Combining experimental and observational studies to estimate individual treatment effects: applications to customer journey optimization

TOTTE HARINEN, Toyota Research Institute, USA
RUMEN ILIEV, Toyota Research Institute, USA
ANG LI, UCLA, USA
SCOTT MUELLER, UCLA, USA
CHI ZHANG, UCLA, USA

In current industry practice, decision-makers commonly use the results of isolated, one-off experiments and observational studies to determine key aspects of the customer journey. For example, incentives might be allocated by conducting a single A/B test and finding subgroups with high enough conditional average treatment effects. Similarly, new product features might be introduced to segments for whom they were found to perform well in an A/B test or in an observational study. Given this, a natural question is whether it is possible to combine information from experiments and observational studies to estimate the effects of interventions in a way that gives us more insights compared to any of these studies taken in isolation. In this paper, we show that the answer to the question is "yes". We demonstrate how recent advancements in counterfactual logic allow us to combine experiments, observational studies and causal diagrams to estimate treatment effects at an individual level. Focusing on common customer journey decisions such as how to target incentives and personalize products, we demonstrate the additional insights afforded by the counterfactual approach in a number of simulation setups.

Additional Key Words and Phrases: probability of causation

1 INTRODUCTION

2 NECESSARY AND SUFFICIENT CAUSES

An intervention is both necessary and sufficient for an outcome if and only if:

- The outcome would not occur without the intervention (necessity)
- The outcome will occur if the intervention occurs (sufficiency)

Interventions that are both necessary and sufficient are interesting for industry practitioners because they are not superfluous in the sense that the outcome would occur even in the absence of the intervention, and they are powerful enough to bring about the outcome.

As a concrete yet hypothetical example, suppose a car manufacturer wants to encourage drivers to charge their car every night. To achieve this outcome, they devise an intervention where a charging reminder is sent via a mobile app every time the car approaches home. For this manufacturer, it would be useful to know whether the intervention is necessary and sufficient. If the intervention is not necessary, then those who charge their car every night would also do so in the absence of the intervention, and it might be better not to bother them with charging reminders. If the intervention is not sufficient, then those who received the reminder might still fail to charge their car at night, and the intervention would fail to bring about the desired behaviour.

Luckily, recent advances in causal inference make it possible to calculate the bounds for the probability that an intervention is necessary and sufficient. [1–6] The methods for bounding the probability of necessity and sufficiency (PNS) combine experimental and observational information in order to understand what would have happened in the absence of the intervention. Apart from special circumstances, the methods do not provide a point estimate of PNS;

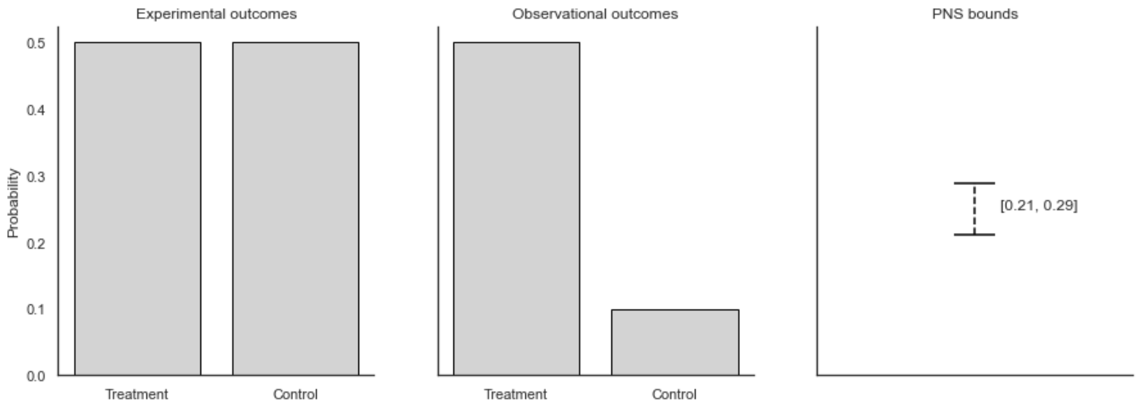

Fig. 1. The outcomes observed in the experimental and observational studies and the resulting PNS bounds. Note that we find no average treatment effect in the experiment, which normally would result us judging that the intervention is ineffective. However, the outcomes of the observational study reveal that the intervention is necessary and sufficient for some individuals.

instead, they reveal that the PNS lies within certain limits. Despite this, the information provided by PNS calculation can be highly useful in practice, as we will show when we discuss the examples below.

3 CASE STUDIES

We simulated data from pairs of experimental and observational studies where each study had one treatment and one control arm with an outcome drawn from a binomial distribution. For each of the four outcomes, we varied the probability of success in 11 evenly spaced steps from 0 to 1, resulting in $11^4 = 14641$ pairs of experimental and observational studies. However, since some combinations of values are not possible[1], we ended up with 11137 pairs of experiments and observational studies. The data for the case studies below are drawn from this set.

