Article

Bayesian network enables interpretable and state-of-the-art prediction of immunotherapy responses in cancer patients

Hideki Hozumi,[1] Hideyuki Shimizu[2, †]

[1]Keio University School of Medicine, Tokyo 160-8582, Japan.
[2]Department of AI Systems Medicine, M&D Data Science Center, Tokyo Medical and Dental University, Tokyo 113-8510, Japan

[†] Correspondence: h_shimizu.dsc@tmd.ac.jp (H.S.)

Running title

Bayesian network for precision medicine in lung cancer immunotherapy

Abstract

Immune checkpoint inhibitors, especially PD-1/PD-L1 blockade, have revolutionized cancer treatment and brought tremendous benefits to patients who otherwise would have had a limited prognosis. Nonetheless, only a small fraction of patients responds to immunotherapy, and the costs and side effects of immune checkpoint inhibitors cannot be ignored. With the advent of machine and deep learning, clinical and genetic data has been used to stratify patient responses to immunotherapy. Unfortunately, these approaches have typically been "black-box" methods that are unable to explain their predictions, thereby hindering their clinical and responsible application. Herein, we developed a "white-box" Bayesian network model that achieves accurate and interpretable predictions of immunotherapy responses against non-small cell lung cancer (NSCLC). This Tree-Augmented naïve Bayes model (TAN) precisely predicted durable clinical benefits and distinguished two clinically significant subgroups with distinct prognoses. Furthermore, Our state-of-the-art white-box TAN approach achieved greater accuracy than previous methods. We hope our model will guide clinicians in selecting NSCLC patients who truly require immunotherapy and expect our approach will be easily applied to other types of cancer.

Keywords

Bayesian network, Tree Augmented naïve Bayes, non-small cell lung cancer, immune checkpoint inhibitors

Abbreviations

AUC: Area under the curve
BN: Bayesian network
DCB: Durable clinical benefit
HC: Hill-climbing method
ICIs: Immune checkpoint inhibitor
ML: Machine learning
MCMC: Markov Chain Monte Carlo
NB: Naïve Bayes
NN: Neural network
NSCLC: Non-small cell lung cancer
TAN: Tree-Augmented naïve Bayes
TMB: Tumor mutational burden

Structured Abstract

Background

Immune checkpoint inhibitors have revolutionized cancer treatment. Given that only a small fraction of patients responds to immunotherapy, patient stratification is a pressing concern. Unfortunately, the “black-box” nature of most of the proposed stratification methods, and their far from satisfactory accuracy, has hindered their clinical application.

Method

We developed a “white-box” Bayesian network model, with interpretable architecture, that can accurately predict immunotherapy response against non-small cell lung cancer (NSCLC). We collected clinical and genetic information from several independent studies, and integrated this via the Tree-Augmented naïve Bayes (TAN) approach.

Findings

This TAN model precisely predicted durable clinical benefit and distinguished two clinically significant subgroups with distinct prognoses, achieving state-of-the-art performance than previous methods. We also verified that TAN succeeded in detecting meaningful interactions between variables from data-driven approach. Moreover, even when data have missing values, TAN successfully predicted their prognosis.

Interpretation

Our model will guide clinicians in selecting NSCLC patients who genuinely require immunotherapy. We expect this approach to be easily applied to other types of cancer. To accelerate the uptake of personalized medicine via access to accurate and interpretable models, we provide a web application (https://pred-nsclc-ici-bayesian.shinyapps.io/Bayesian-NSCLC/) for use by the researchers and clinicians community.

Funding

KAKENHI grant from the Japan Society for the Promotion of Science (JSPS) to H.S (21K17856).

Research in context

Prior evidence

Many studies have advocated the use of biomarkers, such as Programmed Death Ligand-1 (PD-L1) and Tumor Mutational Burden (TMB), to estimate the therapeutic effect of immune checkpoint inhibitors in cancer and utilize them in personalized medicine. Because such single factors are insufficient, many artificial intelligence (AI)-based prediction models have been developed. However, most of these have been “black box” models, in that they lack interpretability. Furthermore, most of them are unable to handle missing data, which is problematic because it is challenging in clinical settings to acquire all of the necessary input information.

Added valued

To address this, we developed an interpretable graphical Tree-Augmented naïve Bayes (TAN) model, and demonstrated its state-of-the-art “white box” performance. It achieved good predictive performance, even when some of the data were missing, and identified relationships between variables that were consistent with previous reports.

Implications

We present the first evidence of a specialized graphical “white-box” model that achieves state-of-the-art performance in immunotherapy, providing strong support for the applicability of interpretable AI models in clinical decision-making. Research using larger datasets will further improve stratification and precision medicine.

Introduction

Lung cancer is the most prevalent cancer and the leading cause of cancer-related death in men and women worldwide[1]. Non-small cell lung cancer (NSCLC) accounts for nearly 85 percent of all lung cancers, and its five-year survival rate remains dismal, ranging from 68% in patients with stage IB cancer to as low as 10% in patients with stage IVA-IVB cancer[2]. Since the invention of immune checkpoint inhibitors (ICIs), many patients have gained tremendous benefits, with improved life expectancy[3]. For instance, nivolumab, an inhibitor of the programmed cell death 1 (PD-1)/ligand (PD-L1) pathway, increased the 2-year survival rate of patients with stage IIIB/IV cancer from 16% to approximately 30%[4].

The decision to administer ICIs to NSCLC patients has been based primarily on the expression level of PD-L1 on the surface of cancer cells, referred to as PD-L1 score[5]. In most cases, patients with higher PD-L1 scores are deemed suitable candidates for ICIs. Nonetheless, numerous studies have demonstrated that not all patients with higher PD-L1 scores respond to ICIs, and even patients with lower PD-L1 scores respond to ICIs[5–7]. According to a meta-analysis of randomized controlled trials, PD-L1 expression alone was insufficient to predict immunotherapy response[7]. In support of this, the PD-L1-based predictive ability was reported to be 0.646 (based on the area under the curve, AUC)[8], indicating that other factors must determine immunotherapy benefits. Further, immunotherapy can have devastating side-effects, particularly immune-related adverse events such as pancreatitis and interstitial pneumonia[9]. The use of ICIs in patients who do not respond to treatment may thus eventually reduce their life expectancy. It is therefore urgent to elucidate the factors other than PD-L1 score that determine the response and prognosis of patients under immunotherapy[10].

Studies to identify factors for stratifying NSCLC patients on ICI treatment have focused on the tumor mutational burden (TMB). Tumors with high TMB would contain many neoantigens and generally respond well to ICIs[10]. However, the predictive ability of TMB was 0.601, based on AUC[8]. Rather than relying on a single indicator (such as PD-L1 score or TMB) to predict immunotherapy response, methods combining multiple factors have emerged. For example, LIPI[11] and EPSILoN[12] integrate clinical data such as clinical stage, performance status, and smoking status. The ratio of neutrophils to lymphocytes, a predictor of rapid progression[10], has been incorporated into these methods. Despite the use of multiple variables, prediction of immunotherapy response rate has remained inadequate, with AUC values of 0.606 and 0.666 for LIPI and EPSILoN, respectively[13]. This evidence indicates that classical approaches cannot provide satisfactory predictions concerning immunotherapy.

In recent years, machine learning (ML)-based methods have been applied to unravel the factors determining the efficacy of ICI treatment for NSCLC. For one example, the AUC of a

neural network model that integrated several factors (TMB, PD-L1 score, mutant-allele tumor heterogeneity, and immune-related pathways) was as high as 0.80 in a test cohort[14]. Another study integrated PD-L1 score and CT images achieved an AUC of 0.76[15]. Other ML methods, such as LightGBM, XGBoost, and regression analysis have also been investigated[16]. Although they harness various types of information (PD-L1 score, radiological images, and clinical features) as input, the AUC remains below 0.80 even for their best models, indicating that it remains challenging to predict responses in ICI therapy. Moreover, ML methods, including neural networks, often lack transparency due to the complexity arising from neural connections and mathematical abstractions[17–19], making it potentially impossible to explain their predictions. This "black-box" nature has retarded the clinical application of established models. Predictive models with higher accuracy and accountability are therefore necessary for the appropriate use of ICIs in NSCLC patients.

