Causal inference in DotA 2 when estimated through randomized data

Stelios Avgousti
Supervisor(s): Rickard Karlsson, Jesse Krijthe
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Strategy games could be considered as an amazing playground for using Causal inference methods. The complex nature of the data and the built-in randomization help with testing causal inference in a scenario where in reality it would be hard and expensive. Randomized data in coherence with causal inference is well documented and tested, but not regarding the strategy game of interest DotA 2. To evaluate the quality of causal inference using randomized data for predictions in the game, the average causal effect estimand is used. The calculation of the average causal effect of certain events between different intervals and their comparison in addition to the calculation of the statistical Independence between variables of concern comprise the bulk of the research. The calculations allow for logical deductions and statistical correlations between values to reach a conclusion. The final verdict being that causal inference with randomized data is helpful for predicting events in DotA 2 but the amount of data and existing complex biases can be deceiving and can heavily influence the results.

1 Introduction

Strategy games and their popularity has grown tremendously in recent years. This paper is interested in using Machine learning and Causal inference for analyzing data and predicting outcomes in a famous MOBA (Multiplayer Online Battle Arena) game, called DotA 2. Firstly what is causal inference? Causal inference is the science of determining cause and effect between phenomena and the way we reason about it. For example its usually a comparison between the expected outcome when an Action A is taken versus an expected outcome when an Action A is withheld, without necessarily observing that action. If the two outcomes are different then one says that A has a causal effect on the outcome [1]. Secondly, what is DotA? Defense of the Ancients 2 or DotA 2 is a video game from 2013 that features several 5 vs 5 game modes [2]. It means that 10 players in teams of 5, battle each other in an online arena with the winning condition being that each team has to destroy the base of the other.

Using the plethora of data in the Open DotA API, and their randomized nature along with causal inference method, it will be shown how some events in the game cause others to happen, for example "How does a "Hero" being picked impact a team winning?". This for example can be calculated as a probability from existing data, but how useful would it be to answer this question by using randomized data and causal inference? "Hero" refers to the avatar a player chooses and in DotA 2 there are 123 of them. Randomization in this case comes solely from the physical randomization of the "Hero" assignment values to each player before the start of the game.

While investigating thoroughly for existing and relevant work, it was found out that there's abundant sources of information concerning causal inference. Mainly the 'What-if' book by Miguel A.Hernan and James M.Robins has important information, which concerns various data analysis approaches to estimate the causal effect of interest under a particular set of assumptions [1]. In addition, there are many examples of how to use causal inference with randomized data-sets, so the question isn't about the feasibility of using causal inference and randomization. What we don't know is how useful it is to use the randomized data to predict other events.

The main research question that needs to be answered in this paper is "How and If, matches with instances of randomization can be useful for predicting events using causal inference in DotA 2". From this question two different important sub questions arise which will be mentioned in section 2. These questions are answered using the randomized data sets that DotA 2 provides and with the use of a simple causal inference metric commonly known

as "Average Causal Effect". The paper contributes to the understanding of causal effects when estimated on randomized data, as randomization in real life is expensive and difficult. In addition it will be shown how the causal effect is impacted by time and the complexity of the data. The interest increases when we think on how to apply the same methods to real-life examples that include a sense of randomization, like for example a randomized medical trial. The rest of the paper includes a formal problem description in 2 , the methodology that was applied to come to a conclusion in 3, the assumptions that had to be made, the experiment along with the results 4, and finally we finish with 3 sections mainly, a responsible research section 5, discussions 6 and the conclusions/ future work 7.

2 Problem Description

This paper has the purpose of not only providing better analytics to Dota 2 casual and professional players but the game also provides a playground for machine learning algorithms as the data is complex enough and time-dependent.

2.1 Research Question

By stating the research question in the introduction, the paper continues by dividing it into sub-questions. The sub-questions have the purpose of guiding the thought process to answer the research question. Now the division is as such: "How does the selection of hero influence the causal effect on a team winning when estimated through randomized data" and "how do the causal effects compare over time?". The first sub-question guides the paper through the way of applying causal inference methods to Dota 2, and the second sub-question to comparing the different causal effects through time and seeing how impactful randomization is to estimate these effects. Both of these questions contribute to the final verdict of the paper, but they are not concretely answered.

2.2 Literature Survey / Background

Every research needs a starting point. And every starting point needs some good literature to support it. Firstly the "What If" book by Miguel A.Hernan and James M.Robins [1] is crucial to get an idea of the different causal inference methods there are. Secondly, "Causal Inference for the Brave and the True" [3] serves as an initial introduction to causal inference and randomized experiments. Thirdly, the Journal "Causal Inference Using Potential Outcomes" by Donald B Rubin [4] and 'Causal Inference' by Kosuke Imai [5] provide guidelines for performing the experiments, in addition to the things to look out for when applying causal inference. The "Exploring the Role of Randomization in Causal Inference" dissertation by Ding[6] has insightful information about related causal inference methods that we are interested in. These articles also include the fundamental information needed to apply causal inference in any setting.

Now, a way to understand how Dota 2 data is structured and find a way to retrieve data that conforms to the experiment can be found online in articles and web pages mainly, [7] [8] [9] that show how Python can be used to retrieve matches. Now information about Dota 2 and estimating win probabilities using causal inference can be found at [2] but not so much to do with randomization. Lastly, insightful information about other applications of causal inference can be found at [10][11][12], or for how a good design for a causal inference experiment is crucial can be found at [13].

3 Methodology

The framework to follow, the assumptions to make, and the way to estimate the causal effect are found in this section. The report follows an approach from one of the first people to study causal inference, Neyman [4]. In addition, the Chi-squared test is used to answer the second sub-question in section 2. This methodology section focuses on the former as the Chi-squared test will be seen in the results.

3.1 Potential Outcomes Framework

In this paper the causality framework of *Potential Outcomes* often called *Neyman-Rubin causal model* [5] is used. Potential Outcomes Framework refers to the key idea that any causal inference is based on both actual and counterfactual outcomes [5]. An example of that would be a TV watcher that has watched an ad about a product. We are interested in the causal effect of that ad on the user, but we are also interested in the effect that would exist if the TV-watcher didn't watch the ad. In the example above, the treatment variable is whether the ad is being watched or not, and the potential outcomes would be the effect that the ad had on the watcher. In the context of this paper, the potential outcomes framework is applied to Dota 2 by assigning treatment variables and potential outcomes to in-game data. Each of the potential outcomes corresponds to the particular value of the treatment variable [5]. These need to be well defined to continue with causal inference. In this case, it means that the treatment variable and outcome have a concrete value, and are not ambiguous. The problem is that, while applying a treatment variable, only the observed outcomes value is known and not the value that would've been there if the treatment variable was not applied. Not knowing the latter introduces the fundamental problem of causal inference [5]. Because of this, it is required to do certain assumptions about the data to be able to calculate the causal effect.

The treatment variable when it comes to this experiment is a binary choice of a "Hero". The variable will have the value of 1 if the "Hero" exists in the game, and 0 if the "Hero" doesn't exist in any team for that particular match. "Hero" is the chosen avatar of the player that will represent him in the game. For each experiment, a singular "Hero" is chosen out of 122, for which the causal effect is calculated. Then the outcome variable is again a binary choice between either a game being won or a game being lost. This means that for each game there are 2 outcome variables with 2 possible outcomes.

3.2 Assumptions

In many causal inference reports and articles, certain assumptions are commonly made, often implicitly, and sometimes without too much thought [4]. These assumptions are under SUTVA (stable unit treatment value assumption) [4] which itself comprises 3 sub-assumptions. First is the no interference assumption, which formally states that the treatment status of one unit does not affect the potential outcome of another unit [4]. Moreover, it is assumed that no simultaneity exists [5], that is that the ordering between the treatment variable and the outcome is fixed, more clearly that the treatment affects the outcome and not the other way around. Lastly, it is needed to assume that there's only one version of treatment across all units [5]. The last assumption means that if for example medical surgery is the treatment, and it's always performed by the same doctor, if a different surgeon operates then it is considered a different treatment [5].

