Discovery of Genes Positively Modulating Treatment Effect Using Potential Outcome Framework and Bayesian Update

Young Keun Lee

Samsung Medical Center, Sungkyunkwan University School of Medicine

Jisoo Kim

Sungkyunkwan University School of Medicine

Sung Wook Seo (✉ sungwseo@gmail.com)

Samsung Medical Center, Sungkyunkwan University School of Medicine

Research Article

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

Posted Date: November 23rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1069469/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

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

Young Keun Lee¹, Jisoo Kim², Sung Wook Seo^{1,2*}

¹Department of Orthopedic Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

²Institute of Biomedical AI, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University School of Medicine, Seoul, Korea

* Corresponding Author: sungwseo@gmail.com

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 an 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 number of data. 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 are 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

1. 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-source 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 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. 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 impute the potential outcome from samples matched with similar covariate profiles(4-6). 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(7).

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(8). 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.

2. Materials and Methods

2.1 Pseudo-counterfactual assumption and similarity weighted Monte Carlo

Assume that individual i with variable X_i is treated with the binary treatment variable T_i ($T_i = 1$, 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(9).

Here, we define $Y(T_i = 1, X_i)$ as a function of $f(X_i)$ and $Y(T_i = 0, X_i)$ as a function of $g(X_i)$. 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, 6). 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 by the similarity $sim(X_i, X_j)$ between the factual and pseudo-counterfactual. Using the similarity weighted Monte Carlo method, we could estimate $ITE(X_i)$ as follows:

$$\sum_j W(X_i, X_j) = 1$$

$$ITE(X_i) \approx f(X_i) - \sum_j g(X_j) \cdot W(X_i, X_j) \quad (1)$$

2.2 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(10). 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.

Since the beta distribution is a conjugate prior for the binomial distribution, the posterior distribution after observing the score s after N_j trials can be defined as follows:

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

2.3 Update Win probability using similarity weighted Monte Carlo

In ideal settings where all individuals $j(j \in \{1, 2, \dots, N_j\})$ in the counterpart group that is 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(11). Other methods use a transformed one-dimensional score, that is, a regression function, such as a propensity function(12).

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:

$$sim(i, j) = e^{-|P_i - P_j| / k\tau}$$

$$W(i, j) = \frac{e^{-|P_i - q_j| / k\tau}}{\sum_j^N e^{-|P_i - q_j| / k\tau}} \quad (3)$$

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:

$$\begin{cases} s_j = 1, \text{ if } Y_i(T = 1) > Y_j(T = 0) \\ s_j = 0, \text{ if } Y_i(T = 1) < Y_j(T = 0) \end{cases} \quad (4)$$

$$\text{Posterior } Pw_i \sim \text{Beta} [\alpha + s_j \cdot W(i, j), \beta + (1 - s_j) \cdot W(i, j)]$$

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

$$\text{Posterior } Pw_i \sim \text{Beta} \left[\alpha + \sum s_j \cdot W(i, j), \beta + \sum (1 - s_j) \cdot W(i, j) \right] \quad (5)$$

We defined the observed clinical covariates as V_i . Genetic alteration is represented by simple binary values (e.g., each existing alteration is given a value of 1; if not, it is given a value of 0). The individual i has the genetic variables G_i as follows:

$$G_i = [g_1, g_2, g_3, \dots, g_N]$$

$$g_1, g_2, g_3, \dots, g_N \in \{0, 1\}$$

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 $W^G(i, j)$.

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

$$\text{Posterior } Pw_i \quad (6)$$

$$\sim \text{Beta} \left[\alpha + \sum s_j \cdot W^V(i, j) \cdot W^C(i, j), \beta + \sum (1 - s_j) \cdot W^V(i, j) \cdot W^C(i, j) \right]$$

2.4 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 $P(\text{"win"}|g_k)$.

$$\begin{aligned} P(\text{"win"}|g_k) &\propto P(g_k|\text{"win"}) \cdot P(\text{"win"}) \\ k &\in \{1, 2, 3, \dots, n\}, n \in \mathbb{N} \end{aligned} \quad (7)$$

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. 6). Observing individual i 's win probability $Pw_i(g_k = 1)$ updates the prior distribution of $P(\text{"win"}) = \text{Beta}(\alpha_0, \beta_0)$ and the posterior is as follows:

$$\text{Posterior } P(\text{"win"}|g_k) \sim \text{Beta}[\alpha_0 + Pw_i(g_k = 1), \beta_0 + (1 - Pw_i(g_k = 1))] \quad (8)$$

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

$$\begin{aligned} &\text{Posterior } P(\text{"win"}|g_k) \\ &\sim \text{Beta}[\alpha_0 + \sum_{i=1}^N Pw_i(g_k = 1), \beta_0 + (N_k - \sum_{i=1}^N Pw_i(g_k = 1))] \end{aligned} \quad (9)$$

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.

$$\begin{aligned} &\text{Posterior } P(\text{"win"}|g_k) \\ &\sim \text{Beta}[\alpha_0 + \sum_{i=1}^N \frac{Pw_i(g_k = 1)}{\hat{e}(V_i)\hat{e}(G_i)}, \beta_0 + (\sum_{i=1}^N \frac{1 - Pw_i(g_k = 1)}{(1 - \hat{e}(V_i))(1 - \hat{e}(G_i))})] \end{aligned} \quad (10)$$

$$\hat{e}(V_i) = P(T|V_i), \hat{e}(G_i) = P(T|G_i)$$

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 $P("win"|g_k = 0)$ should be lower than the lower bound of it.