With this in mind, we harnessed Bayesian theory and developed an interpretable AI model with state-of-the-art predictive power about immunotherapy. Specifically, we utilized Bayesian network (BN)-based models that capitulate causal relationships in the form of a graphical model[20], allowing us to avoid the black-box problems prevalent in other ML methods[17]. We demonstrate that a Tree-Augmented naïve Bayes (TAN) model predicts the durable clinical benefit (DCB) of patients treated with ICIs with comparable or even better accuracy than conventional ML methods, stratifying patients in a clinically significant manner. It achieved robust predictive ability, even with limited information. This data-driven approach can be used to further elucidate the factors determining immunotherapeutic responses. We anticipate that our interpretable and state-of-the-art approach will expand the knowledge of immunotherapy and be readily applicable to other types of cancer.

Methods

Public cohorts

The cBioPortal (http://www.cbioportal.org)[21] was accessed to retrieve clinical and mutation data for NSCLC patients. We chose two studies examining the effects of ICIs on NSCLC patients[8,22] to use as a dataset for this study. The inclusion criteria and clinical and mutation information for the two cohorts are explained in the original papers[8,22].

The characteristics of our dataset are shown in Table 1, including age (<65 years or not), gender, smoking status, and histopathological information. We excluded 25 samples, comprising mostly those with unspecified histological data (described only as "NSCLC"), and a few categorized as "Large Cell Neuroendocrine Carcinoma". We obtained mutation data to prepare the "frequency-based" and "evidence-based" gene-sets. For the patients in these cohorts, we also analyzed progression-free status.

Model construction

The characteristic of TAN lies in its structural constraints that each explanatory variable can be connected with one node other than an objective variable.

A complete undirected graph with nodes and edges is constructed to estimate this structure. In this stage, one node is wholly connected to every other node. Each variable is described as $X_1, \cdots, X_n$, and mutual information values are given to each edge. The edge weights between two nodes (X_i, X_j) are given by Equation 4:

$$I\left(X_i, X_j \middle| C\right) = p(X_i, X_j, C) log \frac{p(X_i, X_j|C)}{p(X_i|C)p(X_j|C)} \cdots (4)$$

To obtain the constrained structure of TAN from this complete graph, a structure with the highest total weights under the constraint is used as an estimated structure. To transform the given undirected graph tree into a directed one, a root variable is randomly chosen, and the direction of the edges is set to outward from the root variable[20,23,24]. The data were then randomly split into training (2/3) and test data (1/3) (Figure 1b).

The training data were used to construct the models and to learn the conditional probability between each node. ROC curves were constructed from the test-data predictions. We constructed the model using the bnlearn (4.7.1) R package, and used the ROCR package (1.0-11) for evaluation.

Model evaluation

Model-averaging methods were adopted to measure the reliability of the connections between nodes in the network, by performing multiple structural estimations using the hill-climbing method[23]. In the Bayesian network, it is important to measure the confidence level

for a particular graph feature (the graph edge). This confidence level (in terms of relative frequencies) referred to as arc strength[23,25,26], is defined as the number of times an internode connection appears while generating multiple graphs; frequencies >85% are considered strong[23].

We adopted two model-averaging methods for evaluating the node connections of our model. The first is the bootstrap approach, which applies nonparametric bootstrapping to generate multiple networks, and estimates the arc strength[23,25].
Algorithm 1 provides the specific method.

Algorithm 1

1. For $b = 1,2,\ldots,B$:
    1.1. Sample a data set D_b from the original data D via nonparametric bootstrapping.
    1.2. Learn the Bayesian network, $\mathcal{G}_b = (V, E_b)$ from D_b.
2. Estimate the arc strength, defined as follows:

$$\widehat{p_i} = \hat{P}(a_i) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{N}_{\{e_i \in E_b\}}$$

where $\mathbb{N}$ is equal to 1 if $e_i \in E_b$ and 0 otherwise.

The second model-averaging method is the random generation of multiple graphs from a uniform distribution, using the MCMC algorithm (Algorithm 2). We randomly sampled one graph for every 50 graphs generated and measured arc strength from 500 sampled graphs[26].

Algorithm 2

1. For $b = 1,2,\ldots,B$:
    1.1. Sample a data set D_b from the original data D via parametric or nonparametric bootstrapping.
    1.2. Learn the Bayesian network, $\mathcal{G}_b = (V, E_b)$ from D_b.
2. Estimate the arc strength, defined as follows:

$$\widehat{p_i} = E(e_i|D) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{N}_{\{e_i \in E_b\}} P(\mathcal{G}_b|D)$$

The robustness of the TAN structure estimation was evaluated by examining whether the connections between nodes determined to be significant by these model-averaging methods were also present in the structure of the TAN.

Inference with limited evidence

To estimate the conditional probability of an event using only the limited evidence available, we used the cpquery function of the bnlearn package (4.7.1). In this method, logic sampling, or an approximate inference, enables it to obtain the probability[27]. First, a new data set is created by randomly extracting data that matches the specified evidence from the whole data set. In our case, patient profiles or genetic mutation information were specified. By repeating this method, one million random samples are generated, and an approximate probability is returned based on them.

Survival analysis

We conducted survival analysis using the survival package (3.3-1). $P < 0.05$ was considered statistically significant.

Data availability

All clinical and mutation information is available from the cBioPortal database, (http://www.cbioportal.org), and the specific explanation of each cohort can be obtained in the original papers[8,22].

We provide a web application (https://pred-nsclc-ici-bayesian.shinyapps.io/Bayesian-NSCLC/) using the shiny package (1.7.2), providing both frequency- and evidence-based models.

Code availability

The R code for training the NB and TAN models, and for validation and scoring via ROC and survival analysis are available at GitHub (https://github.com/Hideki-Hozumi/Prediction-with-bayesian-network.).

Results

Manual curation of clinical information related to immunotherapy

To develop a state-of-the-art explanatory model, we first retrieved data for immunotherapy-receiving cancer patients from cBioPortal (http://www.cbioportal.org), which offers clinical data with mutational information[21]. Specifically, two previously published studies[8,22] examining the effect of ICIs on NSCLC patients were selected and used as a dataset for this study: the Hellman cohort, comprising of 75 NSCLC patients who underwent immunotherapy[22], and the Rivzi cohort, of 240 NSCLC patients treated with immunotherapy[8]. In total, our dataset includes 315 patients (Figure 1a). The characteristics of our dataset are shown in Table 1.

Among the available clinical information, we set the objective variable as DCB, which is defined in the revised RECIST guideline (version 1.1) as partial response/stable disease lasting >6 months[28]. We focused on DCB because the follow-up criteria for overall survival and progression-free status were inconsistent between the two cohorts[8,22]. Given that DCB has been used to assess the efficacy of immunotherapy for various tumors such as melanoma[29] and lung cancer[8], we believe that predicting DCB is of clinical value for stratifying the patients in our study.

We used the three known clinical risk factors of NSCLC: age (< 65years of age or not[30]), gender, and smoking status[31]. Our model incorporated histopathological information, because the pathological subtype substantially affects the prognosis[32]. We excluded 25 samples for which there was insufficient histopathological information (Figure 1a).

