

# Causal Measures for Prognostic and Predictive Biomarkers

Yasutaka Chiba

Clinical Research Center, Kinki University Hospital, Osaka, Japan

Email: [chibay@med.kindai.ac.jp](mailto:chibay@med.kindai.ac.jp)

**How to cite this paper:** Chiba, Y. (2018) Causal Measures for Prognostic and Predictive Biomarkers. *Open Journal of Statistics*, 8, 241-248.  
<https://doi.org/10.4236/ojs.2018.82014>

**Received:** November 11, 2017

**Accepted:** March 30, 2018

**Published:** April 2, 2018

Copyright © 2018 by author and Scientific Research Publishing Inc.  
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).  
<http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

Researchers conducting randomized clinical trials with two treatment groups sometimes wish to determine whether biomarkers are predictive and/or prognostic. They can use regression models with interaction terms to assess the role of the biomarker of interest. However, although the interaction term is undoubtedly a suitable measure for prediction, the optimal way to measure prognosis is less clear. In this article, we define causal measures that can be used for prognosis and prediction based on biomarkers. The causal measure for prognosis is defined as the average of two differences in status between biomarker-positive and -negative subjects under treatment and control conditions. The causal measure for prediction is defined as the difference between the causal effect of the treatment for biomarker-positive and biomarker-negative subjects. We also explain the relationship between the proposed measures and the regression parameters. The causal measure for prognosis corresponds to the terms for the biomarker in a regression model, where the values of the dummy variables representing the explanatory variables are  $-1/2$  or  $1/2$ . The causal measure for prediction is simply the causal effect of the interaction term in a regression model. In addition, for a binary outcome, we express the causal measures in terms of four response types: always-responder, complier, non-complier, and never-responder. The causal measure for prognosis can be expressed as a function of always- and never-responders, and the causal measure for prediction as a function of compliers and non-compliers. This enables us to demonstrate that the proposed measures are plausible in the case of a binary outcome. Our causal measures should be used to assess whether a biomarker is prognostic and/or predictive.

## Keywords

Causal Inference, Interaction, Potential Outcome, Response Type

---

## 1. Introduction

There are some cases in which the responses to a treatment differ remarkably among individuals. A cause of this remarkable difference may be a biomarker. Therefore, it is important for personalized medicine to investigate the relationship between a biomarker and clinical significance (efficacy and/or safety). For example, Baselga *et al.* [1] demonstrated that, in women with HER2-positive metastatic breast cancer undergoing first-line therapy, patients with tumors harboring a *PIK3CA* mutation had worse progression-free survival (PFS) compared with those with *PIK3CA* wild-type tumors, regardless of treatment group. Brugger *et al.* [2] demonstrated that, for the use of erlotinib maintenance treatment for advanced non-small-cell lung cancer, the *EGFR*-mutated subgroup derived a much greater benefit compared with the wild-type subgroup in terms of hazard ratio for PFS.

As seen in the above examples, in a randomized clinical trial with two treatment groups, researchers can investigate whether a particular biomarker is prognostic and/or predictive. According to Ballman [3], a prognostic biomarker informs us about likely outcomes independently of the treatment received, such as the above *PIK3CA* mutation, and a predictive biomarker informs us that the effect of a treatment depends on whether the subject is positive for that biomarker, such as the above *EGFR* mutation. Researchers can investigate these questions using regression models including explanatory variables representing the treatment group, biomarker, and treatment  $\times$  biomarker interaction term. A biomarker is said to be predictive if the interaction term is not 0. If the interaction term is 0 but the term for the biomarker is not, then the biomarker is said to be prognostic. The measure for the prognosis is not clearly defined when the interaction term is not 0.

To understand this more clearly, consider the following regression model:

$$E(Y | X = x, Z = z) = \alpha + \beta x + \gamma z + \delta xz, \quad (1)$$

where  $X$  is the assigned treatment ( $X = t$  if the subject is assigned to the treatment group and  $X = c$  if the subject is assigned to the control group),  $Z$  is the biomarker ( $Z = p$  if the subject is positive and  $Z = n$  if the subject is negative), and  $Y$  is the outcome. Note that we can substitute  $E(Y | X = x, Z = z)$  with  $\Pr(Y = 1 | X = x, Z = z)$  in the case of a binary outcome ( $Y = 1$  if a response is observed and  $Y = 0$  if no response is observed). We can also replace  $E(Y | X = x, Z = z)$  by  $\log \Pr(Y = 1 | X = x, Z = z)$  for the risk ratio, and  $\log(\text{hazard})$  for the time-to-event.

