Hindawi
International Journal of Mathematics and Mathematical Sciences
Volume 2022, Article ID 7835758, 12 pages
https://doi.org/10.1155/2022/7835758

Research Article

The Effectiveness of the M_{GK} Measure against the Odds Ratio in the Epidemiological Study

Bruce Masonova Solozafy Bemena,[1] **André Totohasina**,[2] **and Daniel Rajaonasy Feno**[3]

[1]*School Mixed of Sambava, B. P. 90, Sambava, Madagascar*
[2]*Department of Mathematics and Informatics Application, University of Antsiranana Madagascar, B. P. 0, Antsiranana, Madagascar*
[3]*Department of Mathematics and Informatics Application, University of Toamasina, B. P. 591, Toamasina, Madagascar*

Correspondence should be addressed to Bruce Masonova Solozafy Bemena; brucebemena@gmail.com

Received 8 June 2022; Revised 7 July 2022; Accepted 8 July 2022; Published 21 September 2022

Academic Editor: Chin-Chia Wu

Copyright © 2022 Bruce Masonova Solozafy Bemena et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In epidemiology, the rule of association is used to determine the factors at the origin of diseases; implicative statistical analysis is thus a necessary tool in epidemiology too. Epidemiologists have more often chosen the so-called odds ratio measure in their studies of the quantification of the implicit link between an exposure and disease. In order to obtain good results, we need to be sure that the odds ratio measure is really the most relevant measure available. Therefore, it is necessary to study the mathematical properties of the odds ratio. This paper proposes a comparative study of the behaviour and mathematical properties of the odds ratio measure, the measure of Guillaume–Khenchaff (M_{GK}), and the normalised odd-ratio measure. We have chosen the M_{GK} measure because the literature considers it to be a good measure for extracting implicit association rules according to its mathematical properties. The result in this paper concerns only the study of probabilistic data.

1. Introduction

In epidemiology, in order to search for the etiological factors (i.e., causes) of diseases, the study is interested in the association between an exposure E (or risk factor) and a disease D. In order to quantify such an association, we are interested in the study of the measure of association. In parallel to this, we have noted that several thesis works and publications, such as those of Gail et al. [1]; Antony [2]; Rao and Miller [3]; Axame et al. [4]; Adwar et al. [5]; Agresti [6]; Bland and Altman [7]; and Sedgwick [8], have often used the Odd-Ratio measure in their studies. Several epidemiologists also affirm in their work Held [9]; Bouyer [10] that the Odd-Ratio measure is very effective in case-control studies.

Furthermore, the literature attests that the measure of Guillaume–Khenchaff (M_{GK}) measure is considered to be a good measure for extracting implicit association rules according to its mathematical properties studied in Buchanan and Shortliffe [11], Guillaume [12], Totohasina et al. [13], Wu et al. [14], Totohasina et al. [15], Rakotomalala et al. [16], Rakotomalala [17], and Ralahady and Totohasina [18]. According to the works of Guillaume [12]; Totohasina [19]; and Feno [20], the M_{GK} measure is inspired by the Loevinger measure with its capacity to identify zones of attraction and repulsion and guided by the desire to overcome the disadvantages of the confidence measure, in particular the inconvenience of selecting rules located in the repulsion zone between the premise X and the consequent Y of a rule of association $X \longrightarrow Y$. In his work, Guillaume [12] asserted that M_{GK} verified the principles of [21]. Independently of Guillaume [12], Wu et al. [14]have used M_{GK} for extracting positive and negative rules from a data mining context. They insisted especially on the fact that this measure allows the extraction of negative rules of the forms $X \longrightarrow \overline{Y}$, $\overline{X} \longrightarrow Y$, and $\overline{X} \longrightarrow \overline{Y}$. About this last bilateral negative rule, let us recall that the M_{GK}- based algorithm depends only on the favoring component M^f_{GK}, which is fortunately implicative; that is, $M^f_{GK}(\overline{Y} \longrightarrow \overline{X}) = M^f_{GK}(X \longrightarrow Y)$

[13, 19]. This mathematical property makes M_{GK} more relevant and compatible with using the language of implication as it is naturally desired by researchers on interpreting valid association rules.

However, a comparative study between the M_{GK} and odds ratio measures is already done in 2017 in our own Master work Masonova [22]: in fact, odds ratio has a flaw in the face of an intuitive situation, as it does not take a fixed value at logical implication. This causes the difficulty to define a minimum threshold and to interpret the association rules. Following the study carried out by [23, 24], we propose to deepen the comparative study of M_{GK} and the normalised measure of odds ratio according to the 22 properties proposed by [25].

In this article, we will take as reference the work of [25]. In her thesis work, she identifies 22 properties for evaluating the performance of quality measures, which we will review and evaluate on the three measures M_{GK}, odds ratio, and standardised odds ratio to reassure epidemiologists on the choice of measure to use.

In the following, our work is divided into four sections. Section 2 presents the different tools and concepts used in our work. Section 3 presents our results and discusses the comparison of the M_{GK}, odds ratio, and normalised odds ratio measures across the different studies. Section 4 concludes and discusses some perspectives.

2. Materials and Methods

2.1. Notation and Definition. We present in this section some notations and definitions around data mining and the measurement of association rules.

Let $\mathscr{A}$ be a set of m items or attributes also called variables and ε a set of n transactions or entities defined on the set of attributes. A subset X of $\mathscr{A}$ is called a pattern or an itemset.

According to Ganter and Wille [26], in Formal Concept Analysis, a formal context is a triplet $\mathbb{K} = (\varepsilon, \mathscr{A}, \mathscr{R})$, where ε and $\mathscr{A}$ are the finite sets and $\mathscr{R}$ is a binary relation from ε to $\mathscr{A}$.

Definition 1. Let us consider a binary context $\mathbb{K} = (\varepsilon, \mathscr{A}, \mathscr{R})$, where ε and $\mathscr{A}$ are the finite sets. Such a context will be called a data mining context.

Let $\mathbb{K} = (\varepsilon, \mathscr{A}, \mathscr{R})$ be a context of the data mining; X and Y are two patterns of $\mathbb{K}$.

(i) For a pattern X, its dual X' is defined by $X' = \{e \in \varepsilon | \forall x \in X, x\}$: X' is also called the extension of the pattern X.

(ii) We define the uniform probability P on the finite discrete probability space $(\varepsilon, P(\varepsilon))$ as follows: for any event E of $P(\varepsilon)$: and consider $P(E) = (|E|/|\varepsilon|)$, where $|E|$ designates the cardinality of the event E.

Definition 2. According to Feno [20], an association rule is an implication of the form $X \longrightarrow Y$ expressing the fact that the attributes in X tend to appear with those in Y.

Definition 3. According to Feno [20] and Totohasina [27], a quality measure or interest measure of the rules is a function μ of the set of association rules with values in $\mathbb{R}$ such that for any association rule $X \longrightarrow Y$ the value $\mu(X \longrightarrow Y)$ depends exclusively on the four parameters n, $P(X'), P(Y')$, and $P(X' \cap Y')$, where P designates the uniform discrete probability on the probability space $(\varepsilon, P(\varepsilon))$ and n denotes the cardinality of ε $(n = |\varepsilon|)$.

Definition 4. Let $X \longrightarrow Y$ be an association rule. According to Totohasina [19], a quality measure μ is said to be normalised if it verifies the following five conditions:

(i) $\mu(X \longrightarrow Y) = -1$, if $P(Y'|X') = 0$.

(ii) $-1 < \mu(X \longrightarrow Y) < 0$, if $0 \neq P(Y'|X') < P(Y')$; that is, X and Y are negatively dependent (in partial repulsion).

(iii) $\mu(X \longrightarrow Y) = 0$, if $P(Y'|X') = P(Y')$ (i.e., X and Y are independent).

