The Effect of Sleep Quality on Academic Performance in Undergraduate Students

Jonathon A. Carl JAC16@WILLIAMS.EDU
Computer Science, Mathematics
jac16

1. Introduction

Exam scores are a key metric used to determine the relative academic success of an undergraduate student. Exam scores largely determine a student's GPA, which has a strong impact on their ability to find a desirable career path. As a result, college students looking to attain a high GPA are likely to build habits that drive success in their classes. In this paper, I draw causal conclusions between sleeping habits and academic performance.

To explore the impact of sleep habits on academic performance, I used Flueckiger et al. (2014)'s dataset, which contains survey data from a cohort of 72 first year undergraduate students at the University of Basel. In particular, my causal question of interest is: *Does reported sleep quality impact performance on final exams?* What follows is a thorough analysis of this question. First, I produced a clean dataset through data-processing. Next, I performed graph elicitation via causal discovery and background knowledge. After proper graph elicitation, I leveraged the front-door model (Primal IPW, Dual IPW, and Augmented Primal IPW) to find the average causal effect of sleep quality on exam performance.

My results reveal that under faithfulness, linearity, and edge presence assumptions, undergraduate students with a high sleep rating are more likely to pass their final exam compared to undergraduate students with a low sleep rating. These results provide a basis for students to value their sleep quality more when it comes to academic performance.

2. Preliminaries

Let $\mathcal{G} = (V, E)$ be a graph over a set of vertices V and edges E. $\mathcal{G}$ is a directed acyclic graph (DAG) if $\mathcal{G}$ contains only directed edges and is acyclic (i.e there are no paths such that $V_i \to \cdots \to V_i$). Denote $\mathcal{G}(M \cup U)$ to be a DAG over a set of measured variables M and unmeasured variables U.

Define an acyclic directed mixed graph (ADMG) $\mathcal{G}(V)$ to be the latent projection of the DAG $\mathcal{G}(M \cup U)$ which contains both directed ($\to$) and bi-directed ($\leftrightarrow$) edges. A graph $\mathcal{G} = (V, E)$ is said to be an ADMG if

1. $\mathcal{G}$ contains only directed and bi-directed edges ($\to$, $\leftrightarrow$)
2. $\mathcal{G}$ has at most 2 two edges between each pair of vertices (one directed and one bi-directed)
3. There are no directed cycles ($V_i \to \cdots \to V_i$)

Causal models from a DAG $\mathcal{G}(M \cup U)$ can be interpreted as tuple consisting of $\mathcal{G}(M \cup U)$ itself and a non-parametric structural equations model with independent errors (NPSEM-

Encoding	Type	Variable Name
SQ (treatment)	Binary: 0/1	Sleep Quality
PhysAct	Continuous	Minutes engaged in exercise
LGA	Multinomial: 1 (not at all) - 4 (completely)	Learning Goal Achievement
HSG	Multinomial: 1 (lowest) - 6 (highest)	High School Grades
Exam (outcome)	Binary: 0 (fail)/ 1 (pass)	Final Exam Success

Table 1: Variables processed from Flueckiger et al. (2014)'s dataset used for causal analysis

IE) equipped with a do-operator (Pearl, 2009). That is, each variable $V_i \in V$ is determined as a function of its parents and an error term ϵ. This induces a distribution $p(V)$ for the DAG $\mathcal{G}(M \cup U)$ that factorizes as

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)),$$

where $\text{pa}_{\mathcal{G}}(V_i)$ denotes the parents of V_i in $\mathcal{G}$. Under this interpretation, a directed edge $V_i \to V_j$ may be interpreted as saying that V_i is potentially a direct cause of V_j. Conditional independencies in $p(V)$ can be read off from the DAG via d-separation. That is,

$$X \perp\!\!\!\perp Y \mid Z_{\text{d-sep}} \implies X \perp\!\!\!\perp Y \mid Z_{\text{in } p(V)}.$$

To facilitate structure learning, I will restrict my analysis to the set of *faithful* distributions where $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep}} \iff (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$. In addition, I make the simplifying assumption that the relations between my variables are linear.

