0.11	0.70 ± 0.12	0.73 ± 0.06
	Intermediate	0.81 ± 0.07	0.61 ± 0.09	0.65 ± 0.02
	Poor	0.85 ± 0.08	0.69 ± 0.13	0.64 ± 0.05
Age	< 65	0.86 ± 0.06	0.72 ± 0.09	0.71 ± 0.02
	65-75	0.87 ± 0.15	0.69 ± 0.17	0.65 ± 0.02
	> 75	0.83 ± 0.09	0.73 ± 0.11	0.72 ± 0.04
Treatment	Everolimus	0.87 ± 0.08	0.76 ± 0.08	0.70 ± 0.03
	Nivolumab	0.84 ± 0.06	0.67 ± 0.10	0.73 ± 0.02
Overall		0.86 ± 0.04	0.74 ± 0.06	0.71 ± 0.09

NOTE. CM-025, all available variables in the model; CM-025 (sub), common variables between CM-025 and IMDC from CM-025; IMDC, common variables between CM-025 and IMDC from IMDC.

Abbreviations: CM-025, CheckMate 025; IMDC, International Metastatic Renal Cell Carcinoma Database Consortium; MSKCC, Memorial Sloan Kettering Cancer Center.

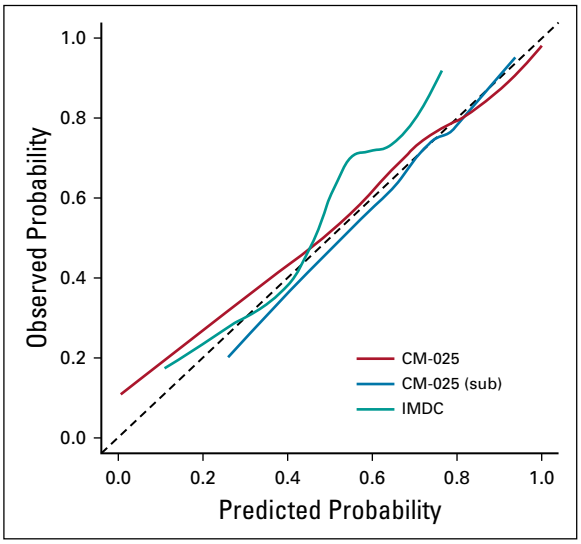

FIG 3. Probability calibration performance. CM-025, all available variables in the model; CM-025 (sub), common variables between CM-025 and IMDC from CM-025; IMDC, common variables between CM-025 and IMDC from IMDC. CM-025, CheckMate 025; IMDC, International Metastatic Renal Cell Carcinoma Database Consortium.

Consistent with the study findings for OS and AE, progression and response in year 1 were also prognostic of TRAE (Appendix).

DISCUSSION

The study focused on developing and validating a proof-of-concept BNM for predicting OS, AE, and TRAE in patients with mRCC.[3] The study also identified prognostic variables that may be useful in clinical practice. The results showed that BNMs could serve as interpretable multivariate predictive models that perform on par with more commonly used black box machine learning approaches. An important differentiation between BNM and the other machine learning models we used is that missing data can be handled intrinsically by the model instead of requiring a prior distinct imputation step, although careful thought needs to be given to whether the mechanism of handling missingness is justified. In our case, CM-025 data contained negligible amounts of total missing data (approximately 3%[16]).

The ability of models to identify prognostic variables over different periods of time may be contingent on changing variability with time. Clinically, being able to predict a patient's 3-month OS is easier than predicting their 24-month OS. AUCs for OS exhibited a logarithmic decline over time (ie, by approximately 26% between years 1 and 2 and by approximately 4% between years 2 and 3). However, adding indicators of treatment response and disease progression in year 1 in the BNM increased the mean AUC for 12-month OS from 0.74 to 0.85. This suggests that the classification of long-term OS may be possible with fair accuracy if baseline data and 12-month follow-up are accounted for in the model. On the other hand, the AUCs for AE increased over time. This is consistent with the general observation that most AE occur within the first 12 months after randomization. It is possible that AE in years 2 and 3 may be present only in patients with poor baseline risk, making them easier to classify.

