

# When accurate prediction models yield harmful self-fulfilling prophecies

**Wouter A.C. van Amsterdam, MD, PhD\***

*Department of Data Science and Biostatistics  
Julius Center of Health Sciences and Primary Care  
University Medical Center Utrecht, Utrecht, the Netherlands  
University of Utrecht, Utrecht, the Netherlands  
Heidelberglaan 100, 3584 CX Utrecht, the Netherlands  
corresponding author*

W.A.C.VANAMSTERDAM-3@UMCUTRECHT.NL

**Nan van Geloven, PhD**

*Department of Biomedical Data Sciences  
Leiden University Medical Center, Leiden, the Netherlands*

**Jesse H. Krijthe, PhD**

*Pattern Recognition & Bioinformatics  
Delft University of Technology, Delft, the Netherlands*

**Rajesh Ranganath, PhD**

*Courant Institute of Mathematical Science, Department of Computer Science  
Center for Data Science  
New York University, New York City, USA*

**Giovanni Cinà\*, PhD**

*Department of Medical Informatics  
Amsterdam University Medical Center, Amsterdam, the Netherlands  
Institute for Logic, Language and Computation  
University of Amsterdam, Amsterdam, the Netherlands  
Pacmed, Amsterdam, the Netherlands*

G.CINA@AMSTERDAMUMC.NL

Word count: 3516

---

\* these authors contributed equally

## Abstract

**Objective** Prediction models are popular in medical research and practice. By predicting an outcome of interest for specific patients, these models may help inform difficult treatment decisions, and are often hailed as the poster children for personalized, data-driven healthcare. Many prediction models are deployed for decision support based on their prediction accuracy in validation studies. We investigate whether this is a safe and valid approach.

**Materials and Methods** We show that using prediction models for decision making can lead to harmful decisions, even when the predictions exhibit good discrimination after deployment. These models are *harmful self-fulfilling prophecies*: their deployment harms a group of patients but the worse outcome of these patients does not invalidate the predictive power of the model.

**Results** Our main result is a formal characterization of a set of such prediction models. Next we show that models that are well calibrated *before* and *after* deployment are useless for decision making as they made no change in the data distribution.

**Discussion** Our results point to the need to revise standard practices for validation, deployment and evaluation of prediction models that are used in medical decisions.

**Conclusion** Outcome prediction models can yield harmful self-fulfilling prophecies when used for decision making, a new perspective on prediction model development, deployment and monitoring is needed.

**Keywords:** Prognosis, Deployment, Monitoring, Decision Support Techniques, Causal Inference

## 1. Introduction

Clinicians and medical researchers frequently employ outcome prediction models (OPMs): statistical models that predict a certain medical outcome based on a patient’s characteristics [1]. Researchers develop OPMs to provide information to clinicians so they may use this information in difficult treatment decisions (e.g. Salazar et al. [2]). In some cases, clinicians will treat patients with a bad expected outcome more aggressively, for example by giving cholesterol lowering medication to patients with a high predicted risk of a heart attack [3, 4]. In other cases, for instance when the treatment is burdensome or scarcely available (e.g. ventilator machines on the intensive care during a pandemic), clinicians may reserve treatment for patients with a good predicted outcome.

Many such OPMs are added to the protocol of care by designing specific thresholds for specific actions [3]. If the predicted outcome is above or below the threshold a certain action is taken, e.g. the patient receives a more aggressive therapy. The basis for including an OPM in a care protocol is generally predictive accuracy in validation studies [5]. In these validation studies, the OPM may or may not have been used to inform treatment decisions.

At first, it may seem that this approach is beneficial since giving more information should lead to better treatment decisions. However, implementing a prediction model for treatment decisions is an intervention that changes treatment decisions and thus patient outcomes. Whether this change in treatment policy improves patient outcomes is not determined by prediction accuracy in a validation study [6]. For instance, in cases where a certain patient subpopulation historically received suboptimal care, an accurate OPM will predict a worse outcome for these patients compared to similar patients outside of the subpopulation. If clinicians decide to withhold effective treatments (e.g., due to scarcity or perceived futility) to this underserved subpopulation based on the OPM’s prediction of a bad outcome, the implementation of the OPM perpetuated biases or caused harm to these patients, despite its accuracy. Moreover, the implementation of this harmful new policy brought about the scenario predicted by the OPM, as in a *self-fulfilling prophecy*. One concrete example where clinicians treat patients with a bad expected outcome less aggressively is in small cell lung cancer. Prognostic scores for small cell lung cancer patients, such as the Manchester score developed in [7] are specifically intended to not over-treat patients with a bad predicted outcome because this is expected to be futile [8, 9].

In this article we address the following questions: 1) Under what conditions is a new policy based on an OPM going to be harmful, meaning that it leads to worse outcomes than before using the model? 2) In what circumstances would such a harmful policy go undetected by measures of discrimination or calibration? In what follows we provide a formalization of the case where patients with a high predicted probability of the

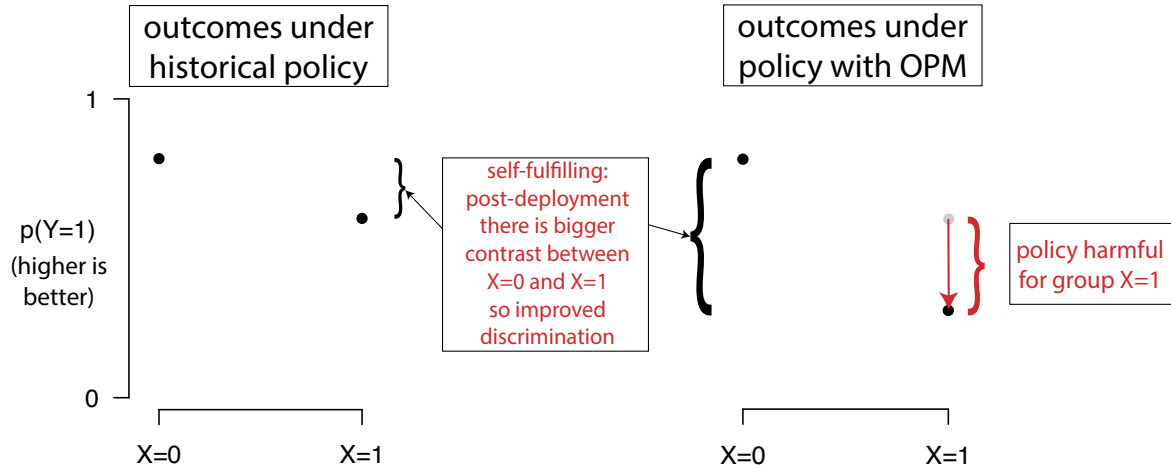


Figure 1: Some outcome prediction models yield harmful self-fulfilling prophecies when used to guide treatment decisions, meaning the new policy harms a subgroup of patients but the prediction model has good discrimination post-deployment.

outcome get treatment, where the outcome may be preferable (e.g. 1-year survival) or undesirable (e.g. a heart attack). Specifically, we examine the setting where a new OPM is supposed to ‘personalize’ an existing treatment policy by considering additional features. Section 2 provides notation and definitions, Section 3 presents the main results concerning OPMs that are harmful and self-fulfilling prophecies. We first show that even in a simple setup with a binary covariate, a non-trivial subset of OPMs yields harmful self-fulfilling prophecies. This means that such models cause harm but exhibit good discrimination on post-deployment data, meaning that naively interpreting this as a successful deployment leads to harmful policies. Next, perhaps surprisingly, we show that when an OPM is well calibrated on both 1) the historical data and 2) a validation study where the model is used for treatment decisions, the OPM is not useful for decision making.

Based on our results, several common practices in building and deploying OPMs intended for decision making need revision: 1. Developing OPMs on observational data without regard of the historical treatment policy is potentially dangerous, because the change in treatment policy between pre- and post-deployment is what determines the effect of the model on patient outcomes. 2. Implementing a personalized outcome prediction model is not always beneficial, even if the model is very accurate. 3. When monitoring discrimination prospectively after deployment, sometimes good discrimination means a harmful model and sometimes a beneficial one.