If the dummy variables in (1) are set to  $t = p = 1$  and  $c = n = 0$ , the interaction term,  $\delta$ , is

$$\begin{aligned} \delta = & \{E(Y | X = t, Z = p) - E(Y | X = c, Z = p)\} \\ & - \{E(Y | X = t, Z = n) - E(Y | X = c, Z = n)\} \end{aligned} \quad (2)$$

It is guaranteed that  $\delta$  is a plausible measure for prediction because, under

randomization, it indicates the difference in the effects of the treatment between biomarker-positive and -negative subjects. If  $\delta = 0$ , the term for the biomarker,  $\gamma$ , is

$$\begin{aligned}\gamma &= E(Y | X = c, Z = p) - E(Y | X = c, Z = n) \\ &= E(Y | X = t, Z = p) - E(Y | X = t, Z = n)\end{aligned}\quad (3)$$

As this is a common difference between biomarker-positive and -negative subjects under both treatment and control conditions,  $\gamma$  is a plausible measure for the prognosis. However, if  $\delta \neq 0$ ,

$$\gamma = E(Y | X = c, Z = p) - E(Y | X = c, Z = n), \quad (4)$$

and  $\gamma$  only concern the control condition, not the treatment condition. This makes the plausibility of  $\gamma$  as a measure for the prognosis questionable. When  $\delta \neq 0$ , in addition to the value of  $\gamma$ , we need to calculate

$$\gamma + \delta = E(Y | X = t, Z = p) - E(Y | X = t, Z = n). \quad (5)$$

Consequently, it is not obvious that any of the terms can be used to measure the prognosis.

The main aim of this article is to define causal measures for assessing whether a biomarker is prognostic and/or predictive. We also explain how the proposed causal measures are related to regression parameters. In addition, we express the proposed causal measures in terms of the response type in the potential outcome framework [4] [5] for a binary outcome. This enables us to demonstrate that the suggested prognosis and prediction measures are plausible.

## 2. Methods

We define causal measures for prognosis and prediction by adopting the potential outcome framework. This corresponds to the outcome that would occur if the subject were assigned to a specific value for the treatment. We denote  $Y(x)$  as the potential outcome if a subject is assigned to  $X = x$ . Our work is based on the following three assumptions: the stable unit treatment value assumption (SUTVA) [5], which states that there is only a single version of each treatment level and no interference among subjects; the consistency assumption [6] that  $Y(x) = Y$  for all subjects, such that the value of  $Y$  that would have been observed if  $X$  had been set to its actual value is equal to the value of  $Y$  that was observed; and the exchangeability assumption [7], which states that  $Y(x)$  is independent of  $X$ . The exchangeability assumption is often made when analyzing randomized trials. We cannot know the values of both  $Y(t)$  and  $Y(c)$  for a particular individual, because  $Y(t)$  is observed but  $Y(c)$  is not if s/he is assigned to the treatment group ( $X = t$ ). Conversely,  $Y(t)$  is not observed and  $Y(c)$  is observed if s/he is assigned to the control group ( $X = c$ ).

To demonstrate that the proposed causal measures are plausible measures for prognosis and prediction in the case of a binary outcome, we define subjects as having one of the following four response types (e.g., [8] [9]):

- Always-responder: Would be a responder regardless of which group s/he was assigned to; *i.e.*,  $(Y(t), Y(c)) = (1, 1)$ .
- Complier: Would be a responder if assigned to the treatment group but a non-responder if assigned to the control group; *i.e.*,  $(Y(t), Y(c)) = (1, 0)$ .
- Non-complier: Would be a non-responder if assigned to the treatment group but a responder if assigned to the control group; *i.e.*,  $(Y(t), Y(c)) = (0, 1)$ .
- Never-responder: Would be a non-responder regardless of which group s/he was assigned to; *i.e.*,  $(Y(t), Y(c)) = (0, 0)$ .