As previously noted, the model was validated with data from applicable IMDC centers. Comparison of variables between IMDC and CM-025 showed large imbalances in OS, Karnofsky performance status, bone metastases, and BMI, suggestive of a larger fraction of poor-risk individuals in IMDC, and by extension the real world. Notably, prognostic

TABLE 3. Top 10 Prognostic Variables Ranked by AUC for Outcomes in Year 1

	12-Month OS		AE in Year 1		TRAE in Year 1	
Rank	**Variable**	**AUC**	**Variable**	**AUC**	**Variable**	**AUC**
1	MSKCC risk group	0.69	Sodium, serum	0.57	Actual treatment	0.63
2	IMDC/Heng risk group	0.66	MSKCC risk group	0.57	*Subsequent radiotherapy in year 1*	0.55
3	Karnofsky score	0.65	*Progression in year 1*	0.56	Tumor size at baseline	0.53
4	Hemoglobin (g/L)	0.64	Karnofsky score	0.56	Smoking status	0.53
5	Albumin, serum	0.63	Albumin, serum	0.55	Karnofsky performance status	0.53
6	Number of sites of metastases	0.62	Hemoglobin	0.55	Bone metastases	0.53
7	*Progression in year 1*	0.62	Platelet count	0.55	*Subsequent immunotherapy in year 1*	0.52
8	Sodium, serum	0.60	Calcium, serum	0.54	Alkaline phosphatase	0.52
9	Platelet count	0.59	IMDC risk score	0.54	*Response in year 1*	0.52
10	*Response in year 1*	0.58	*Response in year 1*	0.54	Liver metastases	0.52

NOTE. To calculate univariable AUCs, only a single predictor was used for classification using the BNM. Postbaseline variables are italicized.

Abbreviations: AE, adverse event; AUC, area under the receiver operating characteristic curve; BNM, Bayesian network model; IMDC, International Metastatic Renal Cell Carcinoma Database Consortium; MSKCC, Memorial Sloan Kettering Cancer Center; OS, overall survival; TRAE, treatment-related adverse events.

variables such as health-related quality-of-life scores were not available in IMDC. However, the data set was found to be suitable for testing the BNM. The analysis indicated that the overall generalizability of the model was quite high–external validation AUC of 0.71 showed a 4% decrease from cross-validation AUC on CM-025 for OS.

This study has some limitations. First, one primary concern about machine learning approaches is the reliance on black box models that sacrifice transparency and interpretability for greater prediction accuracy.[12] Model transparency and interpretability are essential to clinical research, and domain knowledge is critical for the development and evaluation of patients' risk stratification models. Causal constraints were applied during the final step of machine learning as this had the potential to reduce predictive accuracy in the estimated BNM. Second, although the model was validated across clinically important subgroups, with largely consistent overall validation, patients age at least 75 years or with favorable IMDC risk at baseline had relatively poor external AUCs. The poor generalizability of the model to these two subgroups is possibly due to variable imbalances between CM-025 and IMDC or overfitting and may be explored in future research. Third, we discretized survival outcomes, which may lead to a loss of information. Approaches to BN learning using weighted censored instances may be useful for modeling survival as a censored time-to-event outcome.[17] Our choice to discretize continuous predictors was to avoid distributional assumptions, similar to prior work with BNs.[18-23] Although this may not be ideal because of loss of information, it was negligible for prediction in our case as BNM performance was on par with the other machine learning methods where covariates were not discretized. Last, different types of serious AE were grouped together under two categories (AE and TRAE). However, it is plausible that different sets of prognostic variables apply to different types of AE.

In conclusion, the study developed a proof-of-concept BNM for predicting OS, AE, and TRAE using CM-025 data. The model was then validated using IMDC data, and the performance across clinically important subgroups was highlighted. The practical utility of BNM for interpretable machine learning analysis, multivariate prediction, and handling missing data was also addressed. The results indicate that the model performs on par with black box machine learning models and outperforms IMDC/Heng risk score for classifying OS (with a 10% difference in external AUC for patients with all IMDC risk factors available and a 28% difference in patients with one or more missing risk criteria—assuming that IMDC/Heng risk score cannot be calculated). Future work could extend the BNM to allow for extrapolations beyond the 3-year horizon as the results suggest that long-term predictions with information of 12-month follow-up may be possible. The model could also be used as a clinical risk prediction tool for supporting clinical decision making after prospective validation to determine feasibility for use in the clinic, reproducibility, and accuracy.[19,24] Although combination therapy with nivolumab in the first-line setting is increasingly becoming common, this model would be useful in regions where these combinations are not available or funded and where nivolumab is used for second-line therapy. Nonetheless, we anticipate increased use of transparent machine learning models for leveraging bigger data sets to inform health policy and decision making for innovative therapies in the future.