## 2. Notation and definitions

We assume a binary treatment  $T$ , a binary outcome  $Y$  and a binary feature  $X \in \mathcal{X} = \{0, 1\}$ . We denote the outcome obtained with setting treatment  $T$  to  $t$  as  $Y_t$ . An OPM is a function trained on historical data to predict the outcome of interest. We use  $\pi_i(X)$  to denote a policy for assigning treatment, possibly conditional on  $X$ , with an index  $i$  to indicate what policy we are referring to. Throughout the paper  $\pi_0$  will be used to indicate the *historic treatment policy* that was in place in the data in which the OPM was developed.

We assume the historical policy is constant and deterministic, meaning that it is always equal to 0 or 1 (i.e. patients were always treated or never treated). Next we define what it means to craft a policy based on an existing OPM. We will be concerned only with *threshold-based policies*, namely policies that assign

treatment based on a threshold  $\lambda \in \mathbb{R}$ . In our setup, policies assign treatment to patients only if the expected outcome is above  $\lambda$ , which could mean either a desirable outcome (e.g. 1-year survival) or undesirable (e.g. a heart attack).

**Definition 1 (Policy informed by OPM)**

Let  $f : X \rightarrow [0, 1]$  be an OPM and  $\lambda \in \mathbb{R}$  a threshold. We call  $\pi_f$  a policy informed by  $f$  and define it as follows

$$\pi_f(x) = \begin{cases} 1 & f(x) > \lambda \\ 0 & f(x) \leq \lambda \end{cases} \quad (1)$$

Such policies describe the post-deployment scenario, when the OPM influences treatment assignment. This deployment will change some of the (conditional) probability distributions compared to pre-deployment. We distinguish probabilities pre- and post-implementation using subscripts:  $p_i(\cdot)$  with  $i \in \{0, f\}$  respectively. We now present the first key idea of this paper, namely the special class of OPMs whose predictions are realized upon implementation. We consider as a metric of discrimination the popular ‘Area under the ROC-curve’ (AUC [10]).

**Definition 2 (Self-fulfilling OPM)** Let  $f : X \rightarrow [0, 1]$  be an OPM,  $\lambda \in \mathbb{R}$  a threshold and let  $\pi_f$  be the policy informed by  $f$ . Let  $AUC(\pi_i)$  denote the AUC of this OPM on data generated with the historic policy ( $\pi_0$ ) or with the policy defined by  $f$  ( $\pi_f$ ). We call the pair  $(f, \lambda)$  self-fulfilling if the AUC remains the same or increases post-deployment, namely iff:

$$AUC(\pi_f) \geq AUC(\pi_0) \quad (2)$$

Finally, we specify what we mean with an OPM being harmful in comparison with the status quo.

**Definition 3 (Harmful OPM)** Let  $f : X \rightarrow [0, 1]$  be an OPM,  $\lambda \in \mathbb{R}$  a threshold, let  $\pi_0$  denote the historic treatment policy and let  $\pi_f$  be the policy informed by  $f$ . We write the expected outcomes under the different policies as

$$p_i(Y = 1|X) = \mathbb{E}_{T \sim \pi_i(X)} p(Y_T = 1|X) \quad (3)$$

where  $i = 0$  denotes the historical distribution and  $i = f$  the distribution under  $\pi_f$ . We call  $f$  harmful for the group with  $X = x$  with  $p(X = x) > 0$  if the expected outcome of the group<sup>1</sup> is worse under the new policy compared to the old policy, namely when  $Y = 1$  is preferable iff

$$p_f(Y = 1|X = x) < p_0(Y = 1|X = x) \quad (4)$$

or when  $Y = 0$  is preferable iff

$$p_f(Y = 1|X = x) > p_0(Y = 1|X = x) \quad (5)$$

When a policy informed by an OPM is both harmful and self-fulfilling we have a worst-case scenario where the new policy is causing harm to a subgroup but this, perhaps counter-intuitively, does not result in a decrease in AUC post-deployment.

### 3. Results

We now move to the main results, whose proofs can be found in Appendix B. The setting where a new OPM is supposed to ‘personalize’ an already existing treatment policy by considering more features is encoded as follows: the new OPM considers a feature  $X$  that was previously ignored by the historical policy, specifically  $\pi_0$  is constant and deterministic. In addition, the new policy  $\pi_f$  is not constant but varies with  $X$ .

---

1. Note that this is different from a model being marginally harmful, i.e. applying  $\pi_f$  leads to worse outcomes on average. However, we will later see that in our setup with binary  $X$ , one of the two groups has the same outcomes pre- and post-deployment so an OPM that is harmful to a subgroup will also be marginally harmful.

### 3.1. Harmful models may have good discrimination post-deployment

We state our main observation as an informal theorem.

**Theorem 4 (Informal main result)** *Let  $\pi_f$  be the policy informed by the OPM  $f$  using a threshold  $\lambda$ . Assume that: i) the historical policy  $\pi_0$  is constant and deterministic ii) the new policy  $\pi_f$  is not constant, i.e. not always equal to 1 or 0 and iii) the marginal distribution of  $X$  is the same pre and post deployment:  $p_i(X) = p(X)$  for  $i \in \{0, f\}$ .*

*Under these assumptions, a non-trivial subset of OPMs will demonstrate good post-deployment discrimination because they yield self-fulfilling prophecies, and at the same time their deployment harmed patients.*

The theorem is exemplified by Figure 1. We proceed to characterize the contours of the subset of self-fulfilling and harmful OPMs.

**Proposition 5 (Self-fulfilling)** *Suppose the assumptions of Theorem 4 hold. Furthermore assume that the joint probabilities of  $X$  and  $Y$  are non-deterministic both pre- and post-deployment:*

$$0 < p_i(Y = 1, X = x) < 1, \forall x \in \mathcal{X} \quad (6)$$

*Then the following two statements are true: i) if the treatment effect is always positive, namely  $\forall x \in \mathcal{X} : p(Y_1 = 1|X = x) \geq p(Y_0 = 1|X = x)$ , then  $(f, \lambda)$  is self-fulfilling; ii) if the treatment effect is always negative, meaning  $\forall x \in \mathcal{X} : p(Y_1 = 1|X = x) < p(Y_0 = 1|X = x)$ , then  $(f, \lambda)$  is not self-fulfilling.*

**Remark 6** *Proposition 5 gives sufficient conditions for an OPM yielding a self-fulfilling prophecy. When  $Y = 1$  is preferable, meaning the new policy treats only those with a favorable predicted outcome (e.g. under resource scarcity), this happens when the treatment effect is beneficial for all values of  $X$ . When instead  $Y = 0$  is preferable, meaning the ‘treat high-risk patients’-setting, the sufficient condition is that treatment is detrimental for all values of  $X$ . Treatments that are always detrimental are less likely to be used in practice as most often treatments are approved for use after they are proven to be beneficial on average with an RCT. In this case of ‘treat high risk’, self-fulfilling prophecies may still occur when the treatment is detrimental to a subgroup of patients.*

**Remark 7** *Proposition 5 does not depend on the OPM’s discrimination in the historical data, meaning that models with ‘good’ discrimination (i.e. high AUC) and ‘bad’ discrimination (low AUC) are equally susceptible to yielding self-fulfilling prophecies under the conditions of the proposition.*

Now we know when OPMs are self-fulfilling and thus have good post-deployment discrimination, but can these self-fulfilling OPMs also be harmful? Proposition 8 indicates that they can:

**Proposition 8 (Harmful)**

*Under the assumptions of Theorem 4, when  $Y = 1$  is preferable,  $f$  is harmful for the group with  $X = x$  iff*

1.  $\pi_0(x) = 1$  and  $\pi_f(x) = 0$  and  $p(Y_1 = 1|X = x) > p(Y_0 = 1|X = x)$  or
2.  $\pi_0(x) = 0$  and  $\pi_f(x) = 1$  and  $p(Y_1 = 1|X = x) < p(Y_0 = 1|X = x)$

*When  $Y = 0$  is preferable, the inequality signs reverse.*

The conditions of this Proposition indicate that, as one would expect, removing the treatment from this group is harmful iff  $p(Y_1 = 1|X = x) > p(Y_0 = 1|X = x)$  (assuming  $Y = 1$  is preferable), i.e. if the effect of the treatment was positive for this group. Conversely, adding treatment to group with  $X = x$  is damaging iff  $p(Y_1 = 1|X = x) < p(Y_0 = 1|X = x)$  (when  $Y = 1$  is preferable), meaning that the treatment decreases the outcome for the group.

