Causal Inference to Ascertain Causes of Metastasis in Melanoma

Wesley Maddox

Causal inference methods were performed on the TCGA clinical datasets. First, the data was collected via merging of related patient data. Then, an algorithm was used to create a causal DAG. Next, the IDA algorithm was used to guess the minimum bounds for causal effects on the probability of metastasis. Finally, marginal structural models and instrumental variables were used to verify the results from the IDA algorithm.

I. Introduction

UNDERSTANDING the causes of metastasis in cancer is an important task in cancer medicine. Despite decades of research focusing solely on the genetic events implicated in the metastasis pathway, the actual causes of metastasis remain a mystery [1], [2]. In cancer medicine, the current understanding of metastasis is that it results from the growth of tumor cells that had become detached from the primary tumor. This definition is a fairly precise one, and recent work has showed that there is a strong genetic component.

Understanding the causes of metastasis is an important task in cancer medicine for several reasons. As of 2006, metastasis remains the primary cause of death from cancer, causing approximately nine tenths of all cancer deaths [3]. Looking at the statistics of solely bone metastasis, it is estimated that 350,000 people each year die from these events. A similar issue affecting the importance of metastasis is in regards to pain. Metastases, especially those in the brain and bone, are rarely silent in nature and quite often cause intractable pain in the bones themselves [4]. For both the clinician and the patient, understanding the causes of metastasis is an important problem.

Skin cutaneous melanoma (SKCM) is a cancer affecting the cells of the melanin producing cells in the skin. It is a fairly common type of cancer with approximately 70,000 new cases diagnosed each year. Unlike many other cancer types, it has a fairly high survival rate with only approximately 8,000 deaths each year [5]. Similarly, many melanomas are only diagnosed after metastasis because many just stay on the skin or other pigmented tissues and never metastasize. Much of the research regarding this disease has recently focused on attempting to understand and predict metastasis within this type of cancer. In the TCGA project, most of the research has been done on cases that have metastasized. This allows for a fairly interesting dataset which has several interesting observations that are related to metastasis, as well as enough observations that do have metastatic tumors.

With regards to SKCM, metastases are particularly interesting for several reasons. First, these metastases often spread throughout the body in multiple different places. Often these metastases spread to the bone, causing immense pain for the patient, or to other vital organs where they can actually result in death [3], [5]. Secondly, melanoma that does not metastasize often has minimal negative effects, and can linger on the skin undiscovered. Research to ascertain the causes of the switch between a melanoma that does not metastasize and a melanoma that does metastasize is thus relevant and clinically motivated.

Causal inference techniques are increasingly being utilized in research as an alternative and enhancement of traditional statistical techniques. Causal inference is used to gain more detailed knowledge about treatments and effects, rather than just simply measuring associations [6], [7]. Causal inference techniques, though strongly developed, are actually used much less frequently than other statistical measures. For example, while there are many studies that have developed measures of mutational importance to phenotypes in cancer [8]–[12], none of them attempt to address the problem from a causal standpoint. In the field of cancer biology, this fact makes this attempt to use different causal inference techniques novel.

In this project, the primary dataset for causal inference comes from the TCGA project's collection of clinical and new tumor event records for skin cutaneous melanoma (SKCM) [5]. The dataset was quickly assembled by adding together the varying clinical data files. After adding together the different datasets, the end result was a dataset that included 504 records for 71 different clinical observations about the patient. This results in a very high dimensional observational dataset that may drastically increase the challenges for accurately utilizing causal inference techniques in order to understand the causes of metastasis.

Section II describes how the causal graph was generated. Section III gives a very quick introduction to causal inference and then moves into how the causal effects were estimated. Section IV states the results from the project. Finally, Section V discusses the results and concludes the paper.

II. Development of Causal Graph

Much of the work surrounding causal inference depends on the development of a causal directed acyclic graph (DAG) that correctly describes the causal situation in the real data. Recently, several algorithms have been utilized in order to solve this problem and to attempt to infer these DAGs solely from observational data. In this project, I utilized the PC algorithm originally developed by Peter and Clark [13], and later implemented and refined by Kalisch and then Maathuis [14], [15].

