Estimating the effect of 'diverse' team compositions on Dota 2 game outcomes using Inverse Probability Weighting

Christof Goedhart
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

Estimating the effect of 'diverse' team compositions on Dota 2 game outcomes using Inverse Probability Weighting

Christof Goedhart[1]
Supervisor(s): Rickard Karlsson[1] , **Jesse Krijthe**[1]
[1]EEMCS, Delft University of Technology, The Netherlands
c.n.goedhart@student.tudelft.nl, R.K.A.Karlsson@tudelft.nl, J.H.Krijthe@tudelft.nl

Abstract

Commonly, when researchers are figuring out the effect of a putative cause, additional variables influence the cause and the effect. These are called confounders, and they obfuscate causal relationships. Inverse Probability Weighting is a method that can be applied to remove confounding and show a causal effect. This study aims to determine if Dota 2 game outcomes can be predicted based on team composition diversity and if 'Inverse Probability Weighing is a helpful tool for this. First, metrics to assess team diversity were determined, and the two confounders, "Player skill" and "Hero skill", to the 'is-diverse' treatment and game outcome were identified. Next, a dataset with Dota 2 games was gathered containing all the relevant variables to measure the confounders. Finally, inverse probability weighting was applied to measure the Average Treatment Effect of the putative cause.

The results suggest that team diversity significantly impacts the game outcome. However, the results were close in most cases when comparing the results with data that was not corrected for by confounders. A possible explanation may be that the confounders didn't have enough influence on some diversity metrics since they were too vague. For example, the most complex diversity metric, which covered the most team-composition characteristics, showed a clear difference between the average treatment effect with inverse probability weighting being applied and not applied.

1 Introduction

Imagine a scenario where a researcher is conducting an experiment to measure the effect of 'smoking' on life expectancy. He notices that people who smoke have a reduced life expectancy. Can the researcher then conclude that smoking reduces life expectancy? It might seem intuitive, but one cannot know for sure. It could also be the case that smokers are more likely to engage in other unhealthy behaviour, such as drinking or not doing sports, which also influence the life expectancy. Because of this, one can overestimate the effect that smoking has on life expectancy. In this scenario, we call unhealthy behaviour a confounding factor. A confounder can be defined as a variable associated with both the putative cause and with its effect [13]. These confounding factors don't just arise in health-related experiments; they can influence any experiment where one tries to measure the effect of a treatment. Therefore, dealing with them is important to say something about cause and effect.

In the previous example you can also see a challenge in the study of causal inference, namely that we don't know what would have been the outcome if someone in the experiment that smokes had never smoked before or the inverse, in short: there is a missing data problem. In a study by P. Dinga and F. Li they state the following about it: "The potential outcomes framework is a main statistical approach to causal inference, in which a causal effect is defined as a comparison of the potential outcomes of the same units under different treatment conditions. Because for each unit at most one of the potential outcomes is observed and the rest are missing, causal inference is inherently a missing data problem." [5]. There are multiple ways to exclude confounding including randomization, restriction and matching, which are all applicable at the time of study design [9]. However, this is not always possible, so researchers need to rely on statistical models to adjust for confounders [9]. Inverse Probability Weighting is one of those statistical models whose major use is with missing data. The goal is to see if Inverse Probability Weighting can be applied to correct for confounding factors in a complex game like Dota 2.

Dota 2 is a MOBA (Multiplayer Online Battle Arena), where two teams of 5 players each, face off against each other in real time. Games take about 40 minutes, on average [7]. Each team is a combination of 5 of the 123 possible heroes, hence every game is unique. Due to the nature of the game there are a lot of complex causal relationships, which makes it an interesting case for the field of 'Causal Inference'. A predominant question that often arises in the context of Dota 2 is the following: What team compositions are good? What combinations of heroes increase your chances of winning? Heroes in Dota 2 can be put into specific categories, for example, each hero has an attack

type: ranged or melee, and also have an attribute which is strength, agility, and intelligence. Aside from that, a hero can belong to a multitude of roles: “engage”, “pusher”, “support”, etc. The general consensus among players is that if you have a team with “more capabilities”, or a team that is more “diverse” then your team should be better overall.