These assumptions are more clear when we show how they are applied in the context of this experiment. First, it is noteworthy to say that even though the problem of missing data

exists, the randomization will still generate missing values of the counterfactual outcomes but randomization ensures that the missing values occurred by chance [1]. For this reason, the average causal effect can be estimated despite the missing data. The no interference assumption indeed holds as each "Match" of DotA 2 doesn't influence any other matches, in addition to the assignment of the treatment value. Now, no simultaneity does indeed hold as the treatment as a "Hero" being picked and the effect as "a game being won" has already been defined. The game being won can not possibly influence the hero that was picked, because of the time difference. The hero is always picked before the game, in consequence, the assumption holds. Moreover, it needs to be ensured that the probability of the value of treatment for a match is constant across all Heroes [5] and will imply statistical Independence between the treatment and the potential outcomes.

3.3 Estimation of Average Treatment Effects

Now to estimate these causal effects an approach that was first developed by Neyman[5] will be followed. A problem that randomization solves but not completely it is that it's hard to control all factors that influence the potential outcomes when assigning a treatment value. Because of the above and the fundamental causal inference problem, an estimator for calculating the average treatment effect needs to be used which hopefully, will also be unbiased for a binary treatment variable and outcome.

How is the average treatment effect calculated? Firstly one needs to define the unit treatment effect for each game, and that is to look at the difference between the 2 potential outcomes. The unit treatment effect is defined as $t_i = Y_i(1) - Y_i(0)$ where t_i is the treatment effect for that unit, $Y_i(1)$ is the outcome given the treatment, meaning the outcome given the "Hero" was in the game, and $Y_i(0)$ is the outcome given control, meaning the outcome given that the "Hero" wasn't in the game. Unfortunately, as mentioned in the above sections, only one of those values is visible and to estimate the causal effect, SATE (sample average treatment effect) needs to be calculated. SATE is defined as such $\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)$ which needs both of these values and where $\sum_{i=1}^{n}$ is over all the games in our population. To bypass this problem it was decided to go with the difference-in-means estimator which represents the difference in the average outcome between the treatment and control groups. The difference-in-means estimator is as such $\hat{t} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)Y_i$ where T is defined above, n_0 is the number of matches the in control group, n_1 is the number of matches in the treatment group and n is the total number of both the control and treatment group. The difference-in-means estimator is unbiased for the estimation of SATE (sample average treatment effect) [5].

Continuing, the sample variance regarding the potential outcomes $Y_i(1)$ and $Y_i(0)$ is calculated in each experiment for each "Hero" and its also calculated for a single "Hero" throughout the updates of the game. By calculating it, it will be known how spread our data is, that is, the more spread out the potential outcomes are the bigger the variance. Moreover, by calculating the variance one can show if getting more data correlates with having more accuracy. In addition to this, calculating the variance gives good information on how to split the relative control and treatment groups (if it's possible given the experiment) so that the resulting variance is minimized [5]. An estimate of the variance is calculated using the equation $\frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$ where $S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n} T_i(Y_i - \overline{Y}_1)^2$ and $S_0^2 = \frac{1}{n_0 - 1}\sum_{i=1}^{n}(1 - T_i)(Y_i - \overline{Y}_0)^2$ and finally $\overline{Y}(t) = \sum_{i=1}^{n} Y_i(t)/n$. This way of estimating the variance is only an approximation as to calculate the true variance the sample covariance between the 2 potential outcomes is needed, something that is impossible because we never observe the two potential outcomes together, moreover, this estimator for the difference-in-means

variance is on average conservative [5]. This means that the estimator is less likely to be giving out wrong results but it's also less likely to find a statistically significant result. Now a calculation for the potential outcomes sample variance for a singular Hero is given in section A figure 7 which correctly shows that with more data, the sample variance decreases thus increasing accuracy. Lastly by computing the sample variance for each hero, then by taking its square root the uncertainty for each point estimate of the average causal effect is retrieved. The uncertainty tells us how much the value fluctuates.

To do all of this, one requires many amounts of data that will be acquired from the Open Dota API[8]. The API will allow the retrieval of many matches of DotA 2 that conform to randomization and with that and the help of Python, they will be retrieved, defined, and analyzed by comparing the causal effects for the different treatment assignment values and outcomes. When retrieving data it needs to be ensured that the data all comes from a certain DotA 2 Game-mode called "All random" as this ensures the randomization of the assignment treatment variable.

4 Experiment and Results

The experiment that is conducted, includes the calculation of the sample average causal effect of a "Hero" on a team winning using the "Difference-in-means" estimator. This effect was calculated for all DotA 2 Heroes except 'Techies', for different game updates (includes temporal changes), and in addition, the values are compared using the Chi-squared test to see if indeed time and updates affect each Hero.

4.1 Data gathering, Data filtering

Taking all of the above assumptions into consideration, the required data is then retrieved. The data was directly taken from the OPEN-DOTA-API with the use of Python get requests. Then all the games are filtered based on the game mode that the game was played in. Finally, the data is split according to which game patch they were from in order to compare causal effect and variance values later on. The patches that were taken into account were taken from the DotA 2 Wiki [14] and include 3 big intervals. The first significant interval is from 23rd February 2022 until the 4th of April 2022 and includes big patches 7.31 and 7.31B along with minor changes and amounts up to n_{p0} =6.729 games. The second interval is between the 4th of April 2022 and the 4th of May 2022 which by itself doesn't include big patches but has several small updates and amounts up to n_{p1} =5.134 games. It was chosen to see how small almost irrelevant changes can impact the causal effect. The third significant interval is between the 4th of May 2022 and until the 4th of June 2022 and includes big patch 7.31C and several minor changes and amounts up to n_{p2} =4.770 games. All 3 intervals together amount up to a total of n_{ptotal} =16633. Code for retrieval and analysis can be found at the relevant Github repository [1].

4.2 Results

The results section is split into two parts. The first part includes the results from calculating the average causal effect for all heroes on all n_{ptotal} games and the calculation of the average causal effect depending on the 3 different patches mentioned above with n_{p0}, n_{p1}, n_{p2} respectively. Information about the game updates was taken directly from the DotA 2 Wiki[14]. The second part includes the deployment of the Chi-squared test on 2 different

[1] https://github.com/stelios34S/Causal-inference-in-DotA-2-when-estimated-through-randomized-data

criteria, the Independence between the choice of "Hero" and the game outcome, and the Independence between the game patches and the game outcome for **each** "Hero". The results will be discussed in detail in section 6.

4.2.1 Average Causal Effect

The way the average causal effect is calculated was mentioned in section 3 and was strictly followed to reach the results. The results for the average causal effect are in section A and are referenced in the following text (sorted on the increasing average causal effect). In all the results uncertainty for each value is shown as a color on the points.

- Average Causal effect for Interval 23/2 - 4/4. The results for this interval can be found in section A figure 3 and includes the average causal effects on winning for all "Heroes". This update interval includes 2 big in-game updates specifically 7.31B and 7.31C. The 2 updates include a lot of changes done to the game something that should be reflected when comparing the values with the newest patch. Mainly the 2 patches introduced new a "Hero" and rebalanced a good amount of "Heroes" of the game. Other changes include Quality of Life changes, item changes, and bug fixing.

- Average Causal effect for Interval 4/4 - 4/5. The results for this interval can be found in section A figure 4 and includes the average causal effects for all "Heroes". This interval doesn't have any concrete "Hero" changes but includes many smaller patches that tweak small aspects of the game (bug fixing). This update interval is picked to show how indirect or insignificant changes may or may not change the causal effect.