However, this method also makes several flaws. It bears noting that no algorithm can ever learn the exact structure of the causal graph from the data itself due to the fact that multiple different graphs can have the same likelihood score and cannot otherwise be differentiated. Similarly, the

PC algorithm also assumes that any correlations between data points that are not linked in the graph are solely due to the interactions of the other variables in the graph (conditional independence). Next, we must also assume that all of the variables in the data are Gaussian random variables that follow the conditional independence assumption above. The final primary assumption is the sufficiency condition, which states that there are no hidden variables related to the data that could potentially have a causal influence on any of the variables in the data set [16], [17]. This assumption is the largest one, but must be assumed for this project to even occur, and its potential invalidity is discussed in the Discussion section.

The PC algorithm is relatively complex, and its exact details are out of the scope of this paper. However, it consists of two separate parts. The first step of the algorithm is often called the PC_{pop} algorithm and maximizes the probability of the undirected skeleton of the causal graph for a certain amount of time. This probability is determined by iteratively determining different potential neighbors and testing them for conditional independence. This allows for the construction of the skeleton (undirected edges) of the causal graph. The second step of the algorithm determines the direction of the edges via testing independence of triples that are defined as pairs of nodes with a common neighbor. The end result of this algorithm is an equivalence class of DAGs that have estimated directed edges [18].

Although the algorithm itself is fairly complex, the results are actually fairly easy to interpret. The output of this algorithm is a causal graph that has been tested for independence. Similarly, in building an implementation of this algorithm in R, Maathuis showed its accuracy on both simulated data derived from a multivariate normal distribution and real-world gene expression data from a model organism [17], [19].

III. Causal Inference on a Causal Graph

After generating the estimated causal graph, I then performed several different methods for finding the causal effects of different variables in the causal graph on metastasis.

A. Definition of Causal Effect

First, it is necessary to define what a causal effect actually is. Perhaps the most simple measure of any causal effect, the average causal effect (ACE), can be described as:

$$ACE = P(Y^{a=1} = 1) - P(Y^{a=0} = 0)$$

Here, we treat Y as the outcome variable, and $A = a$ as the treatment variable (with two levels). In a population with possible different levels, this can be written in terms of expectations, such that:

$$ACE = E(Y^{a=1} = 1) - E(Y^{a=0} = 1)$$

Here, one should note the different notation being used here - this is a counterfactual variable and not necessarily an observed variable. This counterfactual expectation measures not the expected value of Y given the treatment A, but rather the expectation of the entire population if the entire population was given the treatment [6].

This treatment of a causal effect is just one of multiple methods of defining causal effects. Another notation, developed by Judea Pearl, includes the usage of do-calculus [17].

At its most basic level, causal inference requires three assumptions: consistency, positivity, and exchangeability. In order to do any sort of causal inference, it is first necessary to assume consistency, which as stated by Hernan and Robins is "the outcome for every treated individual ...[is]... his outcome if he had been treated and the outcome for every untreated individual ...[is]... his outcome if he had remained untreated." This means that under consistency it is possible to assume that the versions of treatment are same, and it allows relation of counterfactual to the real value [6]. In the SKCM datset, consistency of all treatments is required to be assumed in order for any meaningful analysis to be done. It does seem like this is a justified assumption, as the questionnaires that were used to generate the data are fairly straightforward.

Next, the positivity condition requires that there must be a treatment level greater than zero for all treatment levels being analyzed. While this does seem trivial at first glance, it actually can become problematic in the real world [6]. In the SKCM dataset, the data was first cleaned in order to give the conditions for positivity.

The exchangeability condition requires that, in a population, the probability of the treatment is independent of the outcome (ie individuals who were untreated would have the same outcomes as the treated if they had been treated instead of those individuals who were treated). It cannot be expected to hold in observational studies in its most broad form, but most causal inference methods allow for the more narrow conditional exchangeability to hold (the same outcomes would result for the untreated being treated given confounding and effect modifying variables). Again, for any real meaningful analysis to be done, it was necessary to assume this condition also holds for the SKCM dataset.

B. IDA Method

On purely observational data, it may actually be impossible to get completely accurate causal effects. This is mostly due to the fact that there are always different confounding variables and that, like the trite statistics adage, an associational effect does not actually estimate the true causal effect [6]. However, different methods exist for estimating the causal effect on purely observational data.