Genetic covariates were determined in two ways. First, genes with mutation rates higher than 10% in our dataset were incorporated (hereafter, the "frequency-based geneset"); these include *TP53, KRAS, TTN, KMT2C, SMARCA4, STK11,* and *KEAP1*. Second, based on a literature survey, we identified six genes (*KRAS*[33] *STK11*[34,35]*, TP53*[36]*, EGFR*[37]*, ALK*[37]*,* and *ROS1*[37]) associated with NSCLC-patient ICI-treatment responses or prognosis (hereafter, “evidence-based geneset”). We categorized "Deletion", “in-frame deletion (IF del)”, “frameshift deletion (FS del)”, “splice mutation (Splice)” and "Missense" modifications as "Mutation" since they would likely impair the original function of the gene[38,39].

We attempted to decipher the relationships underlying DCB by combining clinical characteristics with mutation data. For this purpose, we randomly divided the dataset into training and test data (Figure 1a), using the former to build a model and the latter solely for evaluation[40]. Receiver operating characteristic (ROC) analysis was performed to evaluate model performance. Survival analysis was conducted to verify the model’s ability to predict prognosis in addition to DCB (Figure 1b). We describe the model construction procedure in the following section.

Tree-Augmented naïve Bayes model robustly and interpretability predicted DCB

We harnessed a Bayesian network graphical model to achieve accurate and interpretable predictions of DCB. Bayesian networks graphically represent multivariate probability distributions[20], and are broadly applied in various biomedical tasks, including gene network feature selection[41], signaling network prediction[42], and predicting hematological malignancy type[43]. Naïve Bayes (NB) networks, the simplest type of Bayesian network, but generally achieve favorable prediction accuracy[44]. Based on Bayes' theorem (Equation 1), NB assumes that all covariates are equally important without distinction and are conditionally independent given a class value (Equation 2)[44].

$$p(C|X_1,\cdots,X_n) = \frac{p(X_1,\cdots,X_n|C)p(C)}{p(X_1,\cdots,X_n)} \Leftrightarrow Posterior = \frac{Prior \times Liklihood}{Evidence} \cdots (1)$$

$$p(C|X_1,\cdots,X_n) \propto p(C)\prod_{k=1}^{n} p(X_k|C) \cdots (2)$$

The probability associated with a parent node (objective variable) is described as $p(C)$, and the probability is updated to $p(C|X_1,\cdots,X_n)$ when explanatory information from child nodes $(X_1,\cdots,X_n)$ is provided. In terms of its network structure, arrows (directed edges) extend from one node (a parent node or objective variable) to all other nodes (child nodes) (Supplementary Figure 1a). Despite its simple design and assumptions, NB achieves much better classification than expected, and is used in medical data analysis[45]. Nonetheless, in its original form, it depends relatively heavily on the assumption that the covariates are statistically independent, hampering its application to real-world biomedical data.

To address this, we utilized Tree-Augmented naïve Bayes models (hereafter, "TAN"; Equation 3):

$$p(C|X_1,\cdots,X_n) \propto p(C)\cdot p(X_1|C)\cdot p(X_2|X_k,C)\cdot\cdots\cdot p(X_n|X_l,C) \cdots (3)$$

TAN alleviates the conditional independence between features, while keeping the directed acyclic graph simpler than in conventional neural network models (Supplementary Figure 1b, 1c). TAN does not assume conditional independence, partially allowing dependent relationships between variables (Equation 3)[46]. Therefore, because it can express a greater number of states, TAN must outperform NB model. Indeed, it has been applied in numerous biomedical tasks, including risk stratification in pulmonary hypertension[47], and mammography[48], achieving high accuracy. Here, we used NB and TAN to establish predictive

328 models with higher accuracy and interpretability, and compared their ability to predict DCB in
329 NSCLC patients.
330 First, we constructed frequency-based models, using clinical data and seven genes from
331 the frequency-based geneset (*TP53, KRAS, TTN, KMT2C, SMARCA4, STK11,* and *KEAP1*) as
332 covariates. The structure of frequency-based NB is shown in Figure 2a. TAN structure was
333 estimated using a training dataset (Figure 2b). For the NB model, the area under the curve
334 (AUC) was 0.686 for the training dataset and 0.726 for the test dataset. (Figure 2c); for the
335 TAN model, these values were 0.836 and 0.728, respectively (Figure 2d). These results
336 indicate that the TAN model has comparable or greater predictive accuracy than the NB
337 model.
338 We next constructed evidence-based NB (Figure 2e) and TAN (Figure 2f) models, using
339 clinical information and six genes from the evidence-based geneset (*KRAS, STK11, TP53,*
340 *EGFR, ALK*, and *ROS1*) as covariates, using the same approach used for the frequency-
341 based models. Using the test dataset, the NB and TAN AUCs were 0.712 (Figure 2g) and
342 0.823 (Figure 2h), respectively, suggesting that TAN outperformed NB.
343 These lines of evidence demonstrate that the optimized TAN model outperforms NB, and
344 robustly predicts DCB via frequency- and evidence-based approaches.
345 Its performance is comparable to that of other cutting-edge methods[15,49,50], while retaining
346 explainability.
347
348 **Optimized TAN yields a robust graphical structure**
349 We next evaluated the robustness of the structural estimation of our model in immunotherapy.
350 We statistically generated multiple directed acyclic graphs (DAGs): significant edges
351 (internode connections) were detected when it appeared in >85% of the graphs.
352 We used two model-averaging methods[25,51] to determine if the relationships identified by
353 our methodology (Figure 2b for the frequency-based model and Figure 2f for the evidence-
354 based model, respectively) were sufficiently robust. We performed bootstrap sampling[25] and
355 Markov chain Monte Carlo (MCMC) methods to randomly constructed DAGs from a uniform
356 distribution, as previously reported[51]. This revealed several significant connections (Table 2,
357 Figure 3a, 3b for the frequency-based model; Table 3, Figure 3c, 3d for the evidence-based
358 model). These results demonstrate that the model-averaging methods produce similar
359 architectures, indicating that our method robustly discovers crucial relationships governing
360 the immunotherapy response.
361
362 **Our TAN model stratifies and inferences even with limited clinical information**

Patient stratification is crucial to the development of personalized medicine[52]. We thus evaluated our model's applicability to the stratification of NSCLC patients. We obtained the progression-free status of the patients in our dataset from the cBioPortal database[8,22]. Our models identified two distinct and clinically significant groups based on binary prediction (Figure 4a for the frequency-based model and Figure 4b for the evidence-based model).

Importantly, the optimized TAN model can handle missing data and calculate conditional probabilities. For instance, it can predict whether a tumor will respond to immunotherapy, even if all we know about a particular NSCLC patient is that they have *TP53* and *STK11* mutations; the estimated response probability is 0.163, indicating that this patient would not benefit substantially from ICIs (Figure 5). This speculation is consistent with established evidence[33]. In contrast, for a young female patient with a *KMT2C* mutation, but no *STK11* and *TP53* mutations, the estimated ICI response probability is 0.592, indicating that ICI treatment would be valuable. Previous models, including those based on ML and deep learning methods, cannot adequately handle missing data, requiring all of the necessary information[53]. Given that data acquisition can be laborious, particularly in clinical settings, our model may help clinicians in decision-making, especially in data-limited situation.

Discussion

Most prior attempts to predict immunotherapy responses have used ML-based approaches[14–16], which are complex “black-box” systems that cannot handle missing data. As input, they require all of the necessary clinical and molecular information to be provided. Such data are often difficult to obtain, especially in hospitals with limited resources, hampering the clinical application of these models.

Transparency and clinical validation are necessary to achieve trustable medical AI[17]. Therefore, we sought to develop an interpretable and robust model that predicts NSCLC patient responses to immunotherapy. We used clinical information, selected mutation data based on the frequency and evidence-based approaches, and established optimized TAN models. Our approach is comparable with, or even superior to, several cutting-edge ML-based methods[15,49,50], while retaining explainability. It provides clinically informative predictions even when data are limited (Figure 5), as is quite common in clinical settings. Furthermore, because this model only computes conditional probabilities based on Bayes’ theorem[25], it is possible, if necessary, to control which nodes should have (or should not) have connections, using a “white list” (or “black list”) based on expert knowledge.