(iv) $0 < \mu(X \longrightarrow Y) < 1$, if $1 \neq P(Y'|X') > P(Y')$ (i.e., if X favours Y or X and Y are partially attracted).

(v) $\mu(X \longrightarrow Y) = 1$, if $P(Y'|X') = 1$ (or if X totally implies Y).

Definition 5. Let X and Y be two patterns of a context of the data mining.

Guillaume [12], Feno [20], and Totohasina [27] define the measure M_{GK} as

$$M_{GK}(X \longrightarrow Y) = \begin{cases} \dfrac{P(Y'/X') - P(Y')}{1 - P(Y')}, & \text{if } X \text{ favours } Y, \\ \dfrac{P(Y'/X') - P(Y')}{P(Y')}, & \text{if } X \text{ disadvantages } Y. \end{cases} \tag{1}$$

Definition 6. Let X and Y be two patterns of a data mining context.

$$M_{GK}(X \longrightarrow Y) = \begin{cases} M^{f}_{GK}(X \longrightarrow Y), & \text{if } X \text{ favours } Y, \\ M^{d}_{GK}(X \longrightarrow Y), & \text{if } X \text{ disadvantages } Y. \end{cases} \tag{2}$$

Definition 7. Let X and Y be two patterns of a data mining context.

Tan et al. [28] defined the odds ratio (OR) measure as

$$\begin{aligned}\mathrm{OR}(X \longrightarrow Y) &= \frac{P(X' \cap Y')P\left(\overline{X'} \cap \overline{Y'}\right)}{P\left(\overline{X'} \cap Y'\right)P\left(X' \cap \overline{Y'}\right)} \\ &= \frac{P(Y'/X')\left(1 - P(X') - P(Y') + P(Y'/X')P(X')\right)}{\left(P(Y') - P(Y'/X')P(X')\right)\left(1 - P(Y'/X')\right)}.\end{aligned} \tag{3}$$

Definition 8. According to Feno [20] and Totohasina [27], let X and Y be two patterns of a data mining context.

X favours Y, if and only if Y favours X.

Proof.

$$\begin{aligned}X \text{ favours } Y &\Leftrightarrow P(Y'/X') > P(Y') \\ &\Leftrightarrow P(Y' \cap X') > P(X')P(Y') \\ &\Leftrightarrow P(X'/Y') > P(X') \\ &\Leftrightarrow Y \text{ favours } X.\end{aligned} \tag{4}$$

□

2.2. Normalisation of a Measure according to an Affine Homeomorphy. The birth of the normalisation of a quality measure was started by Totohasina in his work [19] with the aim of having a unifying vision of the measures in the literature of binary data mining.

2.2.1. Reference Situations

Definition 9. Let X and Y be patterns of a data mining context $\mathbb{K} = (\varepsilon, \mathscr{A}, \mathscr{R})$ and P the uniform discrete probability on the probability space $(\varepsilon, P(\varepsilon))$ [19]. For the sake of consistency with the principle of duality in formal concept analysis, we characterize it by the properties of their respective patterns X' and Y' as events of the tribe $P(\varepsilon)$:

(i) X and Y are incompatible, if their extensions are incompatible, that is, if $P(X' \cap Y') = 0$ (i.e., $P(Y'/X') = 0$), where $X' = \{a \in \varepsilon / X(a) = 1\}$ is the extension of X.

(ii) X and Y are negatively dependent (or X and Y are mutually unfavourable), if $P(X'/Y') < P(X')$ (which is equivalent to $P(Y'/X') < P(Y')$).

(iii) X and Y are positively dependent (or X and Y favour each other), if $P(X'/Y') > P(X')$ (which is equivalent to $P(Y'/X') > P(Y')$).

(iv) X logically (totally) implies Y, if $X' \subseteq Y'$, that is, $P(Y'/X') = 1$ [27].

Thus, the quantities $P(Y'/X') - P(Y')$ and $P(X'/Y') - P(X')$ measure the deviations from independence of the two patterns X and Y which are, respectively, noted $EI(X \longrightarrow Y)$ and $EI(Y \longrightarrow X)$.

In general, these two indicators of the degree of statistical dependence are not equal, despite the mutuality of attraction or repulsion depending on whether the link is positive or negative.

Nevertheless, the notions of positive and negative dependence are linked, as shown by the following lemmas [27].

Lemma 1. *Let X and Y be two patterns.*

(1) The following three conditions are equivalent: (i) X disfavours Y; (ii) X favours Y; and (iii) X favours Y.

(2) The following four conditions are equivalent: (i) X favours Y; (ii) X disfavours Y; and (iii) X favours Y and X disfavours Y [27].

Note that the two quantities $P(Y'/X')$ and $P(X'/Y')$ are increasing functions of the number of examples $|X' \cap Y'|$; the marginals $P(X')$ and $P(Y')$ are remaining constant. Moreover, the literature already suggests the following five principles [27].

The three Piatetsky–Shapiro principles say that a measure of interest of an association rule must be zero in case of statistical independence of the premises and consequences, a strictly increasing function of the number of examples, the other parameters being fixed, and a strictly decreasing function of the cardinal of the dual of its premise or decreasing of the cardinal of the dual of its consequent; the other parameters are kept constant [27].

The fourth principle of Major and Mangano: a measure of interest of an association rule must be a strictly increasing function of its coverage (i.e., the cardinal of the intersection of the two extensions), once its confidence is kept constant above a previously fixed minimum value [27].

The fifth principle of Totohasina [27] that corrects the symmetrical character of the Piatetsky–Shapiro index: A measure of quality of interest of an association rule must be nonsymmetrical.

In view of the mathematical objectives of normalisation and the five principles mentioned, [19] introduced the definition of a normalised quality measure as follows.

Definition 10. Let X and Y be patterns of a binary data mining context $\mathbb{K} = (\varepsilon, \mathscr{A}, \mathscr{R})$, P the uniform discrete probability on the probability space $(\varepsilon, P(\varepsilon))$ [19], ε the set of transactions from ε to $\mathscr{A}$, $\mathscr{A}$ the set of attributes called items or patterns, $\mathscr{R}$ is the binary relation from ε to $\mathscr{A}$, μ a probabilistic quality measure, and $X \longrightarrow Y$ an association rule. A quality measure of an association rule is said to be normalised, if it verifies the following five conditions:

(i) $\mu(X \longrightarrow Y) = -1$, if $P(Y'/X') = 0$, that is, X' and Y' are two incompatible events: we say that the two patterns X and Y are then incompatible.

(ii) $-1 \leq \mu(X \longrightarrow Y) < 0$, if $0 < P(Y'/X') < P(Y')$, that is, X' and Y' are two independent events: we say

that the two patterns X and Y are then negatively dependent (in partial repulsion).

(iii) $\mu(X \longrightarrow Y) = 0$, if $P(Y'/X') = P(Y')$, that is, the two events X' and Y' are independent: we say that the two patterns X and Y are then independent.

(iv) $0 < \mu(X \longrightarrow Y) \leq 1$, if $0 \neq P(Y'/X') > P(Y')$, that is, X favours Y or the two patterns X and Y attract each other partially.

(v) $\mu(X \longrightarrow Y) = 1$, if $P((Y'/X') = 1$, it is said that the pattern X is then totally included in Y.

The distribution of the values of a normalised measure is represented schematically, as shown in figure 1 [27].

According to the definition mentioned, it is very easy to show that the measure M_{GK} is defined by $M_{GK}(X \longrightarrow Y) = \begin{cases} M^f_{GK}(X \longrightarrow Y) \\ M^d_{GK}(X \longrightarrow Y) \end{cases}$ where $M^f_{GK}(X \longrightarrow Y) = P(Y'/X') - P(Y')/1 - P(Y')$, if X favours Y and $M^d_{GK}(X \longrightarrow Y) = P(Y'/X') - P(Y')/1 - P(Y')$, if X disfavours Y, is well normalised and continuous.