In this report it will be determined if Dota 2 game outcomes can be predicted based on team composition diversity, and if Inverse Probability Weighing is a useful tool for this. Questions that come to mind are: ’How does one model “team diversity”? or ’What additional variables does one need to take into account in the previously described causal relationship?’.

2 Methodology

To come to a conclusion, the following steps are performed.

2.1 Causality analysis

First, the game is analysed to identify how it works and what the basic human intuition is on what is perceived to be a ’good’ team and what is perceived to be a ’bad’ team. A good understanding of the game is important to understand the game’s casual relationships and model them. Since the aim is to measure the effect that the treatment “diverse team composition” has on “winning a game” the confounders that influence this relationship are identified. When selecting possible confounders, it’s kept in mind that they need to be established before the game starts, since otherwise they cannot have an effect on establishing a team composition.

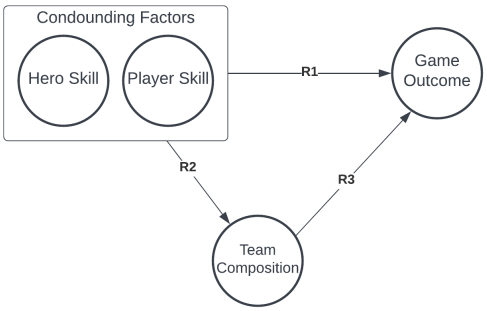

Figure 1: Causality Graph

Two main confounders are identified:

- **Player skill:** It seems sensible that players with a higher skill level than their opponents, have a greater chance of winning (R1 in figure 1). It also seems intuitive to assume that a team with a higher skill level is better at drafting a ’diverse’ team composition since players generally assume a certain distribution over the multiple roles is better (R2 in figure 1).
- **Hero skill:** The intuition behind ’Hero skill’ as a confounding factor is that if a player has more skill with a specific hero, it is more likely that he picks that hero (R2 in figure 1). The skill with a hero also influences the game result since if everybody on a team is better at playing their heroes than their opponent, they have a bigger chance of winning (R1 in figure 1).

2.2 Team Diversity metrics

A specific metric for Team Diversity needs to be specified. Specifying what a “Diverse Team” means in a Dota 2 context can be complex. In this research, a discrete approach is used, meaning that a team can be diverse or not; there is no in-between. See section 5.4 for the continious altenative. To look at what makes the team diverse, the characteristics of a single hero are considered first. Dota 2 categorizes heroes in the following way [1]:

- **Attack-type:** Ranged, Melee
- **Attribute:** Intelligence, Agility, Strength
- **Role:** Carry, Nuker, Initiator, Disabler, Durable, Escape, Support, Pusher, Jungler

For example: ’Gyrocopter’ is a ranged, agility hero with the roles of Carry, Nuker and Disabler. While ’Ogre Magi’ is a melee, intelligence hero with the roles of Support, Durable, Initiator, Nuker and Disabler. With these hero characteristics, team diversity metrics are created. See Table 1.

Table 1: Diversity Metrics

Nr.	Metric condition
01	Team contains 3 or less carry roles
02	Team contains each of the Attributes at least once
03	Team contains less than 4 melee, and less than 4 ranged heroes
04	Team contains more than 12 non-support and non-carry roles
05	Metric 01 & Metric 04
06	Metric 01 & Metric 02 & Metric 03 & Metric 04

2.3 ATT & Inverse Probability Weighting

The goal is to calculate the average treatment effect (ATE) [6]. Where Y_1 is the potential outcome with the treatment and Y_0 is the potential outcome without the treatment.

$$ATE = E[Y_1 - Y_0] \tag{1}$$

Before it is calculated, the data is corrected by the previously identified confounders using Inverse Probability Weighting (IPW). In the paper “Constructing Inverse Probability Weights for Marginal Structural Models” from 2008 [4], the following is said: “The method of inverse probability weighting can be used to adjust for measured confounding and selection bias under the four assumptions of consistency, exchangeability, positivity, and no misspecification of the model used to estimate weights”[4]. The mentioned assumptions need to hold for the data to be able to conduct the experiment. Finally, to use IPW, a model is needed to calculate the probability of data that is missing [10]. In this case, logistic regression is applied since a discrete treatment is used. A logistic regression model uses a sigmoid function, so the output is always a probability value between 0 and 1. In

