BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Discovery of genes positively modulating treatment effect using potential outcome framework and Bayesian update

Young Keun Lee[1], Jisoo Kim[2] and Sung Wook Seo[1,2*]

Abstract

Background: The recent explosion of cancer genomics provides extensive information about mutations and gene expression changes in cancer. However, most of the identified gene mutations are not clinically utilized. It remains uncertain whether the presence of a certain genetic alteration will affect treatment response. Conventional statistics have limitations for causal inferences and are hard to gain sufficient power in genomic datasets. Here, we developed and evaluated a C-search algorithm for searching the causal genes that maximize the effect of the treatment.

Methods: The algorithm was developed based on the potential outcome framework and Bayesian posterior update. The precision of the algorithm was validated using a simulation dataset. The algorithm was implemented to a cBioPortal dataset. The genes discovered by the algorithm were externally validated within CancerSCAN screening data from Samsung Medical Center.

Results: Simulation data analysis showed that the C-search algorithm was able to identify nine causal genes out of ten. The C-search algorithm shows the discovery rate rapidly increasing until the 1500 data instances. Meanwhile, the log-rank test shows a slower increase in performance. The C-search algorithm was able to suggest nine causal genes from the cBioPortal Metabric dataset. Treating the patients with the causal genes is associated with better survival outcome in both the cBioPortal dataset and the CancerSCAN dataset which is used for external validation.

Conclusions: Our C-search algorithm demonstrated better performance to identify causal effects of the genes than multiple log-rank test analysis especially within a limited number of data. The result suggests that the C-search can discover the causal genes from various genetic datasets, where the number of samples is limited compared to the number of variables.

Keywords: Causal inference, Potential outcome framework, Genomics, Treatment modulators, Bayesian

Introduction

Identifying genomic sequences and analyzing data is a major focus in cancer studies [1]. An understanding of the causal relationship between therapeutic effect and genomic variances among tumors will allow individualized treatment and reduce unnecessary treatment.

There are open-access, open data sources, such as the cBioPortal for Cancer Genomics. Although a large amount of data is readily accessible to researchers, most of the identified gene mutations are not clinically utilized [2].

Conventional statistical analysis of genetic data consists of a series of single-statistic tests. The cumulative probability of false positives increases as the number of genes increases. To deal with multiple-testing problems, the false discovery rate (FDR) is used, which is "expected

*Correspondence: sungwseo@gmail.com
[1] Department of Orthopedic Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

type I errors among the total number of rejected null hypotheses" [3]. Despite the approach, the dimension of the data significantly affects the statistical power of the test. In addition, conventional statistics only draw an association; therefore, distinguishing causal relationships from spurious associations is a challenge [4].

To draw causal effects from observational data, Rubin introduced a potential outcome framework [5, 6] Individual levels of treatment effect are derived from a comparison of two potential outcomes. however, observing the exposed and unexposed outcomes at the same time is impossible. One of the methods to overcome this fundamental problem of causal inference is to compute the potential outcome from samples matched with similar covariate profiles [4, 7, 8]. However, due to the curse of dimensionality in cancer genomics, enough sample size may not be available to match the exposed and unexposed within a genetic subset [9].

We developed an algorithm called C-search that can estimate potential outcomes using the similarity-weighted Monte Carlo method. We adopted Bayesian posterior update, which allows us to estimate the uncertainty of our decision boundary from small datasets without losing power [10]. This system was used to identify the causal genes that maximize the effect of the treatment. In this study, we compared the performance of causal gene discovery between the conventional statistical method and our C-search using a simulation dataset and an open-source gene dataset.

Materials and methods

Pseudo-counterfactual assumption and similarity weighted Monte Carlo

Assume that individual i with variable X_i is treated with the treatment variable T_i. T_i is a binary variable ($T_i = 1$, if treated; $T_i = 0$, if not). There are two potential outcomes: $Y(X_i, T_i = 1)$, and $Y(X_i, T_i = 0)$. The causal effect of the treatment can be drawn from the comparison between both [6].

Here, we define $f(X_i)$ as the outcome of an individual i with variable X_i when treated, and $g(X_i)$ as the oucome of the individual when not treated. Then, individual treatment effect (ITE) of the individual i, $ITE(X_i)$ can be written as follows:

$$ITE(X_i) = f(X_i) - g(X_i)$$

However, we can observe only one potential outcome at most [4, 8]. Therefore, we should infer the counterfactuals from an untreated data pool. We call them pseudo-counterfactuals because they are not identical to the factual.

Draw an individual j with variable X_j from the untreated data pool. Weight function $W(X_i, X_j)$ is defined as the probability of similarity $sim(X_i, X_j)$ between the factual and pseudo-counterfactual [11]. Using the similarity weighted Monte Carlo method [12], we could estimate $ITE(X_i)$ as follows:

$$\begin{aligned} &\sum_j W(X_i, X_j) = 1 \\ &ITE(X_i) \approx f(X_i) - \sum_j g(X_j) \cdot W(X_i, X_j) \end{aligned} \tag{1}$$

Measurement of the difference in survival outcome using Win probability (Pw_i)

Survival outcome Y includes survival time and survival events. Measuring individual differences in survival outcomes is difficult because they are right-censored. One of the most well-established outcome measures for survival difference is the bi-partite ranking system, such as the Wilcoxon – Mann – Whitney statistics [13]. Adopting this concept, we assumed the comparison in outcome between two individuals i and j as a Bernoulli trial. If i lives longer than j, i will win a score. ITE for the survival outcome can be defined as the win probability (Pw_i : the chance that the treated individual i wins over its untreated counterpart), which follows a binomial distribution.

