Journal of Mathematics Research; Vol. 14, No. 1; February 2022
ISSN 1916-9795 E-ISSN 1916-9809
Published by Canadian Center of Science and Education

Relationship between Event Prevalence Rate and Gini Coefficient of Predictive Model

Fei Han[1,2], Ian Stockwell[1,3,4]

[1] The Hilltop Institute, University of Maryland, Baltimore County, Baltimore, USA

[2] Computer Science and Electrical Engineering Department, University of Maryland, Baltimore County, Baltimore, USA

[3] Department of Information Systems, University of Maryland, Baltimore County, Baltimore, USA

[4] Erickson School of Aging Studies, University of Maryland, Baltimore County, Baltimore, USA

Correspondence: Fei Han, The Hilltop Institute, University of Maryland, Baltimore County, Baltimore, USA

Received: November 13, 2021 Accepted: January 6, 2022 Online Published: January 14, 2022

doi:10.5539/jmr.v14n1p46 URL: https://doi.org/10.5539/jmr.v14n1p46

Abstract

Predictive models are currently used for early intervention to help identify patients with a high risk of adverse events. Assessing the accuracy of such models is a crucial part of the development process. To measure the predictive performance of a scoring model, quantitative indices such as the K-S statistic and C-statistic are used. This paper discusses the relationship between Gini coefficients and event prevalence rates. The main contribution of the paper is the theoretical proof of the relationship between the Gini coefficient and event prevalence rate.

Keywords: predictive model, Gini coefficient, prevalence rate, discrimination

1. Introduction

Risk prediction models are currently being used in the health care sector to help care providers identify high-risk patients in order to implement diversionary interventions (Morgan et al., 2019; Henderson et al., 2021). Prediction model evaluation is a crucial step in the development of such models and is normally focused on discrimination and calibration (Steyerberg et al., 2010). Model discrimination refers to the ability of the model to discriminate between patients with and without the event of interest. Commonly used measures for evaluating model discriminative ability for binary events are receiver operating characteristic curves, concordance statistic (C-statistic), and precision-recall plots. Geometrically, the C-statistic is equal to the area under the receiver operating characteristic curve. C-statistic can also be interpreted as the probability that a randomly selected patient who had the event will have a higher predicted probability of having the event than a randomly selected patient who did not have the event. In ideal discrimination, in which predicted probabilities of the patients with the event are all higher than predicted probabilities of patients without the event, the C-statistic is equal to 1. Calibration, on the other hand, refers to the agreement between observed events and predictions. K-S statistics and Net Benefit can be used to evaluate model calibration but may not be appropriate in some circumstances (Morgan et al., 2019). As more and more machine learning algorithms are used in predictive models, especially in ensemble models, calibration may not be necessary for models that are only used for ranking, or when the predicted score cannot be interpreted as probability (Morgan et al., 2019).

Concentration curves and associated Gini coefficients are widely used tools for analyzing economic inequality. See Cowell (2011) and Jackson (1992). Concentration curves are also appropriate for measuring predictive model performance (Morgan et al., 2021). Concentration curves display the relationship between the accumulative true positive rate and the accumulative population proportion when patients are ranked in descending order by predicted risk score. Concentration curves provide more insight than receiver operating characteristic curves, especially in the case of low-prevalence events when interventions are prioritized to only the highest risk individuals (Keya et al., 2020). For example, from Figure 1, below, one can easily see that the top 10% riskiest patients include around 50% of patients who have the outcome. Thus, if health care professionals provide outreach to the top 10% riskiest patients, then half of all patients who will experience the event will have been contacted. The Gini coefficient is defined as two times the area between the concentration curve and the diagonal line and gives a summarized measure of the model discrimination. A larger Gini coefficient means better model discrimination. As with C-statistic, Gini coefficients are rank order statistics; that is, if the risk score values change while the relative ranking of individuals within the population remains unchanged, then both the C-statistic and Gini score will remain unchanged.

Although concentration curves and Gini coefficients are valuable in model discrimination evaluation, they are affected by the prevalence rate of the event of interest. As far as the authors know, this paper is the first one that rigorously proves the relationship between Gini coefficients and event prevalence rates. The main contribution is the mathematical theoretical proof of the relationship between Gini coefficients and event prevalence rates through introducing a parametric equation. This formula provides an upper bound of Gini coefficient for evaluating predictive model performance.