3. Results

3.1 Simulation data analysis

The C-search algorithm was validated with the simulation data (Supplementary 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.

The power of log-rank survival analysis depends on the number of data (13). 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)

3.2 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 2,433 patients with primary breast cancer(14). Clinical data consisted of age, chemotherapy, radiation therapy, sex, survival lifetime, and survival events. Among

the 2,433 records, 964 records that have missing on clinical data were omitted. A total of 1,469 patients were included in the analysis. (Supplementary 2)

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 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).

3.3 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.

A total of 10 genes were shown to be correlated with positive treatment outcomes: *EGFR*, *CLK3*, *PTEN*, *CDH1*, *GATA3*, *KRAS*, *RBI*, *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, 3c). 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).

3.4 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)

3.5 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(15). 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. (Supplementary 3) 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 *RBI* 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).

4. Discussion

Biotechnological breakthroughs in gene profiling have led to an increased focus on individualized precision therapy(16). 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.

There are studies on genetic assays to predict prognosis or response to treatment for breast cancer (17). BluePrint uses 80 genes to determine the sensitivity to adjuvant treatment. (18) Prosigna utilizes the PAM50 test, which identifies gene signatures specific to breast cancer subtypes (luminal A/B, HRE2, basal-like) (19). 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 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.

In the genomic data set, some of the genes rarely mutated and accounted for a smaller subset of patients. 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 (20, 21). The C-search algorithm updates the gene's win probability with a Bayesian update, thus, reflecting its uncertainty in the analysis.

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 (22-24).

PRKACG is a gene that encodes the protein kinase A subunit C γ , 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 (25).

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 (12, 26) 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.

5. 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 may be applied to various types of datasets where the number of samples is limited compared to the number of variables.

List of abbreviations

FDR False Discovery Rate

ITE Individaul Treatment Effect

IPW Inverse Propensity Score Weighting

Declarations

Ethics approval and consent to participate

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

Availability of data and materials

The cBioPortal dataset analyzed during the current study is available from the cBioPortal repository, https://cbiportal-datahub.s3.amazonaws.com/brca_metablic.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.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.2016R1E1A1A0191433, No.2021R1A2B5B0208627111).

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.

Acknowledgements

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

Figure legends

Figure 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 number of data. The log-rank test shows a slower increase in performance.

Figure 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 causal genes are assigned to the other gene group. All Kaplan–Meier survival curves are adjusted with propensity score matching; [29] 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.

Figure 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 causal genes are assigned to the other gene group. All Kaplan–Meier survival curves are adjusted with propensity score matching; [29] 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.

Figure 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.

Figure 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.

References

1. Carrasco-Ramiro F, Peiró-Pastor R, Aguado B. Human genomics projects and precision medicine. *Gene Therapy*. 2017;24(9):551-61.
2. Chia SKL. Clinical application and utility of genomic assays in early-stage breast cancer: key lessons learned to date. *Curr Oncol*. 2018;25(Suppl 1):S125-S30.
3. Noble WS. How does multiple testing correction work? *Nature Biotechnology*. 2009;27(12):1135-7.
4. Rubin DB. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*. 1978;34-58.
5. Stuart EA. Matching Methods for Causal Inference: A Review and a Look Forward. *Statist Sci*. 2010;25(1):1-21.
6. Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association*. 1986;81(396):945-60.
7. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in medicine*. 2002;21(19):2917-30.
8. Van de Schoot R, Kaplan D, Denissen J, Asendorpf JB, Neyer FJ, Van Aken MA. A gentle introduction to Bayesian analysis: Applications to developmental research. *Child development*. 2014;85(3):842-60.
9. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*. 1974;66(5):688.
10. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. 1945;1(6):80-3.
11. De Maesschalck R, Jouan-Rimbaud D, Massart DL. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*. 2000;50(1):1-18.
12. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
13. 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.
14. Pereira B, Chin SF, Rueda OM, Volland HK, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun*. 2016;7:11479.
15. Shin H-T, Choi Y-L, Yun JW, Kim NKD, Kim S-Y, Jeon HJ, et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nature communications*. 2017;8(1):1377-.
16. Tsimberidou AM, Fountzilas E, Nikanjam M, Kurzrock R. Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer Treat Rev*. 2020;86:102019.

17. 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.
18. Krijgsman O, Roepman P, Zwart W, Carroll JS, Tian S, de Snoo FA, 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.
19. Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW, et al. Comparison of PAM50 Risk of Recurrence Score With Oncotype DX and IHC4 for Predicting Risk of Distant Recurrence After Endocrine Therapy. *Journal of Clinical Oncology*. 2013;31(22):2783-90.
20. Spiegelhalter DJ. Statistics. The future lies in uncertainty. *Science*. 2014;345(6194):264-5.
21. Aitken C, Mavridis D. Reasoning under uncertainty. *Evidence Based Mental Health*. 2019;22(1):44-8.
22. 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.
23. 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.
24. Kopp F, Wagner E, Roidl A. The proto-oncogene KRAS is targeted by miR-200c. *Oncotarget*. 2014;5(1):185-95.
25. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, et al. Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer. *Journal of Clinical Oncology*. 2006;24(26):4236-44.
26. 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.

Figures

Figure 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 number of data. The log-rank test shows a slower increase in performance.

Figure 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 causal genes are assigned to the other gene group. All Kaplan-Meier survival curves are adjusted with propensity score matching; [29] 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.

Figure 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 causal genes are assigned to the other gene group. All Kaplan-Meier survival curves are adjusted with propensity score matching; [29] 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.

Figure 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.

Figure 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [20211119Supplementary.docx](#)