- Average Causal effect for 4/5 - 4/6. The results for this interval can be found at section A figure 5. This includes the average causal effects for all "Heroes". The update interval includes 1 significant update 7.31C [14]. The update mainly takes care of the balancing of "Heroes" but also includes some other game balancing updates(items, neutral creeps). Other than that the interval also includes several smaller patch notes that are less significant. This interval should show a difference in the causal effect between the previous 2 patch intervals.

- Average Causal Effect for all games. The results for this interval can be found in section A figure 6 and includes the average causal effect for all "Heroes". This is calculated with n_{ptotal} =16633 games and should reflect that the values we get, are the average of the 3 intervals combined. In addition, this should show the state of the game in regards to the average causal effect for each "Hero", when estimated on all the data.

4.2.2 Independence Test

The Chi-squared test is a statistical test that can be performed to check if there is a statistical correlation between 2 variables [15]. As mentioned in the introduction of this section, we have performed the test for 2 different cases using the standard Chi-square test method provided by SciPy. Both tests were able to be done just by adapting the data that was retrieved for calculating the Average causal effect. The first test is included in this section and the second test is in section B.

- First test was to check if the outcome for each Game is independent of the selection of "Hero", something that's done just by taking all the wins and losses for each "Hero" and throwing everything in a contingency table. Then Scipy calculates the p-value

which tells us if the variables are indeed independent or not. Now the hypothesis is that hero selection and game outcome are independent of each other, and the Hypothesis holds if and only if the p-value calculated from the contingency table is more than the critical chi-square value which was set as 0.05 because by convention, people often use the 5% value [16]. It turns out that the Hypothesis is rejected, as the p-value calculated equals 4.093358722572413e-216 which is smaller than the critical 0.05%, thus making the game outcome dependent on the hero selection. Something that was expected. Why it was expected is discussed in section 6.

- Second test was checking for each different "Hero" if game updates (the intervals that were established) are statistically independent of the game outcome. This is different for each "Hero" as different patches influence different parts of the game thus making the outcome for the game dependent on some heroes, at least statistically. This is done by taking all the games for each hero depending on the intervals. For example a hero might have Interval 1: [23 wins, 50 losses], Interval 2: [30 wins, 40 losses], Interval 3: [40 wins, 50 losses]. Then Scipy would receive the table like this [[23 wins, 50 losses],[30 wins, 40 losses],[40 wins, 50 losses]] and return the p-value for each experiment. Once again we use the critical value of 5% to decide whether, for that "Hero", the outcome of the game is dependent on the update intervals. The results from this can be found in section B table 1 and include the 122 tests done that cover all heroes except 'Techies'.

5 Responsible Research

Does the research outcome have ethical implications? Is the research reproducible? These questions are answered in the following section.

5.1 Data Ethics

The data, as mentioned in section 4 was taken directly from the Open Dota Api which provides Dota 2-related data including advanced match data extracted from match replays[8]. The API provides a way to retrieve replays and match data for specific players but there's practically no way to connect the player-specific id or IGN (in-game name) to any real details. In addition, in this research, the data that is retrieved include only details about the games themselves and nothing about the players. This is done using the most recent match id something that can also be seen from the GitHub repository. The data set is also available to be posted and shared with the public.

5.2 Reproducability

To retrieve the data that are needed, the opendota python script is called which will save all the matches retrieved into a .csv file (comma-separated values). To analyze the data the main python script is called which will calculate the average causal effect, the data variance for each hero for all games, and the 3 different patches, all saved in separate .csv files. From there the results and the plots are calculated. By just using Python and an IDE one can recreate the experiment just by importing the necessary libraries. The only thing that is left is the actual data set to experiment on, something that is readily available in the GitHub repository. The results of the experiments are like discussions, as they were derived from the comparison of the absolute results. Relevant code can be found at the Github repository [2].

[2]https://github.com/stelios34S/Causal-inference-in-DotA-2-when-estimated-through-randomized-data

5.3 Scientific Integrity

Proper credit was given to sources that worked as inspirations for the experiments, and ideas taken were specified in detail.

6 Discussion

In this section, a discussion regarding the results and some design choices will be conducted. Research Findings will be compared to the expected outcomes via various DotA 2 websites and will also be placed in a broader context to understand what influenced the findings and how (negatively and positively). More concrete findings from calculating the average causal effect will also be used to discuss the results of the Pearsons Chi-squared test.

- Why was the difference-in-means estimator picked for the sample average causal effect (SATE)? Firstly there are several different ways of calculating the average causal effect. One is to calculate the population average treatment effect (PATE) which is different from the latter as it's done on the whole population. Of course, this was impossible as we only obtain a sample data set. The Conditional average treatment effect (CATE) could also be calculated, but that is done if there's more interest in seeing how treatment effect varies as a function of pre-treatment covariates[5]. For example, the causal effects of exposure to a cooking ad may differ between a chef and a beginner cook. Now even though other estimators weren't considered, the reason is that several papers suggested using the difference-in-means estimator as it is unbiased for a binary outcome problem. It is also noted that for randomized experiments with binary outcomes, all test statistics are equivalent to the difference-in-means estimator[6], which is something that applies to this experiment.

- How does the average causal effect compare to the actual win rate in-game overall? For this, look at figure 6 in section A and take into account 2 different heroes, "Meepo" and "Abaddon", the former being a really difficult "Hero" to play and the latter a really easy one. Difficult as characterized by the DotA 2 website in terms of complexity [17] and as characterized by the DotA 2 community. When estimating the causal effect for "Meepo" on winning the game, a value of -12% can be seen, which translates to having a -12% chance of winning the game when "Meepo" is in the team. Now "Meepo" has a 54.90% win rate as shown by DotaBuff [18]. Now looking at "Abaddon", a value of +3% can be seen, and when looking at DOTABUFF we can see a win rate of 54.50%. So, when the causal effect is looked at alone, one would say that a hard hero would result in a lesser chance to win the game than an easy hero.
  Now when looking at the win rate we can see that the difficult hero's win rate is almost the same as the easy ones. From both statistics, it can be deduced that because the causal effect is calculated on fully randomized data, where the player choice (player skill) is not a factor, it doesn't properly indicate how good or bad a hero is for winning the game. A player playing a hard hero at random would have a harder time winning than playing an easy hero at random, but that is not true if the player actively picked that Hero. In turn, the win rate is almost the same for the 2 heroes but the causal effect for "Meepo" is negative and quite high compared to "Abaddon", something that reinforces the above statements.

- How does the average causal effect for the heroes compare to the buffs (making a hero stronger) or nerfs (making a hero weaker) that the "Heroes" have received? We chose to analyze "Bane" and "Sand King". Bane's average causal effect between intervals

can be seen below.