interpretation of $Y = 1$ (and policy)	$\pi_0$	$AUC(\pi_f) - AUC(\pi_0)$	OPM deployment was
undesirable (treat high risk patients)	0 (treat no one)	>0 (self-fulfilling)	harmful
	0 (treat no one)	<0 (not self-fulfilling)	beneficial
	1 (treat everyone)	>0 (self-fulfilling)	beneficial
	1 (treat everyone)	<0 (not self-fulfilling)	harmful
desirable (treat low risk patients)	0 (treat no one)	>0 (self-fulfilling)	beneficial
	0 (treat no one)	<0 (not self-fulfilling)	harmful
	1 (treat everyone)	>0 (self-fulfilling)	harmful
	1 (treat everyone)	<0 (not self-fulfilling)	beneficial

Table 1: Overview of when OPM deployment was harmful, based on three pieces of information that are available post-deployment. This table excludes the trivial case where nothing changes post-deployment (see Theorem 11).  $\pi_0$ : historical treatment policy (either treat everyone or treat no one in our setup);  $AUC(\pi_f)$ : AUC in distribution post deployment;  $AUC(\pi_0)$ : AUC in distribution pre deployment; OPM: outcome prediction model

**Remark 9 (harmful OPMs are marginally harmful)** *Under the assumptions of Theorem 4, OPMs that are harmful for one subgroup are also harmful on average, as the other subgroup’s treatment policy and outcomes do not change.*

Taking together Proposition 5 on when OPMs yield self-fulfilling prophecies and Proposition 8 on when OPM deployment is harmful, we reach the perhaps surprising conclusion of Theorem 4: even in the simple setup of binary treatment and binary  $X$ , some OPMs are both self-fulfilling prophecies, and thus demonstrate good post-deployment discrimination, and harm a patient subgroup when deployed. We present an example based on realistic medical assumptions in Appendix A. In Table 1 we list the cases in which OPM deployment is harmful, based on three pieces of information that are available post-deployment: i) is  $Y = 1$  preferable or undesirable? ii) was the historical policy ‘treat everyone’ or ‘treat no one’? and iii) did the AUC of the OPM increase post-deployment compared to the AUC pre-deployment (i.e. is the OPM self-fulfilling)? Finally, we note that the performance of the OPM on the historical data does not feature in the assumptions or statement of Proposition 8. This entails, contrary to what some may expect, that a high performance on historical data, including external validation, provides no guarantee on whether the OPM-driven policy will be beneficial.

### 3.2. OPMs that are calibrated pre- and post-deployment are not useful for treatment decisions

Monitoring discrimination post-deployment and naively interpreting good post-deployment discrimination as a safe deployment is thus not a good strategy, as self-fulfilling prophecies have good post-deployment discrimination but can still be harmful depending on the context. We now turn to another key metric of OPMs predicting the risk of an outcome: *calibration* [11, 12, 13] and investigate how post-deployment calibration relates to harmful policies. We use the following definition of calibration:

**Definition 10** *Let  $p(X, Y)$  be a joint distribution over feature  $X$  and binary outcome  $Y$ , and  $f : X \rightarrow [0, 1]$  an OPM.  $f$  is calibrated with respect to  $p(X, Y)$  if, for all  $\alpha \in [0, 1]$  in the range of  $f$ ,  $\mathbb{E}_{X, Y \sim p(X, Y)}[Y | f(X) = \alpha] = \alpha$ .*

We distinguish two distributions  $p_i(Y = 1 | X)$  on which an OPM can be calibrated depending on the treatment policy indicated with  $i \in \{0, f\}$ . Theorem 4 states that harmful OPMs can have good pre- and post-deployment discrimination, but can they also have good calibration? The following theorem shows that OPMs that are calibrated pre- and post-deployment do not lead to better treatment decisions.

**Theorem 11** *Let  $f$  be an OPM that is calibrated on historical data and  $\pi_f$  be non constant. Such OPM is calibrated on the deployment distribution iff for every  $x \in \mathcal{X}$ :*

$$\pi_0(x) = \pi_f(x) \text{ or } p(Y_1 = 1|X = x) = p(Y_0 = 1|X = x) \quad (7)$$

Note that this entails that for all  $x \in \mathcal{X}$  either the treatment policy does not change, or it changes where it is irrelevant because for that value of  $X$  the treatment effect is zero. Both cases imply the implementation of the OPM is inconsequential. This may seem counterintuitive, but an OPM being calibrated both before and after deployment means the distribution has not changed, so the policy remains the same or the policy was changed where it is irrelevant (i.e. no treatment effect). So an OPM that is calibrated on the development cohort, which remains calibrated post deployment is not a useful OPM.

## 4. Related work

Previous work noted that prediction accuracy does not equal value for treatment decision making [6, 14, 15]. Here we exactly characterize a set of prediction models that yield harmful self-fulfilling prophecies. The idea that model deployment changes the distribution and affects model performance was noted in several lines of previous work. Several authors noted that model performance may degrade over time due to the effect of deployment of the model [16, 17], but we study the case where model performance does *not* degrade but the implementation of it still caused harm. In fact, degraded discrimination may indicate benefit of the deployment. Perdomo et al. [18] and Liley et al. [19] study the setting of performing successive model updates, each time after deploying the previous model for decision making. Perdomo et al. [18] study when over successive deployments predictive performance stabilizes or reaches optimality, and Liley et al. [19] study both model stability and the effect of model deployment on outcomes. Our work may be seen as a special case of these works with only a single model deployment and no model update, but we add new insights as we describe exactly *when* a single model deployment leads to harm and good post-deployment discrimination. Several groups have studied out-of-distribution generalization and its connections to causality and invariance [20, 21, 22] with the aim of removing a model’s dependency on *spurious correlations*. Again our work differs as we are interested in characterizing model performance following a very specific distribution change (a treatment policy change induced by a prediction model) that is particularly relevant in health care, and our main concern is the effect of this policy change on outcomes. Finally, current guidelines on prediction model validation and deployment focus on discrimination and calibration only, not on these newer invariance metrics [5, 15].

## 5. Discussion

We showed how OPMs can be harmful self-fulfilling prophecies, meaning they lead to patient harm when used for treatment decision making, but retain good discrimination after deployment. Moreover, we showed that when a model is well calibrated before and after deployment it is not useful for treatment decision making. The upshot of these findings is not only that harmful and self-fulfilling policies exist, but also that in some scenarios it is even *desirable* to see worse discrimination after deployment, since this signals a beneficial new policy in terms of patient outcomes. These results cast doubt on the adequacy of current practice for the evaluation of predictive models post deployment, when these models are used for decision making.

In recent years, the United States Food and Drug Administration is developing protocols on regulating artificial intelligence based software for medical applications. Their guiding principles explicitly include a total product life-cycle approach, where post-deployment monitoring and certain potential model updates are foreseen and described during initial approval, both with the aim to ensure post-deployment safety for example under dataset shifts, but also to avoid the need for re-approval after each model update. Though their guiding principles on ‘good machine learning practice’ [23] and ‘Predetermined Change Control Plans’ [24] both mention post-deployment monitoring for safety, the intended monitoring seems to center mostly around predictive performance, which our results demonstrate to be insufficient to protect against harmful



self-fulfilling prophecies. Requiring explicit monitoring of changes in patient outcomes over time and changes in treatment policy may in some cases be warranted. Though monitoring patient outcomes in important pre-determined patient subgroups before and after deployment may detect harmful model deployments, before-after comparisons are plagued by well known biases such as potential concurrent changes in policies or general time-trends in outcomes. The best experiment to demonstrate the safety of deploying an OPM is to conduct a cluster randomized controlled trial, where some care-givers are randomly selected to have access to the OPM and others are not. The difference in average outcomes of patients between the care-givers with and without access determines whether using the OPM led to better patient outcomes. How to pre-specify safe model updates in a total product life-cycle approach after a cluster randomized trial in light of our self-fulfilling prophecy framework is left for future work.

Some limitations remain, encoded in the assumptions of our formal results. The setting we describe is kept simple on purpose, a choice that helps in pinpointing the problem but limits somewhat the applicability of this theory to real-world use cases. The extension of our results to other feature types (continuous or categorical  $X$ ), non-threshold based policies, or to a  $\pi_0$  that is not constant (i.e. varies with  $X$ ) or is non-deterministic, is left to future work. Other more complex use cases worth investigating might display policies that are harmful for subgroups identified by variables not included in the list of predictors of the model. The continuation of this line of work entails the re-evaluation of the metrics to monitor and assess a model’s effectiveness, and given that model deployments for decision support are interventions, this will benefit from using the language of causal inference.