AFFILIATIONS

[1]Cytel, Toronto, Ontario, Canada
[2]University of Toronto, Toronto, Ontario, Canada
[3]University of Calgary, Calgary, Alberta, Canada
[4]Parexel, Billerica, MA
[5]Parexel, London, United Kingdom
[6]Bristol Myers Squibb, Princeton, NJ
[7]Bristol Myers Squibb, Uxbridge, United Kingdom

CORRESPONDING AUTHOR

Alind Gupta, PhD, Cytel Inc, 1 University Ave, Floor 3, Toronto, Ontario M5J 2P1 Canada; e-mail:alind.gupta@outlook.com.

PRIOR PRESENTATION

Presented in part as a poster at the 2019 Annual Meeting of the Society for Medical Decision Making, October 23, 2019, in Portland, OR.

SUPPORT

Supported by Bristol Myers Squibb.

AUTHOR CONTRIBUTIONS

Conception and design: Alind Gupta, Paul Arora, Darren Brenner, Jacqueline Vanderpuye-Orgle, Devon J. Boyne, Warren Stevens, Samuel Wagner, John Borril, Elise Wu
Administrative support: Paul Arora, Darren Brenner
Collection and assembly of data: Mark Edmondson-Jones, Shaan Dudani
Data analysis and interpretation: Alind Gupta, Paul Arora, Jacqueline Vanderpuye-Orgle, Devon J. Boyne, Mark Edmondson-Jones, Elena Parkhomenko, Daniel Y. C. Heng, Samuel Wagner, John Borrill, Elise Wu
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Alind Gupta
Research Funding: Bristol-Myers Squibb

Paul Arora
Employment: Cytel
Leadership: Lighthouse Outcomes

Darren Brenner
Employment: Cytel

Jacqueline Vanderpuye-Orgle
Consulting or Advisory Role: Bristol-Myers Squibb
Research Funding: Bristol-Myers Squibb

Devon J. Boyne
Employment: Cytel

Elena Parkhomenko
Consulting or Advisory Role: AstraZeneca, Bristol-Myers Squibb

Warren Stevens
Consulting or Advisory Role: Medicus Economics
Research Funding: Medicus Economics

Daniel Y. C. Heng
Consulting or Advisory Role: Pfizer, Novartis, Bristol-Myers Squibb, Janssen, Astellas Pharma, Ipsen, Eisai, Merck
Research Funding: Pfizer, Novartis, Exelixis, Bristol-Myers Squibb, Ipsen

Samuel Wagner
Employment: Bristol-Myers Squibb
Stock and Other Ownership Interests: Bristol-Myers Squibb
Travel, Accommodations, Expenses: Bristol-Myers Squibb

John Borrill
Employment: Bristol-Myers Squibb
Stock and Other Ownership Interests: Bristol-Myers Squibb
Travel, Accommodations, Expenses: Bristol-Myers Squibb

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We thank Bill Malcolm, Murat Kurt, Flavia Ejzykowicz, David Berger, and Heddy Bartell from Bristol-Myers Squibb for valuable feedback on the manuscript and Marek Druzdzel (University of Pittsburgh) for input on analysis.