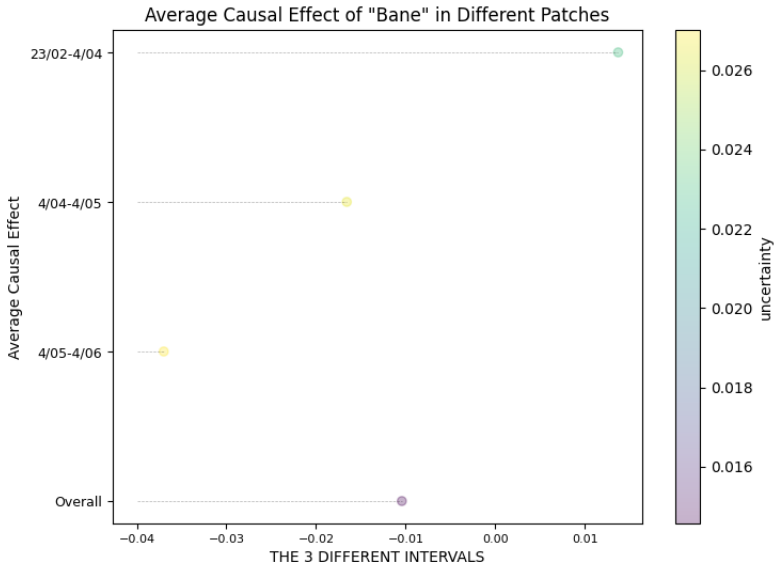


Figure 1: Average Causal Effect for Bane

From the figure above it can be seen that Bane's average causal effect starts as positive at 0.014 which translates to almost 1.5% more chance of winning the game. Without knowing entirely what Bane's average causal effect was before the first patch we consider (23/02-4/04), it's still known that Bane's attributes and statistics were changed in patch 7.31 and thus the calculations already include that. So, in other words, a 1.5% more chance to win the game when Bane's in the game is observed when patch 7.31 was applied. Then when looking at the causal effect for the second interval, it's observed that it becomes negative at -1.5%. No changes were made directly to "Bane" but still resulted in a 3% decrease in the average causal effect. One explanation is that either the minor changes that occurred in the second interval could somehow affect Bane (if for example there was a bug making Bane stronger, or a bug making other heroes weaker?), or just by time 'Bane' became less relevant than other Heroes at winning. Finally, we know that in patch 7.31C "Bane" was nerfed (making the hero weaker)[14]. This is nicely shown by the decrease in the average causal effect from the second interval to the third where "Bane's" average causal effect sits at -3%. Moreover, it's important to notice how the uncertainty of the average causal effect decreases depending on how many games it's calculated on. For the first interval which is the largest in terms of size, the uncertainty is lower than the rest, something that just makes sense as more data gives more accuracy. The lowest uncertainty can be seen when the Average causal effect is calculated on all the games. Now the next example result to take a look at is "Sand King" where the average causal effect can be seen below.

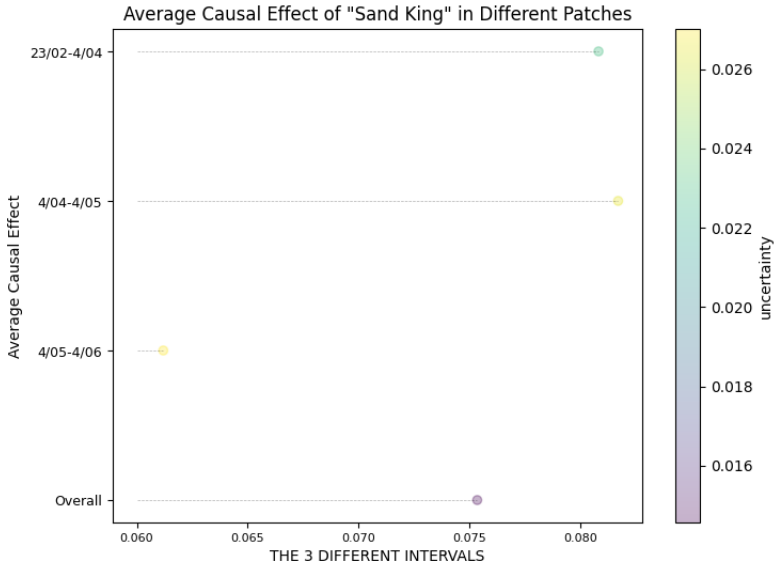


Figure 2: Average Causal Effect for Sand King

The first interval reworks Sand King's abilities, some positively and some negatively, resulting in the average causal effect being 0.08%. So initially Sand King sits on a positive causal effect and it continues into the second interval by being almost the same, something that makes sense as no changes were made to Sand king in that interval. The third interval correctly shows a decrease in the average causal effect as Sand King was nerfed thus making him weaker with an average causal effect of 0.06%. From this, a correlation between nerfs and decreasing the average causal effect can be understood. Sadly, this is not definite for every "Hero", as we've seen above that the causal effect might decrease without having changes to the "Hero" directly.

- Is the outcome of the game dependent on the selection of hero statistically? The outcome is indeed dependent on the selection, as mentioned in section 4 by using the Chi-squared test the relationship status between the 2 variables was confirmed. Now, why is it the case? People that play DotA 2 or MOBAs, in general, would be able to tell you immediately that a game is won or lost from just the draft phase (the "Hero" selection phase). Of course, one could also just calculate the probability of winning a game depending on the team composition. Both of these reasons suggest that selecting the right "Hero" has a huge impact on the game outcome. Now the importance of this test came from the fact that if the test using randomized data gives back the correct result, then the randomized data can also be used to test for different kinds of variables, such as checking how update intervals affect the game outcome. This is something that is discussed next.

- Why are game outcomes dependent on the update interval for some heroes and some not? Or in other words, why is the game outcome sometimes dependent on the update

intervals? To discuss the results of the tests 2 specific heroes are chosen, which produce different results. Firstly take a look at "Dragon Knight", at table 1 in section B. It can be seen that the game outcome for this hero is dependent on the interval. Using the test resulted in a p-value of 0.00188. This means that for "Dragon Knight", the game won or lost had most to do with the patches themselves and not with the Hero. This is reinforced by the fact that the Hero was changed 2 times, once in patch 7.31 and once in 7.31C. This can also be seen from figure 3 and figure 5 in section A which show the drastic change of the average causal effect for Dragon Knight between the patches. In 7.31 Dragon knight had an average causal effect of -0.03% and in the final patch 7.31C 0.08%. Taking both results into account (chi test results, and the causal effect) it can be assumed that Dragon Knight was worse for winning in the first interval and became better in the third (because of balancing and buffs) thus the game outcome being dependent.
Secondly, once again look at "Sand King". The game outcome is independent of patch intervals for "Sand King" and results in a p-value of 0.89. As discussed above, the average causal effect for Sand King is 0.08% in Patch 7.31 and goes down to 0.06% in 7.31C resulting in a difference of only 0.02% in comparison to Dragon Knight which there was a difference of 0.11%. From this, it is understood that the game outcome is only dependent on the intervals if there are drastic changes done to the "Hero", or changes that increase or decrease the average causal effect by a significant margin. If we would see drastic changes in the heroes' win rate between patches, it would strongly suggest that our tests were correct and accurate, but the win rate is only available between versions (and not patches) of the game thus not observable for this specific case.

- Interesting Observations. See the top 5 and bottom 5 heroes regarding their causal effect. Of course, this changes between intervals but not necessarily. Take a look at section A figure 6. The top 5 heroes are Sniper, Silencer, Necrophos, Bloodseeker, and SandKing. The bottom includes Tinker, Invoker, Meepo, Broodmother, and Earth Spirit. But what common qualities do the top heroes have and what do the bottom have? These can be attributed to many factors but the significant one that we can say for sure is that the top 5 heroes are more famous and easy to play in comparison with the bottom five. For example "Sniper" is one of the most beginner-friendly heroes in comparison to "Meepo" or "Earthspirit" which are amongst the hardest Heroes to play. This shows how the bias can be seen in the average causal effect for hard heroes where it's significant.
  Another interesting observation about the uncertainty of the values can be seen at figure 3, figure 4, figure 5, and figure 6 at section A. The uncertainty of the average causal effect of the heroes that have a positive effect is more than those with a negative average causal effect. The color between the 2 graphs in each interval might be deceiving, as the yellow color in the top figures is roughly close to 0.025 and the blue to green color in the bottom figures is close to 0.0235-0.0230. It should be noted that the only Hero that the uncertainty is zero, is "Techies", as no data exist thus, the value of 0 average causal effect being definite.

