Characterization and reduction of variability in selection based on effect-size using association measures in cohort study of heterogeneous diseases.

Venkateshan Kannan[1], Kristina Alexandersson[2], and Jesper Tegner[1]

[1]Computational Medicine Unit, Department of Medicine, Solna, Karolinska Institutet, SE-17176, Stockholm, Sweden
[2]Division of Insurance Medicine, Department of Clinical Neuroscience, Karolinska Institutet, SE-17177 Stockholm, Sweden

Abstract

Epidemiological studies employ pairwise measures of association to quantify dependencies among diseases and exposures. The reliable use of these measures to draw conclusions about the underlying association strengths requires that the measures have no undue dependencies that systematically distort their values. These conditions are particularly relevant when multiple pairwise associations are compared and ranked in a cohort study, as in the case of construction of disease networks. Following an empirical approach and using disease diagnoses data from a large cohort of 5.5 million patients as a test set, we develop a comprehensive methodology to characterize the variability of measures in selection of disease pairs based on effect-size. Specifically, we define putative bias variables, and examine the distribution of the measure scores conditioned upon it. This procedure reveals systematic bias in widely used measures such as relative risk and correlation coefficient. In addition, we devise a novel measure family using a stochastic model for differential rate of development of diseases. We demonstrate marked reduction in above-mentioned bias with an appropriate choice from this measure family.

1 Introduction

Population cohort is used in epidemiological and medical research to infer important relationships between diseases and related factors. These relationships are deduced based on statistical measures [1, 2] employed to quantify the nature and degree of association.

The validity of inferences made using such measures therefore depends on its consistency and robustness. In qualitative terms, a pairwise association measure should represent the 'true interaction' between a pair of diseases, which, among other things, must be independent of those quantities that do not involve their joint co-occurrence in the cohort. For example, the measure should be independent of the minimum (or maximum) of the prevalences of the two diseases. This must not be mistaken to mean that the definition of the measure should not involve such quantities, which they most often do. The requirement is that, when applied to large number of disease pairs, there be no systematic dependence of the measure on such quantities. Naturally, this condition is of even greater significance when comparison and rankings of multiple disease pairs with dissimilar prevalences are involved as in the case of disease networks [3, 4, 5].

The straightforward approach to investigate this potential problem would be to simulate several pairs of diseases with a similar 'interaction' but differing with respect to the other quantities (those independent of joint co-occurrence) and determining if there is a systematic dependence of the measure on these quantities. The difficulty with this approach is that there is no objective definition of 'interaction' of a pair of diseases; indeed, the measure is trying to precisely capture this, leading to circularity in the problem statement. And yet, this phenomena does not arise from an artifice in semantics; its origins lie in our lack of understanding of how a pair of related diseases interact.

For example, in the study of [6], several comorbidity measures are compared on simulated data, with the strength of association determined by the $z-$score of the co-incidence assuming binomial distribution. This of course is a specific choice for setting association strengths and evaluating the different measures using that as a benchmark. *A priori*, there is no justification for assuming this form for the relationship between strength of association and co-occurrence numbers.

In observational studies, where statistical associations are only cautiously interpreted as potential causal relations [7], this lack of understanding is accepted. Nonetheless, what is less appreciated is how similar values of association obtained by applying a single measure may reflect different extents of association, even when other sources of bias such as selection type or confounder effect is absent.

Although there have been studies describing the differences in measure properties and suggested procedures for domain-specific selection [8, 9], to the best of our knowledge, there has been no systematic approach to examine the bias in association measures in the context of diseases. Therefore, a large cohort data become useful in investigation of measure bias using the empirical distribution of disease associations. Critically, the difficulty of accurate simulation of disease associations discussed earlier is directly overcome with an empirical dataset consisting of prevalence and co-occurrence of a heterogeneous set of diseases. This permits translation of our abstract conception of bias into a more well-defined form characterized by systematic over- or under- estimation of as-

sociations depending on underlying putative bias variables. This also provides an unambiguous method to compare the bias of different measures.

Here, we use the ICD-10 coded diagnoses from in-patient care, disability pension and causes of death for a cohort of all 5.5 million working-aged people living in Sweden in December 1994, followed prospectively for 13 years. With this cohort serving as test data, we develop a systematic methodology to characterize the biases of standard measures of associations between diseases. With the aim of understanding the impact of bias on effect-size calculations, we quantify the bias using three distinct indices. The ensuing analysis provides insight into the sources of bias which guides our approach to devise new measures. Using a stochastic model of disease development with differential rate of diagnosis, we derive a novel measure family. We find that a suitable measure from this family provides the best performance among all the considered measures in terms of having the least overall bias.

2 Results

We introduce the framework for characterizing bias by considering two commonly used measures for pairwise associations, relative risk (RR) [10, 11] and $\phi-$correlation [12, 13]. RR and the family of measures similar to it such as odds ratio [14, 15], hazard ratio, and Yule's Q [16] are based on relative probabilities of occurrence of diseases in different conditions. To consider these measures in more general terms, we define the 2×2 contingency table (Table 1) for a disease pair A and B, where the top-left entry p represents the number of individuals having both diseases, q, the number having B but not A and similarly for the entries on the second row. $N = p + q + r + s$ is the total number of individuals.