## References

- [1] Ewout W Steyerberg. *Applications of prediction models*. Springer, 2009 (cit. on p. 2).
- [2] Ramon Salazar et al. “Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer”. In: *Journal of Clinical Oncology* 29.1 (Jan. 2011). 384 citations (Crossref) [2021-08-06] Publisher: Wolters Kluwer, pp. 17–24. ISSN: 0732-183X. DOI: [10/d2zq5b](https://doi.org/10.1200/JCO.2010.30.1077). URL: <https://ascopubs.org/doi/10.1200/JCO.2010.30.1077> (visited on 08/06/2021) (cit. on p. 2).
- [3] Donna K. Arnett et al. “2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines”. en. In: *Circulation* 140.11 (Sept. 2019). ISSN: 0009-7322, 1524-4539. DOI: [10.1161/CIR.0000000000000678](https://doi.org/10.1161/CIR.0000000000000678). URL: <https://www.ahajournals.org/doi/10.1161/CIR.0000000000000678> (visited on 09/04/2023) (cit. on p. 2).
- [4] Kunal N. Karmali et al. “Blood pressure-lowering treatment strategies based on cardiovascular risk versus blood pressure: A meta-analysis of individual participant data”. eng. In: *PLoS medicine* 15.3 (Mar. 2018), e1002538. ISSN: 1549-1676. DOI: [10.1371/journal.pmed.1002538](https://doi.org/10.1371/journal.pmed.1002538) (cit. on p. 2).
- [5] Michael W. Kattan et al. “American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine”. eng. In: *CA: a cancer journal for clinicians* 66.5 (Sept. 2016), pp. 370–374. ISSN: 1542-4863. DOI: [10.3322/caac.21339](https://doi.org/10.3322/caac.21339) (cit. on pp. 2, 7).
- [6] Wouter A. C. van Amsterdam et al. *Decision making in cancer: Causal questions require causal answers*. arXiv:2209.07397 [cs, stat]. Sept. 2022. DOI: [10.48550/arXiv.2209.07397](https://doi.org/10.48550/arXiv.2209.07397). URL: <http://arxiv.org/abs/2209.07397> (visited on 03/15/2023) (cit. on pp. 2, 7).
- [7] T. Cerny et al. “Pretreatment prognostic factors and scoring system in 407 small-cell lung cancer patients”. en. In: *International Journal of Cancer* 39.2 (Feb. 1987), pp. 146–149. ISSN: 0020-7136, 1097-0215. DOI: [10.1002/ijc.2910390204](https://doi.org/10.1002/ijc.2910390204). URL: <https://onlinelibrary.wiley.com/doi/10.1002/ijc.2910390204> (visited on 01/26/2024) (cit. on p. 2).



- [8] Raphael Haggmann, Alfred Zippelius, and Sacha I. Rothschild. “Validation of Pretreatment Prognostic Factors and Prognostic Staging Systems for Small Cell Lung Cancer in a Real-World Data Set”. en. In: *Cancers* 14.11 (May 2022), p. 2625. ISSN: 2072-6694. DOI: [10.3390/cancers14112625](https://doi.org/10.3390/cancers14112625). URL: <https://www.mdpi.com/2072-6694/14/11/2625> (visited on 01/26/2024) (cit. on p. 2).
- [9] Roberta Ferraldeschi et al. “Modern Management of Small-Cell Lung Cancer:” en. In: *Drugs* 67.15 (2007), pp. 2135–2152. ISSN: 0012-6667. DOI: [10.2165/00003495-200767150-00003](https://doi.org/10.2165/00003495-200767150-00003). URL: <http://link.springer.com/10.2165/00003495-200767150-00003> (visited on 01/26/2024) (cit. on p. 2).
- [10] J. A. Hanley and B. J. McNeil. “The meaning and use of the area under a receiver operating characteristic (ROC) curve”. eng. In: *Radiology* 143.1 (Apr. 1982), pp. 29–36. ISSN: 0033-8419. DOI: [10.1148/radiology.143.1.7063747](https://doi.org/10.1148/radiology.143.1.7063747) (cit. on p. 4).
- [11] Ana Carolina Alba et al. “Discrimination and calibration of clinical prediction models: users’ guides to the medical literature”. In: *Jama* 318.14 (2017), pp. 1377–1384 (cit. on p. 6).
- [12] Yingxiang Huang et al. “A tutorial on calibration measurements and calibration models for clinical prediction models”. In: *Journal of the American Medical Informatics Association* 27.4 (2020), pp. 621–633 (cit. on p. 6).
- [13] Ben Van Calster et al. “Calibration: the Achilles heel of predictive analytics”. In: *BMC medicine* 17.1 (2019), pp. 1–7 (cit. on p. 6).
- [14] Andrew J Vickers and Elena B Elkin. “Decision curve analysis: a novel method for evaluating prediction models”. In: *Medical Decision Making* 26.6 (2006), pp. 565–574 (cit. on p. 7).
- [15] Karel G.M. Moons et al. “Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration”. en. In: *Annals of Internal Medicine* 162.1 (Jan. 2015), W1. ISSN: 0003-4819. DOI: [10/gfrkkz](https://doi.org/10/gfrkkz). URL: <http://annals.org/article.aspx?doi=10.7326/M14-0698> (visited on 05/19/2021) (cit. on p. 7).
- [16] Matthew C Lenert, Michael E Matheny, and Colin G Walsh. “Prognostic models will be victims of their own success, unless...” en. In: *Journal of the American Medical Informatics Association* 26.12 (Dec. 2019), pp. 1645–1650. ISSN: 1527-974X. DOI: [10.1093/jamia/ocz145](https://doi.org/10.1093/jamia/ocz145). URL: <https://academic.oup.com/jamia/article/26/12/1645/5559573> (visited on 08/31/2023) (cit. on p. 7).
- [17] Matthew Sperrin et al. “Explicit causal reasoning is needed to prevent prognostic models being victims of their own success”. en. In: *Journal of the American Medical Informatics Association* 26.12 (Dec. 2019), pp. 1675–1676. ISSN: 1527-974X. DOI: [10.1093/jamia/ocz197](https://doi.org/10.1093/jamia/ocz197). URL: <https://academic.oup.com/jamia/article/26/12/1675/5625126> (visited on 08/31/2023) (cit. on p. 7).
- [18] Juan C. Perdomo et al. *Performative Prediction*. arXiv:2002.06673 [cs, stat]. Feb. 2021. URL: <http://arxiv.org/abs/2002.06673> (visited on 08/28/2023) (cit. on p. 7).
- [19] James Liley et al. “Model updating after interventions paradoxically introduces bias”. en. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. ISSN: 2640-3498. PMLR, Mar. 2021, pp. 3916–3924. URL: <https://proceedings.mlr.press/v130/liley21a.html> (visited on 08/31/2023) (cit. on p. 7).
- [20] Martin Arjovsky et al. *Invariant Risk Minimization*. arXiv:1907.02893 [cs, stat]. Mar. 2020. DOI: [10.48550/arXiv.1907.02893](https://doi.org/10.48550/arXiv.1907.02893). URL: <http://arxiv.org/abs/1907.02893> (visited on 09/04/2023) (cit. on p. 7).
- [21] Yoav Wald et al. “On Calibration and Out-of-Domain Generalization”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 2215–2227. URL: [https://papers.nips.cc/paper\\_files/paper/2021/hash/118bd558033a1016fcc82560c65cca5f-Abstract.html](https://papers.nips.cc/paper_files/paper/2021/hash/118bd558033a1016fcc82560c65cca5f-Abstract.html) (visited on 08/28/2023) (cit. on p. 7).
- [22] Aahlad Puli et al. *Out-of-distribution Generalization in the Presence of Nuisance-Induced Spurious Correlations*. arXiv:2107.00520 [cs, stat]. Feb. 2023. DOI: [10.48550/arXiv.2107.00520](https://doi.org/10.48550/arXiv.2107.00520). URL: <http://arxiv.org/abs/2107.00520> (visited on 09/04/2023) (cit. on p. 7).