3.1 Intervention with zero treatment effect

Consider again the example of charging reminders. Suppose the company runs an A/B test where 1000 randomly selected drivers receive a charging reminder and another 1000 randomly selected drivers serve as the control. Suppose the nightly charging rate, defined as whether the car is charged every night during the 7-day testing period, is exactly 50% in both the treatment and control groups. If the company followed the standard A/B testing paradigm, they would conclude that the intervention is not effective.

Suppose, however, that the company also has data of another sample of drivers who were simply given the possibility of signing up for the charging reminders but who were not forced to do so. Some of these drivers elected to receive the reminders while others preferred not to have them. Comparing these two groups, it turns out that again 50% of those who chose to receive the reminder charged their car every night, but that the proportion of nightly chargers was just 10% for those who chose not to receive reminders. The data from this observational group of drivers can now be combined with the experimental data to calculate the PNS bounds for the charging intervention, which turn out to be [0.21, 0.29] in this example. Figure 1 shows the outcomes of both the experiment and observational study together with the resulting PNS bounds.

[1] [7] discuss the constraints that the relevant probabilities must satisfy.

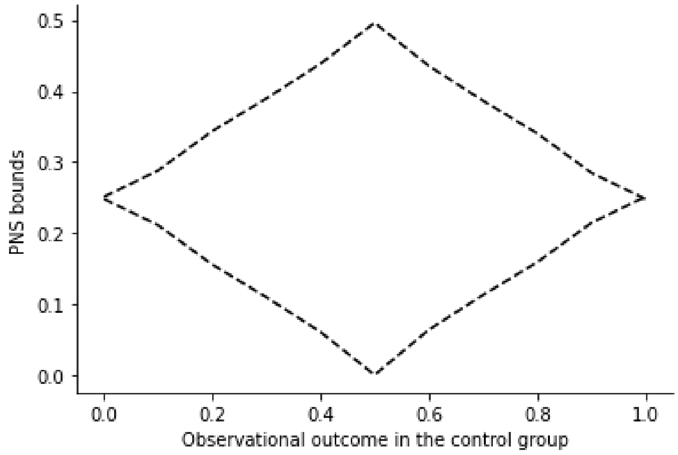

Fig. 2. The upper and lower PNS bounds when the experimental treatment and control outcomes are held fixed at 0.5, the observational treatment outcome is also held at 0.5, and the observational control outcome varies from 0 to 1.

This result is interesting given that the outcome of the A/B test would have led us to conclude that charging reminders are not effective. Our analysis indicates that there must be some individuals for whom the intervention is both necessary and sufficient, given the results of the observational study. While the formal proof for this relationship between the experimental and observational results is beyond the scope of this paper, we can develop some intuition for it by noting that the observational data reveals how drivers would have behaved if left to choose freely between receiving and not receiving reminders. In particular, we learn that those who choose not to receive reminders are relatively poorer decision-makers in this regard, because their nightly charging rate is much lower (assuming that nightly charging is desirable for most people). Enabling charging reminders for these drivers would consequently cause some of them to charge more regularly.

The relationship between experimental and observational results has an interesting property where the PNS bounds get narrower the bigger the difference in the outcomes of the experimental study and the observational study. Figure 2 shows the PNS bounds calculated for a number of simulated datasets where the treatment and control outcomes in the experimental data are kept fixed at 0.5. The treatment outcome in the observational study is also kept at 0.5 while the control outcome varies from 0 to 1. We can see that the PNS bounds are the widest where the outcomes in the observational study are identical to the experiment, and they get narrower as the observational control outcome starts to diverge from that observed in the experiment. For calculating PNS bounds, more confounding is better.

3.2 Interventions with different treatment effects

For another case study of the implications of PNS bounds on decision-making, suppose the car manufacturer tests two different interventions to increase regular charging. The first one is a charging reminder shown on the dashboard of the car. The second intervention is a voice reminder that plays when the car is being parked near a charger. Suppose the manufacturer runs two experiments where they compare each intervention against a control group who receives no reminders. The outcomes in these two experiments are as follows:

- **Dashboard notification**: 70% daily charging in treatment, 60% charging in control