Indeed, in case of incompatibility between two X, Y patterns of $P(\mathcal{F})(X' \cap Y' = \varnothing)$, as $P(Y'/X') = 0$.

According to the definition, at incompatibility, $M_{GK}(X \longrightarrow Y) = P(Y'/X') - P(Y')/P(Y')$.

Therefore, $M_{GK}(X \longrightarrow Y) = 0 - P(Y')/P(Y') = -P(Y')/P(Y') = -1$.

$$\text{Hence } M_{GK}(X \longrightarrow Y) = -1. \tag{5}$$

In the case of independence between the premise and the consequent, $P(Y'/X') = P(Y')$.

According to the definition, at independence, $M_{GK}(X \longrightarrow Y) = P(Y'/X') - P(Y')/P(Y')$.

Therefore, $M_{GK}(X \longrightarrow Y) = P(Y') - P(Y')/P(Y') = 0/P(Y') = 0$.

$$\text{Hence } M_{GK}(X \longrightarrow Y) = 0. \tag{6}$$

Finally, in the case of logical implication $(X' \subset Y')$, $P(Y'/X') = 1$.

From the definition, to the logical implication, $M_{GK}(X \longrightarrow Y) = P(Y'/X') - P(Y')/1 - P(Y')$.

Therefore, $M_{GK}(X \longrightarrow Y) = 1 - P(Y')/1 - P(Y') = 1$.

Hence, $M_{GK}(X \longrightarrow Y) = 1$, which was necessary to show. (7)

2.2.2. Remind the Process of Normalisation by Affine Homeomorphism. Let us recall in passing that, according to Totohasina, the normalised measure μ_n of the measure μ has the expression

$$\mu_n(X \longrightarrow Y) = \begin{cases} x_f \mu(X \longrightarrow Y) + y_f, & \text{if } X \text{ favours } Y, \\ x_d \mu(X \longrightarrow Y) + y_d, & \text{if } X \text{ disfavours } Y. \end{cases} \tag{8}$$

These four coefficients are determined by crossing unilateral limits in reference situations (incompatibility, independence, and logical implication) due to the continuity of the evolution in the two zones: attraction (positive dependence) and repulsion (negative dependence).

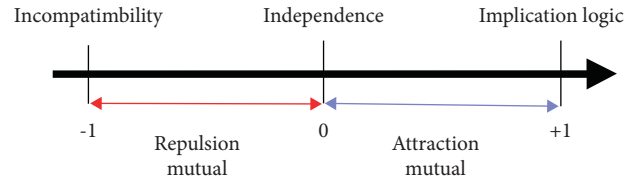

Figure 1: Distribution of continuous normalised MPQ values mutual repulsion and mutual attraction.

Let us put $\mu_{\text{imp}}(X \longrightarrow Y)$ the value of $\mu(X \longrightarrow Y)$ at the logical implication, $\mu_{\text{ind}}(X \longrightarrow Y)$ that of $\mu(X \longrightarrow Y)$ to independence and $\mu_{\text{inc}}(X \longrightarrow Y)$ the value of $\mu(X \longrightarrow Y)$ to incompatibility [27]. In the case where X favours Y, we obtain

$$\begin{cases} x_f \mu_{\text{imp}}(X \longrightarrow Y) + y_f = 1, & \text{logical implication,} \\ x_f \mu_{\text{ind}}(X \longrightarrow Y) + y_f = 0, & \text{independence to right.} \end{cases} \tag{9}$$

In the case where X disfavours Y, we obtain

$$\begin{cases} x_d \mu_{in\,d}(X \longrightarrow Y) + y_d = 0, & \text{independence to left,} \\ x_d \mu_{inc}(X \longrightarrow Y) + y_d = -1, & \text{incompatibility.} \end{cases} \tag{10}$$

This gives the following system of linear equations:

$$\begin{cases} x_f \mu_{\text{imp}}(X \longrightarrow Y) + y_f = 1, \\ x_f \mu_{\text{ind}}(X \longrightarrow Y) + y_f = 0, \\ x_d \mu_{\text{ind}}(X \longrightarrow Y) + y_d = 0, \\ x_d \mu_{\text{inc}}(X \longrightarrow Y) + y_d = -1. \end{cases} \tag{11}$$

2.2.3. Example of Normalisation of Quality Measures. To illustrate the processes of normalisation of quality measures, here are some details of the calculation of the normalised associated with some quality measures. Let $X \longrightarrow Y$ be an association rule of a data mining context [27].

Support: $\text{Supp}(X \longrightarrow Y) = P(X' \cap Y')$ which is such that $\text{Supp}_{\text{imp}}(X \longrightarrow Y) = P(X')$, $\text{Supp}_{\text{ind}}(X \longrightarrow Y) = P(Y')P(X')$, and $\text{Supp}_{\text{inc}}(X \longrightarrow Y) = 0$. So, we get $d'et\, M = P^2 (X')P(Y')P(\overline{Y}') \neq 0$, Therefore, $x_f = 1/P(X')(1 - P(Y'))$, $y_f = P(X')P(Y')/P(X')(1 - P(Y'))$ $x_d = 1/P(X')\ P(Y')$ and $y_d = -1$.

Hence, $Supp_n$

$$= \begin{cases} \dfrac{P(X' \cap Y') - P(X')P(Y')}{P(X')(1 - P(Y'))}, & \text{if } X \text{ favours } Y, \\ \dfrac{P(X' \cap Y') - P(X')P(Y')}{P(X')P(Y')}, & \text{if } X \text{ disfavours } Y. \end{cases} \tag{12}$$

Finally, we find $\text{Supp}_n = M_{GK}$.

Confidence: $\text{Conf}(X \longrightarrow Y) = P(X'/Y')$ which is such that $\text{Conf}_{\text{imp}}(X \longrightarrow Y) = 1$, $\text{Conf}_{\text{ind}}(X \longrightarrow Y) = P(Y')$ and $\text{Conf}_{\text{inc}}(X \longrightarrow Y) = 0$. Therefore, $x_f = 1/(1 - P(Y'))$, $y_f = -P(Y')/(1 - P(Y'))$ $x_d = (1/P(Y'))$ and $y_d = -1$.

$$\text{Hence, Conf}_n = \begin{cases} \dfrac{P(X' \cap Y') - P(X')P(Y')}{P(X')(1 - P(Y'))}, & \text{if } X \text{ favours } Y, \\ \dfrac{P(X' \cap Y') - P(X')P(Y')}{P(X')P(Y')}, & \text{if } X \text{ disfavours } Y. \end{cases} \tag{13}$$

Finally, we find $\text{Conf}_n = M_{GK}$.

Faced with the normalisation procedure, the author himself stated that there are certain measures that resist normalisation with an affine homeomorphy. For this reason, he announced a following theorem called the condition theorem of normalisability of a measure according to his theories.

A measure of quality μ is normalisable if and only if, for any rule $X \longrightarrow Y$, the following conditions are verified:

(i) The quantities $\mu_{\text{imp}}(X \longrightarrow Y)$, $\mu_{\text{ind}}(X \longrightarrow Y)$, and $\mu_{\text{inc}}(X \longrightarrow Y)$ are finite

(ii) The following inequalities are verified $\mu_{\text{imp}}(X \longrightarrow Y) \neq \mu_{\text{ind}}(X \longrightarrow Y)$, $\mu_{\text{imp}}(X \longrightarrow Y) \neq \mu_{\text{ind}}(X \longrightarrow Y)$ [27]

2.2.4. Study of the Normalisability of the OR Measure. The odds ratio measure is defined by

$$\text{OR}(X \longrightarrow Y) = \frac{P(X' \cap Y').P\left(\overline{X'} \cap \overline{Y'}\right)}{P\left(\overline{X'} \cap Y'\right).P\left(X' \cap \overline{Y'}\right)}. \tag{14}$$

Now, according to Definition 9, it is always necessary to express the measure of the quality of the association rules using the quantities: $P(Y'/X'), P(X'), P(Y')$.