All subjects belong to one of these four groups. However, we cannot know which group a subject belongs to, because one of  $Y(t)$  and  $Y(c)$  is unknown. Notably, the outcomes for always- and never-responders do not depend on whether they are assigned to the treatment group or the control group, but the outcomes for compliers and non-compliers do depend on which group they are assigned to.

### 3. Results

#### 3.1. Causal Measures for Prognosis and Prediction

We define the causal measure for prognosis as follows:

$$\frac{[E\{Y(t)|Z=p\} - E\{Y(t)|Z=n\}] + [E\{Y(c)|Z=p\} - E\{Y(c)|Z=n\}]}{2} \quad (6)$$

where  $E\{Y(t)|Z=p\}$  is the expected outcome if all biomarker-positive subjects are assigned to the treatment group, and  $E\{Y(t)|Z=n\}$  is the same in the case of all biomarker-negative subjects. Likewise,  $E\{Y(c)|Z=p\}$  and  $E\{Y(c)|Z=n\}$  indicate the expected outcomes if these subjects are assigned to the control group. Thus,  $E\{Y(x)|Z=p\} - E\{Y(x)|Z=n\}$  is the difference between the outcomes when the biomarker is positive/negative under the same treatment condition. Indeed, as (6) informs us about likely outcomes independent of the treatment received, it is a measure for the prognostic biomarker according to Ballman [3]. It is important to note that (6) treats the differences between biomarker-positive and -negative subjects under treatment and control conditions equally.

We define the causal measure for prediction as follows:

$$E\{Y(t) - Y(c) | Z = p\} - E\{Y(t) - Y(c) | Z = n\}. \quad (7)$$

This represents the difference between the causal effect of the treatment for biomarker-positive and -negative subjects. Therefore, (7) indicates whether the effect of the treatment depends on the positive/negative for the biomarker. Hence, this corresponds to a measure for the predictive biomarker discussed in Ballman [3].

#### 3.2. Relation to Regression Model

Under the three assumptions (SUTVA, consistency, and exchangeability),  $E\{Y(x)|Z=z\}$  can be expressed as follows:

$$E\{Y(x)|Z=z\} = E\{Y(x)|X=x, Z=z\} = E(Y|X=x, Z=z). \quad (8)$$

By substituting this into (1) and setting the dummy variables to  $t = p = 1/2$  and  $c = n = -1/2$ , we obtain the following equations:

$$E\{Y(t)|Z=p\} = \alpha + \frac{1}{2}\beta + \frac{1}{2}\gamma + \frac{1}{4}\delta, \quad (9)$$

$$E\{Y(t)|Z=n\} = \alpha + \frac{1}{2}\beta - \frac{1}{2}\gamma - \frac{1}{4}\delta, \quad (10)$$

$$E\{Y(c)|Z=p\} = \alpha - \frac{1}{2}\beta + \frac{1}{2}\gamma - \frac{1}{4}\delta, \quad (11)$$

$$E\{Y(c)|Z=n\} = \alpha - \frac{1}{2}\beta - \frac{1}{2}\gamma + \frac{1}{4}\delta. \quad (12)$$

These four equations derive that (6) =  $\gamma$  and (7) =  $\delta$ . This implies that (6) can be expressed by one regression parameter regardless of the interaction term being 0 when we use a regression model with dummy variables set to  $t = p = 1/2$  and  $c = n = -1/2$ . Hence,  $\gamma$  can be used as a measure for prognosis. We note that, when we use ordinal dummy variables  $t = p = 1$  and  $c = n = 0$ , we again obtain (7) =  $\delta$ , but (6) =  $\gamma + \delta/2$ , which is different to  $\gamma$  when  $\delta \neq 0$ . This implies that (6) can only be expressed by one regression parameter when  $\delta = 0$ .

### 3.3. Expression by Response Type on a Binary Outcome

As mentioned above, we can replace  $E\{Y(x)|Z=z\}$  by  $\Pr\{Y(x)=1|Z=z\}$ ,  $\log \Pr\{Y(x)=1|Z=z\}$ , or  $\log(\text{hazard})$ . However, in this section, we only discuss the case where  $E\{Y(x)|Z=z\} = \Pr\{Y(x)=1|Z=z\}$ .