The beta distribution is a conjugate prior for the binomial distribution. If we consider the comparison between two individuals i and j as a simple Bernoulli trial, the posterior distribution after observing the score s (or observing s times of winning of i) after N_j trials can be defined as follows:

$$\begin{aligned} &Pw_i \sim p(\text{“}win''|X_i) \\ &Pw_i \propto p(X_i|\text{“}win'')p(\text{“}win'') \\ &p(\text{“}win'') \sim Beta(\alpha_0, \beta_0) \\ &Posterior\ Pw_i \sim Beta\left[\alpha_0 + s, \beta_0 + (N_j - s)\right] \end{aligned} \tag{2}$$

Update Win probability using similarity weighted Monte Carlo

In ideal settings where all individuals $j\left(j \in \{1, 2, \ldots N_j\}\right)$ in the counterpart group that are identical to the individual i, the outcome of j is a good estimator for the counterfactual outcome of i. However, an identical condition is impossible in the observation setting. We use the similarity weighted Monte Carlo method to update Pw_i.

We matched individual i with the individuals in the counterpart data pool and updated the score(s) with the similarity weight $W(i,j)$ calculated from the similarity $sim(i,j)$ between i and j. The $sim(i,j)$ can be the Euclidean distance in the original data space [14]. Other methods use a transformed one-dimensional score, that is, a regression function, such as a propensity function [15].

Here, we defined a basal function—a regression function that approximates the survival state. The covariates of individual i and matched controls j are projected onto the space through the basal functions $P_i(Y|X_i)$ and $P_j(Y|X_j)$. To incorporate the difference onto the similarity weight, we used the Boltzmann probability distribution:

$$\begin{aligned} sim(i,j) &= e^{-|P_i-P_j|/k\tau} \\ W(i,j) &= \frac{e^{-|P_i-P_j|/k\tau}}{\sum_j^N e^{-|P_i-P_j|/k\tau}} \end{aligned} \tag{3}$$

k is a constant and τ is the annealing temperature. These hyperparameters represent degrees of freedom.

Let s_j be the score from a single comparison between i and j. The posterior distribution after observing a single comparison between i and j can be written as follows:

$$\begin{aligned} &\begin{cases} s_j = 1, \ if \ Y_i(T=1) > Y_j(T=0) \\ s_j = 0, \ if \ Y_i(T=1) < Y_j(T=0) \end{cases} \\ &Posterior \ Pw_i \sim Beta\left[\alpha + s_j \cdot W(i,j), \beta + (1-s_j) \cdot W(i,j)\right] \end{aligned} \tag{4}$$

The Pw_i is estimated from the posterior distribution after observing N_j the number of counterpart individuals.

$$\begin{aligned} Posterior Pw_i \sim Beta\Bigg[&\alpha + \sum_{j=1}^{N_j} s_j \cdot W(i,j), \beta \\ &+ \sum_{j=1}^{N_j} (1-s_j) \cdot W(i,j)\Bigg] \end{aligned} \tag{5}$$

We defined the observed clinical covariates as V_i. Genetic alteration is represented by simple binary values (e.g., for each genetic profile $g_1, g_2, g_3, \ldots, g_{N_g}$, existing alteration is given a value of 1; if not, it is given a value of 0). The individual i has the genetic variable G_i that consists of a set of genetic profile.

$$\begin{aligned} &G_i = \left[g_1, g_2, g_3, \ldots, g_{N_g}\right]_i \\ &g_1, g_2, g_3, \ldots, g_{N_g} \in \{0,1\} \end{aligned}$$

Individual	Clinical covariates	Genetic covariates	T	$Y(T=0)$	$Y(T=1)$
i	V_i	G_i	T_i	$Y_i(T=0)$	$Y_i(T=1)$

To estimate Pw_i, we may use the similarity weight calculated from the basal function using clinical covariates and/or genetic covariates. We denote the weight of the basal function of clinical covariates as $W^V(i,j)$ and the weight using the basal function of the genetic covariates as $W^G(i,j)$. Using Eq. 3, $W^V(i,j)$ and $W^G(i,j)$ can be written as follows:

$$\begin{aligned} sim^V(i,j) &= e^{-|P_i(Y|V_i)-P_j(Y|V_j)|/k\tau} \\ W^V(i,j) &= \frac{sim^V(i,j)}{\sum_j^{N_j} sim^V(i,j)} \\ sim^G(i,j) &= e^{-|P_i(Y|G_i)-P_j(Y|G_j)|/k\tau} \\ W^G(i,j) &= \frac{sim^G(i,j)}{\sum_j^{N_j} sim^G(i,j)} \end{aligned} \tag{6}$$

The win probabilities of individual i using both weights are as follows:

$$\begin{aligned} Posterior \ Pw_i \sim Beta\Bigg[&\alpha + \sum_{j=1}^{N_j} s_j \cdot W^V(i,j) \cdot W^G(i,j), \\ &\beta + \sum_{j=1}^{N_j} (1-s_j) \cdot W^V(i,j) \cdot W^G(i,j)\Bigg] \end{aligned} \tag{7}$$

Causal gene suggestion

To find a single gene (g_k) effect on the treatment effect, we assumed that each gene has an independent win probability $P(\text{“}win''|g_k)$.

$$\begin{aligned} &P(\text{“}win''|g_k) \propto P(g_k|\text{“}win'') \cdot P(\text{“}win'') \\ &k \in \{1,2,3,\ldots,n\}, n \in \mathbb{N} \end{aligned} \tag{8}$$

We used the similarity weighted Monte Carlo method to estimate $P(\text{“}win''|X_i, g_k = 1)$ or individual i's win probability $Pw_i(g_k = 1)$(Eq. 7). Observing individual i's win probability $Pw_i(g_k = 1)$ updates the prior distribution of $P(\text{“}win'') = Beta(\alpha_0, \beta_0)$ and the posterior is as follows:

$$\begin{aligned} &Posterior P(\text{“}win''|g_k) \sim Beta\left[\alpha_0 + Pw_i(g_k = 1), \beta_0\right. \\ &\left. + (1 - Pw_i(g_k = 1))\right] \end{aligned} \tag{9}$$

Sampling N_k-number of individuals with $g_k = 1$, the posterior can be as follows:

$$\begin{aligned} &Posterior\ P(\text{“}win''|g_k) \\ &\sim Beta\left[\alpha_0 + \sum_{i=1}^{N_k} Pw_i(g_k = 1), \right. \\ &\left. \beta_0 + \left(N_k - \sum_{i=1}^{N_k} Pw_i(g_k = 1)\right)\right] \end{aligned} \tag{10}$$