- [23] FDA. “Good Machine Learning Practice for Medical Device Development: Guiding Principles”. en. In: *FDA* (Oct. 2021). Publisher: FDA. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles> (visited on 01/22/2024) (cit. on p. 7).
- [24] FDA. “Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles”. en. In: *FDA* (Oct. 2023). Publisher: FDA. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles> (visited on 01/22/2024) (cit. on p. 7).
- [25] K. Breur. “Growth rate and radiosensitivity of human tumours—II: Radiosensitivity of human tumours”. en. In: *European Journal of Cancer (1965)* 2.2 (June 1966), pp. 173–188. ISSN: 0014-2964. DOI: [10.1016/0014-2964\(66\)90009-0](https://doi.org/10.1016/0014-2964(66)90009-0). URL: <https://www.sciencedirect.com/science/article/pii/0014296466900090> (visited on 09/20/2021) (cit. on p. 11).
- [26] John Muschelli. “ROC and AUC with a Binary Predictor: a Potentially Misleading Metric”. en. In: *Journal of classification* 37.3 (Oct. 2020). Publisher: NIH Public Access, p. 696. DOI: [10.1007/s00357-019-09345-1](https://doi.org/10.1007/s00357-019-09345-1). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7695228/> (visited on 11/24/2023) (cit. on p. 12).

## Appendix A. Hypothetical example of a harmful self-fulfilling prophecy

We now give a full-fledged hypothetical example based on realistic assumptions that would result in an OPM yielding a policy that is both harmful and self-fulfilling.

Consider the problem of selecting a subset of end-stage cancer patients for palliative radiotherapy. Such treatment has severe side-effects and thus domain experts advise to attempt to reduce over-treatment in the population of cancer patients. To comply with this advice, a medical center needs to decide which patients will not be eligible anymore for the therapy.

The medical center decides to give the therapy to patients with the longest expected overall survival, under the assumption that these patients would be those for whom the side-effects are justifiable. To support this policy, researchers built an OPM to predict the probability of 6-months overall survival based on pre-treatment tumor growth rate using historical patient records from the medical center. Fast-growing tumors are more aggressive so these patients have a shorter survival overall. The medical center decides to use this model to allocate the therapy and tests the model’s discrimination post deployment. Based on this we have the following facts:

1.  $X = 1$ : fast growing tumor,  $X = 0$ : slow-growing tumor;
2.  $\pi_0(X) = 1$ , the historical policy was treating everyone;
3.  $p(Y_0 = 1|X = 0) - p(Y_0 = 1|X = 1) > 0$ , with radiotherapy, patients with fast growing tumors live shorter

A model with a good fit to the data will predict that patients with slow-growing tumors have a higher probability of 6-months survival. We also assume that the new policy is non-constant and favors those with highest predicted outcome, which means that the new policy will be ‘treat patients with slow growing tumors but not those with fast growing tumors’:

$$\pi_f(X) = 1 - X$$

However, it is well known that fast-growing tumors respond better to radiotherapy than slow growing tumors [25]. Based on this we add the following two assumptions:

1.  $p(Y_0 = 1|X = 0) - p(Y_1 = 1|X = 0) = 0$ , radiotherapy is not effective against slow growing tumors;
2.  $\delta := p(Y_0 = 1|X = 1) - p(Y_1 = 1|X = 1) < 0$ , radiotherapy *is* effective for fast growing tumors.

This means that the antecedent of Proposition 5 is satisfied, meaning that  $f$  yields a self-fulfilling prophecy in combination with any threshold  $\lambda$  such that the resulting policy is non-constant. Removing the therapy from the group  $X = 1$  will worsen their outcomes by  $\delta$ , separating the two groups even more and resulting in higher AUC post-deployment.

Moreover, according to the first case of Proposition 8, the OPM is harmful because the new treatment policy leads to worse outcomes for the group with fast growing tumors ( $X = 1$ ). So the OPM-based policy treats exactly the wrong patients: those who do not benefit from treatment still receive it, those who would benefit from treatment do not, but paradoxically it has good discrimination before and after deployment.

## Appendix B. Proofs of main results

### B.1. Proof of Proposition 5.

#### Proof

First we give some elementary definitions and equalities. Define

$$\mu_i(x) = p_i(Y = 1|X = x) = (1 - \pi_i(x))p(Y_0 = 1|X = x) + \pi_i(x)p(Y_1 = 1|X = x) \quad (8)$$

So by the law of total probability we can write

$$p_i(Y = 1) = p_i(X = 0)\mu_i(0) + p_i(X = 1)\mu_i(1) \quad (9)$$

By Bayes rule we have:

$$p_i(X = x|Y = y) = \frac{p_i(Y = y|X = x)p(X = x)}{p_i(Y = y)} \quad (10)$$

Filling in the definition of  $\mu_i(x)$  into 10 using the assumption that  $p_i(X = x) = p(X = x)$  we have in particular:

$$p_i(X = x|Y = 1) = \frac{\mu_i(x)p(X = x)}{p_i(Y = 1)} \quad (11)$$

ROC-curves are created by transforming a continuous-valued function to a binary prediction based on a varying *threshold*  $\tau$  and calculating the *sensitivity* and *specificity* for each value of  $\tau$ :

$$\text{sensitivity} = p(f(X) \geq \tau|Y = 1) \quad (12)$$

$$\text{specificity} = p(f(X) < \tau|Y = 0) \quad (13)$$

For each possible threshold, all predictions under the threshold are labeled *negative* and all predictions greater or equal to the threshold *positive*. In the case of a binary  $X$ ,  $f(X)$  only takes two unique values so the ROC-curve is given by just three points:

1. sensitivity = 1, specificity = 0 ( $\tau = -\infty$ )
2. sensitivity = 0, specificity = 1 ( $\tau = +\infty$ )
3. sensitivity = sens, specificity = spec ( $\tau = \max_X f(X)$ )

See Figure 2. We can directly calculate the AUC by dividing the area under the ROC-curve in two adjacent non-overlapping triangles. This gives us the following expression for the AUC (see also [26]):

$$\text{AUC} = \frac{1}{2}\text{sens} + \frac{1}{2}\text{spec} \quad (14)$$

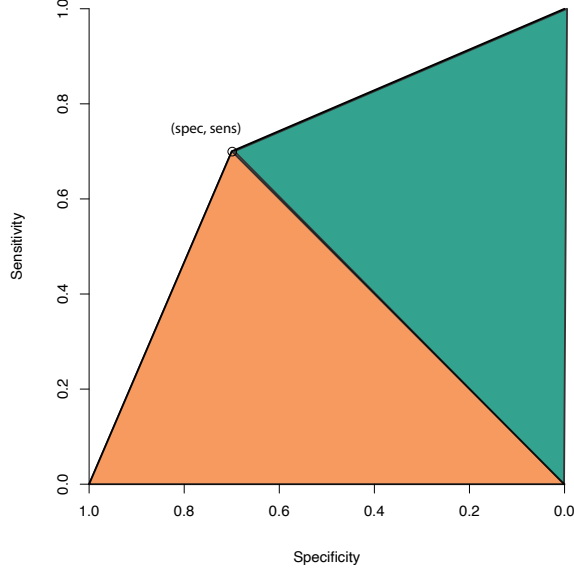
In this binary case, the area-under the ROC curve is thus determined by a single point denoted as (spec,sens). A pair  $(f, \lambda)$  is self-fulfilling when:

$$\text{AUC}(f) - \text{AUC}(0) = \frac{1}{2} (\text{sens}_f + \text{spec}_f - \text{sens}_0 - \text{spec}_0) \geq 0 \quad (15)$$

We structure the proof by first creating an enumeration over all possible scenarios. We assumed  $\pi_f$  is non-constant, which implies that  $f$  varies with  $X$ . Since  $X$  is binary, it must be that either  $f(0) > f(1)$  or  $f(1) > f(0)$ . These cases are symmetric under relabeling of  $X$  so without loss of generality we proceed assuming that  $f(0) > f(1)$  is the case. Since  $\pi_f$  is not constant but  $\pi_0$  is, it must be that either the treatment policy changes for  $X = 0$  but remains the same for  $X = 1$ , or vice versa. This in turn implies that either  $\mu_f(0) = \mu_0(0)$  or  $\mu_f(1) = \mu_0(1)$ .