We selected several genes based on the mutation frequency or previous evidence. *KRAS,* an immunomodulatory oncogenic gene, leads to escape from immunotherapy[34]. Together

with *TP53* or *STK11* mutations, *KRAS* mutation is a potent prognostic factor[33,36]. *STK11* is associated with diminished immunotherapy response[35]. *BRAF* mutation, which are associated with a higher tumor burden, may make tumors vulnerable to immunotherapy[54]. Mutations in driver oncogenes such as *EGFR, ALK,* and *ROS1* in tumors cause a lack of immunogenicity, and thus, a poor response to immunotherapy, regardless of PD-L1 score[55]. Therefore, the expert consensus suggests that patients with these mutations should not be treated with immunotherapy[5].

Our models could also be used to generate intriguing hypotheses for future research. For instance, our inferences based on limited data (Figure 5) are consistent with the findings of recent reports[33,35]. This suggests that, by using more clinical samples with diverse genetic profiles, our approach may reveal new targets in immunotherapy, providing an invaluable resource for both clinical and basic medicine.

Consistent with an earlier analysis of clinical data on the utility of TAN[48], our TAN-based approach provided greater value than the NB model. Because of the small sample, the conventional NB model using hill-climbing methods were unable to construct suitable structures for inference (Supplementary Figure 2), suggesting that our approach is better suited to inference with a small dataset. TAN alleviates the conditional constraints imposed by NB. Here, some of the essential connections in TAN structural learning were also detected via model-averaging using bootstrap sampling and MCMC (Figure 3). Our model-averaging findings obtained using the frequency-based approach (Figure 3a, 3b), for instance, strongly suggest an association between smoking status and *KRAS* mutation, which has been established in a previous report[37]. Other strong connections between nodes inferred by model-averaging method (Figure 3) are expected to reveal immunotherapy-related hidden relationships.

In terms of potential limitations, we cannot rule out selection bias due to the integrated use of public datasets. Although the datasets comprise patients who underwent immunotherapy, it is plausible that the data do not represent a specific population. In addition, the strength of the internode relationships that we estimated may reflect the small sample size, and an analysis employing a larger dataset may reveal additional relationships. We therefore developed a web-based intuitive DCB estimator (https://pred-nsclc-ici-bayesian.shinyapps.io/Bayesian-NSCLC/) that does not require computational expertise. Future analyses with larger clinical samples are likely to overcome these limitations, and provide further support for the validity of this approach.

In summary, these robust models are comparable with, or even superior to, other predictive models for immunotherapy. They can predict meaningful and interpretable connections and inferences, even with a limited number of observations. We hope that this

435 model will guide clinicians in selecting NSCLC patients who require immunotherapy, and
436 expect it to be easily applied to other types of cancer.
437

Acknowledgements

This work was supported by KAKENHI grants from the Japan Society for the Promotion of Science (JSPS) to H.S. (21K17856). We thank our laboratory members for discussion.

Contributions

H.H. and H.S. designed the project. H.H. contributed to formal analyses and interpretation. H.H. wrote the draft version of the manuscript. H.S. supervised the study and edited the manuscript. All authors contributed to the article and approved the final version.

Competing interests

The authors declare no competing interests.

450 **References**

451 1. Ganti, A. K., Klein, A. B., Cotarla, I., Seal, B. & Chou, E. Update of incidence,
452 prevalence, survival, and initial treatment in patients with non-small cell lung cancer in
453 the US. *JAMA Oncol.* **7**, 1824–1832 (2021).
454 2. Duma, N., Santana-Davila, R. & Molina, J. R. Non–small cell lung cancer: Epidemiology,
455 screening, diagnosis, and treatment. *Mayo Clin Proc.* **94**, 1623–1640 (2019).
456 3. Zhang, Y. & Zhang, Z. The history and advances in cancer immunotherapy:
457 Understanding the characteristics of tumor-infiltrating immune cells and their
458 therapeutic implications. *Cell Mol Immunol.* **17**, 807–821 (2020).
459 4. Horn, L. *et al.* Nivolumab versus docetaxel in previously treated patients with advanced
460 non-small-cell lung cancer: Two-year outcomes from two randomized, open-label,
461 phase III Trials (CheckMate 017 and CheckMate 057). *Journal of Clinical Oncology* **35**,
462 3924–3933 (2017).
463 5. Brahmer, J. R. *et al.* The Society for Immunotherapy of Cancer consensus statement on
464 immunotherapy for the treatment of non-small cell lung cancer (NSCLC). *J Immunother*
465 *Cancer.* **6**, 75 (2018).
466 6. Yu, H., Boyle, T. A., Zhou, C., Rimm, D. L. & Hirsch, F. R. PD-L1 Expression in lung
467 cancer. *J Thorac Oncol.* **11**, 964–975 (2016).
468 7. Shen, X. & Zhao, B. Efficacy of PD-1 or PD-L1 inhibitors and PD-L1 expression status in
469 cancer: meta-analysis. *BMJ.* **362**, k3529 (2018).
470 8. Rizvi, H. *et al.* Molecular determinants of response to anti-programmed cell death (PD)-
471 1 and anti-programmed death-ligand 1 (PD-L1) blockade in patients with non-small-cell
472 lung cancer profiled with targeted next-generation sequencing. *J Clin Oncol.* **36**, 633–
473 641 (2018).
474 9. Martins, F. *et al.* Adverse effects of immune-checkpoint inhibitors: Epidemiology,
475 management and surveillance. *Nat Rev Clin Oncol.* **16**, 563–580 (2019).
476 10. Brueckl, W. M., Ficker, J. H. & Zeitler, G. Clinically relevant prognostic and predictive
477 markers for immune-checkpoint-inhibitor (ICI) therapy in non-small cell lung cancer
478 (NSCLC). *BMC Cancer.* **20**, 1185 (2020).
479 11. Mezquita, L. *et al.* Association of the lung immune prognostic index with immune
480 checkpoint inhibitor outcomes in patients with advanced non-small cell lung cancer.
481 *JAMA Oncol.* **4**, 351–357 (2018).
482 12. Prelaj, A. *et al.* EPSILoN: A prognostic score using clinical and blood biomarkers in
483 advanced non-small-cell lung cancer treated with immunotherapy. *Clin Lung Cancer.*
484 **21**, 365-377.e5 (2020).

13. Zhao, Q. *et al.* Three models that predict the efficacy of immunotherapy in Chinese patients with advanced non-small cell lung cancer. *Cancer Med.* **10**, 6291–6303 (2021).
14. Jiang, J. *et al.* Robust prediction of immune checkpoint inhibition therapy for non-small cell lung cancer. *Front Immunol.* **12**, 646874 (2021).
15. Tian, P. *et al.* Assessing PD-L1 expression in non-small cell lung cancer and predicting responses to immune checkpoint inhibitors using deep learning on computed tomography images. *Theranostics.* **11**, 2098–2107 (2021).
16. Ahn, B. C. *et al.* Clinical decision support algorithm based on machine learning to assess the clinical response to anti-programmed death-1 therapy in patients with non-small-cell lung cancer. *Eur J Cancer.* **153**, 179–189 (2021).
17. Quinn, T. P., Jacobs, S., Senadeera, M., Le, V. & Coghlan, S. The three ghosts of medical AI: Can the black-box present deliver? *Artif Intell Med.* **124**, 102158 (2022).
18. Amann, J., Blasimme, A., Vayena, E., Frey, D. & Madai, V. I. & Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak.* **20**, 310 (2020).
19. Castelvecchi, D. Can we open the black box of AI? *Nature* **538**, 20–23 (2016).
20. Friedman, N., Geiger, D. & Goldszmidt, M. Bayesian Network Classifiers. *Mach Learn.* **29**, 131–163 (1997).
21. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
22. Hellmann, M. D. *et al.* Genomic features of response to combination immunotherapy in patients with advanced non-small-cell lung cancer. *Cancer Cell.* **33**, 843-852.e4 (2018).
23. Scutari, M. & Denis, J.-B. *Bayesian Networks : With Examples in R*. *Chapman and Hall/CRC* (Chapman and Hall/CRC, 2021).
24. Cheng, J. & Greiner, R. Comparing Bayesian network classifiers. *CoRR.* **1301**, 6684 (2013).
25. Friedman, N., Goldszmidt, M. & Wyner, A. Data analysis with Bayesian networks: A bootstrap approach. *CoRR.* **1301**, 6695 (2013).
26. Friedman, N. & Koller, D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning.* **50**, 95–125 (2003).
27. Henrion, M. Propagating Uncertainty in Bayesian Networks by Probabilistic Logic Sampling. *Machine Intelligence and Pattern Recognition.* **5**, 149–163 (North-Holland, 1988).
28. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer.* **45**, 228–247 (2009).