Given that an ADMG $\mathcal{G}(V)$ is the latent projection of $\mathcal{G}(M \cup U)$, we have that the marginal distribution $P(V)$ satisfies the global Markov property with respect to the ADMG $\mathcal{G}(V)$. That is, for all disjoint subsets X, Y, Z in the observed variables V,

$$X \perp\!\!\!\perp_{\text{m-sep}} Y \mid Z \text{ in } \mathcal{G}(V) \iff X \perp\!\!\!\perp_{\text{d-sep}} Y \mid Z \text{ in } \mathcal{G}(M \cup U).$$

Given an ADMG $\mathcal{G}$, we can identify a causal effect between a treatment A and outcome Y by constructing a Single World Intervention Graph (SWIG) $\mathcal{G}(a)$. A SWIG executes the $do(A = a)$ operator on $\mathcal{G}$ which encodes conditional independencies between observed and potential outcomes associated with an intervention on a treatment A (Richardson and Robins, 2013). The mechanical process of constructing a SWIG is as follows:

1. Copy over $\mathcal{G}$ into $\mathcal{G}(a)$.
2. Split A into a random vertex A and a fixed vertex a.
3. A inherits all incoming edges from the initial ADMG (including bi-directed) and a inherits all outgoing edges.
4. All descendants of a are converted into potential outcomes (i.e if $Y \in \text{de}_{\mathcal{G}}(a)$ then Y becomes $Y(a)$ in $\mathcal{G}(a)$).

After constructing the SWIG $\mathcal{G}(a)$ for an ADMG $\mathcal{G}$, m-separation is used to identify the causal effect of A on Y. To m-separate A from Y, all paths from the random A to $Y(a)$ in

Variables	Mean	Median	Std. Dev	Min	Max
SQ (treatment)	0.5	0.5	0.51	0	1
PhysAct	331.73	270.81	261.64	33.75	1737.10
LGA	2.57	3	0.64	1	4
HSG	4.68	4.7	0.44	3.4	5.6
Exam (outcome)	0.47	0	0.50	0	1

Table 2: Data distribution from Flueckiger et al. (2014) post-processing for $n = 72$ students

$\mathcal{G}(a)$ must be blocked by some adjustment set Z. In words, Z is the optimal set of variables that potentially confound the effect of A on $Y(a)$. An optimal adjustment set can be open to interpretation, but optimality often conforms to the proximity of a variable $Z_i \in Z$ to the potential outcome $Y(a)$.

3. Methods

3.1 Data Processing

Given Flueckiger et al. (2014)'s data, there was a relatively small amount of data processing to perform. The dataset provided information on 13 variables, but I have restricted my analysis to 5 variables: physical activity (*PhysAct*), learning goal achievement (*LGA*), high school grades (*HSG*), sleep quality (*SQ*) (treatment), and exam success (*Exam*) (outcome). For each of the 72 participants in the study, the dataset contained survey data over a month-long period leading up to their final exam. Because my interest is in the causal effect of sleep quality on exam performance, for each participant I took the mean value of each variable over time. This generated a cross-sectional dataset where observations are at participant level.

Furthermore, I transformed several of the variables to aid in the numerical analysis. First, *SQ* was moved from a multinomial scale (1-4) to a binomial scale (0/1). Specifically, within the cross-sectional distribution of sleep quality, participants whose average sleep quality fell above (below) the median value of 2.93 were recoded with 1 (0). Next, participants' learning goal achievement, *LGA*, was rounded to the nearest integer after taking the mean for each participant. For all variables in this paper, there were no missing data.

3.2 Graph Elicitation

I began my analysis through causal graph elicitation from Flueckiger et al. (2014)'s graph and Tetrad's causal discovery application (Scheines et al., 1998). First, I loaded my processed dataset into a Tetrad data node. Then, I created a knowledge node comprised of three tiers which restricted the edge relationships in the graph: (1) *SQ* and *PhysAct*, (2) *LGA* and *HSG*, and (3) *Exam*. The knowledge tiers were implemented based on background knowledge and Flueckiger et al. (2014)'s graph. The tiers are labeled in increasing order of reachability in the ADMG. That is, tier 1 nodes can reach nodes in tiers 1, 2, and 3; tier 2 nodes can reach nodes in tiers 2 and 3; tier 3 nodes can only reach other tier 3 nodes.