7 Conclusions and Future Work

Randomization offers a lot of benefits when applied correctly, but its not always easy to come across instances of randomization in real life. Even though DotA 2 provides such playground,

where full randomized data are abundant, the results from applying basic causal inference methods (like calculating the average causal effect) might sometimes be deceiving. Take for example the comparison between Meepo and Abaddon at section 6. The fact that Meepo had -12% average causal effect even though his win rate is positive indicates how the biases (if significant) from the randomized data, in this case it could be the player skill, sometimes slip through the calculations. In addition, because of the nature of the game, not all changes to heroes were visible as a change in the average causal effect and not all changes in the average causal effect were justified by the patch notes (see the Bane example, first 2 intervals at section 6). When the randomized data were used to check the Independence relations, the tests gave logical results to Heroes that had received a lot of re balancing and radical change to their average causal effect, in addition to heroes with less radical changes. This means that through randomized data the change between intervals was captured, and see for which ones the change was significant. Now to answer the research question, "How and If, matches with instances of randomization can be useful for predicting events using causal inference in DotA 2". Instances of randomization in DotA 2 with causal inference can be helpful and useful to predicting certain events but with caution as the amounts of data and the biases(if significant) of the randomized data can influence the results.

The experiment is heavily based on the nature and the amount of the data, as these are really important factors as to how accurate our calculations are. One main issue that exists is the difference between the amounts of data for the different intervals. This means that even though sometimes the difference in the average causal effect between intervals is justified by the patch notes, sometimes just the pure difference in amounts of data would result in a less accurate calculation for a certain patch. Moreover, when a factor like player skill is too significant for a Hero, then the randomized data fail to filter out that bias. One possible solution to this would be to take even more amounts of data, but its still not determined that biases will be completely filtered out.

To conclude, randomization and causal inference in DotA 2 is definitely interesting, as temporal changes between intervals were captured and portrayed, by both the average causal effect and the Chi-tests. But as the significance of the biases scales with the amount of data needed not all changes can be captured accurately and not all predictions represent the accurate state of the game at the time. We end by giving out a positive suggestion towards using randomization for predicting events, but one needs caution as big biases and external factors are not always filtered out in addition to almost never knowing how much data is enough.

A Average Causal Effect and Variance

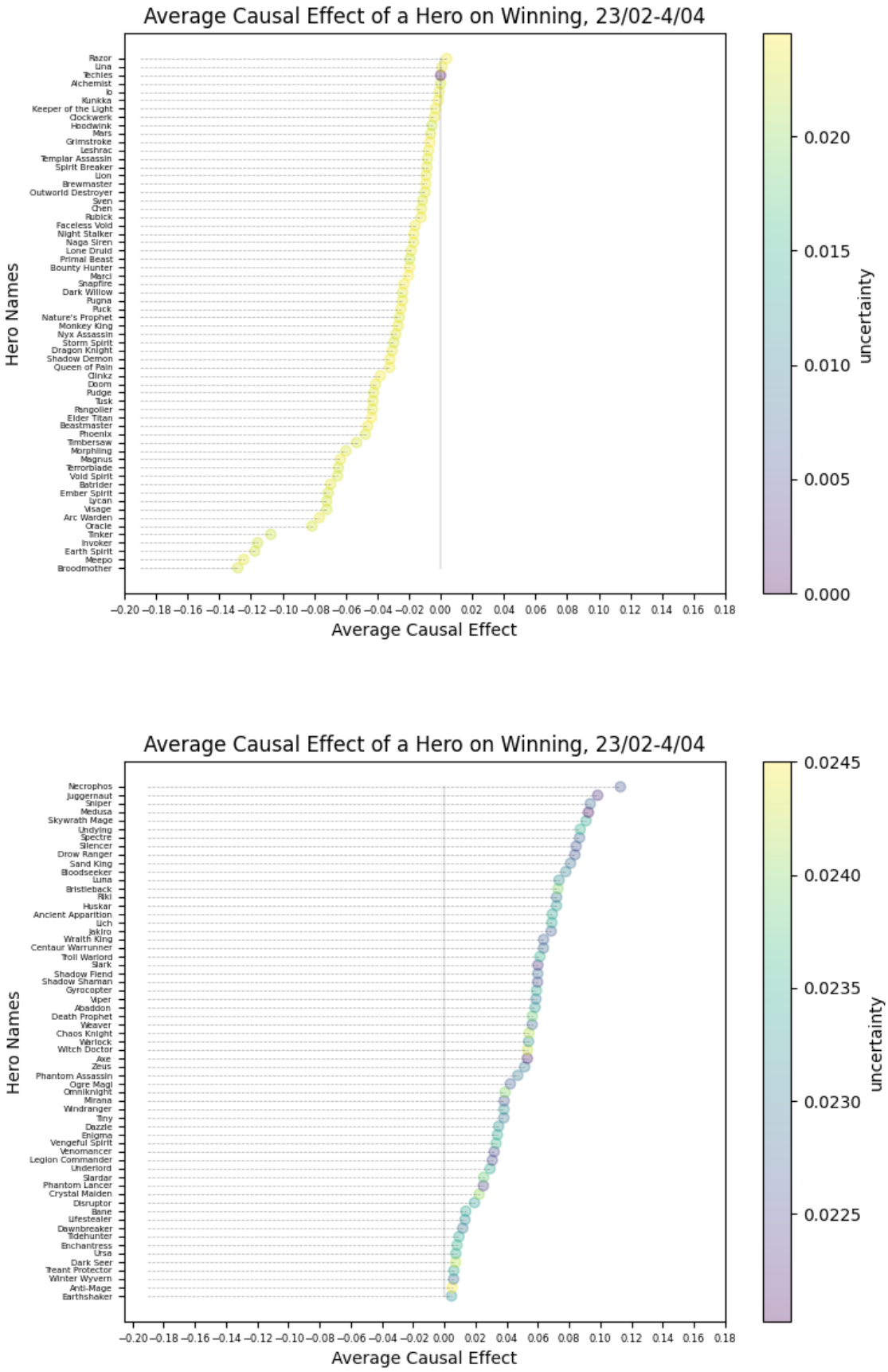
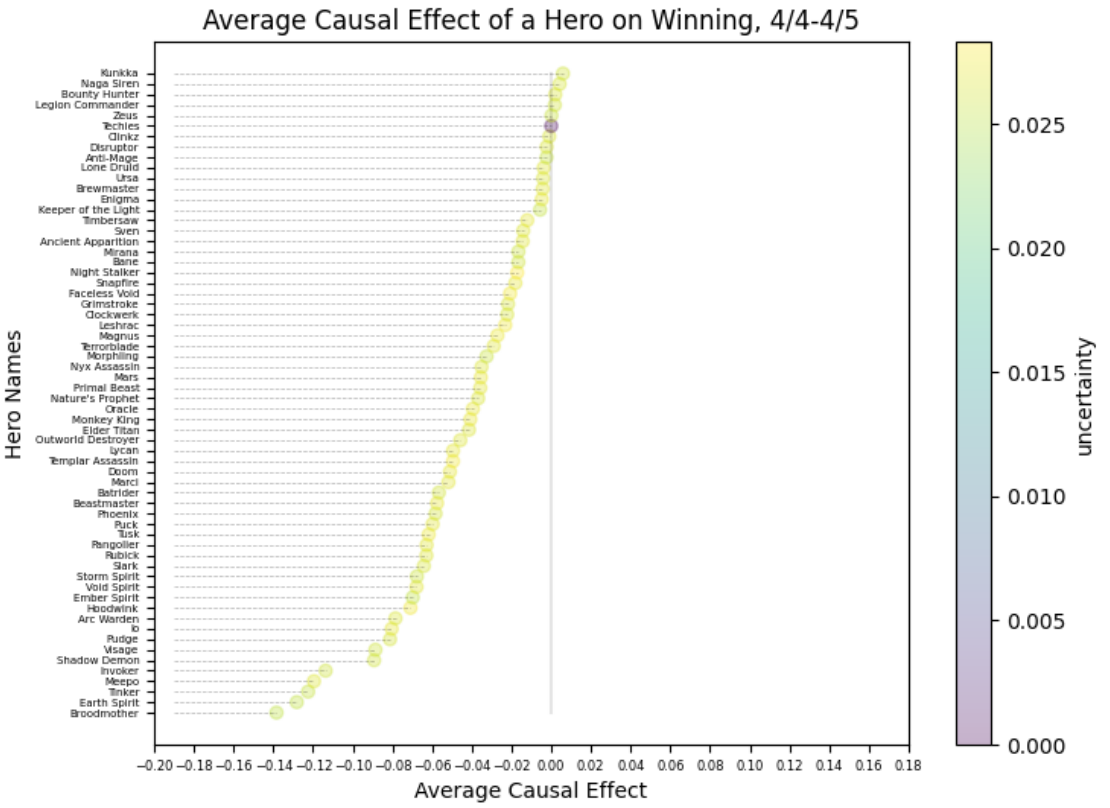