520 29. Kaufman, H. L. *et al.* Durable response rate as an endpoint in cancer immunotherapy:
521 Insights from oncolytic virus clinical trials. *J Immunother Cancer.* **5**, 72 (2017).
522 30. Huo, G., Liu, W. & Chen, P. Inhibitors of PD-1 in non-small cell lung cancer: A meta-
523 analysis of clinical and molecular features. *Front Immunol.* **13**, 875093 (2022).
524 31. Schabath, M. B. & Cote, M. L. Cancer Progress and priorities: Lung cancer. *Cancer*
525 *Epidemiol Biomarkers Prev.* **28**, 1563–1579 (2019).
526 32. Balata, H. *et al.* Predicting survival following surgical resection of lung cancer using
527 clinical and pathological variables: The development and validation of the LNC-PATH
528 score. *Lung Cancer.* **125**, 29–34 (2018).
529 33. La Fleur, L. *et al.* Mutation patterns in a population-based non-small cell lung cancer
530 cohort and prognostic impact of concomitant mutations in KRAS and TP53 or STK11.
531 *Lung Cancer.* **130**, 50–58 (2019).
532 34. Hamarsheh, S., Groß, O., Brummer, T. & Zeiser, R. Immune modulatory effects of
533 oncogenic KRAS in cancer. *Nat Commun.* **11**, 5439 (2020).
534 35. Skoulidis, F. *et al.* STK11/LKB1 mutations and PD-1 inhibitor resistance in KRAS-
535 mutant lung adenocarcinoma. *Cancer Discov.* **8**, 822–835 (2018).
536 36. Xu, F. *et al.* A TP53-associated gene signature for prediction of prognosis and
537 therapeutic responses in lung squamous cell carcinoma. *Oncoimmunology.* **9**, 1731943
538 (2020).
539 37. Chapman, A. M., Sun, K. Y., Ruestow, P., Cowan, D. M. & Madl, A. K. Lung cancer
540 mutation profile of EGFR, ALK, and KRAS: Meta-analysis and comparison of never and
541 ever smokers. *Lung Cancer.* **102**, 122–134 (2016).
542 38. Lea, I. A. *et al.* Genetic pathways and mutation profiles of human cancers: site- and
543 exposure-specific patterns. *Carcinogenesis.* **28**, 1851–1858 (2007).
544 39. Bardelli, A. & Velculescu, V. E. Mutational analysis of gene families in human cancer.
545 *Curr Opin Genet Dev.* **15**, 5–12 (2005).
546 40. Raschka, S. Model evaluation, model selection, and algorithm selection in machine
547 learning. *CoRR.* **1811**, 12808 (2018).
548 41. Lan, Z., Zhao, Y., Kang, J. & Yu, T. Bayesian network feature finder (BANFF): an R
549 package for gene network feature selection. *Bioinformatics.* **32**, 3685–3687 (2016).
550 42. Qin, T., Tsoi, L. C., Sims, K. J., Lu, X. & Zheng, W. J. Signaling network prediction by
551 the Ontology Fingerprint enhanced Bayesian network. *BMC Syst Biol.* **6**, S3 (2012).
552 43. Agrahari, R. *et al.* Applications of Bayesian network models in predicting types of
553 hematological malignancies. *Sci Rep.* **8**, 6951 (2018).
554 44. Webb, G. I. Naïve Bayes. in *Encyclopedia of Machine Learning* (eds. Sammut Claude &
555 and Webb, G. I.) 713–714 (Springer US, 2010).

45. Kononenko, I. Machine learning for medical diagnosis: History, state of the art and perspective. *Artif Intell Med.* **23**, 89–109 (2001).
46. Zheng Fei & and Webb, G. I. Tree Augmented Naive Bayes. in *Encyclopedia of Machine Learning* (eds. Sammut Claude & and Webb, G. I.) 990–991 (Springer US, 2010).
47. Kanwar, M. K. *et al.* Risk stratification in pulmonary arterial hypertension using Bayesian analysis. *Eur Respir J.* **56**, 2000008 (2020).
48. Burnside, E. S. *et al.* Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology.* **251**, 663–672 (2009).
49. He, B. X. *et al.* Deep learning for predicting immunotherapeutic efficacy in advanced non-small cell lung cancer patients: a retrospective study combining progression-free survival risk and overall survival risk. *Transl Lung Cancer Res.* **11**, 670–685 (2022).
50. Ouyang, W. *et al.* A prognostic risk score based on hypoxia-, immunity-, and epithelialto-mesenchymal transition-related genes for the prognosis and immunotherapy response of lung adenocarcinoma. *Front Cell Dev Biol.* **9**, 758777 (2022).
51. Ide, J. S. & Cozman, F. G. Testing MCMC algorithms with randomly generated Bayesian networks. *Machine Intelligence and Pattern Recognition.* **5**, 149–163 (2002).
52. Wang, M., Herbst, R. S. & Boshoff, C. Toward personalized treatment approaches for non-small-cell lung cancer. *Nat Med.* **27**, 1345–1356 (2021).
53. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov.* **9**, e1312 (2019).
54. Negrao, M. v. *et al.* Oncogene-specific differences in tumor mutational burden, PD-L1 expression, and outcomes from immunotherapy in non-small cell lung cancer. *J Immunother Cancer.* **9**, e002891 (2021).
55. Blons, H., Garinet, S., Laurent-Puig, P. & Oudart, J. B. Molecular markers and prediction of response to immunotherapy in non-small cell lung cancer, an update. *J Thorac Dis.* **11**, S25–S36 (2019).