After connecting the data node to the knowledge node, I added a search node to perform causal discovery. The search node implemented the Greedy Fast Causal Inference (GFCI) algorithm (Ogarrio et al., 2016), an optimization of the Fast Causal Inference (FCI) algorithm (Spirtes, 2001). GFCI improves upon FCI by implementing a combined score and constraint based algorithm that works well on smaller sample sizes. A high-level summary of the GFCI algorithm is as follows:

1. Start with an empty graph and begin adding edges between nodes in order to increase the score. In particular, the Conditional Gaussian Bayesian Information Criterion (CG-BIC) score was used for model scoring. CG-BIC relies on parametric assumptions (Andrews et al., 2018). After step 1 in GFCI, we have a set of models of maximum score that remove all unmeasured confounders and selection bias.
2. Use the set of models obtained in (1) as the input for a modified FCI algorithm. Begin with an undirected graph $\mathcal{G}$ and perform conditional independence tests via the Conditional Gaussian Likelihood Ratio Test (CG-LRT) to detect the presence or absence of edges in $\mathcal{G}$. CG-LRT is intended for testing on an amalgam of continuous and discrete variables.
3. All parameters were set to default, except for bootstrapping (50 iterations).

The ADMG discovered by Tetrad was very similar to the DAG proposed by Flueckiger et al. (2014). Using background knowledge, I performed some edge additions and removal in order to obtain the ADMG used for causal identification and estimation in Figure 1.

3.3 Identification

I utilized the front-door criterion—a graphical criterion that gathers data on a strong mediator M which fully mediates the effect of the treatment on the outcome (Pearl, 1995)—to compute the effect of sleep quality on exam success. The front-door criterion can be understood as performing two simultaneous backdoor adjustments: the first finds the effect of A on M by applying $do(A = a)$ on the treatment; the second finds the effect of M on Y by applying $do(M = m)$ on the mediator. The front-door estimate is the product of the estimates obtained by performing backdoor adjustments on A and M. By performing two backdoor adjustments, we guarantee that we account for all variables potentially confounding the effect of A on M and of M on Y.

Given the ADMG $\mathcal{G}$ learned by Tetrad, I constructed two SWIGs—$\mathcal{G}(a)$ and $\mathcal{G}(m)$—to satisfy the front-door criterion with respect to a treatment A and outcome Y. That is, A and $Y(a)$ are m-separated in $\mathcal{G}(a) \iff A \perp\!\!\!\perp Y$ in $\mathcal{G}$; likewise, M and $Y(m)$ are m-separated in $\mathcal{G}(m) \iff M \perp\!\!\!\perp Y$ in $\mathcal{G}$. Given a valid set of mediators which satisfy the front-door criterion, Pearl (1995) defines the following formula as the identifying functional for the average causal effect (ACE) of A on Y:

$$
\begin{aligned}
ACE &\equiv \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')] \\
&\equiv \sum_{M} p(M \mid A = a) \times \sum_{A} p(A) \times \mathbb{E}[Y \mid M, A] - \sum_{M} p(M \mid A = a') \times \sum_{A} p(A) \times \mathbb{E}[Y \mid M, A]
\end{aligned}
$$

3.4 Estimation

Given the identification of a valid set of mediators M, I estimated the ACE with three estimators: Primal IPW (P-IPW), Dual IPW (D-IPW), and Augmented Primal IPW (APIPW) (Bhattacharya et al., 2020). An equivalent representation of the counterfactual mean can be obtained by using a joint propensity score of the treatment and outcome. Bhattacharya et al. (2020) defines the Primal IPW functional for identifying the ACE of A on Y:

$$\begin{aligned} ACE &\equiv \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')] \\ &\equiv \mathbb{E}\left[\mathbb{I}(A=a) \times \frac{\sum_A p(A) \times p(Y \mid A, M)}{p(A) \times p(Y \mid A, M)} \times Y\right] - \mathbb{E}\left[\mathbb{I}(A=a') \times \frac{\sum_A p(A) \times p(Y \mid A, M)}{p(A) \times p(Y \mid A, M)} \times Y\right] \end{aligned}$$

where $\mathbb{I}(A = a)$ is the indicator function which returns 1 if $A = a$ and 0 otherwise. I used Ananke to obtain the ACE via Primal IPW.