In reality, $Pw_i(g_k = 1)$ is not stationary because Pw_i depends on individual variables V_i, G_i. To identify the marginal treatment effect, we need to balance all confounding variables using inverse propensity score weighting (IPW) as follows.

$$\begin{aligned} &Posterior\ P(\text{“}win''|g_k) \\ &\sim Beta\left[\alpha_0 + \sum_{i=1}^{N} \frac{Pw_i(g_k = 1)}{\hat{e}(V_i)\hat{e}(G_i)}, \right. \\ &\left. \beta_0 + \left(\sum_{i=1}^{N} \frac{1 - Pw_i(g_k = 1)}{(1 - \hat{e}(V_i))(1 - \hat{e}(G_i))}\right)\right] \\ &\hat{e}(V_i) = P(T|V_i), \hat{e}(G_i) = P(T|G_i) \end{aligned} \tag{11}$$

For the treatment decision, we need to know whether treating a patient with genetic alteration (g_k) is beneficial with sufficient evidence. We can estimate the posterior distribution of Pw without g_k in the same manner. If g_k has a significant benefit for the treatment, the upper bound of the 95% confidence interval of $P(\text{“}win''|g_k = 0)$ should be lower than the lower bound of $P(\text{“}win''|g_k = 1)$.

Results

Simulation data analysis

The C-search algorithm was validated with the simulation data (Additional file 1), where the 10 causal genes that positively modulated the treatment outcome were hidden among 300 genes. We compared the precision of the C-search algorithm with that of the conventional log-rank survival analysis. Both the C-search algorithm and conventional statistics suggested 10 possible causal genes. The precision of the algorithm was determined by the number of true causal genes among the suggested genes. The performance of the algorithm was evaluated using five-fold cross-validation. To balance the covariate profile, we used propensity score matching. To perform the multiple comparison tests, we set the FDR to 0.05 and controlled it with the Benjamini – Hochberg procedure [16, 17].

The power of log-rank survival analysis depends on the number of data [18]. To demonstrate whether the algorithms are dependent on the number of data, 100, 200, 500, 1000, 1500, 2000, 2500, 3000, 3500, and 4000 simulation data were used for the analysis. As the number of data increases, both algorithms show better causal gene discovery. The C-search algorithm continuously improved its discovery performance until 1500 samples were obtained and then plateaued. Conventional statistics showed a linear improvement according to the number of samples. It required at least 4000 samples to show performance comparable to that of the C-search algorithm (Fig. 1).

Finding positive modulators from open-source data

We also analyzed the cBioPortal Metabric breast cancer data using the C-search algorithm to identify causal genes that are associated with improved outcomes of chemotherapy. The dataset includes gene mutation profiles of 173 genes and clinical data from 2433 patients with primary breast cancer [19]. Clinical data consisted of age, chemotherapy, radiation therapy, sex, survival lifetime, and survival events. Among the 2433 records, 964 records that have missing on clinical data were omitted. A total of 1,469 patients were included in the analysis (Additional file 1).

The C-search suggested nine positive modulators: *PRKCZ, CLK3, CDKN2A, BRAF, KRAS, CASP8, JAK1, PRKACG, and SIK2*. We allocated all patients who had any of them to the causal gene group and those who had not to the other gene group. If they are the true causal genes, treated patients should show a statistically significantly better prognosis compared to the untreated patients in the causal gene group. In addition, among the

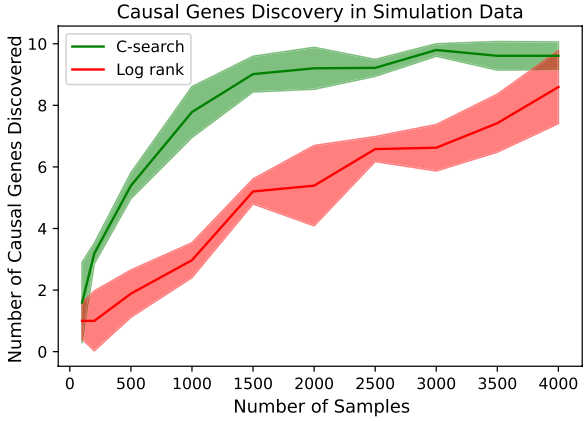

Fig. 1 The number of causal genes discovered by C-search and conventional statistics. The X-axis is the number of samples consisting of the simulation data. The Y-axis is the number of true causal genes among the 10 suggested causal genes that the algorithm discovered. The C-search algorithm shows the discovery rate rapidly increasing until the 1500 data instances. The log-rank test shows a slower increase in performance