this model, the features are the confounders, and the probability of treatment is predicted. This probability value's inverse is the weight assigned in the pseudo population. So a low probability of 0.2 would get a weight of 5, and a high probability of 0.8 would get 1.25. The pseudo population is the original population reweighted. J. M. R. Miguel and A. Hernan mention: "The expected mean of the weights W^a is 2 because, heuristically, in the pseudo-population, all individuals are included both under treatment and under no treatment." [8]. This means the pseudo-population should be around double the size of the original. All this being said the pseudo-population is constructed using IPW and given the previous assumptions, this pseudo-population has the following properties (Where Y is the outcome and A the treatment)[8]:

- The mean of Y^a is the same in both populations
- Unconditional exchangeability (i.e., no confounding) holds in the pseudo-population.
- The counterfactual mean $E[Y^a]$ in the actual population is equal to $E_{pseudo-pop}[Y|A = a]$ in the pseudo-population
- Association is causation in the pseudo-population

Since theoretically, association is now causation. The ATE is calculated on the pseudo-population, and the results for different metrics are compared to see the effect of a diverse team composition on the game outcome.

3 Experiment setup and Results

3.1 Data Collection

The data collected for the experiment is taken from the Open Dota API [3]. It's a publicly available API that tracks match and player data. The 'game explorer' section of the API, which accepts SQL queries, was used to collect the games. This contains fewer games than is available in other sections of the API, but all the games have complete player information. This means that all the game players have their profiles and game statistics public. This is important since access to that is needed to get the confounding factors. Additionally, one should remember that his restriction might introduce some bias in the results if there is a connection between how players play the game and how they go about online data privacy. The collected games also don't go further back than about a month in real-time. This is because data such as a player rank is only accessible at this moment in time and not historically. Going further back in time would make it less accurate. With this list of all the games, all the other player and hero information is requested from other parts of the API.

The final data set consists of 794 teams from 397 games, so half the teams lost, and half the teams won. The games are all in the same game mode for consistency, namely captains mode. The teams in the dataset come from a wide range of skill levels, as seen in figure 2. The team rank is the mean of the estimated individual rankings in a team. Since not every player in Dota 2 has a rank, the game estimates the rank for each player.

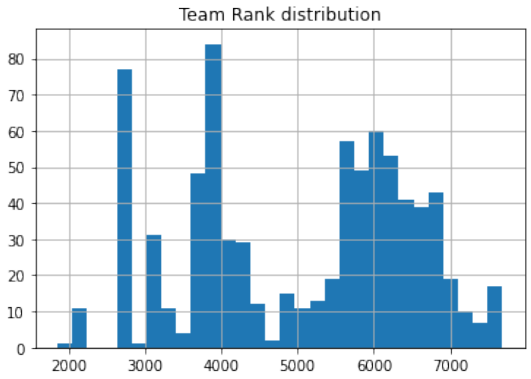

Figure 2: The Team Rank Distribution of all the teams

3.2 Experiment

With the data, the 2 confounders 'player skill' and 'hero skill' need to be represented in some form and it needs to be per team. The game tracks the 'estimated MMR' per player, this can be used as an indicator of the skill of a player. MMR stands for Matchmaking Rank. The average of team MMR is taken to get the MMR per team. With that the team-mmr compared to the other team can be calculated, see the equation below. 'team-mmr-difference' is the first confounder, it follows a normal distribution, see figure 3. It is normalised to the average MMR of the opponent since a high value means little if the opponent's MMR is much higher.

$$\text{team-mmr-difference} = \frac{AVG_t}{AVG_t + AVG_o} \tag{2}$$

where:

AVG_t	average team MMR
AVG_o	average team MMR of the opponent

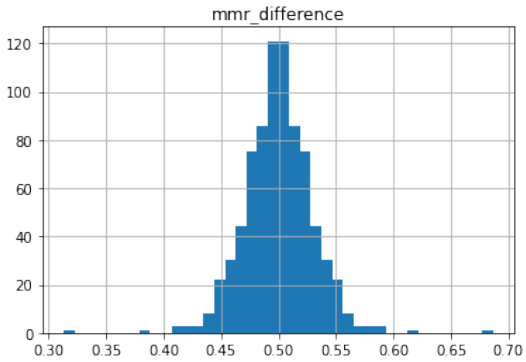

Figure 3: the team-mmr-difference distribution

The second confounder is 'hero skill'. To track that, the attribute 'hero score' in Dota 2 is used, which is used as an indicator to show how proficient someone is with a specific hero. The calculation off hero score takes into account factors such as: win rate, matches played, KDA ratio (Kills Deaths Assists) and ranking division [2]. For each team the Average

Hero score is measured by taking the mean of the hero scores and that is used as an indicator of the 'Hero skill' of a team. The distribution of this confounder can be seen in Figure 4.