I obtained another estimate for the ACE with my own implementation of Dual IPW—an equivalent representation of the counterfactual mean which involves a propensity score of the mediator. Bhattacharya et al. (2020) defines the Dual IPW functional for identifying the ACE of A on Y:

$$\begin{aligned} ACE &\equiv \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')] \\ &\equiv \mathbb{E}\left[\frac{p(M \mid A=a)}{p(M \mid A)} \times Y\right] - \mathbb{E}\left[\frac{p(M \mid A=a')}{p(M \mid A)} \times Y\right] \end{aligned}$$

For my own implementation of D-IPW, I fit a multinomial regression model with the mediator LGA as a function of its parents in the ADMG $\mathcal{G}$ using the statsmodels module in Python (Seabold and Perktold, 2010). I then obtained propensity scores for the original dataset and two copies of the dataset where treatment is manually assigned to 0 or 1 (bad or good sleep quality, respectively). I then used these propensity scores to find the ACE based on the functional defined above. To bound the ACE and account for uncertainty, I computed 95% confidence intervals with 200 bootstraps. A random seed of 0 was set for reproducibility of results.

I used APIPW as another metric to check for the validity of the previous estimates. The functional is rather complicated, so I will not describe it in detail here. For more on this functional, see Bhattacharya et al. (2020)'s paper. At a high level, APIPW is doubly robust and combines the implementations of Primal IPW and Dual IPW. That is, the $\widehat{ACE}_{APIPW}$ is guaranteed to to converge to the true ACE if either (1) a propensity score model $p(A \mid Z)$ and $\mathbb{E}[Y \mid A, M, Z]$ is correctly specified or (2) an outcome regression model $p(M \mid A, Z)$ is correctly specified.

4. Results

4.1 Learned ADMG

The learned ADMG from causal discovery used for analysis of the effect of sleep quality on academic performance is shown in Figure 1. The red bi-directed edges were transformed

from directed edges after graph elicitation with Tetrad. The learned ADMG $\mathcal{G}$ is slightly different than the DAG proposed by Flueckiger et al. (2014). I feel that the relationship between sleep quality, exam, and physical activity is more interrelated than a single-directional edge. For example, number of hours slept may both impact sleep quality and exam performance; the same could be said for sleep quality and physical activity: more sleep increases sleep quality and ability to endure vigorous physical activity. Stress may impact physical activity and exam performance. Sleep duration and stress are unobserved, however, so the red edges must be bidirectional.

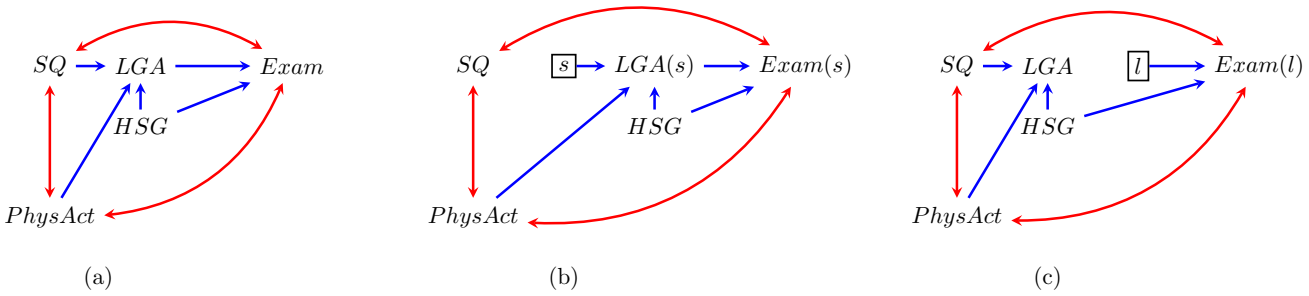

Figure 1: The ADMG $\mathcal{G}$ learned by Tetrad (a); SWIGS $\mathcal{G}(s)$ (b) and $\mathcal{G}(l)$ (c) used for the Primal IPW estimate