2. Main Results

Assume that we have N patients, and every patient is associated with a tuple (S_i, E_i), where $i = 1,2, \dots, N$ and S_i is the predicted risk score for patient i, E_i is the event status for patient i, and

$$E_i = \begin{cases} 1, & if\ patient\ i\ has\ event \\ 0, & otherwise \end{cases}$$

We assume that $S_1 \geq S_2 \geq \dots \geq S_N$ since patients with high-risk scores are of interest in most situations. To facilitate the proof, we introduce some mathematical notations. We denote p_0 as the event prevalence rate, $p_0 = \frac{n}{N}$, where n is the number of patients who have the event of interest. The empirical distribution function of the scores of the event is the accumulative percentage of patients with the event and with scores at least α. It is denoted as

$$D_{n,event}(\alpha) = \frac{1}{n}\sum_{i=1}^{N} I(S_i \geq \alpha \wedge E_i = 1)$$

where I is an indicator function with the definition:

$$I(x) = \begin{cases} 1, & x\ is\ True \\ 0, & x\ is\ False \end{cases}$$

and α is a parameter such that $\alpha \in [L_1, L_2]$, $L_1 = \min(S_i, i = 1,2, \dots, N)$ and $L_2 = \max(S_i, i = 1,2, \dots, N)$. The empirical distribution function for the scores of all patients is denoted as

$$D_{N,all}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} I(S_i \geq \alpha).$$

We first prove a mathematical formula between Gini coefficient and the event prevalence rate for ideal discrimination. The ideal discrimination is defined as a set of tuples $(S_i, E_i), i = 1,2, \dots, N$, such that $S_1 \geq S_2 \geq \dots \geq S_N$ and $\min_i(S_i|E_i = 1) > \max_i(S_i|E_i = 0)$.

When $\alpha \geq \min(S_i|E_i = 1)$, we have

$$y = D_{n,event}(\alpha) = \frac{1}{n}\sum_{i=1}^{N} I(S_i \geq \alpha \wedge E_i = 1) = \frac{1}{n}\sum_{i=1}^{N} I(S_i \geq \alpha), y \in [0,1], \quad (1)$$

since $E_i = 1$ always holds for $\alpha \geq \min(S_i|E_i = 1)$ with the ideal discrimination.

We denote x as

$$x = D_{N,all}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} I(S_i \geq \alpha), x \in [0, p_0]. \quad (2)$$

The range of x in $[0, p_0]$ is derived from the assumption of $\alpha \geq \min(S_i|E_i = 1)$ and that p_0 is the event prevalence rate and with the ideal discrimination. From equation (2) we have

$$Nx = \sum_{i=1}^{N} I(S_i \geq \alpha), x \in [0, p_0]. \quad (3)$$

Plug equation (3) into (1), we have that

$$y = \frac{Nx}{n} = \frac{1}{n/N}x = \frac{1}{p_0}x, \text{ for } x \in [0, p_0]. \quad (4)$$

When $min(S_i) \leq \alpha < min(S_i|E_i = 1)$, we have that

$$y = D_{n,event}(\alpha) = \frac{1}{n}\sum_{i=1}^{N} I(S_i \geq \alpha \wedge E_i = 1) = \frac{1}{n}n = 1, \tag{5}$$

$$x = D_{N,all}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} I(S_i \geq \alpha), x \in (p_0, 1]. \tag{6}$$

In Equation (5), $\sum_{i=1}^{N} I(S_i \geq \alpha \wedge E_i = 1) = n$ because there are n events when the prevalence rate is p_0 and under condition $min(S_i) \leq \alpha < min(S_i|E_i = 1)$. In Equation (6), n< $(\sum_{i=1}^{N} I(S_i \geq \alpha)) \leq N$ since $min(S_i) \leq \alpha < min(S_i|E_i = 1)$ and $S_1 \geq S_2 \geq \ldots \geq S_N$. Based on equations (5) and (6) we have

$$y = 1 \text{ for } x \in (p_0, 1] \tag{7}$$

From equation (4) and (7), we have

$$y = \begin{cases} \frac{x}{p_0} & 0 \leq x \leq p_0 \\ 1 & p_0 < x \leq 1 \end{cases}. \tag{8}$$

From equation (8), it is clear that the concentration curve consists of two segments of lines, which is illustrated in Figure 2, below.

The Gini coefficient (G) is defined as two times the area between the concentration curve and the diagonal line. Hence, we see that $0 \leq G$. By geometry, from Figure 2, we have that

$$G = 2 \cdot Area_{\Delta OAB} = 2\left(\frac{1}{2} \cdot AC \cdot AB\right) = 2\left(\frac{1}{2} \cdot 1 \cdot (1 - p_0)\right) = 1 - p_0.$$

Hence, we have $G = 1 - p_0$ for $p_0 \in (0,1)$. Thus, we have the following result.

***Theorem* 1**. For the predicted risk scores with ideal discrimination, the relationship between Gini score and event prevalence rate is

$$G = 1 - p_0, p_0 \in (0,1),$$

where G is Gini coefficient and p_0 is the event prevalence rate.

Equation in Theorem 1 is for ideal discrimination. However, in reality, not all predicted probabilities of the patients with the event are higher than the predicted probabilities of patients without the event. Thus, we have the following corollary.