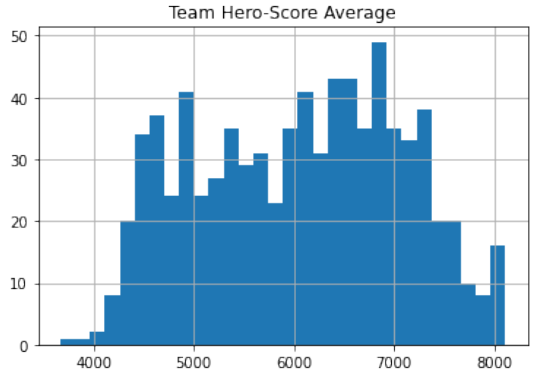

Figure 4: The average hero score distribution

To be able to apply IPW the following assumptions need to be true: exchangeability, positivity and consistency. Exchangability means that there is no unmeasured confounding, this is hard to prove but it can be reasoned about. It is assumed that this is the case here since the 2 major areas of player and hero are covered, more about this in the Discussion section. Positivity is the condition that there are both exposed and unexposed individuals at every combination of the confounders [12]. This holds here since each team has a chance of being qualified as diverse, if they choose the right heroes. That leaves consistency which can be formally defined the following way: "Consistency means that a subject's counterfactual outcome under her observed exposure history is precisely her observed outcome" [4]. In other words there aren't multiple versions of the same treatment, which is the case here since each diversity metric is precisely defined and applied the same way on all the data.

With the assumptions satisfied, IPW is now applied. To compute the probability of treatment given the confounders, logistic regression is used with an 'LBFGS' solver, L2 regularization and no class weighing. The model is trained on the same data-set that the experiment is conducted on. For each team the probability of treatment is estimated, and with the inverse of that probability the pseudo-population is constructed. The size of the psuedo-population is now 1588 (a few decimal places off), double the size of the original population which consisted of 794 teams. The ATE of the different diversity metrics are calculated on the psuedo-population. See Section 3.3.

3.3 Results

Table 2 shows the results of Average Treatment Effect, corrected for confounders with Inverse Probability Weigthing. Table 3 shows the results on the original population without confounder correction. The first column shows the ATE calculated on the original population. To get an understanding of the uncertainty of this result bootstrapping is used to get the mean, standard deviaton and the 95% confidence intervals assuming a normal distribution. The bootstrapping of the ATE was done with 1000 resamples and with replacement, the logistic regression model was fit again for each sample . The metrics that are used can be seen in table 1.

Table 2: Results with correction for confounders

	ATE	Mean	StDev	95% CI	% with treat.
1	5,44	5,31	4,19	(5.05, 5.57)	74,69
2	-1,00	-0.78	5,02	(-1.09, -0.46)	75,82
3	-0,30	-0.33	3,74	(-0.57, -0.1)	66,88
4	4,63	4,70	3,54	(4.48, 4.92)	54.53
5	6,78	6,72	3,69	(6.49, 6.95)	39,04
6	7,72	7,80	4,48	(7.52, 8.08)	19,14

Table 3: Results without correction for confounders

	ATE	Mean	StDev	95% CI
1	6,33	6,19	4,18	(5.93, 6.45)
2	0,00	0.24	4,8	(-0.06, 0.53)
3	0,28	0,26	3,73	(0.03, 0.49)
4	4,83	4,91	3,54	(4.69, 5.13)
5	6,35	6,29	3,70	(6.06, 6.52)
6	5,70	5,76	4,48	(5.48, 6.04)

When looking at these results in general, it can be seen that the metrics vary in their effect on the game outcome. For example, metrics 2 and 3 seem to have a small, almost negligible effect given the confidence interval. However, the remaining diversity metrics 1, 4, 5 and 6 significantly affected the game outcome, with metric 6 being the outlier where a team with this specific treatment has a 7.72% higher chance of winning than a team without it. In general, the ATE results in both tables lay close to the bootstrapping mean, giving it credibility. Another noticeable thing is that most of the ATEs in table 2 lay roughly within 0.5 of the ATEs in table 3, the exception being metric 6.