4.2 Identification and Analysis

From the ADMG $\mathcal{G}$ in Figure 1(a), we can find the SWIG for $\mathcal{G}(s)$ and $\mathcal{G}(l)$ by applying the $do(\cdot)$ operator on SQ and LGA. Both SWIGS are required for the Primal IPW performed by Ananke. Given both $\mathcal{G}(s)$ and $\mathcal{G}(l)$, we have the functional from Bhattacharya et al. (2020) for identifying the ACE:

$$\begin{aligned} ACE &\equiv \mathbb{E}[Exam(SQ=1)] - \mathbb{E}[Exam(SQ=0)] \\ &\equiv \mathbb{E}\left[\mathbb{I}(SQ=1) \times \frac{\sum_{SQ} p(SQ) \times p(Exam \mid SQ, LGA)}{p(SQ) \times p(Exam \mid SQ, LGA)} \times Exam\right] \\ &- \mathbb{E}\left[\mathbb{I}(SQ=0) \times \frac{\sum_{SQ} p(SQ) \times p(Exam \mid SQ, LGA)}{p(SQ) \times p(Exam \mid SQ, LGA)} \times Exam\right] \end{aligned}$$

where $\mathbb{I}(SQ=1)$ is the indicator function which returns 1 if $SQ=1$ and 0 otherwise; the same holds for $\mathbb{I}(SQ=0)$. Using Primal IPW, I obtained an ACE point estimate of 1.580 with a 95% confidence interval of (0.976, 2.711).

For my Dual IPW implementation, I fit the multinomial model for the mediator LGA as a function of its parents ($PhysAct$, HSG, and SQ as seen in Figure 1 (c)). Given $M = \{LGA\}$, we have the functional from Bhattacharya et al. (2020) for identifying the ACE:

$$\begin{aligned} ACE &\equiv \mathbb{E}[Exam(SQ=1)] - \mathbb{E}[Exam(SQ=0)] \\ &\equiv \mathbb{E}\left[\frac{p(LGA \mid SQ=1)}{p(LGA \mid SQ)} \times Exam\right] - \mathbb{E}\left[\frac{p(LGA \mid SQ=0)}{p(LGA \mid SQ)} \times Exam\right] \end{aligned}$$

where $SQ = 1$ (0) indicates good (poor) sleep quality. I used this functional in my own D-IPW to estimate the ACE, where I obtained a point estimate of 1.103 with no confidence intervals; my own D-IPW is sensitive computing confidence intervals with resampling since all values of LGA in the resample must take on integer values between 1-4.

With Ananke's D-IPW, I obtained a point estimate of 1.647 (1.034, 4.090). With Ananke's APIPW, I found a point estimate of 1.647 (0.999, 3.823).