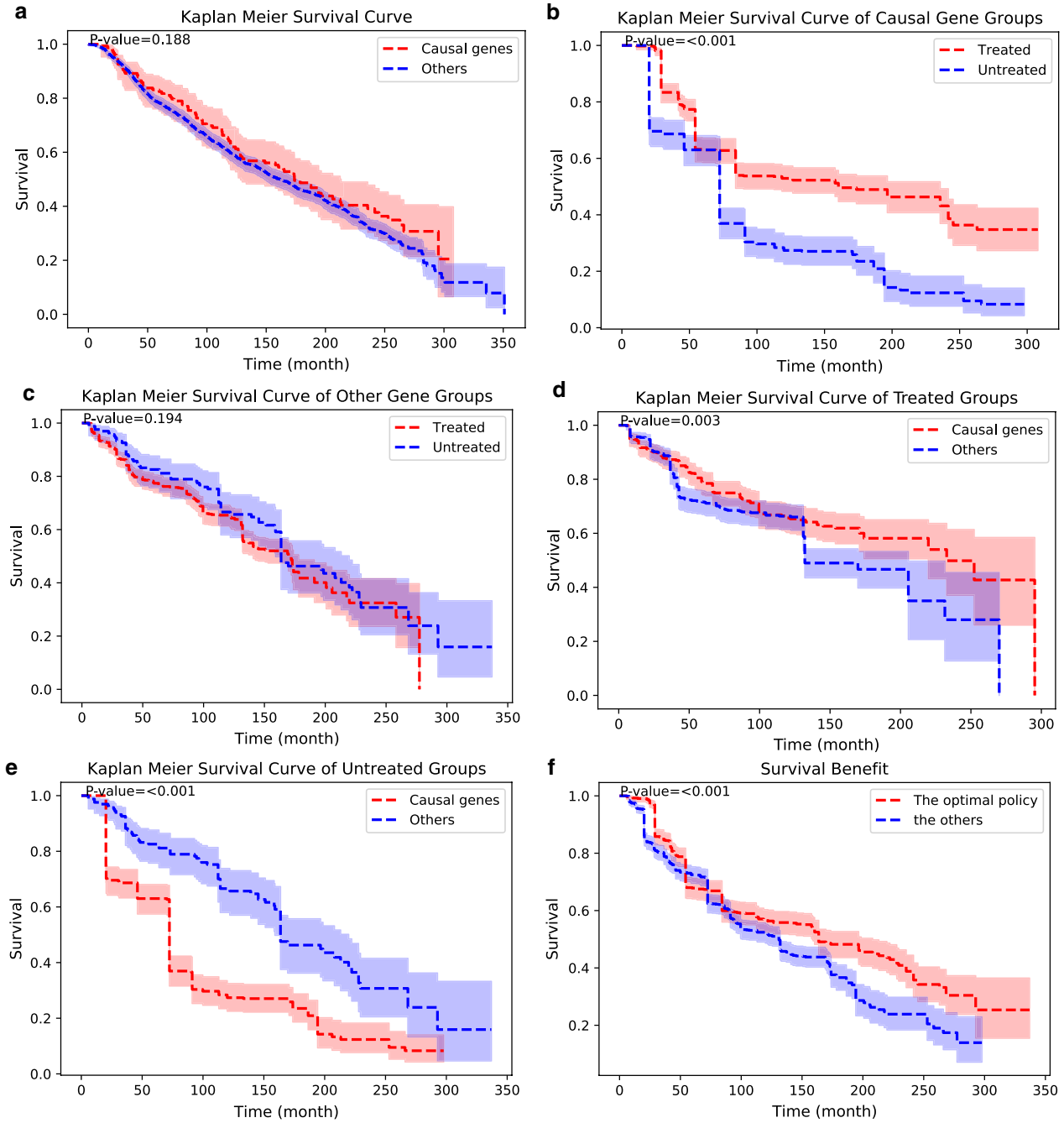

Fig. 2 Discovery of positive modulator genes by C-search in the cBioPortal breast cancer dataset. Nine causal genes are discovered, and patients with causal genes are assigned to the causal gene group. Patients without casual genes are assigned to the other gene group. All Kaplan — Meier survival curves are adjusted with propensity score matching [39]; 95% confidence intervals are depicted, and p-values are noted. **a** Kaplan — Meier survival curves of the causal gene group and the other gene group. **b** Treated and untreated patients are compared in the causal gene group. **c** Treated and untreated patients are compared for the other gene group. **d** The causal gene and other gene group are compared between treated patients. **e** The causal gene and other gene group are compared between the untreated patients. **f** Survival curve following the optimal policy and the other policy is shown

patients who were treated, the causal gene group should show a better prognosis than the other gene group.

There were no overall survival differences between the causal gene group and the other gene group (Fig. 2a). Among patients with causal genes, the treatment group showed a better prognosis than the untreated group (Fig. 2b). Meanwhile, in the other gene group, there were no statistically significant differences between the treated and untreated groups (Fig. 2c). Among the treated patients, the causal gene group showed significantly better survival than the other gene group (Fig. 2d). In the untreated group, the other gene group showed better survival outcomes compared with the causal gene group (Fig. 2e). We define the optimal policy as treating patients with causal genes and not treating patients without causal genes. The other policy is to treat patients in the other gene group and not to treat patients in the causal gene group. The Kaplan – Meier survival curve of the optimal policy showed better survival outcomes than the other policies (Fig. 2f).

Conventional log-rank analysis of open-source data

To discover causal genes using conventional statistics from the cBioPortal dataset, we performed log-rank survival analysis. As the dataset includes gene mutation profiles of 173 genes, multiple log-rank survival analyses were applied for each mutation profile. For each gene mutation, the patients treated were divided into two groups: one group consisted of patients who harbored the mutation and one group of patients without the mutation. Survival outcomes were compared between the two groups. Propensity score matching was used to balance the covariate profile. We set the FDR to 0.01 and controlled it using the Benjamini – Hochberg procedure [16, 17].

A total of 10 genes were shown to be correlated with positive treatment outcomes: *EGFR, CLK3, PTEN, CDH1, GATA3, KRAS, RB1, PRKACG, NEK1*, and *NRAS*. Patients who have any of these are allocated to the causal gene group, and those who do not are assigned to other gene group. Kaplan – Meier survival curves comparing the overall survival differences between both groups are depicted in Fig. 3a. No survival differences were observed between patients with causal genes and without causal genes. In both the causal gene group and the other gene group, treatment was not associated with positive outcomes compared with no treatment (Fig. 3b, c). Meanwhile, in the treated group, the causal gene group showed better survival than the other gene group (Fig. 3d). Among the untreated patients, the causal gene group and the other gene group had no statistically significant differences in survival (Fig. 3e). Following optimal policy demonstrated better survival than following the other policy (Fig. 3f).

Comparison between C-search and conventional log-rank analysis

We compared the Kaplan – Meier survival curve of the following optimal policy, which was determined by the genes that each algorithm discovered. The survival outcome following the C-search policy showed statistically significant better survival outcomes (Fig. 4).

Validation with external data

To determine whether the genes found by the algorithm will act as causal genes in the other dataset, we used CancerSCAN screening data from Samsung Medical Center. CancerSCAN is a custom panel developed by the Samsung Genomic Institute [20]. The usage of the data was approved by the institutional review boards of the participating institutions (Samsung Medical Center 2019-11-127).