When we have a binary outcome,  $E\{Y(t)|Z=z\} + E\{Y(c)|Z=z\}$  can be expressed as follows:

$$\begin{aligned} & E\{Y(t)|Z=z\} + E\{Y(c)|Z=z\} \\ &= \Pr\{Y(t)=1|Z=z\} + \Pr\{Y(c)=1|Z=z\} \\ &= \Pr\{Y(t)=1, Y(c)=1|Z=z\} + \Pr\{Y(t)=1, Y(c)=0|Z=z\} \\ &\quad + \Pr\{Y(t)=1, Y(c)=1|Z=z\} + \Pr\{Y(t)=0, Y(c)=1|Z=z\} \\ &= \Pr\{Y(t)=1, Y(c)=1|Z=z\} + 1 - \Pr\{Y(t)=0, Y(c)=0|Z=z\}, \end{aligned} \quad (13)$$

where  $z = p, c$ . Substituting this into (6) gives

$$\begin{aligned} (6) &= 0.5 \times \left( \left[ \Pr\{Y(t)=Y(c)=1|Z=p\} - \Pr\{Y(t)=Y(c)=1|Z=n\} \right] \right. \\ &\quad \left. + \left[ \Pr\{Y(t)=Y(c)=0|Z=n\} - \Pr\{Y(t)=Y(c)=0|Z=p\} \right] \right). \end{aligned} \quad (14)$$

This is the average of two differences: the differences in the proportions of biomarker-positive and -negative subjects who are always-responders and never-responders. In the first difference, we subtract the proportion of biomarker-negative subjects who are always-responders. However, in the second difference, we subtract the proportion of biomarker-positive subjects who are never-responders. This is because, if the proportion of biomarker-negative subjects who are always-responders is higher than that of biomarker-negative subjects, the proportion of

biomarker-negative subjects who are never-responders would be higher than that of biomarker-positive subjects. Again, always- and never-responders are subjects who would have the same outcome regardless of which group they are assigned to. Therefore, (3) is a plausible causal measure for prognosis. We note that the causal measure version of (4),  $E\{Y(c)|Z=p\} - E\{Y(c)|Z=n\}$ , cannot be expressed as a function of only always- and never-responders.

In the case of a binary outcome,  $E\{Y(t)|Z=z\} - E\{Y(c)|Z=z\}$  can be expressed in a similar way to the above calculation:

$$\begin{aligned}
 & E\{Y(t)|Z=z\} - E\{Y(c)|Z=z\} \\
 &= \Pr\{Y(t)=1|Z=z\} - \Pr\{Y(c)=1|Z=z\} \\
 &= \left[ \Pr\{Y(t)=1, Y(c)=1|Z=z\} + \Pr\{Y(t)=1, Y(c)=0|Z=z\} \right] \\
 &\quad - \left[ \Pr\{Y(t)=1, Y(c)=1|Z=z\} + \Pr\{Y(t)=0, Y(c)=1|Z=z\} \right] \\
 &= \Pr\{Y(t)=1, Y(c)=0|Z=z\} - \Pr\{Y(t)=0, Y(c)=1|Z=z\},
 \end{aligned} \tag{15}$$

where  $z = p, c$ . Substituting this into (7) gives

$$\begin{aligned}
 (7) &= \left[ \Pr\{Y(t)=1, Y(c)=0|Z=p\} - \Pr\{Y(t)=0, Y(c)=1|Z=p\} \right] \\
 &\quad - \left[ \Pr\{Y(t)=1, Y(c)=0|Z=n\} - \Pr\{Y(t)=0, Y(c)=1|Z=n\} \right].
 \end{aligned} \tag{16}$$

This is the difference between two differences; the differences in the proportions of compliers and non-compliers in biomarker-positive and -negative subjects. Again, the outcome for compliers and non-compliers depends on which group they are assigned to. Hence, (7) is a plausible causal measure for prediction.

#### 4. Discussion

In this article, we defined causal measures that indicate whether a biomarker is prognostic and/or predictive. Our measure for prediction is the causal measure version of the interaction term used in ordinal regression analysis. However, we do not use the term for the biomarker in an ordinal regression model, which uses dummy variables set to 0 or 1, as our measure for prognosis. The measure is not calculated by comparing the outcomes of biomarker-positive and -negative subjects in the control group. Our measure takes both the treatment and control conditions into account instead of only the control condition. Specifically, on the difference scale for a binary outcome, this measure can be expressed as a function of only always- and never-responders. The outcomes for these subjects do not depend on which treatment group they are assigned to. Hence, our causal measure is a plausible measure for prognosis.