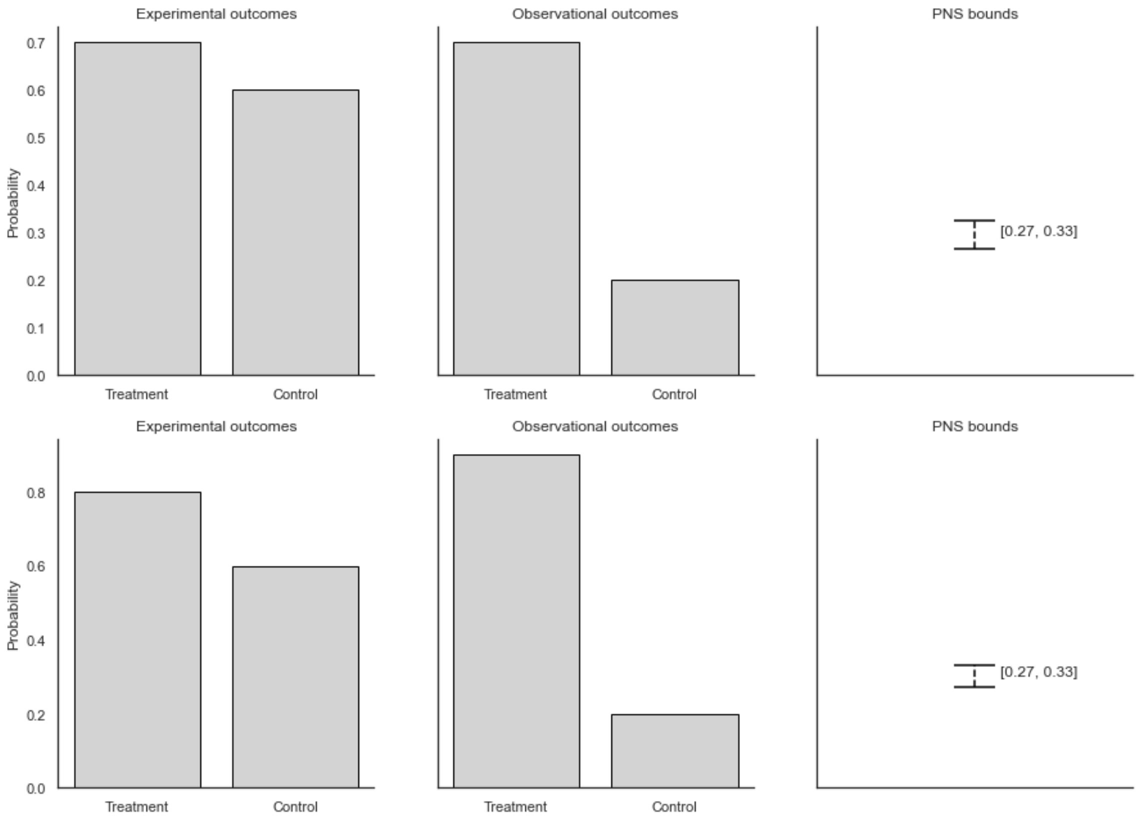

Fig. 3. A comparison of two interventions than end up having the same PNS bounds despite having different experimental treatment effects. The upper panel shows that the first intervention has a 10 ppt treatment effect in an A/B test and a 70% success rate in the observational treatment condition. The lower panel shows that the second intervention has a 20 ppt treatment effect in the A/B test and a 80% success rate in the observational treatment condition. In all of the four studies, the control conditions happen to have the same success rates, namely 60% in the A/B test and 20% in the observational study. These combinations of outcomes result in the same PNS bounds for each intervention.

- **Voice reminder**: 80% daily charging in treatment, 60% in control

Clearly, voice reminders seem more promising as an intervention. If the company were to follow the standard A/B testing framework, they would implement voice reminders, provided that it would not be much more costly to do so. However, suppose the manufacturer also conducts an observational follow up study where drivers could simply select between dashboard notifications, voice reminders and no intervention. Suppose, moreover, that the results of these studies turned out as follows:

- **Dashboard notification**: 70% daily charging for those who opt in, 20% for those who opt out
- **Voice reminder**: 90% daily charging for those who opt in, 20% for those who opt out

As shown in Figure 3, these two combinations of experimental and observational outcomes would result in the same PNS bounds for both treatments (with a lower bound of 27% and an upper bound of 33%). This result might reverse the decision made based on the A/B test results if, for example, the voice reminder intervention was more expensive, or if it was for example deemed as more invasive than the dashboard intervention.

3.3 Different population segments with the same treatment effect

For data scientists working on experimentation, the above two examples surely suggest the possibility of segmenting customers based on their PNS bounds. Consider again the A/B test in which an app notification is sent to increase charging rates, where the experimental data showed a 50% charging rate in both treatment and control. Suppose a data scientist split the data from this experiment further to examine the treatment effect within two segments of customers, such as between urban and rural drivers. If the treatment effect still turned out to be zero within these two segments, the urban/rural customer segmentation would be typically deemed as irrelevant from the point of view of the intervention.