CancerSCAN data consists of 559 breast cancer samples obtained at the Samsung Medical Center from January 2014 to September 2016. Mutation profiles of 81 genes and clinical data on age, chemotherapy, radiation therapy, sex, survival time, and survival events were included (Additional file 1). Among the causal genes suggested by C-search, mutation profiles of *BRAF, KRAS, CDKN2A*, and *JAK1* were found in the CancerSCAN dataset. We assigned patients to the C-search causal gene group who acquired at least one of the mutations. The *CDH1, EGFR, KRAS, PTEN*, and *RB1* are genes suggested by log-rank analysis whose mutation profiles exist in the CancerSCAN dataset. Patients with these mutations are assigned to the conventional statistics causal gene group.

Figure 5a shows the significant survival difference between the treated and the untreated in the C-search causal gene group. The optimal policy suggested by the C-search showed a significantly better survival outcome (Fig. 5b). Meanwhile, among conventional statistical causal gene group, the treated did not demonstrate statistically better survival compared to the untreated (Fig. 5c). The survival outcomes following the optimal policy suggested by conventional statistics and those of the other policies are not statistically different (Fig. 5d).

Discussion

Biotechnological breakthroughs in gene profiling have led to an increased focus on individualized precision therapy [21]. Clinicians want to treat patients who will benefit the most from the therapy while avoiding treatment that will not benefit or even get harm from therapy.

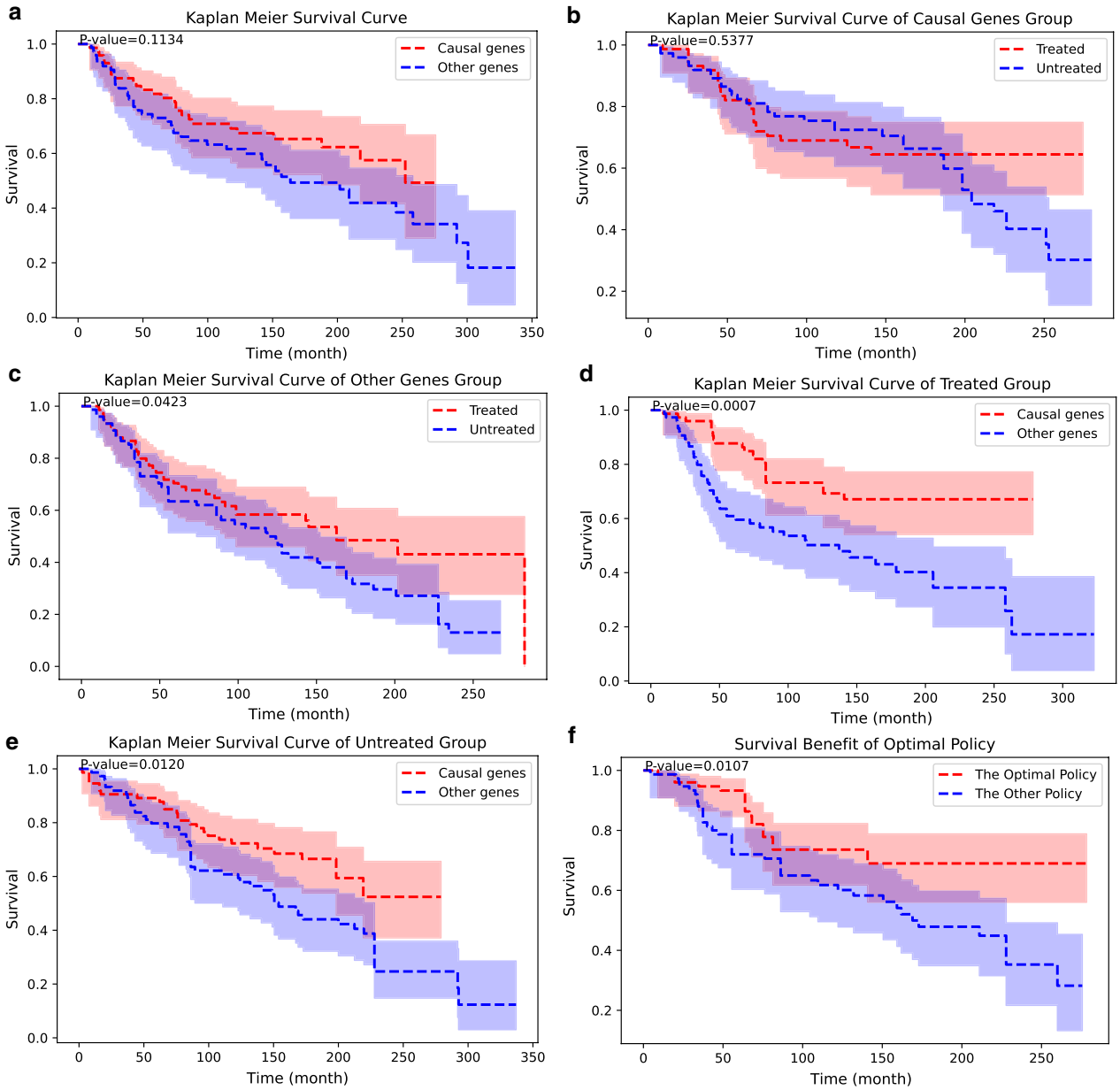

Fig. 3 Discovery of positive modulator genes using conventional log-rank analysis in the cBioPortal breast cancer dataset. Ten causal genes are discovered, and patients with causal genes are assigned to the causal gene group. Patients without casual genes are assigned to the other gene group. All Kaplan — Meier survival curves are adjusted with propensity score matching [39]; 95% confidence intervals are depicted, and p-values are noted. **a** Kaplan — Meier survival curves of the causal gene group and the other gene group. **b** Treated and untreated patients are compared in the causal gene group. **c** Treated and untreated patients are compared for the other gene group. **d** The causal gene and other gene group are compared between the treated patients. **e** The causal gene group and the other gene group are compared between the untreated patients. **f** Survival curve following the optimal policy and the other policy is shown