To provide a proof for the theorem, we enumerate all the subcases based on two factors:

1. for which group does the policy change ( $X = 0$  or  $X = 1$ )?
2. for the group with the policy change, does the outcome under the new policy remain the same (the policy is inconsequential as the treatment effect is zero), increase or decrease (this will be beneficial or detrimental depending on whether  $Y = 1$  is good or bad )


 Figure 2: AUC for a binary predictor  $X$ 

This leads to the following 6 cases:

- policy change for which  $X$ ?

0.  $\pi_f(0) \neq \pi_0(0)$

**effect of policy change:**

$$=: \mu_f(0) = \mu_0(0), \mu_f(1) = \mu_0(1)$$

$$<: \mu_f(0) < \mu_0(0), \mu_f(1) = \mu_0(1)$$

$$>: \mu_f(0) > \mu_0(0), \mu_f(1) = \mu_0(1)$$

1.  $\pi_f(1) \neq \pi_0(1)$

**effect of policy change:**

$$=: \mu_f(0) = \mu_0(0), \mu_f(1) = \mu_0(1)$$

$$<: \mu_f(0) = \mu_0(0), \mu_f(1) < \mu_0(1)$$

$$>: \mu_f(0) = \mu_0(0), \mu_f(1) > \mu_0(1)$$

These 6 combinations cover all possibilities. Since we have that  $f(0) > f(1)$ , by assumption of a non-deterministic  $\pi_f(x) = I_{f(x) > \lambda}$  it must be that for all subcases  $\pi_f(0) = 1$  and  $\pi_f(1) = 0$ . Each of these cases have implications for  $\pi_0$  and, depending on which policy changes,  $p(Y_1 = 1|X = 0) - p(Y_0 = 1|X = 0)$  or  $p(Y_1 = 1|X = 1) - p(Y_0 = 1|X = 1)$ . For instance case  $(0, >)$  specifies that  $\pi_f(0) \neq \pi_0(0)$  so it follows that  $\pi_0 = 0$ . And because  $Y_1(0) = \mu_f(0) > \mu_0(0) = Y_0(0)$  it must be that  $p(Y_1 = 1|X = 0) - p(Y_0 = 1|X = 0) > 0$ , meaning that the treatment increases the outcome for the group with  $X = 0$ .

In the two cases where the outcomes do not change  $((0, =)$  and  $(1, =))$ ,  $(f, \lambda)$  is trivially self-fulfilling as nothing changes in the distribution of  $X, Y$  so the sensitivity and specificity remain the same.

We first prove self-fulfillingness in cases  $(0, >)$  and  $(0, <)$ :

**Case  $(0, >)$  and  $(0, <)$**  We first address case  $(0, >)$ , which gives us this information:

- $\pi_f(0) \neq \pi_0(0)$

- $\mu_f(0) > \mu_0(0)$
- $\mu_f(1) = \mu_0(1)$

Since  $f(0) > f(1)$  we get these sensitivity and specificity:

$$\text{sens}_i = p_i(f(X) \geq \max(f)|Y = 1) = p_i(X = 0|Y = 1) \quad (16)$$

$$\text{spec}_i = p_i(f(X) < \max(f)|Y = 0) = p_i(X = 1|Y = 0) \quad (17)$$

with  $i \in \{0, f\}$ . Plugging this into 15 yields:

$$\begin{aligned} \text{AUC}(f) - \text{AUC}(0) &= \frac{1}{2}(p_f(X = 0|Y = 1) - p_0(X = 0|Y = 1) \\ &\quad + p_f(X = 1|Y = 0) - p_0(X = 0|Y = 0)) \\ &= \frac{1}{2}\left(\mu_f(0) \frac{p(X = 0)}{p_f(Y = 1)} - \mu_0(0) \frac{p(X = 0)}{p_0(Y = 1)} \right. \\ &\quad \left. + (1 - \mu_f(1)) \frac{p(X = 1)}{p_f(Y = 0)} - (1 - \mu_0(1)) \frac{p(X = 1)}{p_0(Y = 0)}\right) \end{aligned}$$

where the first equality is by substitution and rearrangement, and the second by Bayes rule. We can determine the sign of this difference based on the sign of two terms:

$$= \frac{1}{2}\left(p(X = 0) \left( \frac{\mu_f(0)}{p_f(Y = 1)} - \frac{\mu_0(0)}{p_0(Y = 1)} \right) \right. \quad (18)$$

$$\left. + p(X = 1) \left( \frac{1 - \mu_f(1)}{p_f(Y = 0)} - \frac{1 - \mu_0(1)}{p_0(Y = 0)} \right) \right) \quad (19)$$

We write the difference between pre- and post-deployment expected outcome for the group  $X = 0$  as

$$\delta := \mu_f(0) - \mu_0(0) \quad (20)$$

This gives us

$$p_f(Y = 1) = p(X = 1)\mu_f(1) + p(X = 0)\mu_f(0) \quad (21)$$

$$= p(X = 1)\mu_0(1) + p(X = 0)(\mu_0(0) + \delta) \quad (22)$$

$$= p_0(Y = 1) + p(X = 0)\delta \quad (23)$$

where the first step is the law of total probability, the second by the definition of  $\delta$  and the case information  $\mu_f(1) = \mu_0(1)$ , and finally again using the law of total probability. Furthermore

$$p_f(Y = 0) = 1 - p_f(Y = 1) \quad (24)$$

$$= 1 - p_0(Y = 1) - p(X = 0)\delta \quad (25)$$

$$= p_0(Y = 0) - p(X = 0)\delta \quad (26)$$

where the second step is by our previous calculation and the other two just the property of binary outcomes. We can now determine the signs of the two terms in 18.

$$\text{sign}\left[\frac{\mu_f(0)}{p_f(Y=1)} - \frac{\mu_0(0)}{p_0(Y=1)}\right] = \text{sign}\left[\frac{\mu_f(0)p_0(Y=1) - \mu_0(0)p_f(Y=1)}{p_f(Y=1)p_0(Y=1)}\right] \quad (27)$$

$$= \text{sign}[\mu_f(0)p_0(Y=1) - \mu_0(0)p_f(Y=1)] \quad (28)$$

The first equality is cross-multiplying, the second equality is because the product of two probabilities (which are positive by assumption) is always a positive number.

Filling in the definition of  $\delta$  :

$$\text{sign}\left[\frac{\mu_f(0)}{p_f(Y=1)} - \frac{\mu_0(0)}{p_0(Y=1)}\right] \quad (29)$$

$$= \text{sign}[(\mu_0(0) + \delta)p_0(Y=1) - \mu_0(0)(p_0(Y=1) + p(X=0)\delta)] \quad (30)$$

$$= \text{sign}[\delta p_0(Y=1) - \mu_0(0)p(X=0)\delta] \quad (31)$$

$$= \text{sign}[\delta(p_0(Y=1) - \mu_0(0)p(X=0))] \quad (32)$$

$$= \text{sign}[\delta\mu_0(1)p(X=1)] \quad (33)$$

$$= \text{sign}[\delta] \quad (34)$$

In the second equality we remove canceling terms. In the third equality we pull out  $\delta$ . In the fourth equality we use the expansion of  $p_0(Y=1) = p(X=0)\mu_0(0) + p(X=1)\mu_0(1)$ , and for the final equation we note again that  $\mu_0(1)$  and  $p(X=1)$  are positive probabilities so the sign is determined by the sign of  $\delta$ .

Now for the second term of 18:

$$\text{sign}\left[\frac{1 - \mu_f(1)}{p_f(Y=0)} - \frac{1 - \mu_0(1)}{p_0(Y=0)}\right] = \text{sign}\left[\frac{1 - \mu_0(1)}{p_f(Y=0)} - \frac{1 - \mu_0(1)}{p_0(Y=0)}\right] \quad (35)$$

$$= \text{sign}\left[(1 - \mu_0(1))\left(\frac{1}{p_f(Y=0)} - \frac{1}{p_0(Y=0)}\right)\right] \quad (36)$$

$$= \text{sign}\left[\frac{1}{p_f(Y=0)} - \frac{1}{p_0(Y=0)}\right] \quad (37)$$

$$= \text{sign}\left[\frac{p_0(Y=0) - p_f(Y=0)}{p_f(Y=0)p_0(Y=0)}\right] \quad (38)$$

$$= \text{sign}[p_0(Y=0) - p_f(Y=0)] \quad (39)$$

$$= \text{sign}[p_0(Y=0) - p_0(Y=0) + p(X=0)\delta] \quad (40)$$

$$= \text{sign}[p(X=0)\delta] \quad (41)$$

$$= \text{sign}[\delta] \quad (42)$$

The first equality uses the case assumption that  $\mu_f(1) = \mu_0(1)$ . The second equality pulls out the common term  $(1 - \mu_0(1))$ . The third equality follows because  $0 < \mu_0(1) < 1$ . The fourth and fifth equality are cross-multiplying and again using the positive probability property. In the sixth equality we substitute in the definition of  $\delta$ . The seventh equality removes the canceling terms, and the final equality again relies on that  $0 < p(X=0)$ .