588 **Tables**

589 **Table 1. Characteristics of the dataset**

Characteristic	Numbers
Age	
≧65	157
<65	133
Histopathology	
Adenocarcinoma	256
Squamous	34
Gender	
Male	141
Female	149
Smoker	
Ever	231
Never	59
DCB	
Yes	99
No	191
TP53	
Mutation	168
No Mutation	122
KRAS	
Mutation	100
No Mutation	190
STK11	
Mutation	50
No Mutation	240
KEAP1	
Mutation	47
No Mutation	243
TTN	
Mutation	34
No Mutation	256
KMT2C	

Mutation	32
No Mutation	258
SMARCA4	
Mutation	30
No Mutation	260
EGFR	
Mutation	26
No Mutation	264
ALK	
Mutation	13
No Mutation	277
ROS1	
Mutation	7
No Mutation	283

593 **Table 2. Node verification via bootstrapping and MCMC for the frequency-based model**

594

Methods	Connection	Strength	TAN model
Bootstrap	Smoker-*KRAS*	0.956	○
	Smoker-*KMT2C*	0.874	○
	TP53-*KRAS*	0.990	○
	KRAS-Cancer	0.968	○
	TTN-Cancer	0.854	×
	STK11-*KEAP1*	1.000	○
MCMC	Cancer-*KRAS*	1.000	○
	Smoker-Age	0.878	○
	Smoker-*TP53*	0.982	×
	Smoker-*KRAS*	0.982	○
	Smoker-*KEAP1*	0.962	○
	KRAS-*TP53*	1.000	○
	TTN-Cancer	0.978	×
	TTN-*KMT2C*	0.868	×
	TTN-*SMARCA4*	0.974	×
	STK11-*TP53*	0.978	×
	KEAP1-*SMARCA4*	0.970	×
	KEAP1-*STK11*	1.000	○

595

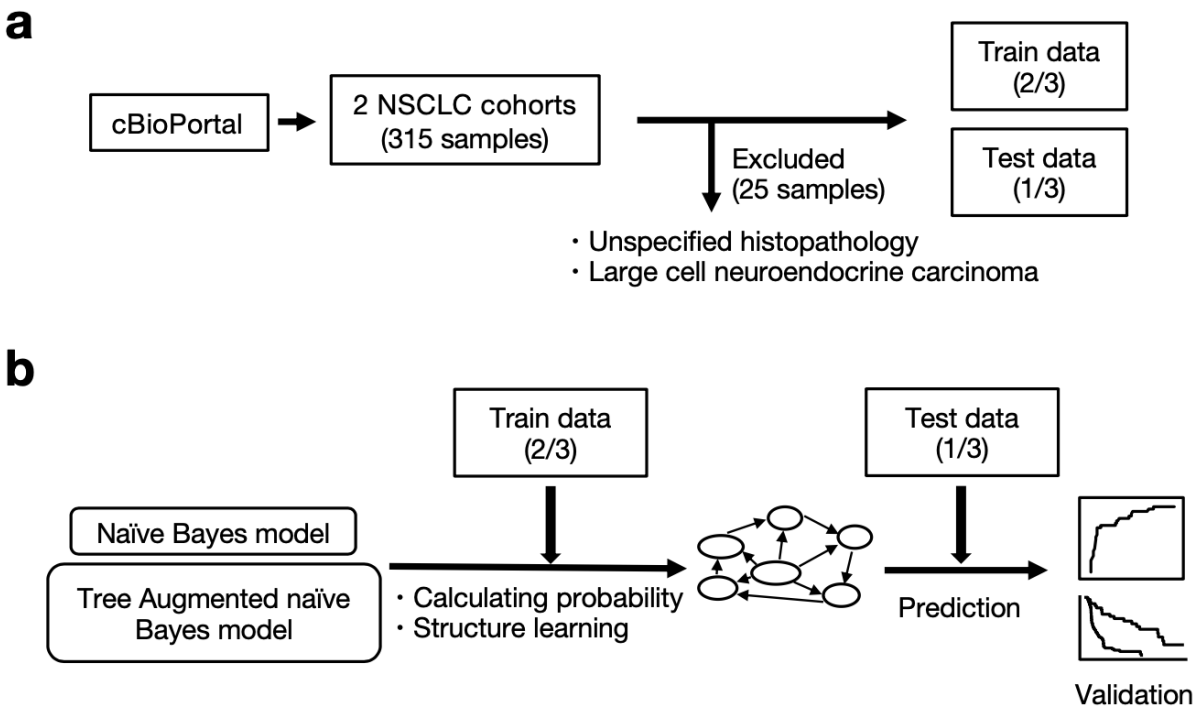
596

Table 3. Node verification via bootstrapping and MCMC for the evidence-based model

Methods	Connection	Strength	TAN model
Bootstrap	Cancer-*KRAS*	0.978	○
	Cancer-*ROS1*	0.880	×
	Smoker-*TP53*	0.942	×
	Smoker-*EGFR*	0.872	○
	KRAS-*TP53*	1.000	○
	STK11-*TP53*	0.847	○
	ALK-*ROS1*	0.956	○
MCMC	Smoker-*KRAS*	0.958	○
	Smoker-*STK11*	0.980	×
	Smoker-*TP53*	0.998	×
	Smoker-*EGFR*	0.908	○
	KRAS-Cancer	1.000	○
	KRAS-*TP53*	1.000	○
	STK11-*TP53*	1.000	○
	ALK-*EGFR*	0.942	×
	ROS1-Cancer	1.000	×
	ROS1-*ALK*	0.970	○

601 **Figures and figure legends**

602



Hozumi et al., Figure 1

603

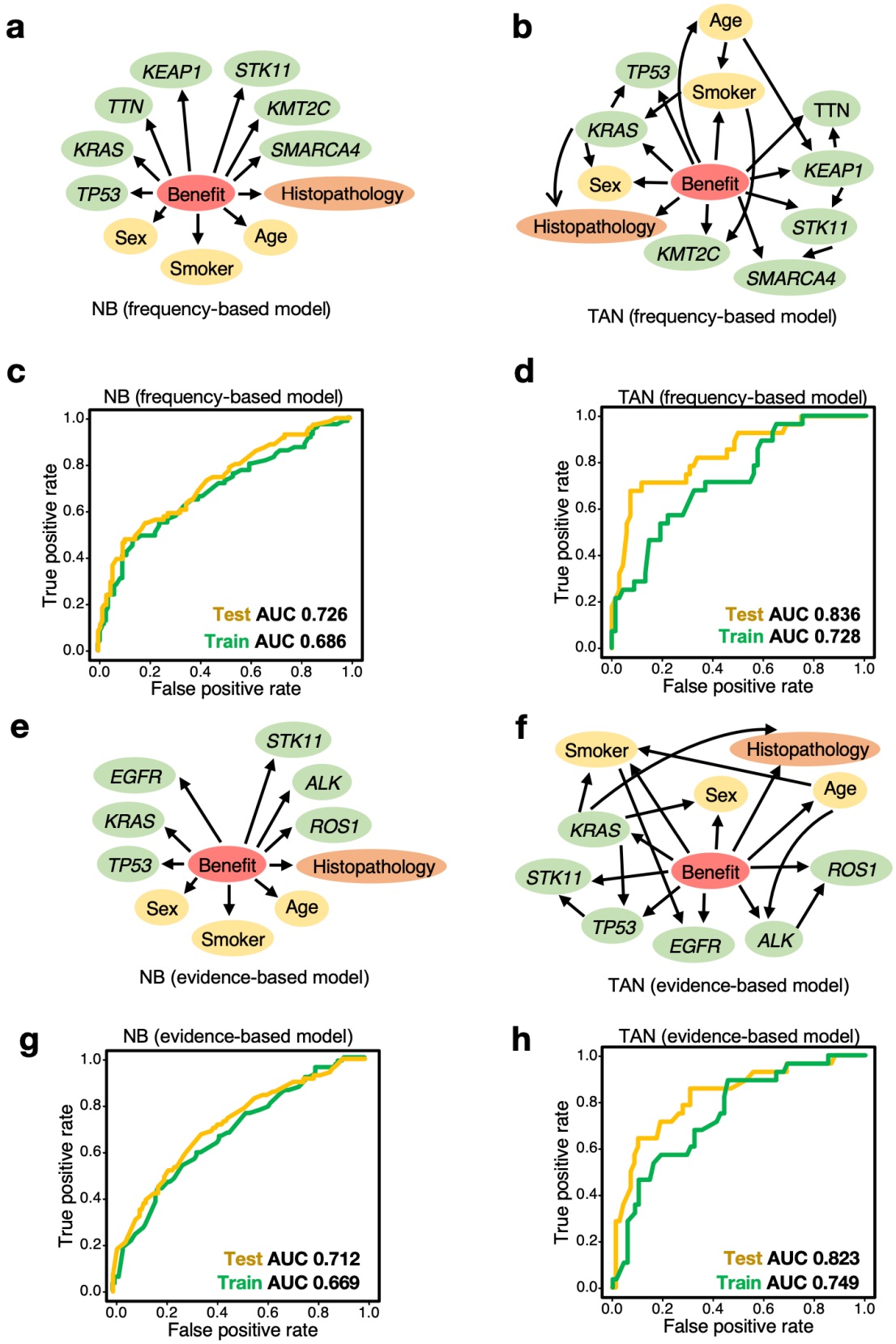
604

605 **Figure 1. Workflow of the study**

606 (**a**) We obtained clinical and genetic data of non-small cell lung cancer (NSCLC) patients from
607 cBioPortal (http://www.cbioportal.org). There were 315 data samples, of which 25 samples
608 were excluded because they had insufficient histopathology data, or because the disease
609 were rare. Two-thirds of the data was used to construct the models (train data), and the rest
610 was used for evaluation (test data).