However, based on the previous examples, we can easily see that the urban/rural segmentation is not necessarily irrelevant from the point of view of the treatment's necessity and sufficiency. Suppose the data from the observational follow-up study is also segmented according to the urban/rural distinction, and it turns out that the observational outcomes are as follows:

- **Urban**: 40% charging for those who opt in; 30% for those who opt out
- **Rural**: 60% for those who opt in; 20% for those who opt out

While the bounds here clearly overlap, it might be interesting for decision-makers to know that they are not the same. This also suggests that deeming segments relevant or irrelevant based on whether they predict differences in average treatment effects may be premature.

REFERENCES

[1] Ang Li and Judea Pearl. 2019. Unit Selection Based on Counterfactual Logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 1793–1799. https://doi.org/10.24963/ijcai.2019/248

[2] Ang Li and Judea Pearl. 2022. Bounds on causal effects and application to high dimensional data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5773–5780.

[3] Ang Li and Judea Pearl. 2022. Unit selection with causal diagram. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5765–5772.

[4] Scott Mueller, Ang Li, and Judea Pearl. 2021. Causes of Effects: Learning individual responses from population data. *arXiv preprint arXiv:2104.13730* (2021).

[5] Scott Allen Mueller. 2021. *Estimating Individualized Causes of Effects by Leveraging Population Data.* University of California, Los Angeles.

[6] Judea Pearl. 2022. Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. 317–372.

[7] Jin Tian and Judea Pearl. 2000. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence* 28, 1 (2000), 287–313.

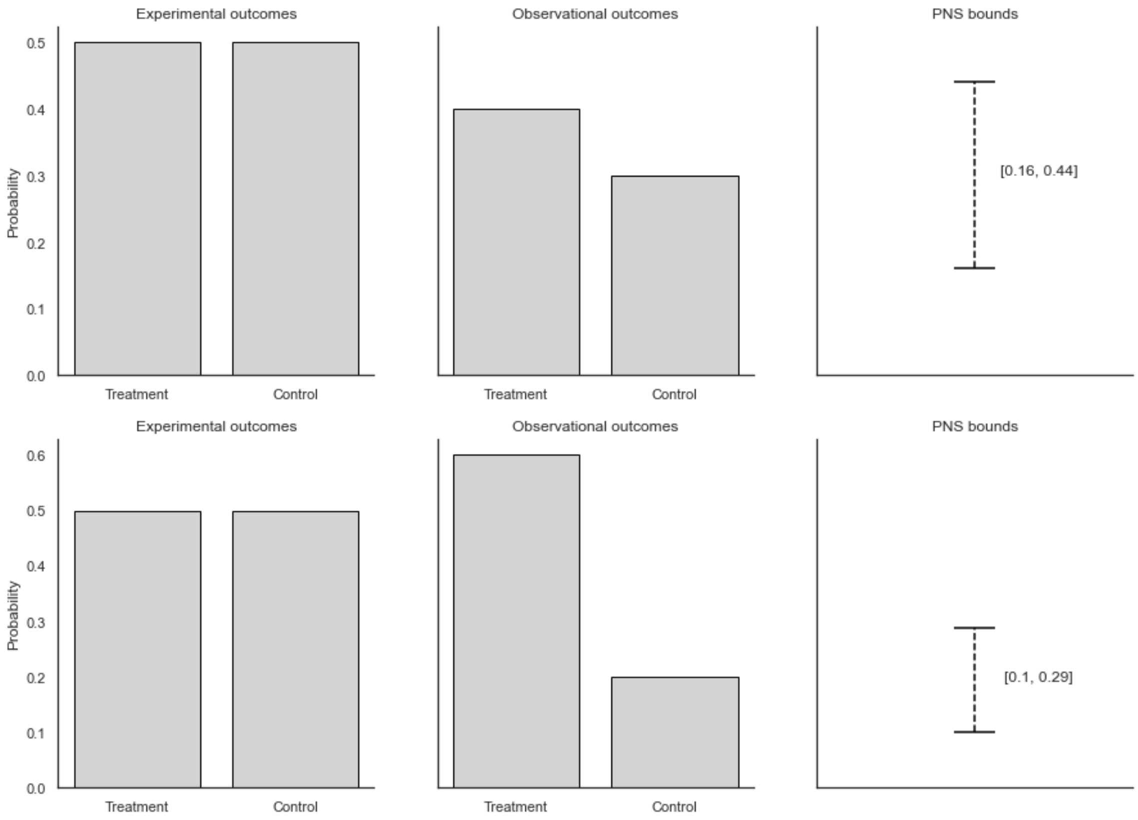

Fig. 4. In this case study, a population characteristic initially appears irrelevant because splitting the population in an A/B test based on that characteristic doesn't result in differing within-segment treatment effects. However, after splitting the data from the observational study based on the population characteristic, we find that the PNS bounds for the intervention are quite different between the segments.