So, to express the measure $\text{OR}(X \longrightarrow Y)$ as a function of $P(Y'/X'), P(X'), P(Y')$, we must use the Bayes theorem: $P(X' \cap Y') = P(X'/Y').P(Y') = P(Y'/X').P(X')$.

After the transformation, we obtain

$$\text{OR}(X \longrightarrow Y) = \frac{P(Y'/X')(1 - P(X') - P(Y') + P(Y'/X').P(X'))}{(P(Y') - P(Y'/X').P(X'))(1 - P(Y'/X'))}. \tag{15}$$

In the case of incompatibility between two X, Y patterns of $P(\mathcal{F})(X' \cap Y' = \varnothing)$, as $P(Y'/X') = 0$ translates the incompatibility:

It is obvious from (14) that

$$\text{OR}_{\text{inc}}(X \longrightarrow Y) = \frac{0.P\left(\overline{X'} \cap \overline{Y'}\right)}{P\left(\overline{X'} \cap Y'\right).P\left(X' \cap \overline{Y'}\right)} = 0. \tag{16}$$

$$\text{Hence, OR}_{\text{inc}}(X \longrightarrow Y) = 0. \tag{17}$$

In the case of independence between the premise and the consequent, we have $P(X' \cap Y') = P(X').P(Y')$.

As $P(Y'/X') = P(Y')$, we obtain at independence $\text{OR}_{\text{ind}}(X \longrightarrow Y) = P(Y')(1 - P(X') - P(Y') + P(Y').P(X'))/(P(Y') - P(Y').P(X'))(1 - P(Y'))$.

Therefore, $\text{OR}_{\text{ind}}(X \longrightarrow Y) = P(Y') - P(Y').P(X') - P(Y')^2 + P(Y')^2.P(X')/P(Y')) - P(Y')^2 - P(Y').P(X') + P(Y')^2.P(X') = 1$.

$$\text{Hence, OR}_{\text{ind}}(X \longrightarrow Y) = 1. \tag{18}$$

Finally, in the case of logical implication $(X' \subset Y')$, so $(X' \cap Y') = X'$.

As $\text{P}(Y'/X') = 1$, we obtain the logical implication: $\text{OR}_{\text{imp}}(X \longrightarrow Y) = 1 - P(X') - P(Y') + P(X')/(P(Y') - P(X'))(1 - 1) = 1 - P(Y')/0$.

Now, the probability between the interval of [0,1] is positive.

So, $\text{OR}_{\text{imp}}(X \longrightarrow Y) = 1 - P(Y')/0^+ = +\infty$.

$$\text{Hence, OR}_{\text{imp}}(X \longrightarrow Y) = +\infty. \tag{19}$$

The relations of equations (16)–(19) show that the measure OR can take very large values in the reference situations. These properties prove that the measure OR is nonnormalised and nonnormalisable by an affine homeomorphy. It is therefore clear that this measure is not normalisable by an affine homeomorphism.

2.3. Normalisation of a Measure according to a Homography Homeomorphy. Thanks to the research collaboration by [23, 24], this measure remains currently normalisable. This time, they used the proper homography according to the following approach.

If one or two of the three values x_{imp}, x_{ind}, and x_{inc} are infinite and in the case where we have two infinite values, it is necessary that x_{ind} is excluded, which leads us to use the following expression to find the four real coefficients, x_f, y_f, x_d, and y_d:

$$\mu_{hn}(X \longrightarrow Y) = \begin{cases} \dfrac{x_f}{x+m} + y_f, & \text{if } X \text{ favours } Y, \\ x_d x + y_d, & \text{if } X \text{ disfavours } Y. \end{cases} \tag{20}$$

These four coefficients are always determined by crossing unilateral limits in reference situations (incompatibility, independence, and logical implication) due to the continuity of the evolution in the two zones: attraction (positive dependence) and repulsion (negative dependence). In the case where X favours Y, x_{imp} can be infinite, $x_{\text{ind}} \neq x_{\text{inc}}$, $(x_{\text{ind}}, x_{\text{inc}}) \in \mathbb{R}^2$ and $x_{\text{ind}} \in \mathbb{R}^*$, then we obtain the system of equations as

$$\begin{cases} \dfrac{x_f}{x_{\text{imp}}} + y_f = 1, & \text{(implication logic)}, \\ \dfrac{x_f}{x_{\text{ind}}} + y_f = 0, & \text{(independence right)}. \end{cases} \tag{21}$$

As $(x_{\text{ind}}, x_{\text{inc}}) \in \mathbb{R}^2$ and $x_{\text{ind}} \in \mathbb{R}^*$, it is therefore sufficient to use the theory in [19] for the left-hand normalisation. We can write the following system of four nonlinear equations with four unknowns:

$$\begin{cases} \dfrac{x_f}{x_{imp}} + y_f = 1, \\ \dfrac{x_f}{x_{ind}} + y_f = 0, \\ x_d x_{ind} + y_d = 0, \\ x_d x_{inc} + y_d = -1. \end{cases} \tag{22}$$

Here, we only need to take $m = 0$ and we have four equations with four unknowns, with the particularity that the coefficient x_{imp} can be infinite. Hence, we have the following proposition.

Proposition 1. *(i) If $(x_{imp}, x_{ind}, x_{inc}) \in \mathbb{R}^3$, with x_{imp}, x_{ind}, and x_{inc} are two distincts, then the system of (16) admits four real solutions*

(ii) If $x_{imp} = \infty$ and $x_{in\,d} \in \mathbb{R}^*$, then the system of (16) has four real solutions such that $y_f = 1$, $x_f = -x_{ind}$, $x_d = 1/x_{ind} - x_{inc}$ and $y_d = -x_d x_{ind}$

(iii) If $x_{inc} = \infty$, $x_{ind} \in \mathbb{R}$ and if $x_{imp} = \infty$, then the system of (16) has four real solutions such that $y_d = -1$, $x_d = (x_{ind} + m)$, $y_f = 1$ and $x_f = -(x_{ind} + m)$

(iv) Otherwise, this system of equations has no solution

It is sufficient to take $\lim_{x_{imp} \longrightarrow \infty} x_f/x_{imp} = 0$.

Let us take advantage of this proposition with odds ratio: $x = (P(X' \cap Y')P(\overline{X'} \cap \overline{Y'})/P(\overline{X'} \cap Y')P(X' \cap \overline{Y'})) = P(Y'/X')(1 - P(X') - P(Y') + P(Y'/X')P(X'))/(P(Y') - P(Y'/X')P(X'))(1 - P(Y'/X'))$ such that $x_{imp} = +\infty$, $x_{ind} = 1$, and $x_{inc} = 0$.

We have $x_f = -1$, $x_d = -1$, $y_f = 1$ and $y_d = 1$.