***Corollary* 1**. For any predicted risk scoring, the relationship between Gini coefficient and event prevalence rate is

$$0 \leq G \leq 1 - p_0, \ p_0 \in (0,1).$$

The inequality $G \geq 0$ is due to its geometric meaning and inequality $G \leq 1 - p_0$ is because there is at least one non-event subject having predicted score larger than the predicted probability of some subject with event. Corollary 1 gives an upper bound of Gini coefficient for any predicted risk scores.

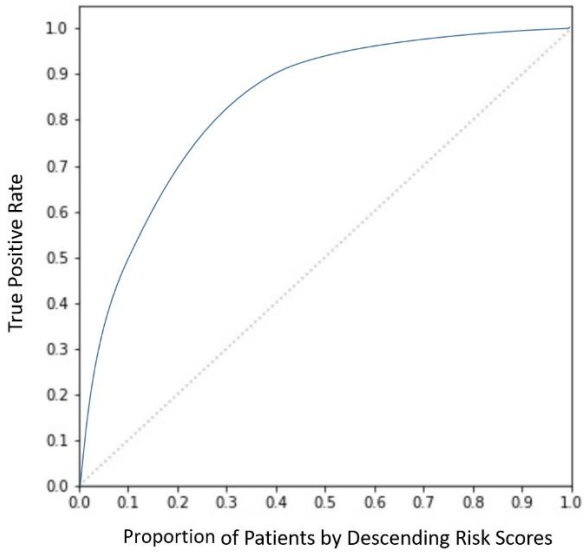

Figure 1. Concentration Curve

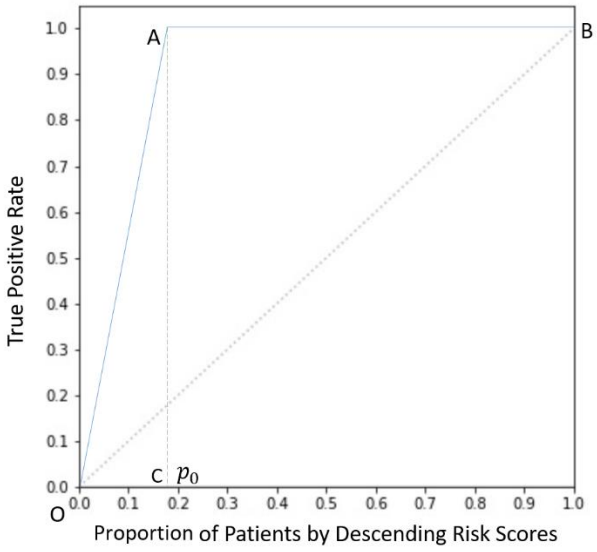

Figure 2. Concentration Curve of Ideal Predicted Scores

3. Discussion

In order to determine the true accuracy of a predictive model, one must know how well a "perfect" model would behave. While it is theoretically possible for any model to achieve a C-statistic of 1, the authors have shown that the same cannot be said for a model's Gini coefficient. As shown in Theorem 1 and Corollary 1, the Gini coefficient of a predictive model with ideal discrimination increases as prevalence rates decrease. For example, if the event prevalence rate is 40%, then the maximum Gini coefficient the predicted risk score could attain is 0.6. However, if the event prevalence rate is 1%, then the maximum Gini coefficient the predicted risk score could reach is 0.99.

These findings suggest two important strategies for determining and comparing the performance of predictive models. First, no model should be benchmarked against a Gini coefficient of 1 because that is only possible for events that do not occur. Second, the relative performance of models trained on events with different prevalence cannot be determined by comparing their Gini coefficients alone. For example, it may not be true that a model predicting an event with 10% prevalence with a 0.6 Gini coefficient performs better than a model predicting an event with 50% prevalence with a 0.4 Gini coefficient, since the upper limit on the Gini coefficient is different in each scenario. Both of these strategies have practical applications in applied predictive modeling and provide insights into the relationship between the Gini coefficient and event prevalence for model evaluation.

References

Cowell, F. (2011). *Measuring inequality*. Oxford University Press.

Henderson, M., Han, F., Perman, C., Haft, H., & Stockwell, I. (2021). Predicting avoidable hospital events in Maryland. *Health Services Research*. https://doi.org/10.1111/1475-6773.13891

Jackson, P. M. (1992). *Current issues in public sector economics*. Macmillan International Higher Education.

Keya, K. N., Islam, R., Pan, S., Stockwell, I., & Foulds, J. R. (2020). *Equitable allocation of healthcare resources with fair cox models*. arXiv preprint arXiv:2010.06820

Morgan, D. J., Bame, B., Zimand, P., Dooley, P., Thom, K. A., Harris, A. D., ... & Liang, Y. (2019). Assessment of machine learning vs standard prediction rules for predicting hospital readmissions. *JAMA network open*, *2*(3), e190348-e190348. https://doi.org/10.1001/jamanetworkopen.2019.0348

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, *21*(1), 128.

Copyrights