The relative risk for disease A, in the presence of B is the ratio of prevalence of A in the subset diagnosed with B to the prevalence in the subset not diagnosed with B. :

$$RR_{A|B} = \frac{p/(p+q)}{r/(r+s)} \tag{1}$$

and likewise, swapping A and B

$$RR_{B|A} = \frac{p/(p+r)}{q/(q+s)}$$

It should be noted that the original definition of relative risk considers disease incidence among sets exposed and not exposed to a given condition. This explains the asymmetry in the above definition, where presence of a disease is considered as an exposure condition. However, we cannot apply the original interpretation directly to the cohort when the order of occurrence of the disease pair in an individual is unknown. Despite that, we can make simplifying assumptions for large cohort, and this was precisely the definition used in previous studies

involving disease networks [3, 13]. As the individual disease prevalences in our cohort is small, we can assume that

$$p/s, q/s, r/s << 1.$$

Further if we assume that

$$p/r, p/q << 1,$$

which effectively implies that prevalence of one disease within the subset of patients having another disease to also be very small, it is easy to show that the reduced expressions for both are identical.

$$RR_{A|B} \sim \frac{pN}{n_A n_B} \tag{2}$$

We find that the expression in the last line above is explicitly symmetric in the two diseases.

A related measure, odds ratio [17] for the same contingency table is given by:

$$OR = \frac{p/q}{r/s} = \frac{ps}{qr} \tag{3}$$

Unlike RR, odds ratio is explicitly symmetric in the two diseases, and further it is easy to show that, in the limit that we are interested in, where prevalence rates are assumed to be small, they converge to the same value.

$$\frac{ps}{qr} = \frac{pN}{n_A n_B} \frac{1}{1 + (p+q+r)/s}(1 + p/r)(1 + p/q) \sim \frac{pN}{n_A n_B}$$

ϕ correlation:

Another common measure used for contingency table is the ϕ correlation and measures that reduce to a similar form include Cohen's (κ) [18] and Kendall's tau-b [19].

The ϕ correlation is obtained by taking the standard correlation between the binary vectors corresponding to the two diseases. For a given disease A, the corresponding vector is of length equal to the number of patients and each entry is 1(0) depending on the disease being present (absent) in that individual.

$$\phi_{A,B} = \frac{p - n_A n_B}{\sqrt{n_A n_B (N - n_A)(N - n_B)}} \tag{4}$$

Equivalently, it can be defined as $\sqrt{\chi^2/N}$ where χ^2 is chi-squared statistic calculated for the contingency table.
Another related measure (based on concordant statistics) is Sommer's D [6], and in the limit discussed earlier, reduces to an asymmetric version of the ϕ correlation coefficient.

$$\text{Sommer's D} = \frac{p - \frac{qr}{N}}{\min\{n_A, n_B\}}$$

2.1 Measure Biases

Let $\mathcal{D}$ be a sample of size N drawn from a population where each element is a vector of binary variables $(A_1, A_2, \cdots A_K)$ such that pairwise associations among these variables are to be determined and compared. For the cohort disease set, A_i's are the indicator variables for the presence or absence of disease i and every data-point represents an individual. For a given pair of binary variables (A_i, A_j), let n_{uv} be the number of pairs of the form (u, v), $u, v \in \{0, 1\}$, and n_{1+} be the marginal on A_i and likewise for A_j . For K sufficiently large, we say that the measure distribution has a dependence on a variable W (W is some function of n_{1+}, n_{+1}, N only) if the distribution of the measure conditioned on pairs of variables for which W lies in some finite interval Δw is different from that of the independent distribution, i.e., $P(\mu|W \in \Delta w) \neq P(\mu)$. When there is a systematic dependence of the distribution statistic (for example, mean or median) on W, we say there is a bias with respect to W.

It is important to note that this definition of bias is empirical. The existence or non-existence of bias in a given measure is dependent on the nature of associations found in the population. A measure may show considerable bias when applied to binary variables representing diseases but may be well-suited (i.e., show no bias) when measuring association between developmental indices of nations. This is qualitatively different from identifying general properties that a measure is required to satisfy [20, 21] although it is generally accepted that identifying the property that is most relevant depends on the specific context [8].

The interval width Δw is chosen so that there are sufficient number of data-points falling within that interval to obtain reliable statistics. Larger the value of K, the smaller Δw can be chosen.

Define $\mathcal{F}$ to be the set of distinct diseases (classified by ICD 10 three-character precision) in our cohort with $K = |\mathcal{F}| > 1400$. We investigate measure bias by considering variables W that are likely to affect the measure characteristics. We first examine potential bias with respect to the expected co-occurrence obtained for a given pair of diseases $A, B \in \mathcal{F}$ under the assumption of independence of co-occurrence $W = n_{AB}^{(0)} = \frac{n_A n_B}{N}$. Observe that there is no *a priori* reason to expect the distribution of the measure to depend on it.