By replacing x, x_f, y_f, x_d, and y_d with their values in the expression: $\mu_n = \begin{cases} (x_f/x + m) + y_f, & \text{if } X \text{ favours } Y \\ x_d x + y_d, & \text{if } X \text{ disfavors } Y \end{cases}$. For $m = 0$, we have

$$\text{OR}_{hn}(X \longrightarrow Y)$$

$$= \begin{cases} \dfrac{-1}{P(Y'/X')(1 - P(X') - P(Y') + P(Y'/X')P(X'))/(P(Y') - P(Y'/X')P(X'))(1 - P(Y'/X'))} + 1, & \text{if } X \text{ favours } Y; \\ \dfrac{P(Y'/X')(1 - P(X') - P(Y') + P(Y'/X')P(X'))}{(P(Y') - P(Y'/X')P(X'))(1 - P(Y'/X'))} - 1, & \text{if } X \text{ disfavours } Y, \end{cases}$$

Hence $\text{OR}_{hn}(X \longrightarrow Y)$

$$= \begin{cases} \dfrac{P(Y'/X') - P(Y')}{P(Y'/X')(1 - P(X') - P(Y') + P(Y'/X')P(X'))}, & \text{if } X \text{ favours } Y; \\ \dfrac{P(Y'/X') - P(Y')}{(1 - P(Y'/X'))(P(Y') - P(Y'/X')P(X'))}, & \text{if } X \text{ disfavours } Y. \end{cases} \tag{23}$$

It comes

$$\mathrm{OR}_{hn}(X \longrightarrow Y) = \begin{cases} \mathrm{OR}^{f}_{hn}(X \longrightarrow Y) \text{ if } X \text{ favours } Y; \\ \mathrm{OR}^{d}_{hn}(X \longrightarrow Y) \text{ if } X \text{ disfavours } Y. \end{cases} \tag{24}$$

In the case of incompatibility between two X, Y patterns of $P(\mathcal{F})(X' \cap Y' = \varnothing)$, as $P(Y'/X') = 0$, we obtain $OR_{hninc}(X \longrightarrow Y) = P(Y'/X') - P(Y')/(1 - P(Y'/X'))(P(Y') - P(Y'/X')P(X')) = 0 - P(Y')/(1-0)(P(Y') - 0)$.

Therefore, $OR_{hninc}(X \longrightarrow Y) = -P(Y')/P(Y') = -1$.

$$\text{Hence, } \mathrm{OR}_{hninc}(X \longrightarrow Y) = -1. \tag{25}$$

In the case of independence between the premise and the consequent, we have $P(X' \cap Y') = P(X').P(Y')$.

As $P(Y'/X') = P(Y')$, *we obtain at independence:* $OR_{hnin\ d}(X \longrightarrow Y) = P(Y'/X') - P(Y')/(1 - P(Y'/X'))(P(Y') - P\ (Y'/X')P(X'))$.

Therefore, $OR_{hnin\ d}(X \longrightarrow Y) = 0/(1 - P(Y'))(P(Y') - P(Y')P\ (X')) = 0$.

$$\text{Hence, } \mathrm{OR}_{hnin\ d}(X \longrightarrow Y) = 0. \tag{26}$$

In the case of logical implication $(X' \subset Y')$, $(X' \cap Y') = X'$.

As $P(Y'/X') = 1$, *we obtain the logical implication:* $OR_{hnimp}(X \longrightarrow Y) = P(Y'/X') - P(Y')/P(Y'/X')(1 - P\ (X') - P(Y') + P(Y'/X')P(X'))$.

Therefore, $OR_{hnimp}(X \longrightarrow Y) = 1 - P(Y')/1 - P(X') - P(Y') + P(X') = 1 - P(Y')/1 - P\ (Y') = 1$.

$$\text{Hence, } \mathrm{OR}_{hnimp}(X \longrightarrow Y) = 1. \tag{27}$$

The relations of equations (25)–(27) show that the measure OR_{hn} *takes particular values in reference situations other than the equilibrium situation. These properties prove that the homography-normalised odds ratio* (OR_{hn}) *is well normalised. Figure 2 summarises the particular values of* OR_{hn} *in the reference situations.*

However, $OR_{hn} \neq M_{GK}$.

2.4. Case-Control Study. According to Held [9], a case-control study examines the degree of association between exposure to a potentially harmful agent and the prevalence of a disease. To do this, a number of people (cases) with a disease (e.g., lung cancer) are first identified. Another group of people (the controls) with the same profile (in terms of age, gender distribution, blood pressure, other medications, and concomitant diseases) as the case group is then selected, except that the controls do not have the disease under study. Finally, the number of people who have been exposed to the toxic agent under investigation (e.g., smoking) is determined for each of the two groups.

2.5. Cross-Tabulation (2×2). According to Held [9], a cross-tabulation table or a contingency table in probability, is a table allowing to compare the distributions (absolute or relative frequencies) of population according to characteristics like the sex (male/female) or the smoking status (smoker/nonsmoker), but also according to variables which can take more than two values. A 2×2 table is used when the responses are dichotomous (can take two values), for example, yes/no or present/absent.

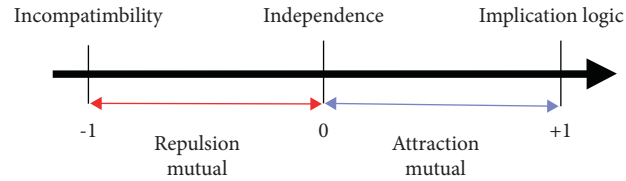

FIGURE 2: Reference situations of the measure OR_{hn} mutual repulsion, mutual attraction.

3. Results and Discussion

3.1. Summary of the Study of the Behaviour of the Three Measures at Reference Situations. In this section, we present the main properties of the measures M_{GK}, odds ratio, and normalised odds ratio homography. Thus, we consider a data mining context $\mathbb{K} = (\epsilon, \mathscr{A}, \mathscr{R})$ and X and Y as two patterns of $\mathbb{K}$.

(1) The results of the following propositions result from the definition of $\mathrm{M}_{GK}, \mathrm{OR}, \mathrm{OR}_{hn}$.

Proposition 2. *(reference situations). According to [20], for any patterns X and Y, we have the following:*

(i) X *and* Y *are incompatible, if and only if* $M_{GK}(X \longrightarrow Y) = -1$

(ii) X *disadvantages* Y, *if and only if* $-1 < M_{GK}(X \longrightarrow Y) < 0$

(iii) X *and* Y *are independent, if and only if* $M_{GK}(X \longrightarrow Y) = 0$

(iv) X *favours* Y, *if and only if* $0 < M_{GK}(X \longrightarrow Y) < 1$

(v) X *logically implies* Y, *if and only if* $M_{GK}(X \longrightarrow Y) = 1$

The above properties express the fact that M_{GK} *takes its values on the interval* $[-1, 1]$ *while reflecting the reference situations.*

Proposition 3. *(reference situations). According to [20], for any patterns X and Y, we have the following:*

(i) X *and* Y *are incompatible, if and only if* $OR(X \longrightarrow Y) = 0$

(ii) X *disfavours* Y, *if and only if* $0 < OR(X \longrightarrow Y) < 1$

(iii) X *and* Y *are independent, if and only if* $OR(X \longrightarrow Y) = 1$

(iv) X *favours* Y, *if and only if* $1 < OR(X \longrightarrow Y) < +\infty$

(v) X *logically implies* Y, *if and only if* $OR(X \longrightarrow Y) = +\infty$

The above properties express the fact that odds ratio (OR) takes its values on the interval $[0; +\infty]$ *while reflecting the reference situations.*

Proposition 4. *(reference situations). According to [29], for any patterns X and Y, we have the following:*

(i) X *and* Y *are incompatible, if and only if* $OR_{hn}(X \longrightarrow Y) = -1$

(ii) X *is unfavourable to* Y*, if and only if* $-1 < OR_{hn}(X \longrightarrow Y) < 0$

(iii) X *and* Y *are independent, if and only if* $OR_{hn}(X \longrightarrow Y) = 0$

(iv) X *favours* Y*, if and only if* $0 < OR_{hn}(X \longrightarrow Y) < 1$

(v) X *logically implies* Y*, if and only if* $OR_{hn}(X \longrightarrow Y) = 1$

The above properties express the fact that the normalised odds ratio homography (OR_{hn}) *takes its values on the interval* $[-1, 1]$ *while reflecting the reference situations.*

(2) So far, we have the following.

Starting from $OR(X \longrightarrow Y)$*, we obtain* $OR_{hn}(X \longrightarrow Y) \in [-1, 1]$.

Let us determine the critical values of OR_{hn}.

For that, it is enough to express OR_{hn} *as a function of* M_{GK}.

We get $OR^{f}_{hn}(X \longrightarrow Y) = P(Y'/X') - P(Y')/P(Y'/X')(1 - P(X') - P(Y') + P(Y'/X')P(X'))$.