There are studies on genetic assays to predict prognosis or response to treatment for breast cancer [22]. BluePrint molecular subtyping profile uses 80 genes to determine the sensitivity to adjuvant treatment [23]. Prosigna Breast Cancer Prognostic Gene Signature Assay utilizes the PAM50 test, which identifies gene signatures specific to breast cancer subtypes (luminal A/B, HRE2, basal-like) [24]. Both studies used genomic profiles to infer the molecular subtypes of cancer and drew a correlation between the subtypes and the prognosis or response to treatment. However, since the correlation does not infer causation, we cannot conclude

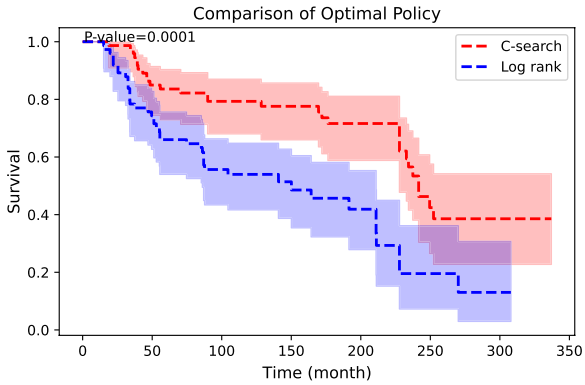

Fig. 4 Survival differences between the optimal policy determined by C-search and conventional log-rank analysis are shown. The C-search's policy shows better outcomes than the others. The Kaplan — Meier survival curve is adjusted with propensity score matching; 95% confidence intervals are depicted, and *p*-values are noted

that the specific genomic profile results in a positive response to treatment. In contrast, our C-search algorithm uses a potential outcome framework to study the causal effect of each gene on treatment.

The estimation of causal effects from observational studies can be done in a number of ways. Methods based on propensity scores match, stratify, and/or inversely weight covariates that affect treatment allocation [25, 26]. *G*-computation implements regression models [27]. Mendelian randomization uses germline gene mutations as instruments to make causal inference [28, 29]. These methods are used to estimate average treatment effect of the target population. To estimate and compare the treatment effect of the patient who have a specific genetic mutation, one must calculate conditional causal effect, which is the average treatment effect of a subgroup of patients with that mutation [30]. However, due to the curse of dimensionality in cancer genomics, some

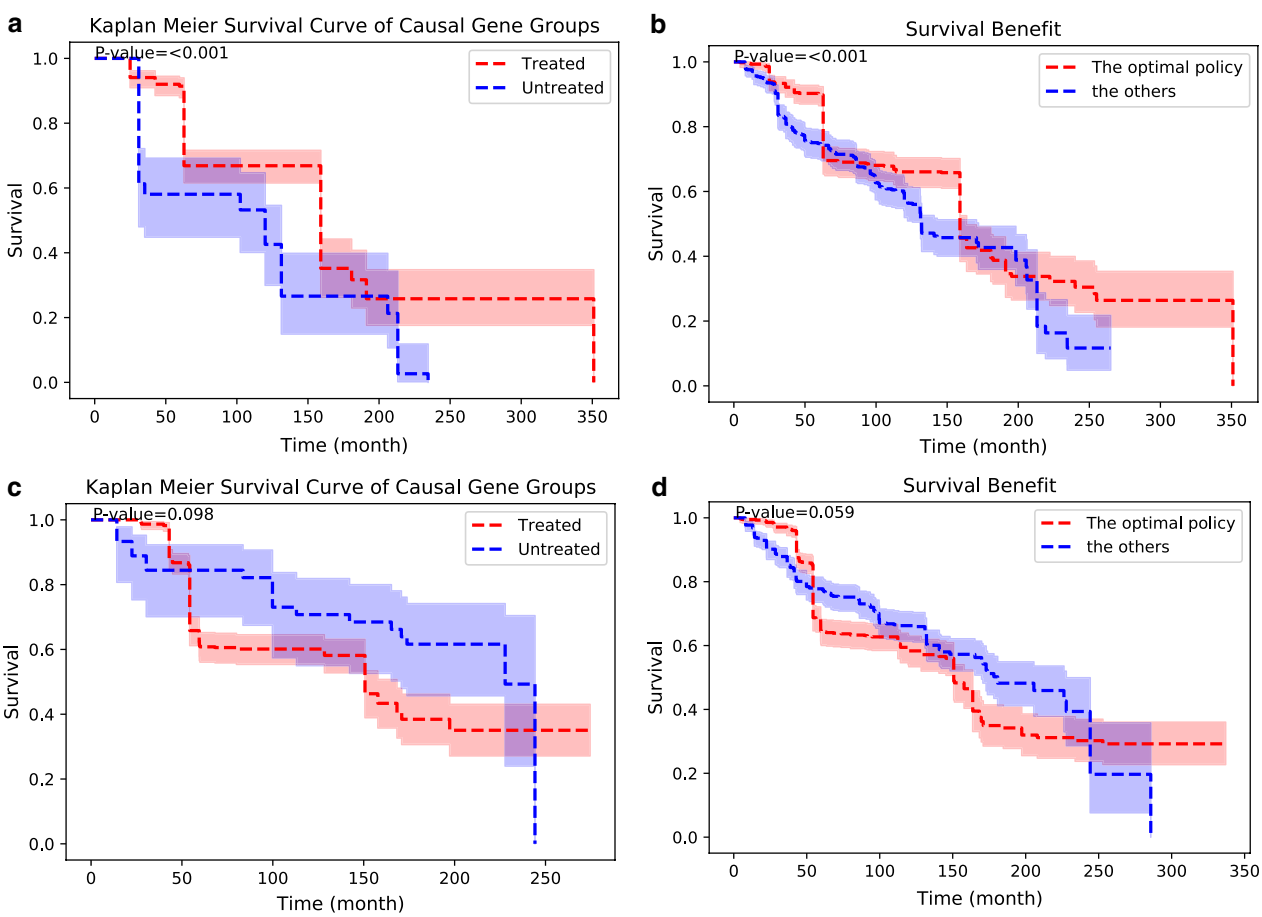

Fig. 5 External validation of the causal genes suggested by C-search and conventional log-rank analysis. **a** Kaplan — Meier survival curve of the treated and the untreated among the C-search causal gene group. **b** Kaplan — Meier survival curve following and not following C-search optimal policy. **c** Kaplan — Meier survival curve of the treated and the untreated among the conventional statistics causal gene group. **d** Kaplan — Meier survival curve following and not following conventional statistics optimal policy