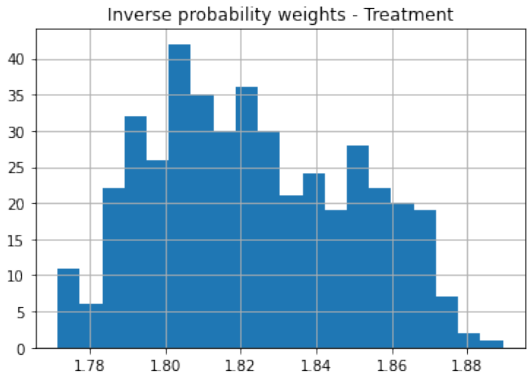

Figure 5: Metric 4 Distribution of the inverse of the probability for assigning treatment to treated teams

Let's look at why, for most results, the difference is negligible between ATE with and without correction for confounders. Taking metric 4 as an example, the logistic regression model estimates the probability of assigning the treat-

ment, for the treated teams, based on the confounders to be high. This can be seen in the weights in Figure 5; they range between 1.75 and 1.85 since they are the inverse of a high probability. In contrast, the probability of not assigning the treatment for the untreated teams is low, reflected in high weights ranging between 2.15 and 2.30.

However, Figure 6 shows an oddity. This figure shows the distribution of the probability of no-treatment for every team in the dataset (both the treated and untreated). The model has found zero cases in the population where the confounders indicate that non-treatment is equally or more likely for a team than treatment. This could be due to a lack of data which makes the model unable to create sufficiently accurate estimates (see Section 5.3), or due to confounders having little effect on the treatment, which in turn also creates an inaccurate logistic regression model (see Section 5.1).

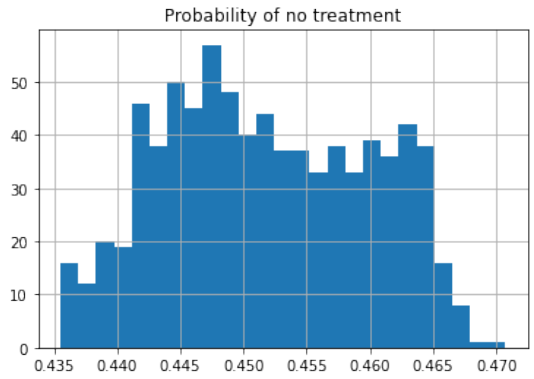

Figure 6: Metric 4 Distribution of the probability for assigning no treatment

Figure 7 shows that the ATE for metric 6 with confounders is still substantially higher, considering the confidence intervals. Figure 8 shows that the ATE distribution from the bootstrapping is symmetric and centred around the estimated ATE and mean. Since the ATE estimate looks reliable, what is the reason for this exception? It could be because the relationship between the confounders and treatment is stronger for this specific metric. Since this metric is the one that uses the most variables for a diverse team composition, the likelihood of players actively influencing this metric, compared to the other metrics, is higher.

4 Responsible Research

4.1 Data Privacy

The data collected from the Open Dota API is all player information from real people. One could look up the accounts of each player in the data set. However, all the available information is related to the game, not to anything outside the game such as date of birth. Dota 2 players can also choose to be anonymous so that their account identifier remains hidden. Although this is good for privacy, it means that significantly fewer data can then be used.