Researchers should use the causal measures defined in this article to determine whether a biomarker is prognostic and/or predictive. When trials are analyzed using a regression model, it is convenient to represent the explanatory variables by dummy variables with values  $-1/2$  or  $1/2$ . When we make this choice of values, we can express the causal measures in terms of the regression parame-

ters.

Recently, some authors [9] [10] [11] [12] have discussed approaches to infer causal effects defined on the basis of the response type. An interesting area for future work is to extend the approaches to a statistical method to determine whether a biomarker is prognostic and/or predictive. Such a work will make it possible to evaluate these causal measures, such as (14) and (16), directly.

## Acknowledgements

The author thanks the reviewers for helpful comments. This work was supported partially by Grant-in-Aid for Scientific Research (No. 15K00057) from Japan Society for the Promotion of Science.

## References

- [1] Brugger, W., Triller, N., Blasinska-Morawiec, M., Curescu, S., Sakalauskas R., Manikhas, G.M., Mazieres, J., Whittom, R., Ward, C., Mayne, K., Trunzer, K. and Cappuzzo, F. (2011) Prospective Molecular Marker Analyses of EGFR and KRAS from a Randomized, Placebo-Controlled Study of Erlotinib Maintenance Therapy in Advanced Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology*, **29**, 4113-4120. <https://doi.org/10.1200/JCO.2010.31.8162>
- [2] Baselga, J., Cortés, J., Im, S.A., Clark, M., Ross, G., Kiermaier, A. and Swain, S.M. (2014) Biomarker Analyses in CLEOPATRA: A Phase III, Placebo-Controlled Study of Pertuzumab in Human Epidermal Growth Factor Receptor 2-Positive, First-Line Metastatic Breast Cancer. *Journal of Clinical Oncology*, **32**, 3753-3761. <https://doi.org/10.1200/JCO.2013.54.5384>
- [3] Ballman, K.V. (2015) Biomarker: Predictive or Prognostic? *Journal of Clinical Oncology*, **33**, 3968-3972. <https://doi.org/10.1200/JCO.2015.63.3651>
- [4] Rubin, D.B. (1978) Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics*, **6**, 34-58. <https://doi.org/10.1214/aos/1176344064>
- [5] Rubin, D.B. (1990) Formal Models of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference*, **25**, 279-292. [https://doi.org/10.1016/0378-3758\(90\)90077-8](https://doi.org/10.1016/0378-3758(90)90077-8)
- [6] Cole, S.R. and Frangakis, C.E. (2009) The Consistency Statement in Causal Inference: A Definition or an Assumption? *Epidemiology*, **20**, 3-5. <https://doi.org/10.1097/EDE.0b013e31818ef366>
- [7] Greenland, S. and Robins, J.M. (1986) Identifiability, Exchangeability, and Epidemiological Confounding. *Epidemiology*, **15**, 413-419.
- [8] Suzuki, E., Mitsuhashi, T., Tsuda, T. and Yamamoto, E. (2013) A Counterfactual Approach to Bias and Effect Modification in Terms of Response Types. *BMC Medical Research Methodology*, **13**, 101. <https://doi.org/10.1186/1471-2288-13-101>
- [9] Chiba, Y. (2017) Sharp Nonparametric Bounds and Randomization Inference for Treatment Effects on an Ordinal Outcome. *Statistics in Medicine*, **36**, 3966-3975. <https://doi.org/10.1002/sim.7400>
- [10] Rigdon, J. and Hudgens, M. (2015) Randomization Inference for Treatment Effects on a Binary Outcome. *Statistics in Medicine*, **34**, 924-935. <https://doi.org/10.1002/sim.6384>
- [11] Ding, P. and Miratrix, L.W. (2017) Model-Free Causal Inference of Binary Experi-

mental Data.

- [12] Chiba, Y. (2017) Stratified Exact Tests for the Weak Causal Null Hypothesis in Randomized Trials with a Binary Outcome. *Biometrical Journal*, **59**, 986-997.  
<https://doi.org/10.1002/bimj.201600085>