subgroups may lack sufficient sample size to draw meaningful conclusions [31]. In addition, when the algorithm handles a smaller dataset, the result drawn from the inference has a considerable amount of uncertainty. It is important to incorporate uncertainty into the prediction of the algorithm [32, 33]. The C-search algorithm updates the gene's win probability with a Bayesian update, thus, reflecting its uncertainty in the analysis. Pseudocode for the algorithm and the computational complexity are shown in Additional file 1.

We demonstrated that the C-search algorithm can identify causal genes from a simulation dataset that includes hidden confounders. Compared to conventional log-rank analysis, C-search requires fewer data to gain sufficient power to find the causal genes. Therefore, the C-search may find more candidate causal genes than conventional association studies using genomic data. When there are not enough data subsets, it is important as we used Bayesian update to estimate the gene's win probability.

Both C-search and log-rank analysis successfully found positive modulators in the Metabric dataset (Figs. 2, 3), yet the gene set found by C-search showed better results than log-rank analysis (Fig. 4). Among the positive regulators that C-search and conventional log-rank analysis found, two genes are found in common in both algorithms: *KRAS* and *PRKACG*. The *KRAS* gene targets several miRNAs to enhance chemotherapy in acute myeloid leukemia, lung cancer, breast cancer, and gallbladder cancer [34–36]. *PRKACG* is a gene that encodes the protein kinase A subunit Cγ, whose role in cancer has not been elucidated. The causal genes suggested by both algorithms contained relatively few overlapping genes. This may be due to intercorrelated genes, at least to some extent; therefore, one of the co-expressed genes may be selected for the set as a predictor and yield comparable results [37].

Our study has several limitations. There were only a few clinical variables available in the open-source dataset; therefore, hidden confounders may affect the performance of the algorithm. IPW may result in bias in this setting [15, 38]. However, the algorithm performance in the simulation data, including 10 hidden confounders, showed better results than log-rank analysis. As there are no golden standard datasets to evaluate the performance of the algorithm to determine the treatment modulating effect of genetic variables, we can only evaluate the algorithm's performance indirectly. Using the CancerSCAN dataset, we were able to externally validate that the causal genes suggested by C-search showed comparable results in the external dataset.

Conclusion

We proposed an algorithm that uses a potential outcome framework and Bayesian updating, inferring the causal effect of the genetic variable on treatment outcomes. The proposed algorithm was shown to find causal genes from the simulation data in a relatively small number of samples compared to the log-rank analysis. It also showed its performance in finding positive treatment modulators from the open-source breast cancer dataset, which is validated with external data. The C-search algorithm may be applied to various types of datasets where the number of samples is limited compared to the number of variables.

Abbreviations

FDR: False Discovery Rate; ITE: Individaul Treatment Effect; IPW: Inverse Propensity Score Weighting.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-022-01852-3.

Additional file 1. How to simulate data generation, demographics of cBioPortal dataset, demographics of CancerSCAN dataset, pseudocode of the C-search Algorithm, computational complexity

Acknowledgements

We would like to thank Editage (www.editage.co.kr) for English language editing.

Authors contributions

Y.L. and S.S. conceived the idea for this paper. Y.L. drafted the manuscript, which was reviewed and revised by S.S.. Y.L., J.K and S.S. build the program and implemented the analysis. All authors read and approved the final manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.2016R1E1A1A01941433, No.2021R1A2B5B02086271).

Availability of data and materials

The cBioPortal dataset analyzed during the current study is available from the cBioPortal repository, https://cbioportal-datahub.s3.amazonaws.com/brca_metabric.tar.gz. The Samsung Medical Center CancerSCAN dataset is not publicly available as it is not approved to be open in public, but it is available upon request from the corresponding author for academic use. The code for the algorithm is available upon request from the corresponding author for academic use. All implementation details are described in the Methods section so that they can be replicated with nonproprietary libraries.

Declarations

Ethics approval and consent to participate

The cbioportal dataset adopted in this research is a publicly available dataset that can be downloaded without restriction from the cBioPortal repository (https://www.cbioportal.org/study/summary?id=brca_metabric). It's impossible to re-identify the data because it's completely anonymized. As a result, the Samsung Medical Center Institutional Review Boards (IRB) ethical approval was waived, and informed consent to participate was not expected in this study. The usage of the CancerSCAN dataset was approved by the institutional review boards of the participating institutions (Samsung Medical Center 2019-11-127).

Consent for publication

Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Orthopedic Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea. [2]Institute of Biomedical AI, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University School of Medicine, Seoul, Korea.

Received: 11 November 2021 Accepted: 11 April 2022
Published online: 27 April 2022