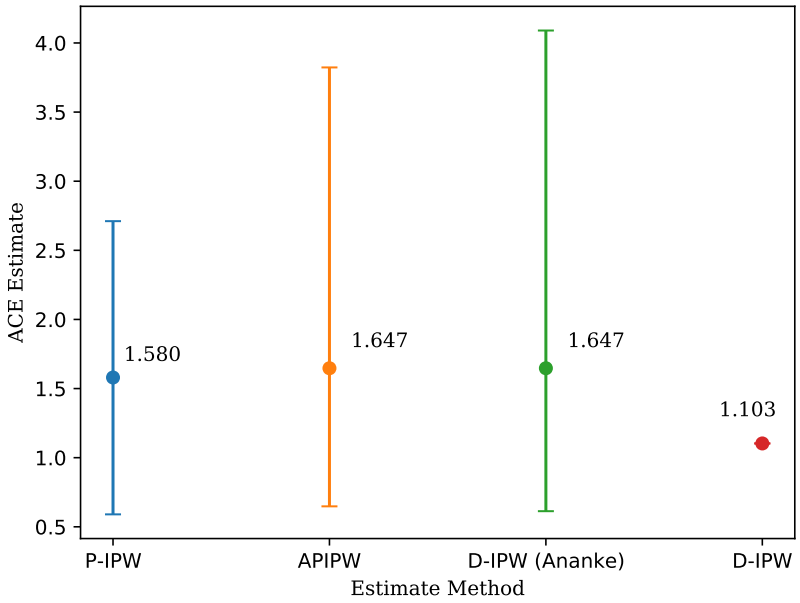

Figure 2: Causal Estimates obtained via Front-Door

4.3 Sensitivity Analysis

After completing analysis on Flueckiger et al. (2014)'s data, I sought to expand my findings onto different undergraduate students. I began inspecting Wang et al. (2014)'s dataset on 48 students from Dartmouth University. My goal was to find the effect of average sleep hours on GPA. After some data processing, I discretized average sleep hours to 0/1 (unhealthy/healthy) and kept GPA as continuous outcome. This dataset had limitations, though, because there were several missing observations for some of the key variables. I performed multiple imputation to predict values for missing data points for GPA and expected grade variables (Westreich et al., 2015). After performing analysis on Flueckiger et al. (2014)'s dataset, I intended to use Wang et al. (2014)'s as a complementary estimate for the effect of sleep on academic performance. However, the datasets contained disparate variables, which made it difficult to obtain comparable point estimates. I obtained a point estimate of 0.279 (-0.293, 1.002) using Ananke's backdoor formula for the Dartmouth dataset. Given there was so much missing data and the confidence interval is wide, I found this estimate to be unreliable. I have included it here to depict my attempt at generalizing my findings to other undergraduate students.

5. Discussion and Conclusion

5.1 Discussion

Through causal estimation via front-door adjustment, I obtained a positive ACE using three different estimators, including my own implementation of D-IPW (Figure 2). While these estimates have rather wide confidence intervals, their interpretations confirm my hypothesis: as sleep quality increases, so too does academic performance. The Primal IPW estimate concludes that a student at the University of Basel was 1.580 times more likely to pass their final exam given that they had good sleep quality. The same holds for APIPW and D-IPW, but for their respective magnitudes found in Figure 2.

My own D-IPW implementation differs significantly in magnitude to Ananke's estimates. I feel that with more work on my algorithm and access to unmeasured confounders (i.e variables like stress, sleep duration), the ACEs will converge and yield stronger results.

5.2 Future Work

There are many relevant datasets which investigate the relationship between health and academic performance. I see two directions this analysis could take: (1) Analyze the effect of stress on academic performance. I hypothesize that there may not be a linear relationship between these two variables. I feel there could be some threshold that follows the idea of moderate arousal. That is, one must have some amount of stress in order to perform well, but too little or too much may be detrimental to performance. (2) Collect more data on the same variables to gain stronger estimates that more accurately quantify the ACE of sleep quality on academic performance.

5.3 Conclusion

In this paper, I examined the causal relationship between sleep quality and exam performance. Causal analysis estimates a positive ACE, suggesting that better sleep quality may impact a student's performance on an exam. This paper demonstrates that, for students who are motivated to maximize their academic performance, sleep quality is an important factor in the process. While my own D-IPW estimation varies largely in magnitude from Ananke's, I find that there is at least a marginal positive effect on exam performance given good sleep quality. However, these findings may be limited by the assumptions stated earlier in the paper: faithfulness, linearity, and edge presences/absences. In future work, further analysis could be performed to confirm the ACE estimates obtained in this paper by including measurement of important unmeasured variables (i.e stress and sleep duration) and improving my own implementation of D-IPW. Estimates may become more precise and students will have a more accurate measure of the effect of sleep quality on academic performance.

References

Bryan Andrews, Joseph Ramsey, and Gregory F Cooper. Scoring bayesian networks of mixed variables. *International journal of data science and analytics*, 6(1):3–18, 2018.

Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

Lavinia Flueckiger, Roselind Lieb, Andrea H Meyer, and Jutta Mata. How health behaviors relate to academic performance via affect: An intensive longitudinal study. *PLoS One*, 9 (10):e111080, 2014.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on probabilistic graphical models*, pages 368–379. PMLR, 2016.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128 (30):2013, 2013.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Peter Spirtes. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR, 2001.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pages 3–14, 2014.

Daniel Westreich, Jessie K Edwards, Stephen R Cole, Robert W Platt, Sunni L Mumford, and Enrique F Schisterman. Imputation approaches for potential outcomes in causal inference. *International journal of epidemiology*, 44(5):1731–1737, 2015.