So, $OR^{f}_{hn}(X \longrightarrow Y) = (1 - P(Y'))M_{GK}(X \longrightarrow Y)/P(Y'/X')(1 - P(X') - P\ (Y') + P(Y'/X')P(X'))$.

In any case, this relation (16), we can write $OR^{f}_{hn}(X \longrightarrow Y)$ *as a function of* $M_{GK}(X \longrightarrow Y)$.

Then, it is possible to obtain a page of critical values from those of $M_{GK}(X \longrightarrow Y)$.

3.2. Comparative Study of the Three Measures with the Binary Data. The two quality measures have varied behaviours with respect to the criteria desired for a good quality measure of the rules. To see the variation between the two formulas and the differences in the sets of values taken by the two measures, consider the data mining context presented in Table 1 formed by five people (A, B, C, D, and E) and six diseases (D_1, D_2, D_3, D_4, D_5, and D_6).

(1) For the M_{GK} *and OR Measures.* The table presents the values taken for the two quality measures for some of the association rules considered.

In Table2 that presents the behaviours of these two quality measures in the reference situations, we noticed that the measures M_{GK} and OR have of the different behaviours. It was noticed that the odds ratio measure does not take fixed values at logical implication, which causes the difficulty of defining a minimum threshold; that is, it is not known from which value taken of this measure can we get a convincing (interesting) value. Thus, it is very difficult to interpret a rule in this interval of $[0, +\infty[$. On the other hand, positive and negative association rules are potentially relevant using the M_{GK} measure. Taking the results in Table 3 as an example, the values of the rules taken by the M_{GK} measure are very accurate and easy to interpret. It lies between the interval [-1, 1].

(2) For the Measures M_{GK} *and* OR_{hn}. The table presents the values taken for the two quality measures for some of the association rules considered.

Table 1: Binary context.

	A	B	C	D	E
M_1	1	1	1	1	0
M_2	0	1	1	0	0
M_3	1	0	1	1	1
M_4	1	1	1	0	1
M_5	0	0	0	1	1
M_6	1	0	0	1	1

Table 2: The behaviour of the quality measures M_{GK} and OR in the reference situations.

Rule	M_{GK} [−1, 1]	OR [0, + ∞]
Incompatibility	−1	0
Repulsion	Negative	Positive
Independence	0	1
Attraction	Positive	Positive
Implication	1	+∞

Table 3: The values taken for the quality measures M_{GK} and OR for some of the association rules considered.

Rule	M_{GK} [−1, 1]	OR [0, + ∞]
$BC \longrightarrow DE$	−1	0
$DE \longrightarrow A$	−(1/4)	(1/3)
$BC \longrightarrow ACD$	0	1
$AC\ D \longrightarrow ABC$	(1/4)	3
$AC\ D \longrightarrow A$	1	+∞

In Table 4 that presents the behaviours of these two quality measures in the reference situations, we noticed that the measures M_{GK} and OR_{hn} have the same behaviours. In Table 5 that presents the values taken by these two quality measures, we noticed on the calculation of the rule $AC\ D \longrightarrow ABC$ that the value taken by the measure OR_{hn} is greater than the value of the measure M_{GK}. If, for example, we fix a minimum threshold of 60%, this rule is rejected by the M_{GK} measure but validated by the OR_{hn} measure. This proves that the M_{GK} measure is more discriminating than the OR_{hn} measure OR_{hn}.

The M_{GK} measure is normalised and normalisable, but the odds ratio measure is nonnormalised, nonnormalisable according to an affine homeomorphy and normalisable according to a homographic homeomorphy. From the results we obtained, we can conclude that the M_{GK} measure is a very efficient and more relevant measure compared to the odds ratio (OR) measure and the odds ratio homography-normalised (OR_{hn}). We advised users to question or abandon their choice of the odds ratio measure and choose the M_{GK} measure for analysis in their epidemiological studies.

3.3. Study on the Relationship between Smoking and Bronchial Carcinoma. For the application of our comparative study of the odds ratio, M_{GK} and the standardised odd-ratio measures in the field of epidemiology, we have chosen the study on the

Table 4: The behaviour of the quality measures M_{GK} and OR_{hn} in the reference situations.

Rule	M_{GK} [−1; 1]	OR_{hn} [−1; 1]
Incompatibility	−1	−1
Repulsion	Negative	Negative
Independence	0	0
Attraction	Positive	Positive
Implication	1	1

Table 5: The values taken for the quality measures M_{GK} and OR_{hn} for some of the association rules considered.

Rule	M_{GK} [−1; 1]	OR_{hm} [−1; 1]
$BC \longrightarrow DE$	−1	−1
$DE \longrightarrow A$	$-(1/4)$	$-(4/5)$
$BC \longrightarrow AC\ D$	0	0
$AC\ D \longrightarrow ABC$	$(1/4)$	$(2/3)$
$AC\ D \longrightarrow A$	1	1

relationship between smoking and bronchial carcinoma. This study is well shown and detailed in the works of Held [9] and Held et al [30]; they have well demonstrated the link between smoking and bronchial carcinoma from the data in Table 6.

Here, we pose the following:

(i) C: All people affected by bronchial carcinoma

(ii) $\overline{C}$: All people who are not affected by bronchial carcinoma

Table 6: The cross-distribution of smoking and bronchial carcinoma (source: Held [9] and Held et al [30]).

	Carcinoma Bronchial (C)	No carcinoma Bronchial ($\overline{C}$)	Total
Smoking (S)	1350	1296	2646
Nonsmoker ($\overline{S}$)	7	61	68
Total	1357	1357	2714

(iii) S: All smokers

(iv) $\overline{S}$: All nonsmokers

To measure the relationship between smoking and bronchial carcinoma, since C and S are the two reasons, we obtain the following probabilities:

(i) $P(C') = 1357/2714 = 0,5000$ and $P(S') = 2646/2714 = 0,9749$.

(ii) $P_{S'}(C) = 1350/2646 = 0,5102$ and $P_{C'}(S') = 1350/1357 = 0,9948$.

It comes:

(1) According to Definition 8 for the favorising case: as $\mathrm{P}_{C'}(S') > \mathrm{P}(C')$, so C favorise S.

(2) So $C \longrightarrow S$

The following values are obtained from the three measurements:

For $M^f_{GK}(C \longrightarrow S) = P_{C'}(S') - P(S')/1 - P \quad (S') = 0.9948 - 0.9749/1 - 0.9749 = 0.7928$.

For

$$\begin{aligned}\mathrm{OR}(C \longrightarrow S) &= \frac{P_{C'}(S')(1 - P(C') - P(S') + P_{C'}(S')P(C'))}{(P(S') - P_{C'}(S')P(C'))(1 - P_{C'}(S'))} \\ &= \frac{0,9948\,(1 - 0,5000 - 0,9749 + 0,9948 \times 0,5000)}{(0,9749 - 0,9948 \times 0,5000)\,(1 - 0,9948)} \\ &= 8,9600.\end{aligned} \tag{28}$$

For

$$\begin{aligned}\mathrm{OR}^f_{hm}(C \longrightarrow S) &= \frac{P_{C'}(S') - P(S')}{P_{C'}(S')(1 - P(C') - P(S') + P_{C'}(S')P(C'))} \\ &= \frac{0.9948 - 0.9749}{0.9948\,(1 - 0.5000 - 0.9749 + 0.9948 \times 0.5000)} \\ &= 0.8884.\end{aligned} \tag{29}$$

3.3.1. Recapitulation. The result is

$$\begin{aligned}M^f_{GK}(C \longrightarrow S) &= 0.7928, \\ \mathrm{OR}(C \longrightarrow S) &= 8,9600, \\ \mathrm{OR}^f_{hn}(C \longrightarrow S) &= 0.8884.\end{aligned} \tag{30}$$

3.3.2. Interpretation. In relation to the results of these three measures, we allow to interpret as follows: <<*If a person has bronchial carcinoma, then it is likely that he is a smoker.*>>.