References

1. Carrasco-Ramiro F, Peiró-Pastor R, Aguado B. Human genomics projects and precision medicine. Gene Ther. 2017;24(9):551–61. https://doi.org/10.1038/gt.2017.77.
2. Chia S. Clinical application and utility of genomic assays in early-stage breast cancer: key lessons learned to date. Curr Oncol. 2018;25:S125–30. https://doi.org/10.3747/co.25.3814.
3. Noble WS. How does multiple testing correction work? Nat Biotechnol. 2009;27(12):1135–7. https://doi.org/10.1038/nbt1209-1135.
4. Rubin DB. Bayesian inference for causal effects: the role of randomization. Ann Stat. 1978;6(1):34–58. https://doi.org/10.1214/aos/1176344064.
5. G.W. Imbens and D.B. Rubin, Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge: Cambridge University Press, 2015. doi:https://doi.org/10.1017/CBO9781139025751.
6. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701. https://doi.org/10.1037/h0037350.
7. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci. 2010;25(1):1–21. https://doi.org/10.1214/09-STS313.
8. Holland PW. Statistics and causal inference. J Am Stat Assoc. 1986;81(396):945–60. https://doi.org/10.2307/2289064.
9. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. Stat Med. 2002;21(19):2917–30. https://doi.org/10.1002/sim.1296.
10. van de Schoot R, Kaplan D, Denissen J, Asendorpf JB, Neyer FJ, van Aken MAG. A gentle introduction to bayesian analysis: applications to developmental research. Child Dev. 2014;85(3):842–60. https://doi.org/10.1111/cdev.12169.
11. Zadeh LA. A note on similarity-based definitions of possibility and probability. Inf Sci. 2014;267:334–6. https://doi.org/10.1016/j.ins.2014.01.046.
12. M.H. Kalos and P.A. Whitlock, Monte Carlo methods. Wiley, 2008. doi:https://doi.org/10.1002/9783527626212.
13. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945;1(6):80. https://doi.org/10.2307/3001968.
14. de Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. Chemom Intell Lab Syst. 2000;50(1):1–18. https://doi.org/10.1016/S0169-7439(99)00047-7.
15. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55. https://doi.org/10.1093/biomet/70.1.41.
16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodol). 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
17. J.D. Storey, False discovery rate. in International Encyclopedia of Statistical Science, Berlin, Heidelberg: Springer Berlin, 2011, pp. 504–508. doi: https://doi.org/10.1007/978-3-642-04898-2_248.
18. Yung G, Liu Y. Sample size and power for the weighted log-rank test and Kaplan-Meier based tests with allowance for nonproportional hazards. Biometrics. 2020;76(3):939–50. https://doi.org/10.1111/biom.13196.
19. Pereira B, et al. The somatic mutation profiles of 2433 breast cancers refine their genomic and transcriptomic landscapes. Nat Commun. 2016;7(1):11479. https://doi.org/10.1038/ncomms11479.
20. Shin H-T, et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. Nat Commun. 2017;8(1):1377. https://doi.org/10.1038/s41467-017-01470-y.
21. Tsimberidou AM, Fountzilas E, Nikanjam M, Kurzrock R. Review of precision cancer medicine: evolution of the treatment paradigm. Cancer Treat Rev. 2020;86: 102019. https://doi.org/10.1016/j.ctrv.2020.102019.
22. Hamdan D, Nguyen TT, Leboeuf C, Meles S, Janin A, Bousquet G. Genomics applied to the treatment of breast cancer. Oncotarget. 2019;10(46):4786–801. https://doi.org/10.18632/oncotarget.27102.
23. Krijgsman O, et al. A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response. Breast Cancer Res Treat. 2012;133(1):37–47. https://doi.org/10.1007/s10549-011-1683-z.
24. Dowsett M, et al. Comparison of PAM50 risk of recurrence score with onco *type* DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. J Clin Oncol. 2013;31(22):2783–90. https://doi.org/10.1200/JCO.2012.46.1558.
25. W. Leite, *Practical propensity score methods using R*. 2455 Teller Road, Thousand Oaks California 91320 : SAGE Publications, Inc, 2017. doi: https://doi.org/10.4135/9781071802854.
26. Seo SW, Kim J, Son J, Lim S. Evaluation of conditional treatment effects of adjuvant treatments on patients with synovial sarcoma using Bayesian subgroup analysis. BMC Med Inform Decis Mak. 2020;20(1):320. https://doi.org/10.1186/s12911-020-01305-9.
27. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol. 2011;173(7):731–8. https://doi.org/10.1093/aje/kwq472.
28. Pingault J-B, O'Reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. Nat Rev Genet. 2018;19(9):566–80. https://doi.org/10.1038/s41576-018-0020-3.
29. Bucur IG, Claassen T, Heskes T. Inferring the direction of a causal link and estimating its effect via a Bayesian Mendelian randomization approach. Stat Methods Med Res. 2020;29(4):1081–111. https://doi.org/10.1177/0962280219851817.
30. Loh W, Cao L, Zhou P. Subgroup identification for precision medicine: a comparative review of 13 methods. WIREs data mining and knowledge discovery, 2019;9(5). doi:https://doi.org/10.1002/widm.1326.
31. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. Stat Med. 2002;21(19):2917–30. https://doi.org/10.1002/sim.1296.
32. Spiegelhalter DJ. The future lies in uncertainty. Science (1979), 345(6194):264–265; 2014, doi: https://doi.org/10.1126/science.1251122.
33. Aitken C, Mavridis D. Reasoning under uncertainty. Evid Based Mental Health. 2019;22(1):44–8. https://doi.org/10.1136/ebmental-2018-300074.
34. Yang G, Xu Q, Wan Y, Zhang L, Wang Z, Meng F. miR-193a-3p enhanced the chemosensitivity to trametinib in gallbladder carcinoma by targeting KRAS and downregulating ERK signaling. Cancer Biother Radiopharm. 2021. https://doi.org/10.1089/cbr.2021.0016.
35. Xiao Y, Deng T, Su C, Shang Z. MicroRNA 217 inhibits cell proliferation and enhances chemosensitivity to doxorubicin in acute myeloid leukemia by targeting KRAS. Oncol Lett. 2017;13(6):4986–94. https://doi.org/10.3892/ol.2017.6076.
36. Kopp F, Wagner E, Roidl A. The proto-oncogene KRAS is targeted by miR-200c. Oncotarget. 2014;5(1):185–95. https://doi.org/10.18632/oncotarget.1427.
37. Hess KR, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J Clin Oncol. 2006;24(26):4236–44. https://doi.org/10.1200/JCO.2006.05.6861.
38. Luo Z, Gardiner JC, Bradley CJ. Applying propensity score methods in medical research: pitfalls and prospects. Med Care Res Rev. 2010;67(5):528–54. doi: https://doi.org/10.1177/1077558710361486.
39. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Stat Med. 2014;33(7):1242–58. https://doi.org/10.1002/sim.5984.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.