In line with our formulation, we partition the set of all disease pairs $\mathcal{D}$ into $M = 20$ mutually exclusive subsets $\mathcal{D}_j, j = 1, 2, \cdots M$.

$$\mathcal{D} = \{\{A, B\}|A, B \in \mathcal{F}\} = \cup_{j=1}^{M} \mathcal{D}_j$$

where $\mathcal{D}_j = \{\{A, B\}|v_{j-1} < n_{AB}^{(0)} < v_j\}$ The intervals v_j are determined by the requirement that each of these partitions contain the same number of pairs $|\mathcal{D}_0| = \cdots |\mathcal{D}_j| = \cdots |\mathcal{D}_M|$.
Having partitioned the data set according to the potential bias variable, we apply the measures within each to identify systematic dependence, if any. We

begin with RR (Eq. 2) and Fig. 1a shows the box plot of the distribution of the measure values within each partition. We find unambiguous systematic bias where low expected co-occurrence leads to higher values of RR. Equally significant is the variation of size of the box representing the boundaries of the 25th and 75th percentile (interquartile range) for the collection of RR values in each partition. This increase is even more pronounced than that of the median, as the expected co-occurrence decreases.

We can explain the large spread by noting that, for lower expected co-occurrences, small fluctuations in co-occurrence numbers leads to wide variations in RR. Under the assumption of the independence of a disease pair, we can show that the distribution of co-occurrences (for a given disease pair) is Poisson with mean being the expected co-occurrence. As the variance of a Poisson distribution is equal to the mean, the coefficient of variation is given by $\sqrt{n_{AB}^{(0)}}/n_{AB}^{(0)} = \frac{1}{\sqrt{n_{AB}^{(0)}}}$, representing an inverse relation with expected co-occurrence .

The same analysis is repeated for the ϕ correlation and Fig. 1b shows the box plot for this case. A systematic bias is once again immediate from inspecting the figure, except that the bias points in the opposite direction: ϕ correlation tends to inflate the associations for higher expected co-occurrences.

We thus find that both RR and ϕ correlation have significant bias with respect to the expected co-occurrence, except that the bias works in opposite directions. Where RR tends to assign higher associations to disease pairs with lower expected co-occurrence, the exact opposite it true for the ϕ correlation. Conversely, higher expected co-occurrences lead to lower RR and higher ϕ correlation. Note that lower (higher) expected co-occurrences arise when one or both diseases have low (high) prevalence.

3 Characterizing Bias

Since both RR and ϕ correlation show bias, we want to compare their magnitude of bias. This requires a scheme to correctly identify and quantify the bias for any given measure.
It has long been recognized in clinical and observational studies and in epidemiology that null hypothesis tests for associations are of only limited use [22, 23]. Specifically, the rejection of the null-hypothesis and the significance level at which it is rejected does not signify the degree of association. Hence effect-sizes are necessary where the strength of association is important [24] as is likely true in most realistic cases.

Selection based on minimum effect-size is equivalent to setting a threshold (assuming everything else to be fixed) for the measure. We want to characterize bias in terms of its impact on selection. For definiteness, assume that we are interested in the subset of pairs comprising the top f fraction of $\mathcal{D}$ (as determined by a given measure). We can equivalently express this in terms of a

threshold ($\theta^{(f)}$) where all pairs with measure values greater than the threshold are selected. If the same threshold is then independently applied to the pairs in every partition for the bias variable, a fraction f_j of pairs within each subset is selected. If there is no bias, then these fractions would be identical; conversely the extent of variation of the fractions across the partitions will be treated as a proxy for the bias.

Fig. 2a shows the selected fractions for RR and ϕ correlation measures when the overall fraction f sought is 0.05 (i.e, we set a threshold such that only 5% of all pairs are greater than that threshold) with expected co-occurrences as the bias variable. As expected the stringency of the selection depends on the partition, with larger fractions being chosen for lower expected co-occurrences in the case of RR (and the opposite for ϕ correlation).

This approach of using fractions to understand bias has a distinct advantage of treating all measures on an equal footing. The fractions represent the effect of the measure on the result of querying the data-set. To capture the extent of variation of fractions across the different partitions, with emphasis on its impact, we define appropriate indices to characterize them. The naive approach of using standard deviation is unsatisfactory because the fractions are not normally distributed. We propose three different indices that helps us better understand the bias.

1) **Interquartile range (IQR)**: IQR makes no assumption about how the data is distributed. This describes the range within which the half of the data around the median is located. This also has the advantage that the metric scales very smoothly and dependably with the bias variation in fractions.