611 (**b**) We developed the naïve Bayes (NB) and Tree-Augmented naïve Bayes (TAN) models and
612 evaluated their predictive accuracy for whether patients will benefit from immunotherapy. We
613 performed survival analyses to compare the two groups based on the binary predictions by
614 TAN model.

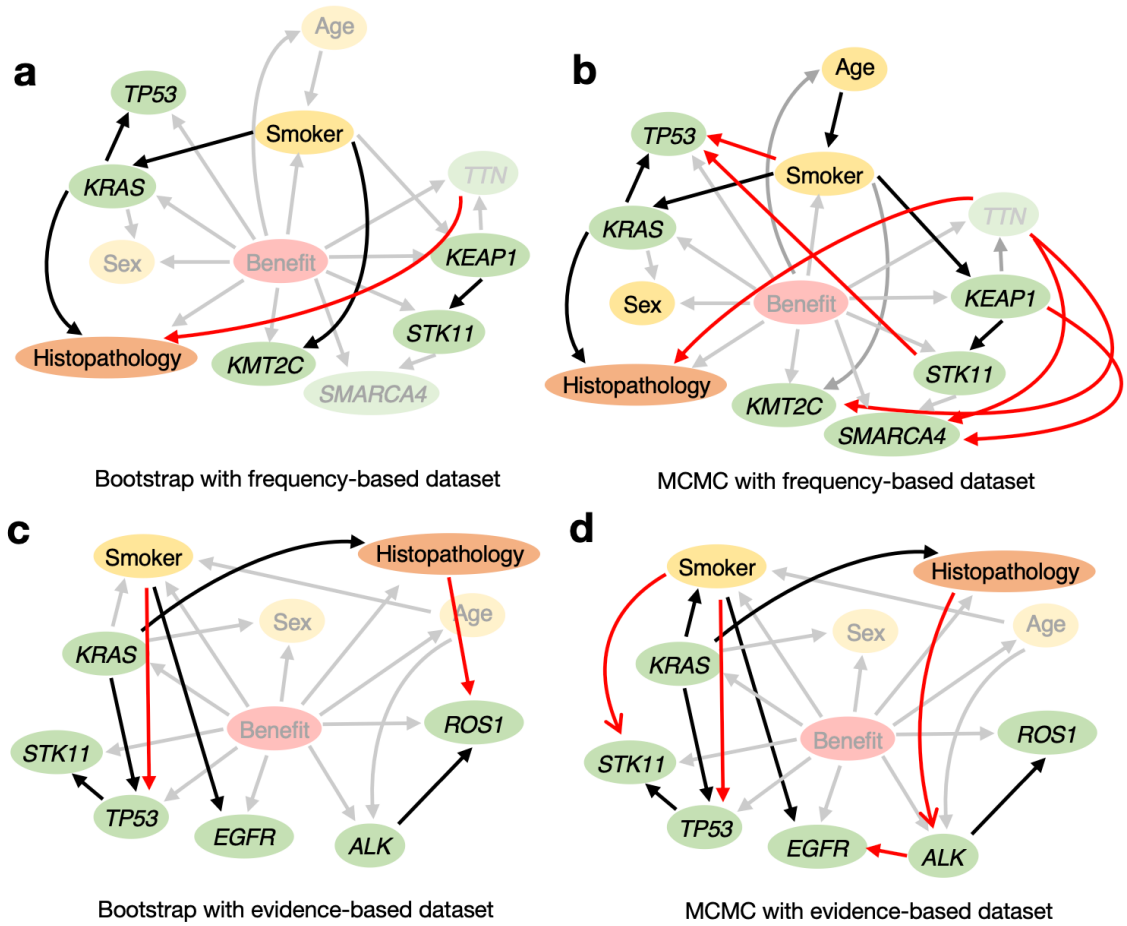
615



Hozumi et al., Figure 2

616

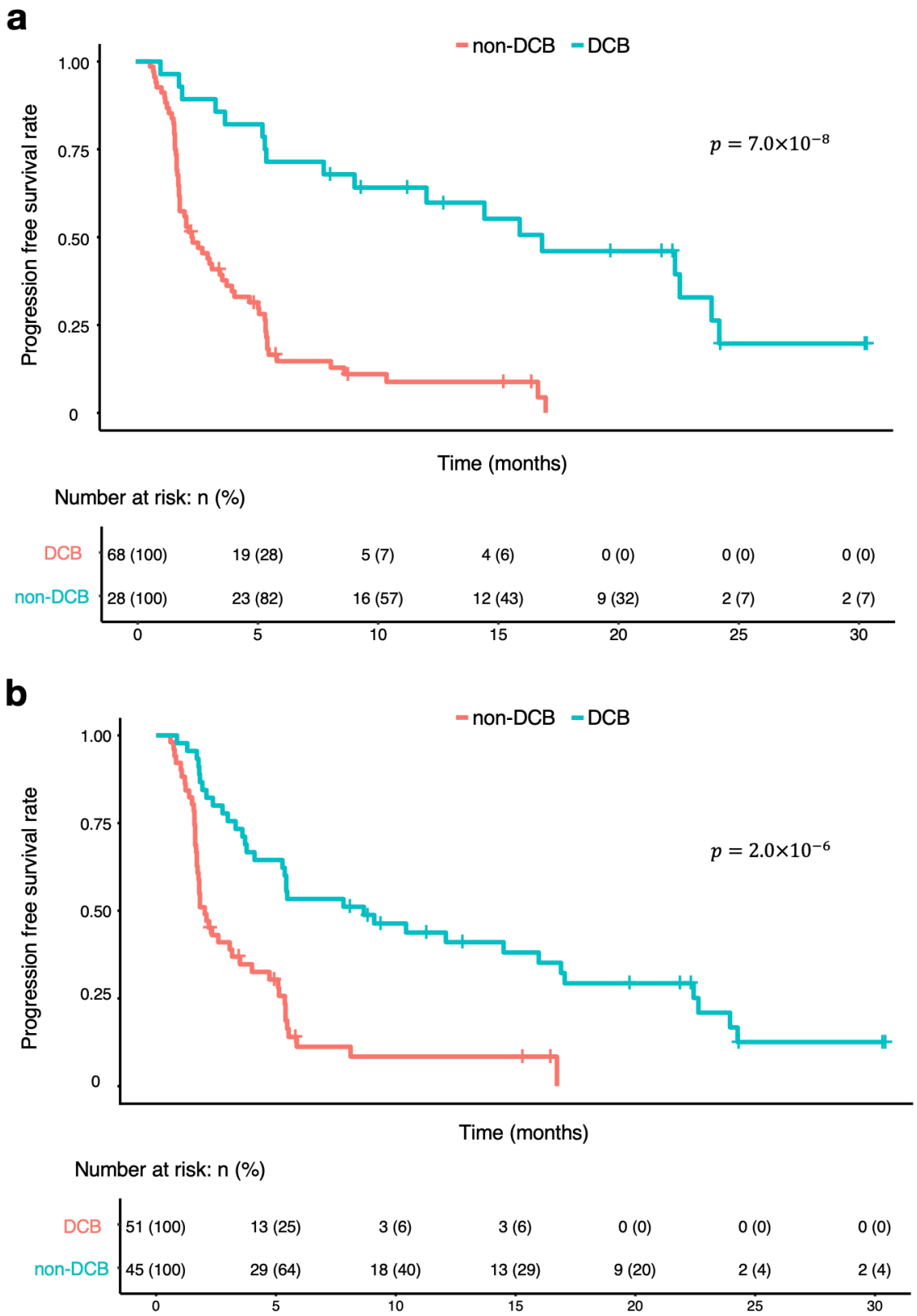
Figure 2. Bayesian network predicted the benefit of immunotherapy with high accuracy. (**a, b**) The naïve Bayes (NB, **a**) and Tree-Augmented naïve Bayes (TAN, **b**) models were trained using the frequency-based dataset. The predictor variable "Benefit" (DCB, shown in red) is defined in the RECIST guideline (version 1.1)[28]. Explanatory variables include patient data (yellow), tumor tissue information (orange), and the frequency-gene dataset (green). (**c, d**) Predictive performance of the frequency-based models (**a, b**) for the test dataset. TAN achieved greater accuracy than NB in terms of the Area Under the Curve (AUC), and was comparable to, or even more accurate than, state-of-the art methods[15,49,50]. (**e, f**) The naïve Bayes (NB, **e**) and Tree-Augmented naïve Bayes (TAN, **f**) models were trained using the evidence-based dataset. (**g, h**) Predictive performance of the evidence-based models (**e, f**) in the test dataset. TAN was more accurate than NB in terms of AUC, and comparable to, or even more accurate than, state-of-the-art methods[15,49,50].