So both terms in 18 have the sign of  $\delta$ . In subcase  $(0, >)$   $\delta$  has positive sign, so

$$\text{AUC}(f) - \text{AUC}(0) > 0$$

and  $(f, \lambda)$  is self-fulfilling.

Immediately it is clear that in subcase  $(0, <)$ ,  $(f, \lambda)$  is not self-fulfilling, as subcase  $(0, <)$  equals subcase  $(0, >)$  in all respects except that instead it has a negative sign for  $\delta$ .



**Case (1, >) and (1, <)** We first address case (1, >), which gives us this information:

- $\pi_f(1) \neq \pi_0(1)$
- $\mu_f(0) = \mu_0(0)$
- $\mu_f(1) > \mu_0(1)$

Again we write the difference between pre- and post-deployment expected outcome as  $\delta$ , this time for the group  $X = 1$ :

$$\delta := \mu_f(1) - \mu_0(1) \quad (43)$$

This gives us

$$p_f(Y = 1) = p(X = 1)\mu_f(1) + p(X = 0)\mu_f(0) \quad (44)$$

$$= p(X = 1)(\mu_0(1) + \delta) + p(X = 0)\mu_0(0) \quad (45)$$

$$= p_0(Y = 1) + p(X = 1)\delta \quad (46)$$

where the first step is the law of total probability, the second by the definition of  $\delta$  and the case information  $\mu_f(0) = \mu_0(0)$ , and finally again using the law of total probability. Furthermore

$$p_f(Y = 0) = 1 - p_f(Y = 1) \quad (47)$$

$$= 1 - p_0(Y = 1) - p(X = 1)\delta \quad (48)$$

$$= p_0(Y = 0) - p(X = 1)\delta \quad (49)$$

where the second step is by our previous calculation and the other two just the property of binary outcomes. We can now determine the signs of the two terms in 18.

The first two steps for the first are the same as in the case (0, >) (see Equation 27), after these steps we substitute in the new definition of  $\delta$ :

$$\text{sign}\left[\frac{\mu_f(0)}{p_f(Y = 1)} - \frac{\mu_0(0)}{p_0(Y = 1)}\right] \quad (50)$$

$$= \text{sign}[\mu_f(0)p_0(Y = 1) - \mu_0(0)p_f(Y = 1)] \quad (51)$$

$$= \text{sign}[\mu_0(0)p_0(Y = 1) - \mu_0(0)(p_0(Y = 1) + p(X = 1)\delta)] \quad (52)$$

$$= \text{sign}[-\mu_0(0)p(X = 0)\delta] \quad (53)$$

$$= \text{sign}[-\delta] \quad (54)$$

In the third equality we remove canceling terms. For the final equation we note again that  $\mu_0(0)$  and  $p(X = 0)$  are positive probabilities so the sign is determined by the sign of  $\delta$ .

Now for the second term of 18:

subcase	$\pi_0$	$\pi_f(0)$	$\pi_f(1)$	CATE(0)	CATE(1)	self-fulfilling
0 =	0	1	0	0		yes
0 <	0	1	0	-		no
0 >	0	1	0	+		yes
1 =	1	1	0		0	yes
1 <	1	1	0		+	yes
1 >	1	1	0		-	no

Table 2: Enumeration of all possible subcases. The first column indicates for which value of  $X$  the treatment policy changes. The second column indicates whether this change improves outcomes for that group ( $>$ ), reduces outcomes ( $<$ ) or is irrelevant ( $=$ ).  $+/-$  indicates the sign of the subgroup treatment effect  $\text{CATE}(x) := p(Y_1 = 1|X = x) - p(Y_0 = 1|X = x)$ ;

$$\text{sign}\left[\frac{1 - \mu_f(1)}{p_f(Y = 0)} - \frac{1 - \mu_0(1)}{p_0(Y = 0)}\right] \quad (55)$$

$$= \text{sign}\left[\frac{(1 - \mu_f(1))p_0(Y = 0) - (1 - \mu_0(1))p_f(Y = 0)}{p_f(Y = 0)p_0(Y = 0)}\right] \quad (56)$$

$$= \text{sign}[(1 - \mu_f(1))p_0(Y = 0) - (1 - \mu_0(1))p_f(Y = 0)] \quad (57)$$

$$= \text{sign}[(1 - (\mu_0(1) + \delta))p_0(Y = 0) - (1 - \mu_0(1))(p_0(Y = 0) - p(X = 1)\delta)] \quad (58)$$

$$= \text{sign}[-\delta p_0(Y = 0) - (1 - \mu_0(1))(-p(X = 1)\delta)] \quad (59)$$

$$= \text{sign}[-\delta(p_0(Y = 0) - (1 - \mu_0(1))p(X = 1))] \quad (60)$$

$$= \text{sign}[-\delta((1 - \mu_0(0))p(X = 0))] \quad (61)$$

$$= \text{sign}[-\delta] \quad (62)$$

The first equality uses cross-multiplication to gather the sum. The second equality follows because we're dividing by a positive number. The third equality is filling in the definition on  $\delta$ . The fourth equality removes canceling terms. The fifth equality factors out  $-\delta$ . The seventh equality is by the law of total probability.

So both terms in 18 have the sign of  $-\delta$ . In subcase  $(1, >)$   $\delta$  has positive sign, so

$$\text{AUC}(f) - \text{AUC}(0) < 0$$

and  $(f, \lambda)$  is not self-fulfilling.

Immediately it is clear that in subcase  $(1, <)$ ,  $(f, \lambda)$  is self-fulfilling, as subcase  $(1, <)$  equals subcase  $(1, >)$  in all respects except that instead it has a negative sign for  $\delta$ .

**Enumerating all the cases** As said, in the two cases where the outcomes do not change ( $(0, =)$ ,  $(1, =)$ ),  $(f, \lambda)$  is trivially self-fulfilling.

Putting all the pieces of information for all subcases together in Table 2 we see that when  $p(Y_1 = 1|X = x) - p(Y_0 = 1|X = x) \geq 0$  (the treatment effect is never negative),  $(f, \lambda)$  is self-fulfilling. Also, when  $p(Y_1 = 1|X = x) - p(Y_0 = 1|X = x) < 0$  (the treatment effect is always negative),  $(f, \lambda)$  is never self-fulfilling. These observations conclude the proof. ■

## B.2. Proof of Proposition 8.

Given that we assumed binary  $T$  and  $X$ , we can write the expected value of the outcome conditional on these two variables with four parameters without making parametric assumptions, marginalizing over other

variables different than  $X$  and  $T$ . For ease of interpretation of our results we write the expected value as a sum:

$$p(Y_{T=t} = 1|X = x) = \alpha + \beta_x x + \beta_t t + \beta_{xt} x t \quad (63)$$

Note that this is not an assumption on the generating process of the outcome  $Y$ , which could have arbitrary form, it is only a formal device to represent the four outcomes of interest, one for each value of  $X$  and  $T$ .

We now proceed to prove the Proposition for the case where higher outcome is better; to obtain a proof for the symmetric case (higher outcome is worse) one needs only to switch the sign in the inequalities 64 and 65, along with their specialization in the subcases.

**Proof** A treatment is harmful for the group with  $X = x'$  iff  $p_f(Y = 1|X = x') < p_0(Y = 1|X = x')$ , where according to definition 3  $p_i(Y = 1|X) = \mathbb{E}_{T \sim \pi_i(X)} p(Y_T = 1|X)$  The proof continues as a case distinction depending on the value of  $x'$ .