3.3.3. Comparative Study of the Three Measures according to the Results Obtained. (1) For the M_{GK} and OR Measures. As

Table 7: Matrix describing the three measures according to the 19 properties.

M	P																			Total
	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
M_{GK}	1	1	1	1	1	0	1	1	0	1	1	1	0	0	0	0	0	1	1	12
OR	0	1	1	1	1	1	1	0	0	1	1	1	0	0	0	0	0	1	1	11
OR_{hn}	0	1	1	0	1	0	1	1	0	1	1	0	0	0	0	0	0	1	1	09

M, measures; P, properties.

the odds ratio measure does not take fixed values at the logical implication, we do not know from which value taken of this measure we can obtain a convincing or interesting value. On the other hand, the M_{GK} measure here has taken the value closer to 1. This easily allows us to say that there is a strong link on the $C \longrightarrow S$.

(2) For the M_{GK} and OR_{hn} Measures. In data mining, despite the volume of data to be explored which leads to obtaining several possible rules, this obliges us to be interested in discriminating measures to filter the rules well and to obtain the most interesting rules.

If, for example, we fix a minimum threshold of 80%, the rule $C \longrightarrow S$ is rejected by the measure M_{GK} but validated by the measure OR_{hn} . This proves that the M_{GK} measure is more discriminating than the OR_{hn} measure.

3.4. Study of Behaviour with Grissa's 22 Properties. We continue our study of the behaviours of the M_{GK}, OR, and OR_{hn} measures using the 22 properties in the work of [25]. The 22 properties are as follows: P_1: intelligibility or comprehensibility of the measure, P_2: ease of setting a threshold for acceptance of the rule, P_3: nonsymmetrical measure, P_4: nonsymmetrical measure in the sense of the negation of the conclusion, P_5: measure evaluating in the same way $(X \longrightarrow Y)$ and $\overline{Y} \longrightarrow \overline{X}$ in the case of logical implication, P_6: increasing measure according to the number of examples, P_8: increasing measure as a function of the size of the learning set, P_8: decreasing measure according to the size of the consequent or the size of the premise, P_9: measure admitting a fixed value in the case of independence, P_{10}: measure admitting a fixed value in the case of logical implication, P_{11}: measure admitting a fixed value in the case of equilibrium, P_{12}: measure admitting identifiable values in the case of attraction between X and Y, P_{13}: measure admitting identifiable values in case of repulsion between X and Y, P_{14}: measure capable of tolerating the first counterexamples, P_{15}: measure invariant in case of dilation of certain numbers, P_{16}: measure capable of differentiating between the rules $X \longrightarrow Y$ and $\overline{X} \longrightarrow Y$ according to an opposition relationship, P_{17}: measure capable of differentiating between the rules $X \longrightarrow Y$ and $X \longrightarrow \overline{Y}$ according to a relation of opposition, P_{18}: measure evaluating in the same way the rules $X \longrightarrow Y$ and $X\overline{X} \longrightarrow \overline{Y}$, P_{19}: measure having a size the random premise, P_{20}: statistical measure, P_{21}: discriminant measure, and P_{22}: robust measure.

3.4.1. Evaluation of the Measures M_{GK}, OR, and OR_{hn} according to the Properties. Reference [25] stated in his work that among the 22 properties proposed, only 19 are studied. Like, the properties P_1: intelligibility or comprehensibility of the measure, P_2: easy setting a threshold for acceptance of the rule, and P_{22}: robust measure have not been retained in this study. Indeed, we think that the first two are subjective and depend on the user's knowledge of statistics and the third property is difficult to study since it requires very advanced calculation tools in order to avoid calculation errors.

For the behaviour of the three measures M_{GK}, OR, and OR_{hn} we study in the following Table 7, the number 1 means that the measure in question verifies the property concerned P_i such that $3 \leq i \leq 21$. And the number 0 means that the measure in question does not validate the property concerned P_i. This work will lead to the construction of a matrix, which we present in Table 7. This study represents the result of the behaviour of the three measures M_{GK}, OR, and OR_{hn} according to their properties.

By studying the mathematical properties of these three measures, we have obtained the following theorems.

According to [31], we define a fixed value c in the case of equilibrium as follows.

Definition 11. Let m be a measure of quality of association rules. A fixed value c is a reference point which is the equilibrium when a rule has as many examples as counterexamples.

(i) If m does not admit a fixed value in the case of equilibrium, that is, if $\forall c \in \mathbb{R}, \exists (X \longrightarrow Y)$, such that $P_X(Y) = P(X)/2$ and $m(X \longrightarrow Y) \neq c$, then $P_{11}(m) = 0$.

(ii) If m admits a fixed value in the case of equilibrium, that is, if $\exists c \in \mathbb{R}, \forall (X \longrightarrow Y)$ such that $P_X(Y) = P(X)/2 \Rightarrow m(X \longrightarrow Y) = c$, then $P_{11}(m) = 1$.

By studying the property P_{11}, we obtain the following theorem.

Theorem 1. *All affine measures and homography normalisable do not admit of a constant value in the case of equilibrium.*

Proof. According to the definition of normalised measures (2.4), according to definition (3.1), and the justification of this theorem in the demonstration of the property P_{11}.

According to [32], we define a discriminating measure as follows. □

Definition 12. A quality measure m is said to be discriminating if it is able to distinguish interesting rules as the size of the training set n increases.

(i) If $\exists\eta \in \mathbb{N}^*, \forall n > \eta, \forall X_1 \longrightarrow Y_1, \forall X_2 \longrightarrow Y_2$ such that $(P_{X_1}(Y_1) > P(Y_1)$ and $P_{X_2}(Y_2) > P(Y_2)) \Rightarrow m(X_1 \longrightarrow Y_1) \approx m(X_2 \longrightarrow Y_2)$, then $P_{21}(m) = 0$ (nondiscriminant).

(ii) If $\forall\eta \in \mathbb{N}^*, \exists n > \eta, \exists X_1 \longrightarrow Y_1, \exists X_2 \longrightarrow Y_2$ such that $(P_{X_1}(Y_1) > P(Y_1)$ and $P_{X_2}(Y_2) > P(Y_2)) \Rightarrow m(X_1 \longrightarrow Y_1) \neq m(X_2 \longrightarrow Y_2)$, then $P_{21}(m) = 1$ (discriminant).

By studying the property P_{21}, we obtain the following theorem.

Theorem 2

(i) A probabilistic quality measure μ is a discriminant measure if, and only if, for any association rule $X \longrightarrow Y$, the following condition is verified at the reference situation: $\mu_{imp} \neq \mu_{ind} \neq \mu inc$.

(ii) All the measures (affine or homography) which can be normalised and normalised are discriminating measures.

Proof

(i) Following the Definition 2 of a discriminant measure and the justification of this theorem in the demonstration of the property P_{21} of supplementary document, Theorem 2 is obvious.

(ii) Following the definition of normalised measures 4 and the justification of this theorem in the demonstration of the property P_{21} . □

3.5. Recapitulation. From the analysis carried out, we claim that the measures OR, OR, and OR_{hn} are all good measures of the quality of association rules. Indeed, they almost all validate half of these properties. According to Table 7, the M_{GK} measure validated 12 out of 19 properties; then, the odds ratio measure (OR) validated 11 out of 19 properties and the odds ratio normalised homography measure (OR_{hn}) validated 09 out of 19 properties. As a consequence, the best measure among the three proposed ones is the M_{GK} measure. Thus, we have well discovered that the tool of homography-normalisation does not improve the totality of the behaviour of the measures, but it simply improves value to the reference situation.