One such method is the IDA algorithm, which was developed by Maathuis [15]. This algorithm describes the possible causal effect among all different DAGs in the set of equivalent DAGs that describe the dataset in question. The term IDA is short for "intervention calculus when the DAG is absent," and has also been verified on both multivariate Gaussian simulated data and on inferring the causal effect of gene deletions on phenotype [19].

Although the exact details of this algorithm are also probably out of the scope of this project, the algorithm works on a fairly simple procedure. For each possible graph in the set of equivalent graphs in the class, it utilizes a simple regression procedure in order to gain an estimate of the causal effect

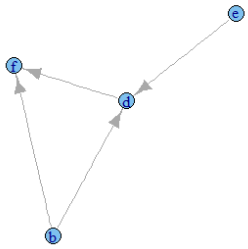

Fig. 1. Simple DAG to illustrate instrumental variables

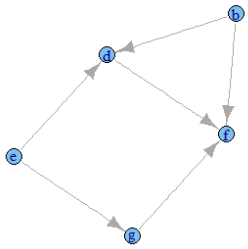

Fig. 2. More Complex DAG to illustrate instrumental variables

of a treatment A on an outcome Y. The exact value that is returned is actually the minimum calculated effect of all of the equivalent graphs. This value can be shown to be the minimum value for the causal effect of A on Y [17], [20].

C. Validation: Other Causal Inference Methods

Although the IDA method is extremely interesting in nature, it does still require validation in order to ensure that it was both estimated properly and that its output does make sense. Thus, after constructing the DAG on the SKCM dataset and running the implementation of IDA on this dataset, I then took the top six nodes that were estimated to have the highest minimum causal effects and then performed modeling using instrumental variable analysis, marginal structural models, and did outcome regression as well.

1) Instrumental Variables

Instrumental variables are a relatively widespread method of performing causal inference in the presence of confounding variables. However, they are also a very useful and widespread method in determining causal effects [21].

A simple instrument can be seen in Figure 1, where e can be viewed as an instrument for understanding the effect of d on f in the presence of the confounding variable b.

One unbiased and fairly accurate estimator can then be described in terms of covariances with respect to the instrument. Thus, the estmated causal effect of d on f is estimated by the following equation:

$$\hat{CE} = \frac{Cov(d, e)}{Cov(f, e)}$$

A more complex case can be in Figure 2 2, where there is now the (front-door) path from e to f via g. Thus, determination of the causal effect of d on f using the instrumental variable b would require conditioning on g. This should be easily accomplished.

2) Marginal Structural Models

Marginal structural models are another method of causal inference on smaller datasets that I also used. A very simple marginal structural model on a population has the form of

$$E(Y^a) = \beta_0 + \beta_1 a$$

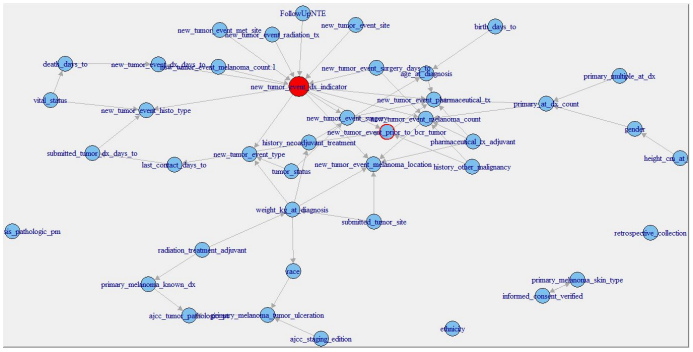

Fig. 3. Full Estimated DAG for the SKCM clinical dataset. The metastasis indicator is shown in red. It should also be noted that several edges were forced in order to reflect the fact that they could only occur in the presence of a metastasis (all nodes that relate to New Tumor Event)

This is regression on a counterfactual and requires a pseudo-population or IP-weighting on the values of a in order to make correct predictions. IP-weighting is another procedure that estimates the weights of the population by finding the probability distribution of the treatment of A and dividing it by the probability distribution of $A|L$, where L is any other set of covariates in the model. It can also be shown that β_1 is a consistent estimator of the ACE, which shows its usefulness [6].