Figure 3: Average Causal Effect for First/Oldest Interval



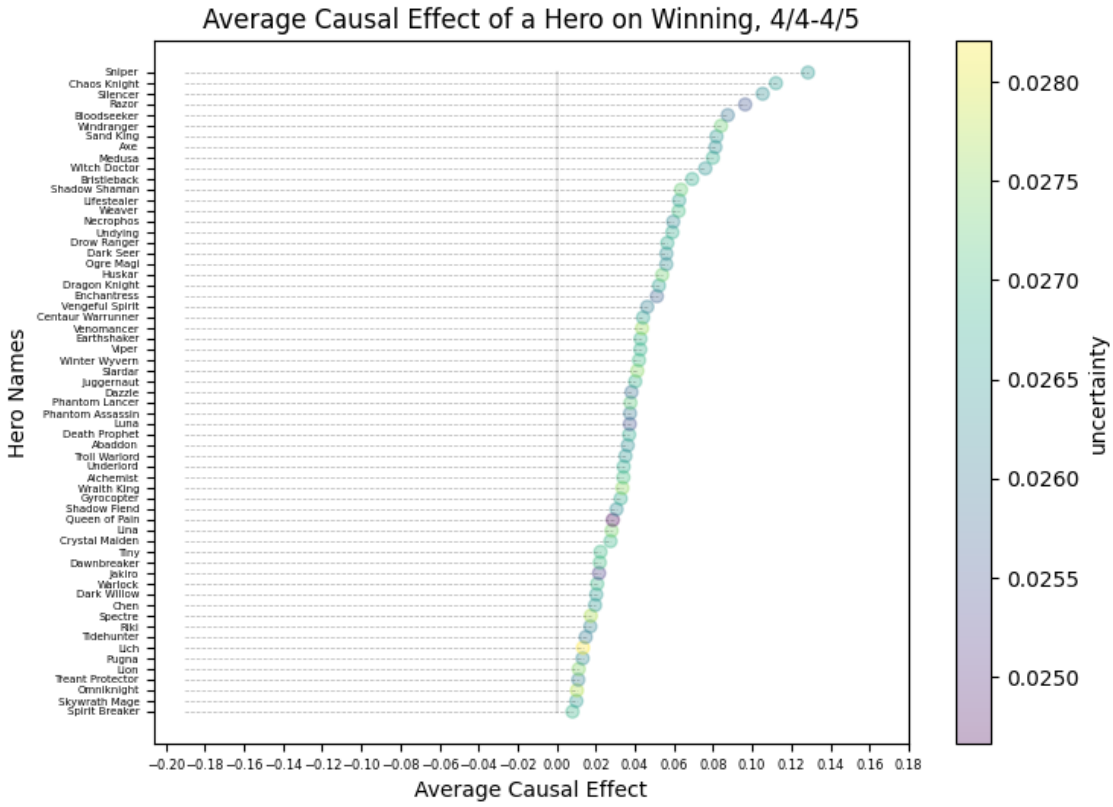


Figure 4: Average Causal Effect for Second Interval

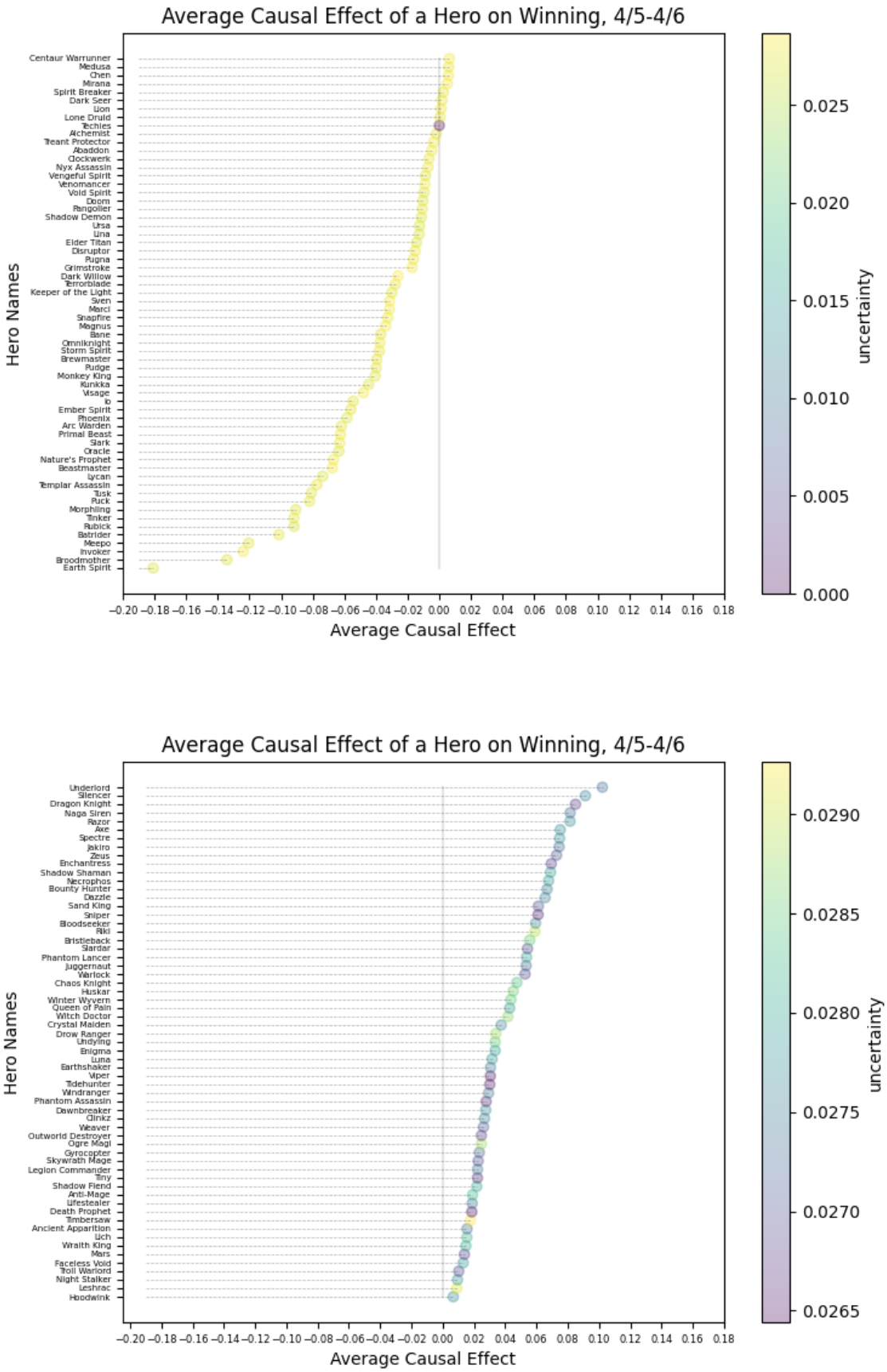


Figure 5: Average Causal Effect for Third Interval

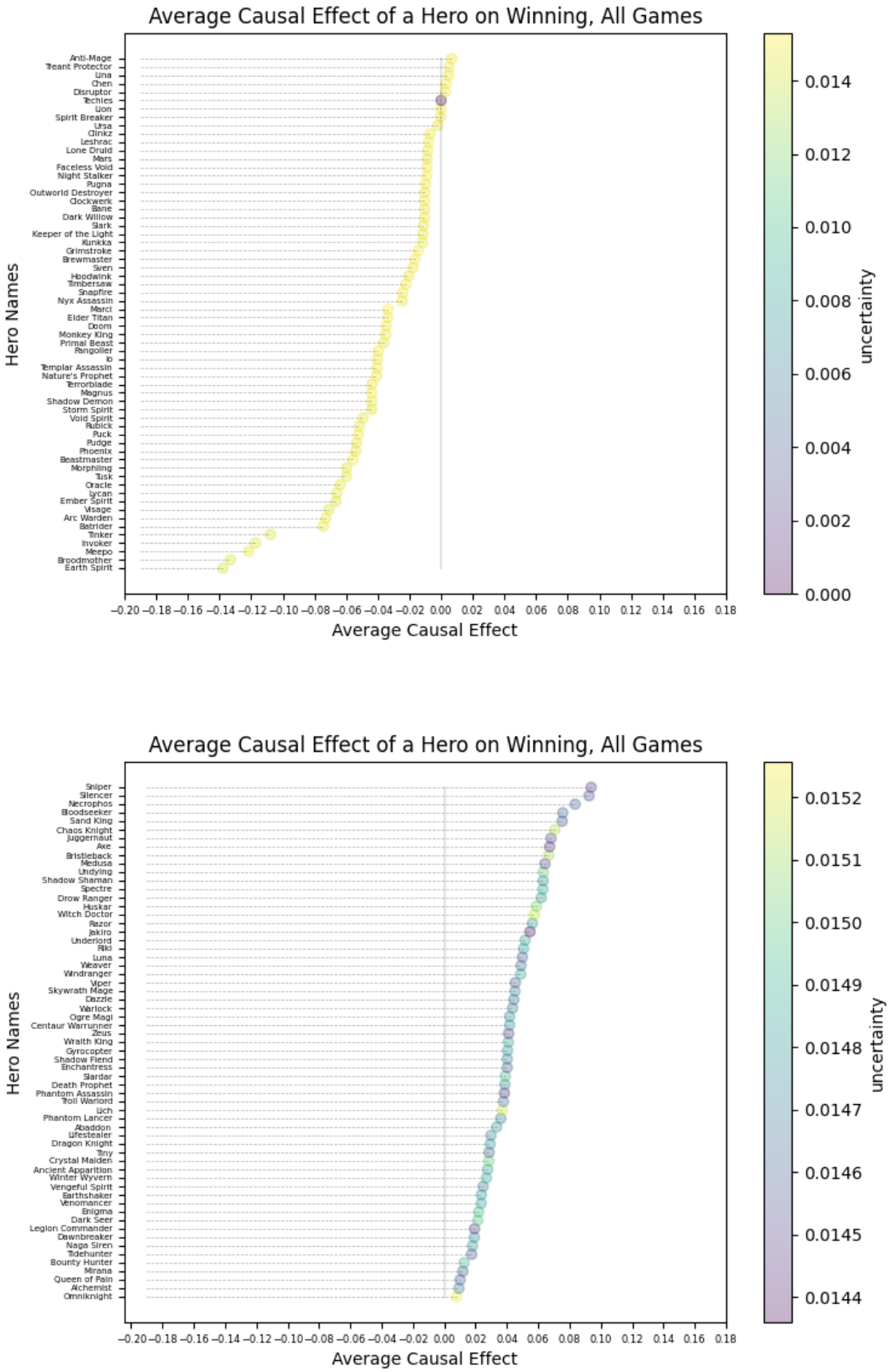


Figure 6: Average Causal Effect for all Games

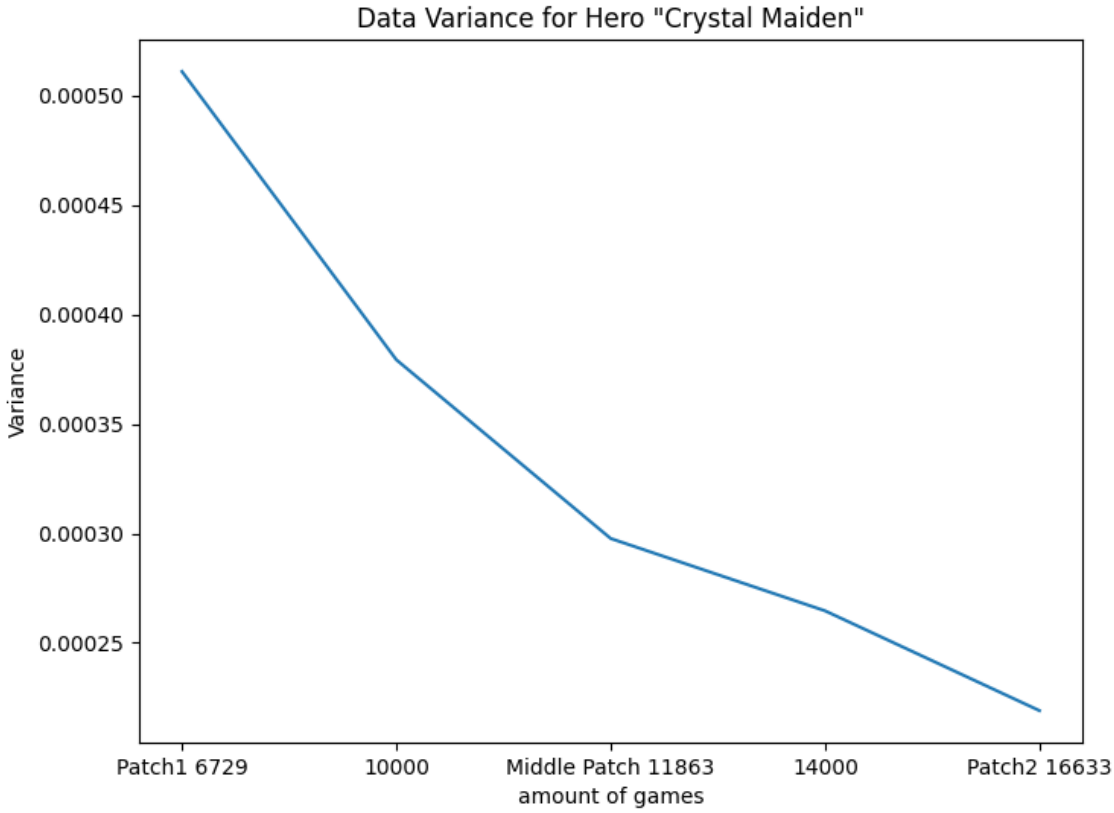


Figure 7: The sample variance concerning potential outcomes for a singular Hero

B Chi Squared Test

Hero Name	Dependence	P-value
Axe	Independent	0.56962043
Bane	Independent	0.466078041
Bloodseeker	Independent	0.670488166
Crystal Maiden	Independent	0.807910852
Drow Ranger	Independent	0.480585061
Earthshaker	Independent	0.389703128
Juggernaut	Independent	0.264885709
Mirana	Independent	0.298505496
Morphling	Independent	0.249879236
Shadow Fiend	Independent	0.582653037
Phantom Lancer	Independent	0.490452424
Puck	Independent	0.279415522
Pudge	Independent	0.415588241
Razor	**Dependent**	0.005679725
Sand King	Independent	0.894363746
Storm Spirit	Independent	0.624388173
Sven	Independent	0.913770552
Tiny	Independent	0.95787503
Vengeful Spirit	Independent	0.281182819
Windranger	Independent	0.261646332
Zeus	Independent	0.159221051
Kunkka	Independent	0.31773053
Lina	Independent	0.527710739
Lion	Independent	0.642733113
Shadow Shaman	Independent	0.888100225
Slardar	Independent	0.534772138
Tidehunter	Independent	0.703605516
Witch Doctor	Independent	0.602075772
Lich	Independent	0.322130367
Riki	Independent	0.297242276
Enigma	Independent	0.577467914
Tinker	Independent	0.709202271
Sniper	Independent	0.239988181
Necrophos	Independent	0.288761339
Warlock	Independent	0.656633689
Beastmaster	Independent	0.91999123
Queen of Pain	**Dependent**	0.031307474
Venomancer	Independent	0.394589098
Faceless Void	Independent	0.508688488
Wraith King	Independent	0.497941806
Death Prophet	Independent	0.672057538
Phantom Assassin	Independent	0.924345522
Pugna	Independent	0.431164912
Templar Assassin	Independent	0.201395855
Viper	Independent	0.788152481
Luna	Independent	0.49886552
Dragon Knight	**Dependent**	0.001882665
Dazzle	Independent	0.488417393
Clockwerk	Independent	0.906820666
Leshrac	Independent	0.741672366
Nature's Prophet	Independent	0.635756376
Lifestealer	Independent	0.221385363
Dark Seer	Independent	0.204170067
Clinkz	Independent	0.123819347
Omniknight	Independent	0.154036055
Enchantress	Independent	0.102849489
Huskar	Independent	0.848348945
Night Stalker	Independent	0.616830558