4. Conclusion and Perspective

This paper has summarised the different comparative studies of the M_{GK}, OR, and OR_{hn} measures. We have seen that the M_{GK} measure gives really more precise and easier to interpret results; on the other hand, the odds ratio measure does not take fixed values at the logical implication, which caused the difficulty to define a minimum threshold: it was very difficult to interpret a rule, and the homography-normalised odds ratio measure (OR_{hn}) is a normalised measure but is not more precise and relevant than the M_{GK} measure. The comparative study of these properties allows us to discover the most relevant measure among the three. From the results we obtained from the comparative studies of the behaviour of the three measures M_{GK}, OR, and OR_{hn} in this paper, there is no doubt that the measure M_{GK} is the most relevant among the three proposed measures. These results answered well the uncertainties of thinking that the odds ratio measure had the best mathematical properties compared to the other measures. In epidemiology, in order to have more reliable, precise, and easier-to-interpret results on the study of the quantification of statistical implication link between an exposure and the disease, we advise analysts and epidemiologists to choose the M_{GK} measure in their work. As a prospect, we propose to extend our comparative study of these measures into the statistical field, more specifically into the study of logistic regression, the study of statistical inference, and so on.

Data Availability

No data were used to support the results of this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are very grateful to the leaders of the Doctoral School "Science, Culture, Society and Development" of the University of Toamasina who welcomed them in its team: Mathematics, Informatics and Applications (MIA) to carry out their research concerning this article.

Supplementary Materials

A supplementary file named "Supplementary material" presents brief reminders and demonstration that check the result of the matrix describing the three measures according to the 19 properties in Table 7. (*Supplementary Materials*)

References

[1] M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis, and W. Wong, *Statistics for Biology and Health*, Springer, New York. NY, USA, 2010.

[2] S. Antony, *Basic Statistics and Epidemiology. in publication data*, British Library Catotoguing, London, UK, 2002.

[3] C. R. Rao and J. P. Miller, *Epidemiology and Medical Statistics*, Library of Congress Cataloging in Publication Data, Amsterdam, Netherlands, 2008.

[4] W. K. Axame, F. N. Binka, and M. Kweku, "Prevalence and factors associated with low birth weight and preterm delivery in the ho municipality of Ghana," *Advances in Public Health*, vol. 2022, Article ID 3955869, 11 pages, 2022.

[5] C. Adwar, S. S. Puleh, I. Ochaba et al., "Factors associated with linkage to care following community-level identification of

hiv-positive clients in lira district," *Advances in Public Health*, vol. 2022, Article ID 4731006, 9 pages, 2022.

[6] A. Agresti, "Generalized odds ratios for ordinal data," *Biometrics*, vol. 36, no. 1, pp. 59–67, 2014.

[7] J. M. Bland and D. Altman, "The odds ratio. Education and debate," *Statistics Notes*, vol. 1, 2000.

[8] P. Sedgwick, "Relative risks versus odds ratios," *BMJ*, vol. 348, 2014.

[9] U. Held, *Qu'est-ce que l'Odds ratio et à quoi sert-il? biostatistique suisse*, Université du Spita, Palakkad, Kerala, 2010.

[10] J. Bouyer, "Epidémiologie environnement et santé publique," Acton vale paris, Paris, France, Tec et Doc, 2003.

[11] G. B. Buchanan and E. H. Shortliffe, *Rule-based Expert Systems*, Springer, Berlin, Germany, 1984.

[12] S. Guillaume, "Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales," Ph. D. thesis, Université de Lyon, Lyon, France, 2000.

[13] A. Totohasina, H. Ralambondrainy, and J. Diatta, "Une vision unificatrice des mesures de la qualité des règles d'association booléennes et un algorithme efficace d'extraction des règles d'association implicative," in *Proceedings of the CARI'04*, pp. 511–518, Hammamet, Tunisie, 2004.

[14] X. Wu, C. Zhang, and S. Zhang, "Efficient mining of both positive and negative association rules," *ACM Transactions on Information Systems*, vol. 22, pp. 381–405, 2004.

[15] A. Totohasina, H. Ralambondrainy, and J. Diatta, "Ion : a pertinent new measure for mining information from many types of data," in *Proceedings of the f TAIMA'05*, pp. 375–380, Hammamet, Tunisie, 2005.

[16] H.-F. Rakotomalala, J. Diatta, and A. Totohasina, "Une mesure de cohésion basée sur la mesure de qualité des règles d'association M_{GK}," *SFC Sciences des données, Actes des 24; mes Rencontres de la Société Francophone de Classification*, pp. 21–24, 2019.

[17] H. F. Rakotomalala, *Classification Hiérarchique Implicative et Cohésitive selon la mesure M_{GK}-Application en didactique de l'informatique. Didactique des mathématiques et de l'informatique*, Université d'Antananarivo, Antananarivo, Madagascar, 2019.

[18] B. B. Ralahady and A. Totohasina, "Experimental study of the valid rules according to the measure M_{GK}," *International Journal of Computer Science & Technology*, 2019.

[19] A. Totohasina, "Normalisation de mesures probabilistes de la qualité des règles," *Journées de statistiques*, pp. 985–988, 2003.

[20] D. R. Feno, "Mesure de qualité des règles d'association : normalisation et caracterisation des règles d'association des bases," Ph. D. thesis, Universitè de La Réunion spécialité : Mathématiques Informatique, Paris, France, 2007.

[21] G. Piatetsky-Shapiro, "Knowledge discovery in real databases," vol. 11A report on the ijcai-89 workshop, p. 2, AI Magazine, Norwich, England, 1991.

[22] S. B. B. Masonova, "Etudes mathématiques des modèles de quelques maladies infectieuses," *ENSET-université d'Antsiranana-Madagascar*, Master's thesis, Master, France, 2017.

[23] R. F. D. Armand and T. André, "Nouvelle vision unificatrice des mesures d'intérêt : une normalisation par homographie," *AAFD SFC Sciences des données, défis mathématiques et algorithmiques*, pp. 287–292, 2016.

[24] R. F. D. Armand and T. André, "Vers une théorie de fonction de normalisation des mesures d'intérêt," *AAFD SFC Sciences des données, défis mathématiques et algorithmiques*, pp. 281–286, 2016.

[25] D. Grissa, *Etude comportementale des mesures d'intérêt d'extraction de connaissances. Ph. D. thesis, Université Blaise Pascal-Clermont-Ferrand II*, Université de Tunis-El Manar, Hammamet, Tunisie, 2013.

[26] B. Ganter and R. Wille, "Formal concept analysis," *Mathematical Foundations*, Springer-Verlag, Berlin, Germany, 1999.

[27] A. Totohasina, "Contribution à létude des mesures de qualité des règles d'associations : normalisation sous cinq contraintes et cas de M_{GK}: propriétés, bases composites des règles et extension en vue d'applications en statistique et en sciences physiques," Ph. D. thesis, Université, d'Antsiranana, Antananarivo, Madagascar, 2008.

[28] P. N. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in *Proceedings of the 8th ACM Intl. Conf. on Knowledge Discovery and Data Mining*, vol. 2, pp. 32–41, Edmonton, Alberta, Canada, July, 2002.

[29] Armand, "Exploration des propriétés des homographies : application à la normalisation des mesures de qualité des règles d'association," Ph. D. thesis, Éditions Universitaires européennes, Saarbrücken, Germany, 2019.

[30] L. Held, C. Rufibach, and C. Seifert, "Einfuhrung in die biostatistik," *Abteilung biostatistik*, Institut fur Sozial-und Praventivmedizin der Universitat Zurich, zurich, Germany, 2009.

[31] J. Blanchard, F. Guillet, H. Briand, and R. Gras, "Ipee : indice probabiliste d'écart á l'équilibre pour l'évaluation de la qualité des règles," in *Atelier Qualité des Données et des Connaissances*, vol. 26–34, Cépaduès-Éditions, Toulouse, Occitanie, 2005.

[32] S. Lallich and O. Teytaud, "Evaluation et validation de mesures d'intérêt des règles d'association," *RNTI–E–1, spécial*, pp. 193–217, 2004.