REFERENCES

1. Marshall HT, Djamgoz MBA: Immuno-oncology: Emerging targets and combination therapies. Front Oncol 8:315, 2018
2. Kaufman HL, Atkins MB, Subedi P, et al: The promise of immuno-oncology: Implications for defining the value of cancer treatment. J Immunother Cancer 7:129, 2019
3. Motzer RJ, Escudier B, McDermott DF, et al: Nivolumab versus everolimus in advanced renal-cell carcinoma. N Engl J Med 373:1803-1813, 2015
4. Borghaei H, Paz-Ares L, Horn L, et al: Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer. N Engl J Med 373:1627-1639, 2015
5. Schachter J, Ribas A, Long GV, et al: Pembrolizumab versus ipilimumab for advanced melanoma: Final overall survival results of a multicentre, randomised, open-label phase 3 study (KEYNOTE-006). Lancet 390:1853-1862, 2017
6. Darvin P, Toor SM, Sasidharan Nair V, et al: Immune checkpoint inhibitors: Recent progress and potential biomarkers. Exp Mol Med 50:165, 2018
7. Kosorok MR, Laber EB: Precision medicine. Annu Rev Stat Appl 6:263-286, 2019
8. Shah P, Kendall FD, Khozin S, et al: Artificial intelligence and machine learning in clinical development: A translational perspective. NPJ Digit Med 2:1-5, 2019
9. Motzer RJ, Mazumdar M, Bacik J, et al: Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma. J Clin Oncol 17:2530-2540, 1999
10. Ko JJ, Xie W, Kroeger N, et al: The International Metastatic Renal Cell Carcinoma Database Consortium model as a prognostic tool in patients with metastatic renal cell carcinoma previously treated with first-line targeted therapy: A population-based study. Lancet Oncol 16:293-300, 2015
11. Voss MH, Reising A, Cheng Y, et al: Genomically annotated risk model for advanced renal-cell carcinoma: A retrospective cohort study. Lancet Oncol 19:1688-1698, 2018
12. Collins GS, Reitsma JB, Altman DG, et al: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMC Med 13:1, 2015
13. Motzer RJ, Tykodi SS, Escudier B, et al: Final analysis of the CheckMate 025 trial comparing nivolumab (NIVO) versus everolimus (EVE) with> 5 years of follow-up in patients with advanced renal cell carcinoma (aRCC). J Clin Oncol 38:617, 2020
14. Koller D, Friedman N, Bach F: Probabilistic Graphical Models: Principles and Techniques. Cambridge, MA, MIT Press, 2009
15. Cella D, Yount S, Brucker PS, et al: Development and validation of a scale to measure disease-related symptoms of kidney cancer. Value Health 10:285-293, 2007
16. Schafer JL: Multiple imputation: A primer. Stat Methods Med Res 8:3-15, 1999
17. Štajduhar I, Dalbelo-Bašić B: Learning Bayesian networks from survival data using weighting censored instances. J Biomed Inform 43:613-622, 2010
18. Park E, Chang HJ, Nam HS: A Bayesian network model for predicting post-stroke outcomes with available risk factors. Front Neurol 9:699, 2018
19. Sesen MB, Nicholson AE, Banares-Alcantara R, et al: Bayesian networks for clinical decision support in lung cancer care. PLoS One 8:e82349, 2013
20. Witteveen A, Nane GF, Vliegen IMH, et al:Comparison of logistic regression and Bayesian networks for risk prediction of breast cancer recurrence. Med Decis Making 38:822-833, 2018
21. Exarchos KP, Exarchos TP, Bourantas CV, et al: Prediction of coronary atherosclerosis progression using dynamic Bayesian networks. 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, Osaka, Japan, 2013
22. Friedman N, Goldszmidt M: Discretizing continuous attributes while learning Bayesian networks. ICML, 1996
23. Chen YC, Wheeler TA, Kochenderfer MJ: Learning discrete Bayesian networks from continuous data. J Artif Intelligence Res 59: 103-132, 2017
24. Sesen MB, Peake MD, Banares-Alcantara R, et al: Lung cancer assistant: A hybrid clinical decision support application for lung cancer care. J R Soc Interf 11:20140534, 2014

APPENDIX
METHODS

Data Preprocessing

For computational tractability for structure learning with Bayesian networks, continuous variables were discretized into three bins (representative of low, intermediate, and high) using *k*-means clustering. The choice of *k*-means with $k = 3$ was guided by exploratory analyses and considerations about interpretability and minimizing the number of parameters for three discretization methods (equal-interval, equal-frequency, and *k*-means) and number of bins ($k = 2, 3, 4,$ or 5) based on prior hyperparameter tuning with other data sets to maximize cross-validation performance, stability of Markov blankets, and model likelihood (Akaike Information Criterion). Both CheckMate 025 (CM-025) and International Metastatic Renal Cell Carcinoma Database Consortium (IMDC) data sets were preprocessed in the same way; *k*-means was run once on CM-025 and identical cutoffs were used to discretize IMDC data. There was no prior variable selection, and regularization was used to minimize the number of parameters to prevent overfitting. Data were not discretized for use with random forest (RF) or logistic regression as these methods do not require it.