2)**MinMax** : A metric that helps us characterize the extremes is $\frac{\max f}{\min f}$. While this measure may be seen as something very narrow in scope, it captures the worst-case scenario of the bias: comparing two pairs that come from two extremes.
3)**Mean Absolute Deviation (MAD)**. This weights all deviations from the median equally. Again, this would be preferred over the mean squared deviation because the underlying distribution need not be normal.
The result of applying the three indices to the fractions shown in Fig. 2a is shown in Fig. 2b. We find that the MinMax score has a largest gap for the two measures (note that MinMax has been reduced by a factor of 100 in all figures), and RR's bias is indeed very high (note the y-axis is logarithmic), suggesting that the effect of the bias on RR is more severe compared to ϕ correlation, when comparing two disease pairs whose expected co-occurrences differ widely. There is not much separating the two measures in terms of IQR or MAD.

3.1 Prevalence Ratio Bias

The ratio of the prevalences of the two diseases as a possible bias for measures has been considered before [25, 20]. We follow a similar approach as with

expected-co-occurrence and partition all pairs based on $W = \frac{n_<}{n_>}$, where $n_<(n_>)$ represents the prevalence of the less (more) prevalent disease in the pair. We use the same threshold on RR and ϕ such that 5% of all pairs are selected overall. The variation of the fractions is shown in Fig. 3a and we find that ϕ correlation shows larger differences in fractions, with distinctly suppressed numbers for low prevalence ratios. RR is relatively more balanced and this fact is also reflected in the three indices shown in Fig. 3b.

3.2 Overcoming Bias

The bias of ϕ correlation with respect to the prevalence ratio can be explained by noting that ϕ has an upper bound that depends on the ratio of prevalences. Assuming disease pairs with prevalence $n_A < n_B$

$$\begin{aligned}
\phi(1,2) &= \frac{n_{AB}/N - (n_A/N)(n_B/N)}{\sqrt{(n_A/N)(n_B/N)(1-n_A/N)(1-n_B/N)}} \\
&<= \frac{\min\{n_A, n_B\}/N - (n_A/N)(n_B/N)}{\sqrt{(n_A/N)(n_B/N)(1-n_A/N)(1-n_B/N)}} \\
&= \frac{(n_A/N)(1-n_B/N)}{\sqrt{(n_A/N)(n_B/N)(1-n_A/N)(1-n_B/N)}} \\
&= \sqrt{\frac{n_A}{n_B}\frac{1-n_B/N}{1-n_A/N}} \\
&< \sqrt{\frac{n_A}{n_B}}
\end{aligned}$$

Thus the maximum possible association between two diseases is not a constant but depends on their prevalence ratio [20]. This suggests that disease pairs with a high disparity in prevalences would systematically have lower values of ϕ correlation and indeed this is what we observed in Fig. 3a.

A quick workaround of this problem is defining a modified $\phi^M = \frac{n_{AB}/N - (n_A/N)(n_B/N)}{\min\{n_A/N, n_B/N\}}$, which has a uniform upper bound of unity, attained when co-occurrence $n_{AB} = \min\{n_A, n_B\}$. Fig. 4a compares the original correlation and the new modified version. We immediately observe the correction offered by ϕ^M to the original correlation goes past the required bias removal: the measure shows a bias in the opposite direction, tending to select greater fraction for more dissimilar prevalence ratios [26]. The three indices in Fig. 4b shows that ϕ^M in fact has higher bias than ϕ.

4 New Measure

We propose a conceptual and systematic approach to define a new measure of association and this leads us to a family of measures parametrized by a constant

γ

$$\phi_\gamma = \left(\frac{n_1 n_2}{N}\right) \frac{\frac{n_{12}}{N} - \frac{n_1 n_2}{N^2}}{\left(\frac{n_1+n_2}{N}\right)^2 - \left(\frac{n_1-n_2}{N}\right)^2 \gamma}. \quad (5)$$

The full derivation of this measure (Eq. 19) is given in the Methods section but the approach and motivation is as follows. We formulate the association between a pair of diseases in terms of the differential rate of development of one disease in the presence or absence of the other. We then use stochastic differential equation to evolve the disease probabilities in time and use the contingency table entries as constrains to estimate these rates. Since these rates are *a priori* unknown, and there are more variables than equations, we further use an additional constraint to choose a unique solution (the parameter γ originates from the constraint equation).

We define our new measure as a member of this family, with $\gamma = 0.994$.

$$\phi^{M2} = \phi_{\gamma=0.994} = \left(\frac{n_1 + n_2}{N\beta}\right) \frac{\frac{n_{12}}{N} - \frac{n_1 n_2}{N^2}}{0.006 * ((n_1/N)^2 + (n_2/N)^2) + 3.988 * \frac{n_1 n_2}{N^2}} \quad (6)$$

The choice of γ is not entirely arbitrary - the justification for its selection is given in the Appendix, along with a discussion of the relationship between the measure properties and γ. The fractions obtained when pairs are partitioned by prevalence ratio for ϕ, ϕ^M and ϕ^{M2} are compared in Fig. 5a. Visual inspection suggests that ϕ^{M2} is the most balanced among them (it increases with decrease in prevalence ratio but unlike ϕ^M, ϕ^{M2} attains a maximum and turns around). The corresponding indices in Fig. 5b confirm our observations that ϕ^{M2} has the least bias for all three indices, and while standard correlation performs comparably well with IQR and MAD, ϕ^{M2} is considerably better with MinMax.