IV. Results

After quickly merging the dataset for the SKCM data downloaded from the TCGA project, I began my analysis by utilizing the implementation of the PC algorithm in the R package pcalg [22]. Since the only tuning parameter is the significance level of the edges, I tried multiple different parameters. However, utilizing a significance level that was too low resulted in a DAG with too few connections. Thus, the utilized significance level was .25, which is fairly high but still resulted in accurate results. The resulting estimated DAG is shown in Figure 3.

My next step after generating the estimated causal DAG was to run the IDA algorithm to estimate lower bounds on the causal effects for all correlates on metastasis. This was done

Name of Variable	Lower Bound of Causal Effect
primary_multiple_at_dx	1.03
history_other_malignancy	.8535
new_tumor_event_histo_type	.4707
primary_melanoma_skin_type	.2456
new_tumor_event_prior_to_bcr_tumor	.197
new_tumor_event_met_site	.1841
race	.09829
vital_status	.09316
retrospective_collection	−.06896
tumor_status	−.00215

TABLE I
ESTIMATED LOWER BOUNDS ON CAUSAL EFFECT ON METASTASIS EVENT AS GENERATED BY THE IDA ALGORITHM

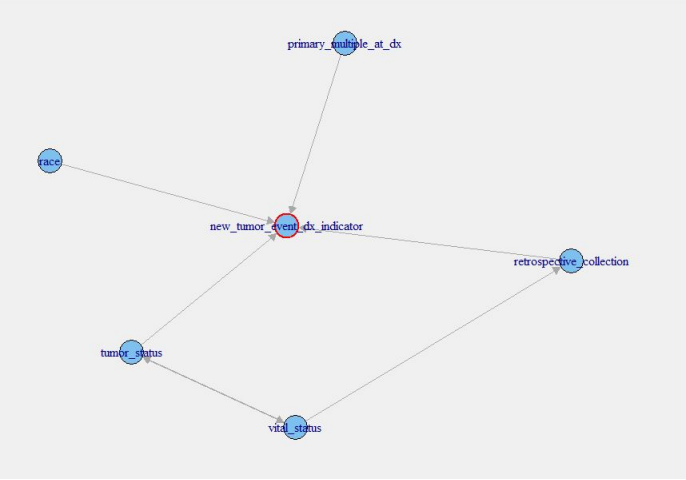

Fig. 4. PC Calculated partial DAG for subset of SKCM data

using the pcalg package in R. The estimated lower bounds for the causal effects for the largest positive nine correlate as well as an two interesting negative correlates are shown in Table I

We can see that many of the correlates that IDA calculated to have a large causal effect are actually metastasis specific. This is a somewhat disappointing result, but significant results can still be seen here. Next, I then chose five of these variables to validate further (primary_multiple_at_dx, race, vital_status, and tumor status). The fact that having a history of other malignancies (tumors) and the skin type of melanoma both have a high lower bound on causal effects is also a significant results.

Next, I utilized the same PC algorithm on the identified subset of clinical correlates in order to better understand this relationship between correlates. This is shown in Figure 4. Next, I used marginal structural models and instrumental variables on this DAG class in order to verify my results. For the marginal structural models, I also utilized IP-weighting to measure a pseudo-population across the other covariates. The results of this causal effect are shown in Table II

Finally, I utilized vital status (binary: 0=alive, 1=dead) as an instrument for estimating the causal effects that both tumor status at recording time and the retrospective collection indicator have on metastasis. Due to the front-door paths in the DAG for tumor status and retrospective collection, when calculating the instrumental variable estimator, it was necessary to condition on the other one in order to block that

Name of Variable	Estimated Causal Effect	Method Used
primary_multiple_at_dx	1	Marginal Structural
race	.537	Marginal Structural
tumor_status	.2798	Instrumental Variable
Retrospective Collection	.618	Instrumental Variable

TABLE II
CALCULATED CAUSAL EFFECTS ON METASTASIS EVENT THROUGH MARGINAL STRUCTURAL MODELS AND INSTRUMENTAL VARIABLES

path. Interestingly enough, the causal effect of tumor status on the event of metastasis is actually fairly small. This result is also verified by the fact that many metastases do not result in death, but simply require surgery to remove both the metastasis and the primary tumor. The retrospective collection indicator also shows that the TCGA Project did not choose a random sample of SKCM patients when collecting data for this survey, but rather a sample of SKCM patients that were biased towards metastasis. Overall, these results do seem to agree with the values calculated by the IDA algorithm even if they are larger in size (expected due to the increased accuracy of the method).