**Case  $x' = 1$ .** For  $x' = 1$  the definition of harmful translates to

$$(\pi_f(1) - \pi_0(1)) (\beta_t + \beta_{xt}) < 0 \quad (64)$$

We consider the possible values of  $\pi_f$  and  $\pi_0$  in subcases. Note that if  $\pi_f(1) = \pi_0(1)$  the above inequality cannot hold since all terms cancel out and the treatment cannot be harmful (because nothing changes for group  $X = 1$ ), so we only consider subcases where these two differ.

**Subcase 1.** We have  $\pi_f(1) = 0, \pi_f(0) = 1$  and  $\pi_0(x) = 1$ . In this scenario, we were treating everyone and with the new policy we withhold treatment from group  $X = 1$ . In this case statement 64 specializes to  $\beta_t + \beta_{xt} > 0$ , meaning that treatment was beneficial and removing it will do damage to group  $X = 1$ .

**Subcase 2.** We have  $\pi_f(1) = 1, \pi_f(0) = 0$  and  $\pi_0(x) = 0$ . In this scenario, we were treating nobody and with the new policy we introduce treatment for group  $X = 1$ . In this case statement 64 specializes to  $\beta_t + \beta_{xt} < 0$ , meaning that treatment is harmful and adding it damages group  $X = 1$ .

**Case  $x' = 0$ .** For  $x' = 0$  the definition of harmful translates to

$$(\pi_f(0) - \pi_0(0)) \beta_t < 0 \quad (65)$$

Again if  $\pi_f(0) = \pi_0(0)$  the above inequality cannot hold since all terms cancel out and the treatment cannot be harmful (because nothing changes for group  $X = 0$ ), so we only consider subcases where these two differ.

**Subcase 1.** We have  $\pi_f(1) = 0, \pi_f(0) = 1$  and  $\pi_0(x) = 0$ . In this scenario, we were treating nobody and with the new policy we introduce treatment from group  $X = 0$ . In this case the statement 65 specializes to  $\beta_t < 0$ , which is what we intended to prove.

**Subcase 2.** We have  $\pi_f(1) = 1, \pi_f(0) = 0$  and  $\pi_0(x) = 1$ . In this circumstance statement 65 specializes to  $\beta_t > 0$ . ■

### B.3. Proof of Theorem 11.

By assumption  $f$  perfectly fits the historical data, so:

$$f(X = x) = p_0(Y = 1|X = x) = \mathbb{E}_{T \sim \pi_0(x)} p(Y_T = 1|X = x).$$

We now prove that  $f$  is calibrated on the deployment distribution generated by  $\pi_f$  iff for all  $x \in \mathcal{X}$ :

$$\pi_0(x) = \pi_f(x) \text{ or } p(Y_1 = 1|X = x) = p(Y_0 = 1|X = x) \quad (66)$$

**Proof** As a shorthand define:

$$\begin{aligned}\mu_i(x) &:= p_i(Y = 1|X = x) \\ &= (1 - \pi_i(x))p(Y_0 = 1|X = x) + \pi_i(x)p(Y_1 = 1|X = x).\end{aligned}$$

$f$  perfectly fits the historical data so:

$$f(X = x) = \mu_0(x), \forall x \in \mathcal{X}. \quad (67)$$

$f$  is calibrated on the post-deployment distribution when for all  $\alpha \in [0, 1]$  in the range of  $f$ ,  $\mathbb{E}_{X, Y \sim p_f(X, Y)}[Y|f(X) = \alpha] = \alpha$ . So if  $f$  is calibrated on both the historic distribution and the post-deployment distribution we have that:

$$\begin{aligned}& \mathbb{E}_{X, Y \sim p_f(X, Y)}[Y|f(X) = \alpha] \\ &= \mathbb{E}_{X, Y \sim p_f(X, Y)|f(X) = \alpha}[Y] \\ &= \mathbb{E}_{X, Y \sim p_f(X, Y)}[Y \mathbb{1}[f(X) = \alpha]] / \mathbb{E}_{X \sim p_f(X)}[\mathbb{1}[f(X) = \alpha]] \\ &= \mathbb{E}_{X, Y \sim p_0(X, Y)}[Y \mathbb{1}[f(X) = \alpha]] / \mathbb{E}_{X \sim p_0(X)}[\mathbb{1}[f(X) = \alpha]]\end{aligned}$$

Where  $\mathbb{1}[\cdot]$  is used for the indicator function. We first show that this holds iff for every  $x \in \mathcal{X}$ ,  $f(x) = \mu_0(x) = \mu_f(x)$ . Note that in the last two equations above, the denominators are the same as  $p_0(X) = p_f(X)$ , so also the enumerators must be the same, so:

$$\begin{aligned}& \mathbb{E}_{X \sim p_0(X)} \mathbb{E}_{Y \sim p_0(Y|X)}[Y \mathbb{1}[f(X) = \alpha]] = \mathbb{E}_{X \sim p_f(X)} \mathbb{E}_{Y \sim p_f(Y|X)}[Y \mathbb{1}[f(X) = \alpha]] \\ \iff & \mathbb{E}_{X \sim p_0(X)} \mathbb{1}[f(X) = \alpha] \mathbb{E}_{Y \sim p_0(Y|X)}[Y] = \mathbb{E}_{X \sim p_f(X)} \mathbb{1}[f(X) = \alpha] \mathbb{E}_{Y \sim p_f(Y|X)}[Y] \\ \iff & \mathbb{E}_{X \sim p_0(X)} \mathbb{1}[f(X) = \alpha] \mathbb{E}_{Y_0, Y_1|X}[(1 - \pi_0(X))Y_0 + \pi_0(X)Y_1] \\ &= \mathbb{E}_{X \sim p_f(X)} \mathbb{1}[f(X) = \alpha] \mathbb{E}_{Y_0, Y_1|X}[(1 - \pi_f(X))Y_0 + \pi_f(X)Y_1]\end{aligned}$$

Since by assumption  $p_0(X) = p_f(X) = p(X)$  we have that

$$\begin{aligned}\iff & \mathbb{E}_{X \sim p(X)} \mathbb{1}[f(X) = \alpha] \mathbb{E}_{Y_0, Y_1|X}[(1 - \pi_0(X))Y_0 + \pi_0(X)Y_1] \\ &= \mathbb{E}_{X \sim p(X)} \mathbb{1}[f(X) = \alpha] \mathbb{E}_{Y_0, Y_1|X}[(1 - \pi_f(X))Y_0 + \pi_f(X)Y_1] \\ \iff & \mathbb{E}_{X, Y_0, Y_1} \mathbb{1}[f(X) = \alpha] ((1 - \pi_0(X))Y_0 + \pi_0(X)Y_1) \\ &= \mathbb{E}_{X, Y_0, Y_1} \mathbb{1}[f(X) = \alpha] ((1 - \pi_f(X))Y_0 + \pi_f(X)Y_1) \\ \iff & \mathbb{E}_X \mathbb{1}[\mu_0(X) = \alpha] \mu_0(X) = \mathbb{E}_X \mathbb{1}[\mu_f(X) = \alpha] \mu_f(X)\end{aligned}$$

Where in the last line we substituted the definition of  $\mu$  and used the assumption that  $f(X) = \mu_0(X)$ . Finally we note that by assumption  $\pi_f(X)$  is non-constant. As  $X$  is binary it must be that  $f$  is injective. This implies that the expectation in the last line is given by the value of  $\mu$  on a single point corresponding with  $\alpha$  which proves that  $\mu_0(X) = \mu_f(X)$ .

Looking at the difference between  $\mu_0(X)$  and  $\mu_f(X)$  we see that:

$$\begin{aligned}\mu_f(X) - \mu_0(X) &= \\ & ((1 - \pi_f(X))p(Y_0 = 1|X) + \pi_f(X)p(Y_1 = 1|X)) - \\ & ((1 - \pi_0(X))p(Y_0 = 1|X) + \pi_0(X)p(Y_1 = 1|X)) \\ &= (\pi_f(X) - \pi_0(X)) (p(Y_1 = 1|X) - p(Y_0 = 1|X))\end{aligned}$$

Hence the difference  $\mu_f(X) - \mu_0(X)$  is zero iff at least one of the last two terms is zero. This means that  $f$  is calibrated on the deployment distribution iff for every  $x$  either  $\pi_f(x) = \pi_0(x)$  or  $p(Y_1 = 1|X) = p(Y_0 = 1|X)$  ■