4.1 Comparison across threshold fractions

We have thus far demonstrated that the new measure has the least bias with respect to prevalence ratio compared to other correlation measures. However, this was done for a specific setting of threshold, such that an overall fraction of 5% of pairs are selected. The next logical step is to find out if that reduction in bias is valid for a wider range of fractions. Indeed, the threshold setting for determining significant pairs would depend on the context of the inquiry.

A comprehensive comparison of the four measures (RR, ϕ, ϕ^M and ϕ^{M2}) is done across five different thresholds (corresponding to overall fraction of selected pairs, 0.1,0.05,0.01,0.005 and 0.001). The basis of comparison is the three indices (IQR, MAD, MinMax), taken one at a time.

Fig. 6a,b and c shows the relative performance of all measures according to IQR, MAD and MinMax respectively when potential bias variable is the prevalence ratio. We find, for example, from the first two sublots for IQR and MAD, that ϕ^{M2} has the least bias for threshold fractions less than or equal to 0.01 but has significantly more bias than RR for higher fraction of selections. For MinMax,

ϕ^{M2} has the least bias for overall fractions less than 0.05 and is only marginally worse than RR when the overall fraction is 0.1. At the other end, either ϕ or ϕ^M have the highest bias for any given overall fraction and any given index.

It is clear that, although the new measure ϕ^{M2} is consistently better than the two correlation measures, RR has lower bias when the selected fractions are higher. There is nothing unexpected about this because the new measure was devised to eliminate the bias in ϕ and ϕ^M only. While it may be tempting to go with RR for higher threshold fractions, we cannot prematurely conclude that until we consider the bias due to the expected co-occurrence as well.

Figures 7a,b and c explore the biases with respect to the expected co-occurrence using IQR, MAD and MinMax indices respectively. We find that, although we had observed in Fig. 6 that RR performed well for overall fractions greater than or equal to 0.05, this is not the case when we examine the bias due to expected co-occurrence. Both MAD and MinMax indices show RR having the highest bias for these thresholds. In addition, the (absolute) magnitude of the bias of RR in Fig 7 is greater than the magnitude of bias of ϕ^{M2} in Fig. 6 for all three indices . Taking together, ϕ^{M2} still comes out as a better choice.

Interestingly, Fig. 7 shows ϕ^M having the least bias for all fractions selected based on MAD and MinMax. Nonetheless, we note that ϕ^{M2} is not far behind in comparison. It is also evident from examining Figs. 7 and 6 that whatever meagre gains can be made with ϕ^M in the case of bias with respect to expected co-occurrence, the cost of greater bias with respect to prevalence ratio is too high a price for the trade-off. Thus, when we consider the bis variables, the thresholds and the indices in totality, it is clear that ϕ^{M2} would be the most preferable measure.

5 Methods

Our approach to the new measure starts with consideration of the relative probabilities to develop one disease following another. In the most general case, these probabilities are independent of each other. Let $\eta_1(t)$ and $\eta_2(t)$) be boolean random variables corresponding to the two diseases (note the slight change in notation from earlier) which take values 1(0) when the disease is present (absent) at time t. We want to obtain a set of relations between the probabilities of occurrence and co-occurrence of the two diseases in the population at the end of time τ assuming that neither disease was present at start. To that end, we assume that the probability to be diagnosed with the disease is given by a Poisson process. However, we assign different, *a priori* unknown, rates to the Poisson process of a given disease depending on whether or not the other disease has already been diagnosed. For example, a given realization would be the following: starting from being disease free, disease 1 is contracted at time t_a following which the Poisson rate for contracting disease 2 is different.

More specifically, if $t_{E1}(t_{E2})$ represent the time point when disease 1(2) was diagnosed,

$$P(t < t_{E1} < t + \delta t | \eta_2(t) = 0) = \lambda_{1P}\delta t \tag{7}$$
$$P(t < t_{E1} < t + \delta t | \eta_2(t) = 1) = \lambda_{1S}\delta t \tag{8}$$

In the first case, the primary rate λ_{1P} determines development of disease 1 in the absence of disease 2, but if disease 2 has been contracted before, then there is the secondary rate λ_{1S}. Likewise, the rate determining development of disease 2 before (after) diagnosis of disease 1 is given by $\lambda_{2P}(\lambda_{1S})$.