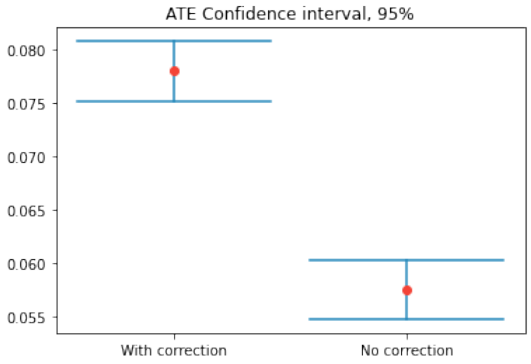

Figure 7: Confidence intervals from metric 6, with and without confounder correction

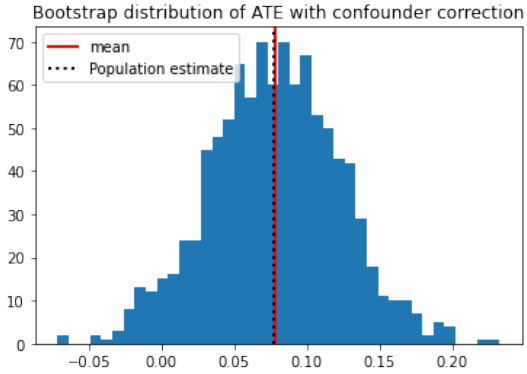

Figure 8: ATE bootstrap distribution of metric 6

4.2 Reproducible Research

The experiment is documented so that another researcher can replicate the steps. Since the API is publicly available, other researchers can gather a data set with all the mentioned features. Any processing of the data can be repeated, and the metrics for team diversity can also be copied. No external package or third-party code was used for performing Inverse Probability Weighting so that it's clear and there are no hidden aspects to the implementation that aren't documented. An outside package was used for logistic regression, but all the used parameters are specified. The dataset, the code for the analysis and code for IPW implementation can be found here: https://github.com/Christof2000/IPW-team-diversity-dota2. Additionally, a request for it can be sent to c.n.goedhart@student.tudelft.nl.

5 Discussion

5.1 Confounders

The results with and without correction do not differ by a big margin. This is interesting since the reasoning that is supplied as to why they are confounding seems logical. One factor that might play a role is that there is a difference between what is confounding and what one can measure. For example, there might be discrepancy between what the 'hero score' shows and what the actual hero skill is. The same could be said

when we talk about 'player skill'. Another thing that could be playing a role is that 'being better at the game' doesn't transfer to the diversity metric. If the player is not actively thinking about 'team diversity' in the hero selection process then the relation between the 'player skill' and the treatment becomes less clear. Another argument that could be made for the previously mentioned 'weak' relationship between skill and the treatment is that the team's skill is calculated relative to the other team. Since all the matches are balanced to a certain extent, and there aren't any big differences, one could ask the question: if my team is a bit better than the other team, will that difference show up in the hero drafting process? A solution to this might be to keep the skill value of the team absolute, but this might negatively influence the relationship between this confounder and the outcome since there is no longer an indication of the opponent's skill.

Another thing to consider is that the game data comes from a wide range of ranks. Therefore, it seems plausible that games on a higher level have players with more knowledge and games on a lower level might have players experimenting more with the game. This phenomenon could mean that when drafting heroes on a lower level, less thought is put into assembling a diverse team composition.

5.2 Exchangability

An assumption before applying IPW is that there is no unmeasured confounding, which is hard to prove since it can only be reasoned about. However, it can be said that the two confounders applied in this case cover two significant areas: player skill and hero skill. There are also confounders which seem likely to be confounding but weren't tracked here. An example is if the players play the game with a group of friends, they can better coordinate the hero selection phase and communicate better in-game, which might give them an advantage. The question then is if this makes our results irrelevant. Although maybe some of the less significant ATEs should be taken with a degree of scepticism, it seems unlikely for untracked confounders to be able to invalidate the best-performing metrics completely. It could reduce the ATE by a bit but there being enough missed confounders to diminish it by 6% is improbable.

5.3 Sample size

The population size of 794 teams used is on the small side. Bigger sample size would give tighter confidence intervals. Still, the current confidence intervals for most metrics deviate enough from a 0% ATE and are tight enough to make conclusions based on them. However, another thing to consider is that in the cases where IPW is applied, a logistic regression model is created based on this same population. That's probably where a bigger sample size could make a substantial difference; as seen in the results in section 3.3, the current model performance is questionable.