Bayesian Network Methods

A Hill-climbing algorithm with Akaike Information Criterion was used for learning Bayesian network structures. Bootstrapping (n = 1,000) followed by model averaging was used to learn the final model structure(Friedman N, et al: 1999; Scutari M, et al: Intelligence Med 57:207-217, 2013) The averaged partially directed graph was converted into a directed acyclic graph using a causal ordering over variables. The treatment node was manually connected to all survival nodes to identify treatment-wise differences for the final model *post hoc*. Bayesian estimation was used for model parametrization—with an uninformative Dirichlet distribution (concentration parameters = 1) as the prior and an imaginary sample size of 1.[14]

Treatment of Missing Values

Missing observations constituted 3% and 17% of the CM-025 and IMDC data sets, respectively. Missing data were treated in two ways—informative, where missing observations were included as an NA category in the model instead of being imputed, and noninformative, where they were marginalized out or imputed. Note that our usage of informative missingness is different from its common usage in survival analysis for estimating the causal effect of interventions in the presence of censoring. There was no difference in cross-validation performance between the two treatments of missingness (data not shown). Missing data for outcomes were not imputed, and observations with a missing outcome were ignored during model fitting and evaluation.

Classification Performance and Calibration

Five independent replicates of 10-fold cross-validation were used to calculate receiver operating characteristic curves and area under the receiver operating characteristic curve (AUC), as well as for model calibration. As classification metrics are defined for binary outcomes, adverse events and treatment-related adverse events were dichotomized as absent (zero events) or present (one or more events). In this paper, AUCs were used to classify if models had (AUC = 0.5), poor ($0.5 < \text{AUC} < 0.6$), fair ($0.6 < \text{AUC} < 0.7$), acceptable ($0.7 < \text{AUC} < 0.8$), excellent ($0.8 < \text{AUC} < 0.9$) or outstanding ($\text{AUC} \geq 0.9$) discrimination. DeLong's test implemented in the *pROC* package was used to compare AUCs (DeLong ER, et al: Biometrics 44:837-845, 1988). For calibration, locally estimated scatterplot smoothing curves were used to fit calibration curves (Austin PC, et al: Stat Med 33:517-535, 2014). Recalibration by linear Platt scaling (Platt J: 10:61-74, 1999) was used to adjust for the difference in mortality between CM-025 and IMDC. Sensitivity analyses were conducted using other machine learning approaches (ie, Lasso regularized logistic regression [Lasso], support vector machine, and RF). All performance metrics for cross-validation and calibration were based on test data only, and observations with missing outcomes, or those who had died or were censored in the previous year, were excluded from evaluation of model performance.

Estimation of Prognostic Value

The prognostic value of individual variables was quantified as univariable (ie, single predictor) AUC estimates using the trained Bayesian network model (BNM). Univariable AUCs were a descriptive measure that used the model trained on the entire CM-025 data set. To calculate univariable AUCs, a single variable X was used for predicting outcomes in the form of the query $P_{BNM}(\text{outcome}|X)$ for all nonoutcome variables in the trained BNM.

Mutual Information

The mutual information between two random variables is defined as a measure of their statistical dependence (Cover TM, et al: 2012). This measure is analogous to (but more general than) correlation.

Software

R statistical computing language was used for all analyses and plots. The *bnlearn* package (Scutari M: Machine Learning arXiv preprint arXiv:0908.3817) (version 4.3) was used for BNM learning and inference. *glmnet*(Friedman J: J Stat Softw 33:1, 2010) (version 2.0.16) was used for Lasso regularized logistic regression. Implementations of support vector machines with a radial basis function kernel (*svmRadial*) and RF (*randomForest*) were from the *caret* package (Kuhn M: J Stat Softw 28:1-26, 2008). For Lasso, support vector machine, and RF, nested cross-validation was used for selecting the optimal value of regularization hyperparameters.