V. DISCUSSION

Utilizing causal inference methods on this set of clinical correlates for SKCM does give some interesting results; however the fact that much of the data does seem to depend on each other does dampen the results. For example, three of the eight highest values of estimated lower bounds on causal effects that were found from the IDA algorithm exist only in cases of metastasis. Although the methods implemented in the study may have worked correctly, this is in reality a reversal of causal effect (metastasis would cause these other variables to change, rather than the reverse) that is simply an artifact of the data.

Although the results for this project were ultimately unsatisfactory in determining interesting possible causes of metastasis in SKCM, there are multiple interesting conclusions that can be drawn from the work done. First, the IDA algorithm is a fairly accurate and useful tool in identifying the causal relationships between different nodes. Although the IDA algorithm only determines a lower possible bound on the causal effect, it did seem to correctly identify the interesting relationship between the Primary Multiple Tumors Status indicator and Metastasis. Similarly, most of the connections in the full DAG did seem to make sense, ie they passed the eyeball test for potentially making sense. Another validation measure that the IDA algorithm passed was in having effects that did survive in other more-established measures for causal inference.

Secondly, there does seem to be an interesting clinical relationship between having multiple primary tumors and eventually having a metastasis event. Looking at the data itself, it does seem like this is a somewhat interesting causal effect. To put it succinctly, having multiple primary tumors at the time of diagnosis seems to be a cause of metastasis at a later time. Further work is necessary to adequately make this conclusion and to determine the validity of the result, and to determine if this relationship is actually just an artifact of the way that the TCGA clinical patient data is structured. Combing through the TCGA data dictionary (available online at: https://tcga-data.nci.nih.gov/docs/dictionary/) makes it seem like this is

not actually an artifact. Since multiple melanomas can be known to develop at the same time, this relationship may just show the increased probabilistic risk for metastasis if a patient has multiple melanomas. For example, a patient with multiple melanomas at different sites on their body would have an increased risk for at least on of these melanomas to metastasize. This fact seems to be understood among clinicians, [3] but has not been empirically shown before.

Finally, the validity of these methods shows that this type of analysis can be extended to include more than just the clinical features of the tumor. The TCGA Project did not just collect the clinical features of these tumors, but also sequenced the DNA of either the metastatic tumor or the primary tumor in the same patients. The sequencing data as well as multiple other data types is also publicly available. An interesting for future work in this direction could be in integrating variants identified either from DNA sequencing (genome) or in mRNA sequencing (expression) into the causal DAG identified. This may actually allow for more accurate identification of driver and passenger mutations in melanoma (and other cancer), which is a large field of current research [23]. This type of integration may also allow for a more accurate understanding of the genetic causes of metastasis. Further research could also attempt to integrate data of multiple types for example the type of clinical data used in this project as well as both genome and expression data collected from the TCGA Project.

ACKNOWLEDGMENTS

The results shown here are in whole based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

REFERENCES

[1] I. J. Fidler, "The pathogenesis of cancer metastasis: the'seed and soil'hypothesis revisited," *Nature Reviews Cancer*, vol. 3, no. 6, pp. 453–458, 2003. [Online]. Available: http://www.nature.com/nrc/journal/v3/n6/abs/nrc1098.html

[2] G. P. Gupta and J. Massagu, "Cancer Metastasis: Building a Framework," *Cell*, vol. 127, no. 4, pp. 679–695, Nov. 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0092867406014140

[3] J.-C. Martinez and C. C. Otley, "The management of melanoma and nonmelanoma skin cancer: a review for the primary care physician," in *Mayo Clinic Proceedings*, vol. 76. Elsevier, 2001, pp. 1253–1265. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0025619611628024

[4] G. R. Mundy, "Metastasis: Metastasis to bone: causes, consequences and therapeutic opportunities," *Nature Reviews Cancer*, vol. 2, no. 8, pp. 584–593, Aug. 2002. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrc867

[5] "Cutaneous Melanoma - TCGA." [Online]. Available: http://cancergenome.nih.gov/cancersselected/melanoma

[6] M. Hernan and J. Robins, *Causal Inference*. Boston: Chapman & Hall.