Broodmother	Independent	0.987495705
Bounty Hunter	**Dependent**	0.027252518
Weaver	Independent	0.594270605
Jakiro	Independent	0.290698566
Batrider	Independent	0.516497167
Chen	Independent	0.53385328
Spectre	Independent	0.154257337
Ancient Apparition	Independent	0.05955837
Doom	Independent	0.559417616
Ursa	Independent	0.917494836
Spirit Breaker	Independent	0.79702482
Gyrocopter	Independent	0.741914592
Alchemist	Independent	0.456548624
Invoker	Independent	0.934172837
Silencer	Independent	0.67766424
Outworld Destroyer	Independent	0.160058359
Lycan	Independent	0.630218348
Brewmaster	Independent	0.672939026
Shadow Demon	Independent	0.091735939
Lone Druid	Independent	0.631604318
Chaos Knight	Independent	0.120099635
Meepo	Independent	0.963373132
Treant Protector	Independent	0.952421426
Ogre Magi	Independent	0.703124498
Undying	Independent	0.385690103
Rubick	Independent	0.100752744
Disruptor	Independent	0.756278119
Nyx Assassin	Independent	0.727067063
Naga Siren	**Dependent**	0.008201527
Keeper of the Light	Independent	0.736840623
Io	Independent	0.10910301
Visage	Independent	0.489987325
Slark	**Dependent**	0.000247593
Medusa	**Dependent**	0.043177485
Troll Warlord	Independent	0.37543456
Centaur Warrunner	Independent	0.306867815
Magnus	Independent	0.430774344
Timbersaw	Independent	0.067795316
Bristleback	Independent	0.913348013
Tusk	Independent	0.5776283
Skywrath Mage	Independent	0.050422044
Abaddon	Independent	0.279169601
Elder Titan	Independent	0.624076053
Legion Commander	Independent	0.732488537
Techies	Independent	
Ember Spirit	Independent	0.775702878
Earth Spirit	Independent	0.213143759
Underlord	**Dependent**	0.049702054
Terrorblade	Independent	0.355946542
Phoenix	Independent	0.995060298
Oracle	Independent	0.409807002
Winter Wyvern	Independent	0.308880409
Arc Warden	Independent	0.789222228
Monkey King	Independent	0.967358968
Dark Willow	Independent	0.276721577
Pangolier	Independent	0.363667491
Grimstroke	Independent	0.982302183
Hoodwink	Independent	0.09860066
Void Spirit	Independent	0.17670058
Snapfire	Independent	0.939806374
Mars	Independent	0.533369268
Dawnbreaker	Independent	0.77165484
Marci	Independent	0.747597894
Primal Beast	Independent	0.572458016

Table 1: Shows the statistical independence between the patch intervals and the game outcome

References

[1] R. J. Hernan MA, *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC, 2020, vol. Chapter 3.a.

[2] A. H. Christiansen, E. Gensby, and B. Weber, "Deployment of causal effect estimation in live games of dota 2," *IEEE Transactions on Games*, pp. 1–1, 2021.