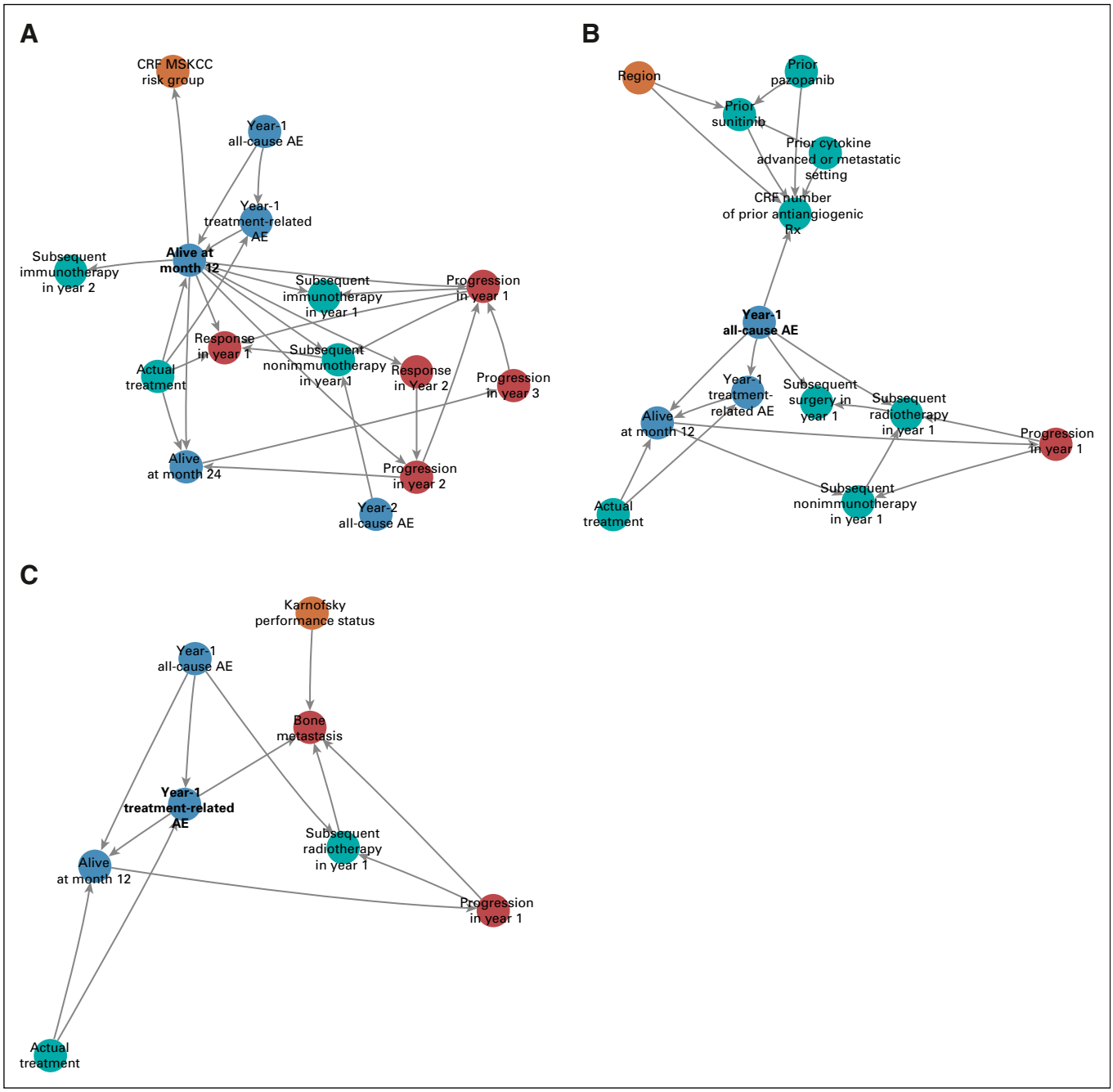

FIG A1. Markov blankets for (A) overall survival, (B) adverse events, and (C) treatment-related adverse events in year 1. AE, adverse event; CRF, case report form; MSKCC, Memorial Sloan Kettering Cancer Center; Rx, prescription.