The conditional probabilities at finite time t:

$$\begin{aligned}
P(\eta_1(t) = 1 | \eta_2(t) = 0) &= \frac{P(\eta_1(t) = 1, \eta_2(t) = 0)}{P(\eta_2(t) = 0)} \\
&= \frac{\int_0^t P(\eta_1(t') = 0, \eta_2(t') = 0) P(t' < t_{E1} < t' + \delta t') P(\eta_2(t'') = 0 | t' < t'' < t) dt'}{P(\eta_2(t) = 0)} \\
&= \frac{\int_0^t e^{-\lambda_{1P}t'} e^{-\lambda_{2P}t'} \lambda_{1P}\delta t' e^{-\lambda_{2S}(t-t')}}{P(\eta_2(t) = 0)} \\
&= \frac{e^{-\lambda_{2S}t} - e^{-(\lambda_{1P}+\lambda_{2P})t}}{P(\eta_2(t) = 0)(\lambda_{1P} + \lambda_{2P} - \lambda_{2S})}
\end{aligned}$$

At the end time point τ :

$$\begin{aligned}
P(\eta_1(\tau) = 1) &= P(\eta_1(\tau) = 1 | \eta_2(\tau) = 0) P(\eta_2(\tau) = 0) + P(\eta_1(\tau) = 1, \eta_2(\tau) = 1) \\
&= \frac{e^{-\lambda_{2S}t} - e^{-(\lambda_{1P}+\lambda_{2P})t}}{(\lambda_{1P} + \lambda_{2P} - \lambda_{2S})} + n_{12}/N \\
&= n_1/N
\end{aligned} \tag{9}$$

where we identify the conditional probabilities at the end point with empirical values from the data : $P(\eta_1(\tau) = 1, \eta_2(\tau) = 1) = n_{12}/N$ and $P(\eta_1(\tau) = 1) = n_1/N$
Likewise, for disease 2:

$$\begin{aligned}
P(\eta_2(\tau) = 1) &= P(\eta_2(\tau) = 1 | \eta_1(\tau) = 0) P(\eta_1(\tau) = 0) + P(\eta_2(\tau) = 1, \eta_1(\tau) = 1) \\
&= \frac{e^{-\lambda_{1S}t} - e^{-(\lambda_{1P}+\lambda_{2P})t}}{(\lambda_{1P} + \lambda_{2P} - \lambda_{1S})} + n_{12}/N \\
&= n_2/N
\end{aligned} \tag{10}$$

We can write the probability of co-occurrence $P(\eta_1(\tau) = 1, \eta_2(\tau) = 1)$ as a sum of two probabilities for two mutually exclusive sets of events, one where disease

1 precedes disease 2, and second where this order is reversed.

$$P(\eta_1(\tau)=1,\eta_2(\tau)=1) = \int_0^\tau P(\eta_1(\tau)=1|\eta_1(t_{E2})=0)P(\eta_1(t_{E2})=0,t(E2)=t_{E2})dt_{E2} + \int_0^\tau P(\eta_2(\tau)=1|t(E1)=t_{E1},\eta_2(t_{E1})=0)P(\eta_2(t_{E1})=0,t(E1)=t_{E1})dt_{E1} \tag{11}$$

where the first(second) factors accounts for cases where diagnosis of disease 1 was made after (before) disease 2. The factors in the integrand above are :

$$P(\eta_1(\tau)=1|t(E_2)=t_{E_2},\eta_1(t_{E_2})=0) = 1-e^{-\lambda_{1S}(\tau-t_{E_2})}$$
$$P(\eta_2(\tau)=1|t(E_1)=t_{E_1},\eta_1(t_{E_1})=0) = 1-e^{-\lambda_{2S}(\tau-t_{E_1})}$$

and similarly:

$$P(\eta_1(t_{E_2})=0,t(E_2)=t_{E_2})dt_{E_2} = e^{-\lambda_{1P}t_{E_2}}e^{-\lambda_{2P}t_{E_2}}\lambda_{2P}dt_{E_2}$$
$$P(\eta_1(t_{E_1})=0,t(E_1)=t_{E_1})dt_{E_1} = e^{-\lambda_{2P}t_{E_1}}e^{-\lambda_{1P}t_{E_1}}\lambda_{1P}dt_{E_1}$$

Plugging this in Eq. (11), the first integral becomes:

$$\begin{aligned}
&\int_0^\tau P(\eta_1(\tau)=1|t(E_2)=t_{E_2},\eta_1(t_{E_2})=0)P(\eta_1(t_{E_2})=0,t(E_2)=t_{E_2})dt_{E_2}\\
&= \int_0^\tau (1-e^{-\lambda_{1S}(\tau-t_{E_2})})e^{-\lambda_{1P}t_{E_2}}e^{-\lambda_{2P}t_{E_2}}\lambda_{2P}dt_{E_2}\\
&= \frac{\lambda_{2P}}{\lambda_{1P}+\lambda_{2P}}(1-e^{-(\lambda_{1P}+\lambda_{2P})\tau}) - e^{-\lambda_{2S}\tau}\frac{\lambda_{2P}}{\lambda_{2P}+\lambda_{1P}-\lambda_{1S}}(1-e^{-(\lambda_{1P}+\lambda_{2P}-\lambda_{2S})\tau})
\end{aligned}$$