[3] M. F. Alves, *Causal Inference for the Brave and True*, 2021.

[4] D. B. Rubin, "Causal inference using potential outcomes," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005. [Online]. Available: https://doi.org/10.1198/016214504000001880

[5] K. Imai, "Causal inference," 2013.

[6] P. Ding, "Exploring the role of randomization in causal inference," 2015.

[7] Essramos. (2018) How to use opendota api. [Online]. Available: https://gist.github.com/essramos/dbac40593b64e2193f2be68232f86b58

[8] (2018) Open dota api. [Online]. Available: https://docs.opendota.com/

[9] S. Khandelwal. (2020, Jul) Access dota game statistics using python. [Online]. Available: https://saket404.github.io/python/access-dota-stats-using-python/

[10] N. van Geloven, S. A. Swanson, C. L. Ramspek, K. Luijken, M. van Diepen, T. P. Morris, R. H. H. Groenwold, H. C. van Houwelingen, H. Putter, and S. le Cessie, "Prediction meets causal inference: the role of treatment in clinical prediction models," *EUROPEAN JOURNAL OF EPIDEMIOLOGY*, vol. 35, no. 7, pp. 619–630, JUL 2020.

[11] M.-A. C. Bind and D. B. Rubin, "Bridging observational studies and randomized experiments by embedding the former in the latter," *Statistical Methods in Medical Research*, vol. 28, no. 7, pp. 1958–1978, 2019, pMID: 29187059. [Online]. Available: https://doi.org/10.1177/0962280217740609

[12] E. Yildiz, J. Safyan, and M. Harper, "User sentiment as a success metric: Persistent biases under full randomization," in *KDD '20: PROCEEDINGS OF THE 26TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING*. Assoc Comp Machinery; ACM SIGMOD; ACM SIGKDD, 2020, pp. 2891–2899, 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), ELECTR NETWORK, AUG 23-27, 2020.

[13] D. B. Rubin, "For objective causal inference, design trumps analysis," *ANNALS OF APPLIED STATISTICS*, vol. 2, no. 3, pp. 808–840, SEP 2008.

[14] "Dota 2 wiki." [Online]. Available: https://dota2.fandom.com/wiki/Patches

[15] "Pearson's chi-squared test," Apr 2022. [Online]. Available: https://en.wikipedia.org/wiki/Pearson27s_chi-squared_test

[16] "Chi-square test for goodness of fit in a plant breeding example," 2022. [Online]. Available: https://passel2.unl.edu/view/lesson/9beaa382bf7e

[17] "Dota 2 heroes," 2022. [Online]. Available: https://www.dota2.com/heroes

[18] "Dotabuff-dota 2 statistics," 2022. [Online]. Available: https://www.dotabuff.com/