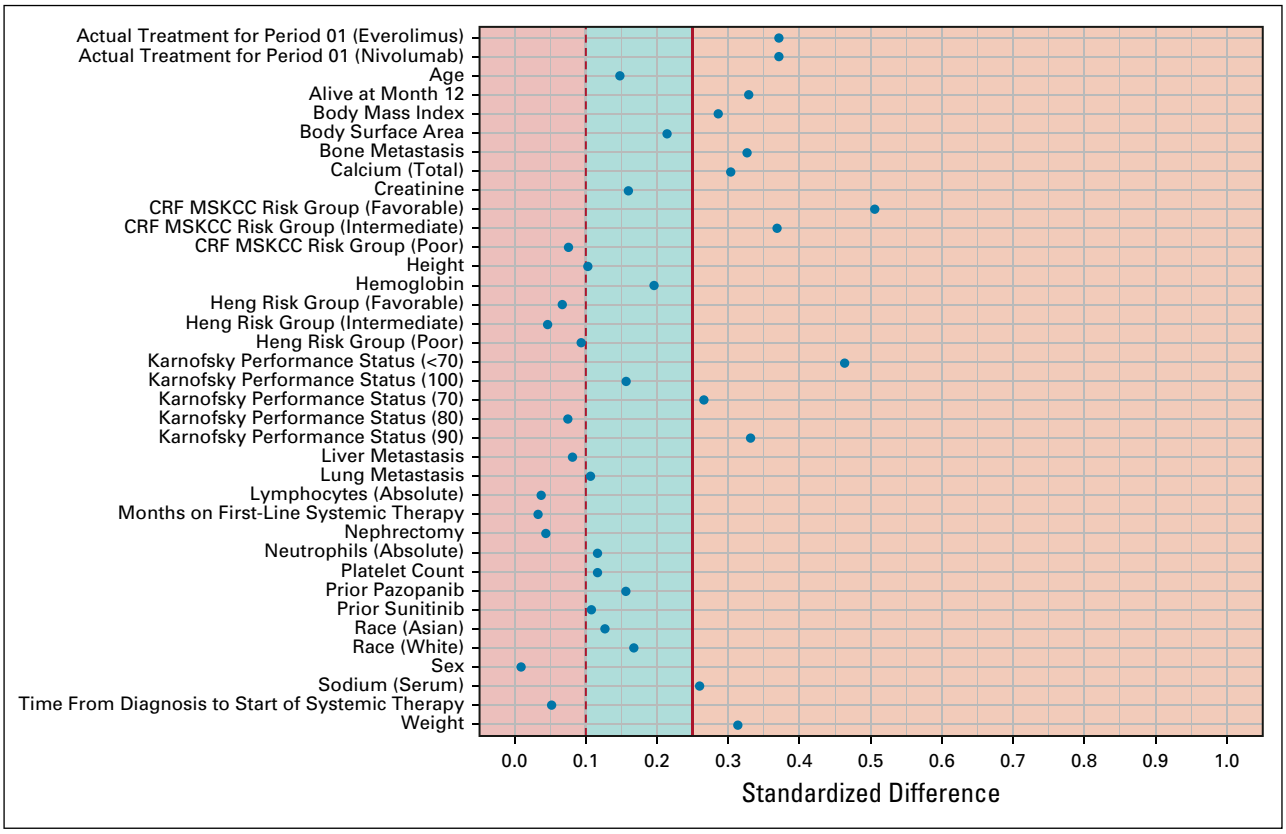

FIG A2. Standardized differences (*d*) between variables in International Metastatic Renal Cell Carcinoma Database Consortium and CheckMate 025. $d < 0.1$ implies no imbalance, whereas $d > 0.25$ implies considerable imbalance. CRF, case report form; MSKCC, Memorial Sloan Kettering Cancer Center.

TABLE A1. AUCs for Classification of OS, AE, and TRAE From Iterated 10-fold Cross-Validation

		AUC (Mean ± Standard Deviation)			
Outcome	**Time**	**BN**	**Lasso**	**SVM**	**RF**
OS	Month 12	0.74 ± 0.06	0.75 ± 0.054	0.76 ± 0.05	0.77 ± 0.05
	Month 24	0.71 ± 0.05	0.72 ± 0.054	0.72 ± 0.05	0.73 ± 0.05
	Month 36	0.68 ± 0.06	0.68 ± 0.059	0.66 ± 0.06	0.68 ± 0.07
AE	Year 1	0.60 ± 0.07	0.60 ± 0.064	0.62 ± 0.06	0.62 ± 0.06
	Year 2	0.68 ± 0.06	0.69± 0.060	0.69± 0.06	0.70± 0.05
	Year 3	0.71 ± 0.06	0.71 ± 0.059	0.68 ± 0.06	0.72 ± 0.06
TRAE	Year 1	0.64 ± 0.06	0.58 ± 0.070	0.56 ± 0.07	0.57 ± 0.05
	Year 2	0.72 ± 0.06	0.73 ± 0.050	0.74 ± 0.05	0.75 ± 0.06
	Year 3	0.71 ± 0.06	0.71 ± 0.053	0.70 ± 0.05	0.72 ± 0.06

NOTE. Means and standard deviations are shown. Only baseline variables were used as predictors.