5.4 Discrete Treatment

In this case, the choice was made to use a discrete measurement of team diversity. Unfortunately, this choice means that information is lost when you convert a variable like the count of 'carry' heroes to a binary value; being close to the threshold has no meaning. A team is either diverse or not. This way of assigning treatment can affect the outcomes since it's not always accurate if a team is close to qualifying as diverse but not quite. A continuous treatment could be made using the same hero characteristics introduced in section 2.2. Although some steps are different, this approach would still be possible in combination with Inverse Probability Weighting.

Instead of estimating predicted probabilities of assigning treatment, conditional densities are used [11]. Furthermore, instead of logistic regression, a linear regression model is used to predict the continuous treatment based on the covariates and then obtain the conditional density of the predicted value [11]. The stabilized weights for a continuous treatment can be formally denoted in the following way [11].

$$\text{Stabilized weights} = \frac{\phi(E[Z])}{\phi(E[Z|X])} \tag{3}$$

where:

ϕ	probability density function
E	Expectation operator
Z	Outcome
X	Vector of covariates

Although the process is a bit more complex than the other variant of IPW, given sufficient data, it could give more accurate results.

6 Conclusion and Future Work

By estimating the effect of diverse team compositions on game outcome, this study established that the effect is significant, depending on the chosen diversity metric. However, the results corrected for confounders with IPW did not differ significantly from those obtained without considering confounders. This is probably because the relationship between the confounders and treatment was hard to measure, as mentioned in the Discussion. Furthermore, the metric that covered the most areas (Metric 6) was the exception, showing a difference between the corrected and non-corrected results. So to apply Inverse probability weighting in a Dota 2 context, the metric needs to be clearly influenceable by the confounders. That's why, in the future, it might be interesting to see the effect of a continuous diversity treatment on the game outcome since this would probably show more clearly how the confounders influence the diversity of a team. Additionally, rerunning the code with more data could improve the accuracy of the logistic regression model and would clear up if the mentioned inaccuracy of the model was due to confounders having little effect on the treatment or due to a lack of data.

References

[1] "Dota 2 heroes." [Online]. Available: https://www.dota2.com/heroes

[2] "Hero rankings." [Online]. Available: https://www.dotabuff.com/pages/rankings

[3] "Open source dota 2 data platform." [Online]. Available: https://www.opendota.com/

[4] S. R. Cole and M. A. Hernán, "Constructing Inverse Probability Weights for Marginal Structural Models," *American Journal of Epidemiology*, vol. 168, no. 6, pp. 656–664, 08 2008. [Online]. Available: https://doi.org/10.1093/aje/kwn164

[5] P. Ding and F. Li, "Causal inference: A missing data perspective," *Statistical Science*, vol. 33, no. 2, pp. 214–237, 2018. [Online]. Available: https://www.jstor.org/stable/26770992

[6] M. Facure, "Causal inference for the brave and true," 2021. [Online]. Available: https://matheusfacure.github.io/python-causality-handbook/landing-page.html

[7] M. Hammes, "What's the average match time in dota 2," 2022. [Online]. Available: https://theglobalgaming.com/gaming/average-match-time-dota-2

[8] J. M. R. Miguel A. Hernán, *Causal Inference: What If*. CRC Press, 2019.

[9] M. A. Pourhoseingholi, A. R. Baghestani, and M. Vahedi, "How to control confounding effects by statistical analysis," *Gastroenterol. Hepatol. Bed Bench*, vol. 5, no. 2, pp. 79–83, 2012.

[10] S. R. Seaman and I. R. White, "Review of inverse probability weighting for dealing with missing data," *Statistical Methods in Medical Research*, vol. 22, no. 3, pp. 278–295, 2013, pMID: 21220355. [Online]. Available: https://doi.org/10.1177/0962280210395740

[11] F. Thoemmes and A. D. Ong, "A primer on inverse probability of treatment weighting and marginal structural models," *Emerging Adulthood*, vol. 4, no. 1, pp. 40–59, 2016. [Online]. Available: https://doi.org/10.1177/2167696815621645

[12] D. Westreich and S. R. Cole, "Invited commentary: positivity in practice," *Am. J. Epidemiol.*, vol. 171, no. 6, pp. 674–7; discussion 678–81, Mar. 2010.

[13] G. Wunsch, "Confounding and control," *Demographic Research*, vol. 16, pp. 97–120, 2007. [Online]. Available: http://www.jstor.org/stable/26347930