Hozumi et al., Figure 3

Figure 3. Evaluating the validity of the structure estimated by TAN through model-averaging methods

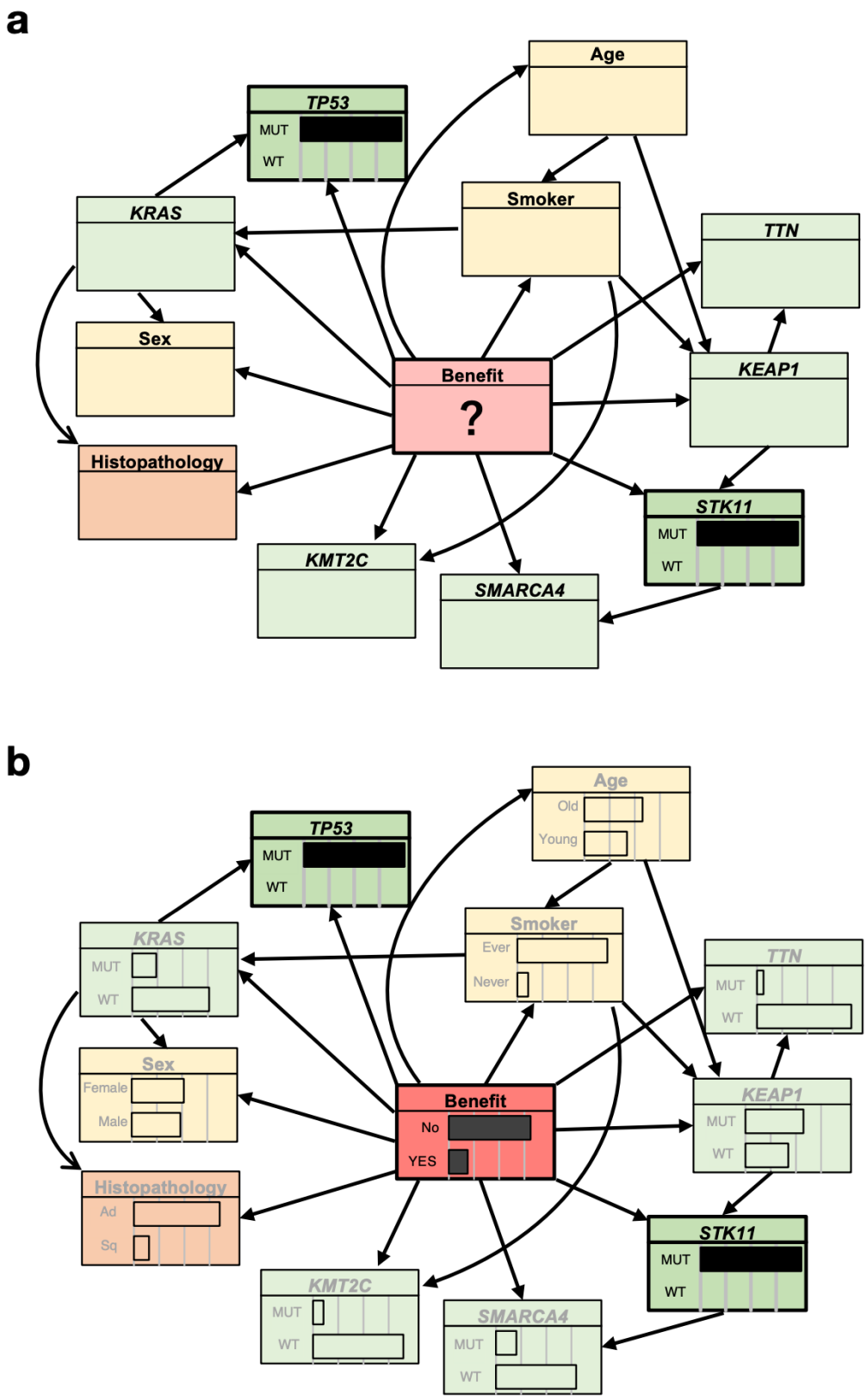
The linkages between nodes estimated by TAN were validated using model-averaging methods (bootstrap and MCMC). (**a, b**) Bootstrap (**a**) and MCMC (**b**) sampling was used to create models using frequency-based datasets; connections considered to be significant in each process are illustrated. Relationships detected by the model-averaging methods but not by TAN are shown in red. Connections detected both by model-averaging methods and TAN are in black. (**c, d**) Bootstrap (**c**) and MCMC (**d**) sampling was used to create models using evidence-based datasets; connections considered to be significant in each process are illustrated. The dependencies among variables estimated by TAN included many of the connections detected by model-averaging methods, indicating the robustness of our models. See also Tables 2 and 3.



Hozumi et al., Figure 4

Figure 4. Our TAN-based interpretable models stratify NSCLC patient prognosis

(**a, b**) We tested whether these TAN models are suitable for stratifying progression-free survival. We classified patients into two groups (durable clinical benefit "DCB", and "non-DCB") based on the binary predictions of the frequency-based (**a**) or evidence-based (**b**) models, and estimated progression-free survival status of the patients in our dataset via the Kaplan-Meier method. The *p*-values shown in this figure are from log-rank tests.



Hozumi et al., Figure 5

652

653 **Figure 5. The optimized TAN model can infer DCB even from limited data**
654 (**a, b**) We investigated whether our model (frequency-based TAN, for instance) could infer the
655 durable clinical benefit (DCB) from limited information. (**a**) In this example, the only information
656 provided to the model was the *TP53* and *STK11* mutations in the patient. (**b**) Using rejection
657 sampling, and approximate inference of the probability distribution of the unknown
658 information, we were able to obtain probabilities for all of the hidden states. From only the
659 information that *TP53* and *STK11* are co-mutated, the model computed a response
660 probability to be 0.163, suggesting that immunotherapy would not be effective for this patient,
661 consistent with the previous reports[33]. The white boxes and DCB status were calculated using
662 our approach.