Abbreviations: AUC, area under the receiver operating characteristic curve; BNM, Bayesian network model; OS, overall survival; TRAE, treatment-related adverse events.

TABLE A2. Top 10 Prognostic Variables Ranked by AUC for Outcomes in Year 2

	24-Month OS		AE in Year 2		TRAE in Year 2	
Rank	**Variable**	**AUC**	**Variable**	**AUC**	**Variable**	**AUC**
1	*Progression in year 1*	0.66	*Subsequent nonimmunotherapy in year 1*	0.77	*Subsequent nonimmunotherapy in year 1*	0.78
2	MSKCC risk group	0.66	*Subsequent radiotherapy in year 1*	0.55	*Subsequent radiotherapy in year 1*	0.61
3	FKSI-DRS	0.65	*Response in year 1*	0.55	*Subsequent nonimmunotherapy in year 2*	0.59
4	Hemoglobin (g/L)	0.65	*Subsequent nonimmunotherapy in year 2*	0.54	*Progression in year 1*	0.57
5	Karnofsky performance score	0.63	*Subsequent surgery in year 1*	0.54	Bone metastasis	0.57
6	IMDC/Heng risk score	0.63	*Response in year 2*	0.54	EQ-5D—activity	0.56
7	EQ-5D—pain	0.61	*Subsequent unknown therapy in year 1*	0.53	AST (U/L)	0.56
8	EQ-5D—activity	0.60	*Subsequent surgery in year 1*	0.53	Triiodothyronine, free	0.56
9	*Subsequent radiotherapy in year 1*	0.60	Time from diagnosis to start of systemic Tx	0.53	Serum sodium (mmol/L)	0.55
10	*Response in year 1*	0.60	Lung metastasis	0.53	Lymphocytes, absolute ($\times 10^9$ c/L)	0.55

NOTE. Postbaseline variables are italicized.

Abbreviations: AE, adverse event; AUC, area under the receiver operating characteristic curve; FKSI-DRS, Functional Assessment of Cancer Therapy–Kidney Symptom Index–Disease-Related Symptoms; IMDC, International Metastatic Renal Cell Carcinoma Database Consortium; MSKCC, Memorial Sloan Kettering Cancer Center; OS, overall survival; TRAE, treatment-related adverse events.

TABLE A3. Top 10 Prognostic Variables Ranked by AUC for Outcomes in Year 3

	36-Month OS		AE in Year 3		TRAE in Year 3	
Rank	**Variable**	**AUC**	**Variable**	**AUC**	**Variable**	**AUC**
1	*Progression in year 1*	0.68	*Progression in year 1*	0.63	*Subsequent nonimmunotherapy in year 1*	0.66
2	FKSI-DRS	0.63	*Subsequent nonimmunotherapy in year 1*	0.62	*Progression in year 1*	0.65
3	Karnofsky score	0.63	*Subsequent nonimmunotherapy in year 2*	0.62	*Subsequent nonimmunotherapy in year 2*	0.62
4	MSKCC risk group	0.62	Hemoglobin	0.60	IMDC/Heng risk group	0.61
5	*Response in year 1*	0.62	Actual treatment	0.59	EQ-5D—anxiety	0.61
6	Hemoglobin (g/L)	0.62	MSKCC risk group	0.59	*Subsequent unknown therapy in year 1*	0.59
7	*Subsequent nonimmunotherapy in year 1*	0.61	IMDC/Heng risk group	0.59	*Subsequent radiotherapy in year 2*	0.58
8	IMDC/Heng risk group	0.60	*Response in year 1*	0.58	*Progression in year 3*	0.58
9	EQ-5D—mobility	0.60	*Subsequent radiotherapy in year 2*	0.57	*Subsequent unknown therapy in year 3*	0.58
10	*Subsequent radiotherapy in year 1*	0.60	Lymphocytes, absolute	0.56	Serum sodium	0.58

NOTE. Postbaseline variables are italicized.

Abbreviations: AE, adverse event; AUC, area under the receiver operating characteristic curve; FKSI-DRS, Functional Assessment of Cancer Therapy–Kidney Symptom Index–Disease-Related Symptoms; IMDC, International Metastatic Renal Cell Carcinoma Database Consortium; MSKCC, Memorial Sloan Kettering Cancer Center; OS, overall survival; TRAE, treatment-related adverse events.