If we make the approximation $\lambda_{i\alpha}\tau << 1$ (W1) for $i=1,2$ and $\alpha = P,S$, then the above reduces to $\lambda_{1S}\lambda_{2P}\tau^2/2$. Correspondingly we for the the second integral in Eq. (11) the approximation $\lambda_{1P}\lambda_{2S}\tau^2/2$. leading to:

$$\frac{n_{12}}{N} = \frac{(\lambda_{1P}\lambda_{2S}+\lambda_{1S}\lambda_{2P})\tau^2}{2} \tag{12}$$

Under the same assumption W1, we have:

$$\frac{e^{-\lambda_{2S}\tau}-e^{-(\lambda_{1P}+\lambda_{2P})\tau}}{(\lambda_{1P}+\lambda_{2P}-\lambda_{2S})} \sim \tau-(\lambda_{1P}+\lambda_{2P}-\lambda_{2S})\tau^2$$
$$\frac{e^{-\lambda_{1S}\tau}-e^{-(\lambda_{1P}+\lambda_{2P})\tau}}{(\lambda_{1P}+\lambda_{2P}-\lambda_{1S})} \sim \tau-(\lambda_{1P}+\lambda_{2P}-\lambda_{1S})\tau^2$$

Plugging the above in Eq. (9,10), we obtain :

$$n_1/N = (\lambda_{1P}\tau) - \lambda_{1P}\frac{\lambda_{1P}+\lambda_{2P}}{2}\tau^2 - \lambda_{2P}\lambda_{1S}\tau^2 \tag{13a}$$
$$n_2/N = (\lambda_{2P}\tau) - \lambda_{2P}\frac{\lambda_{1P}+\lambda_{2P}}{2}\tau^2 - \lambda_{1P}\lambda_{2S}\tau^2 \tag{13b}$$

Eqs. (12, 13) are a set of three equations but with four unknowns $\lambda_{i\alpha}$, $i = 1, 2 \qquad \alpha = P, S$, which we cannot solve without an additional simplifying assumption. This is of course what we would expect, because that extra degree of freedom corresponds to our ignorance of the underlying causal relations between the two diseases.

For example if we use the following ansatz :

$$\lambda_{iS} = \lambda_{iP} q, \qquad i = 1, 2$$

we can solve the equations and $q = \frac{n_{12}/N}{(n_1 - n_{12}/2)/N (n_2 - n_{12}/2)/N}$, which would approximate to the standard definition of relative risk ($n_{12}/\min n_1, n_2 << 1$).

Substituting $\lambda_{iS} = \lambda_{iP} + q_i, \qquad i = 1, 2$, we can rewrite the above relations in terms of q_i which represents the deviation from the situation where the two diseases are unrelated, i.e, $q_i = 0$.

$$\begin{aligned}
\frac{n_1}{N} &= \lambda_{1P}\tau - \frac{(\lambda_{1P}\tau)^2}{2} + \frac{\lambda_{2P} q_1 \tau^2}{2} \\
\frac{n_2}{N} &= \lambda_{2P}\tau - \frac{(\lambda_{2P}\tau)^2}{2} + \frac{\lambda_{1P} q_2 \tau^2}{2} \\
\frac{n_{12}}{N} &= \lambda_{1P}\lambda_{2P}\tau^2 + \frac{\lambda_{1P} q_2 + \lambda_{2P} q_1}{2}\tau^2
\end{aligned} \tag{14}$$

Further simplifying using A1 we have :

$$\lambda_{1P}\tau = \frac{n_1}{N} - q_1 \tau \frac{\lambda_{2P}\tau}{2} \tag{15a}$$

$$\lambda_{2P}\tau = \frac{n_2}{N} - q_2 \tau \frac{\lambda_{1P}\tau}{2} \tag{15b}$$

and plugging into Eq. (14)

$$\begin{aligned}
n_{12}/N &= (n_1/N - \frac{q_1 \tau n_2}{N})(n_2/N - \frac{q_2 \tau n_1}{2}) + \frac{(n_1/N - q_1 n_2 \tau/N) q_2 \tau}{2} + \frac{(n_2/N - q_2 n_1 \tau/N) q_1 \tau}{2}) \\
&= \frac{n_1 n_2}{N^2}(1 + q_1 q_2 \tau^2) + \frac{n_1 q_2 \tau}{2N}(1 - n_1/N) - \frac{n_2 q_1 \tau}{2N}(1 - \frac{n_2}{N}) - q_1 q_2 \tau^2 \frac{n_1 + n_2}{N} \qquad (16) \\
&\sim \frac{n_1 n_2}{N^2} + \frac{n_1 q_2 \tau}{2N} + \frac{n_2 q_1 \tau}{2N}
\end{aligned}$$

where we have arrived at the third expression by dropping terms of order $\frac{n_1 n_2}{N^2} q_i \tau$, $\frac{n_i}{N} q_1 q_2 \tau^2$ and higher. Rewriting the last step,

$$\frac{n_1 q_2 \tau + n_2 q_1 \tau}{N} = 2(\frac{n_{12}}{N} - \frac{n_1 n_2}{N^2}) \tag{17}$$

Eq. (17) relates the two unknowns q_1 and q_2. As we have no other constraint apriori, there is a family of solution to this equation. We have already examined one example above.