[7] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, no. 0, pp. 96–146, 2009. [Online]. Available: http://projecteuclid.org/euclid.ssu/1255440554

[8] D. Ramazzotti, G. Caravagna, L. O. Loohuis, A. Graudenzi, I. Korsunsky, G. Mauri, M. Antoniotti, and B. Mishra, "Efficient inference of cancer progression models," *arXiv preprint arXiv:1408.4224*, 2014. [Online]. Available: http://arxiv.org/abs/1408.4224

[9] I. Korsunsky, D. Ramazzotti, G. Caravagna, and B. Mishra, "Inference of Cancer Progression Models with Biological Noise," *arXiv preprint arXiv:1408.6032*, 2014. [Online]. Available: http://arxiv.org/abs/1408.6032

[10] Stingo, Y. Ni, and V. Baladandayuthapani, "Integrative Bayesian Network Analysis of Genomic Data," *Cancer Informatics*, p. 39, Sep. 2014. [Online]. Available: http://www.la-press.com/integrative-bayesian-network-analysis-of-genomic-data-article-a4392

[11] Q. Zhang, J. E. Burdette, and J.-P. Wang, "Integrative network analysis of TCGA data for ovarian cancer," *BMC Systems Biology*, vol. 8, no. 1, Dec. 2014. [Online]. Available: http://www.biomedcentral.com/1752-0509/8/1338

[12] R. Neapolitan, X. Jiang, D. Xue, A. Brufsky, and S. Khan, "A New Method for Predicting Patient Survivorship Using Efficient Bayesian Network Learning," *Cancer Informatics*, p. 47, Feb. 2014. [Online]. Available: http://www.la-press.com/a-new-method-for-predicting-patient-survivorship-using-efficient-bayes-article-a4064

[13] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000, vol. 81.

[14] M. Kalisch and P. Bhlmann, "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *The Journal of Machine Learning Research*, vol. 8, pp. 613–636, 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1248681

[15] M. H. Maathuis, M. Kalisch, and P. Bhlmann, "Estimating high-dimensional intervention effects from observational data," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3133–3164, Dec. 2009. [Online]. Available: http://projecteuclid.org/euclid.aos/1250515382

[16] C. R. Shalizi, "Advanced data analysis from an elementary point of view," *Preprint of book found at http://www. stat. cmu. edu/ cshalizi/ADAfaEPoV*, 2013. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.371.4613

[17] K. P. Murphy, *Machine learning: a probabilistic perspective*, ser. Adaptive computation and machine learning series. Cambridge, MA: MIT Press, 2012.

[18] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *The Annals of Statistics*, vol. 40, no. 1, pp. 294–321, Feb. 2012. [Online]. Available: http://projecteuclid.org/euclid.aos/1333567191

[19] M. H. Maathuis, D. Colombo, M. Kalisch, and P. Bhlmann, "Predicting causal effects in large-scale systems from observational data," *Nature Methods*, vol. 7, no. 4, pp. 247–248, 2010. [Online]. Available: http://www.nature.com/nmeth/journal/v7/n4/full/nmeth0410-247.html

[20] P. Bhlmann, "Causal statistical inference in high dimensions," *Mathematical Methods of Operations Research*, vol. 77, no. 3, pp. 357–370, Jun. 2013. [Online]. Available: http://link.springer.com/10.1007/s00186-012-0404-7

[21] A. C. Cameron and P. K. Trivedi, *Microeconometrics: methods and applications*. Cambridge ; New York: Cambridge University Press, 2005.

[22] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann, "Causal inference using graphical models with the r package pcalg," *Journal of Statistical Software*, vol. 47, no. 11, pp. 1–26, 2012.

[23] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens *et al.*, "Patterns of somatic mutation in human cancer genomes," *Nature*, vol. 446, no. 7132, pp. 153–158, 2007.