Thus, we need another constraint in order to determine q_i's. We posit two hueristic factors in this regard: one, apriori we would expect q_i's to be close to one another, and the second, maximization of their sum. If we only had the second criterion, then the q corresponding to be the disease with lower prevalence would be 0, and that of the higher prevalent disease very high. Imposing the first criterion alone assumes a symmetry between the diseases, and while that may be reasonable in the absence of any other information, we instead consider a trade-off between them. We propose the minimization of the "energy" function:

$$E = \alpha(q_1' - q_2')^2 - \beta(q_1' + q_2')^2 \tag{18}$$

where $q_i' = q_i\tau$, $\alpha, \beta > 0$; the first term favors the q_i''s being close together and the second maximizing the sum.

We look for solutions of Eqs. (17) that maximizes Eq. (18). This is done by using the Lagrange multiplier technique for finding the extrema given a constraint. $L(q_1', q_2') = E + \rho(q_1' n_2/N + q_2' n_1/N - (n_{12}/N - n_1 n_2/N^2))$

$$\frac{\partial L}{\partial q_1'} = 2\alpha(q_1' - q_2') - 2\beta(q_1' + q_2') + \rho n_2/N$$
$$\frac{\partial L}{\partial q_2'} = 2\alpha(q_2' - q_1') - 2\beta(q_1' + q_2') + \rho n_1/N$$

We can now solve the above constraint equations together with Eq. 17 simultaneously for the three unknows q_1', q_2' and ρ. We give the final expressions for q_i',i=1,2:

$$q_1' = 4\left(\frac{n_1+n_2}{N\beta} + \frac{n_1-n_2}{N\alpha}\right)\frac{\frac{n_{12}}{N} - \frac{n_1 n_2}{N^2}}{\left(\frac{n_1+n_2}{N}\right)^2\frac{1}{\beta} - \left(\frac{n_1-n_2}{N}\right)^2\frac{1}{\alpha}}$$
$$q_2' = 4\left(\frac{n_1+n_2}{N\beta} - \frac{n_1-n_2}{N\alpha}\right)\frac{\frac{n_{12}}{N} - \frac{n_1 n_2}{N^2}}{\left(\frac{n_1+n_2}{N}\right)^2\frac{1}{\beta} - \left(\frac{n_1-n_2}{N}\right)^2\frac{1}{\alpha}}$$

And the sum gives us the desired measure in terms of prevalence and co-occurrence

$$q_{tot}' = q_1' + q_2' = 8\left(\frac{n_1+n_2}{N\beta}\right)\frac{\frac{n_{12}}{N} - \frac{n_1 n_2}{N^2}}{\left(\frac{n_1+n_2}{N}\right)^2\frac{1}{\beta} - \left(\frac{n_1-n_2}{N}\right)^2\frac{1}{\alpha}}.$$

This represents a family of measures parametrized by $\alpha, \beta > 0$. Defining $\gamma = \beta/\alpha$ and skipping the constant multiplicative factor of 8:

$$\phi_\gamma = \left(\frac{n_1+n_2}{N}\right)\frac{\frac{n_{12}}{N} - \frac{n_1 n_2}{N^2}}{\left(\frac{n_1+n_2}{N}\right)^2 - \left(\frac{n_1-n_2}{N}\right)^2\gamma}. \tag{19}$$

For the symmetric case $n_1 = n_2 = n$, we have:

$$q_{tot}' = 0.5\frac{\frac{n_{12}}{N} - \frac{n^2}{N^2}}{\frac{n}{N}}$$

which, except for the factor of 0.5, is very close to what we would get with the original ϕ and modified ϕ^M correlation, Eq. (4), and exact in the limit of vanishing n/N. Although this result is independent of constants α, β, to get reasonable answers for arbitrary ratios n_2/n_1, we require $\alpha/\beta \sim 1$ (see Appendix).

6 Discussion

Our proposed framework to characterize the measure properties across potential bias variables has three key features :(a) bias is characterized in terms of its impact on selection based on effect-size (b) procedure for determining bias is independent of the specific measure and hence comparison between measures is carried out on a neutral platform (c) three indices are devised to capture different aspects of the bias. Although this framework rests on a specific disease cohort data-set, any similarly large cohort would be equally suitable for this task.

The importance of understanding the properties of measures cannot be overstated. Even for randomized controlled trials, conclusions depend on the measure used to characterize the effect of interventions [27]. Although not widely recognized, it is known that most standard measures of association have significant limitations and can give rise to misleading results unless they are interpreted carefully [28]. Despite use of similar measures, the results obtained from studies using different design of experiments cannot be directly compared or combined [29]. While all these issues are certainly very relevant, we should clarify that the measure bias that is highlighted in our work here has different origins, and to the best of our knowledge, there has been no earlier studies that have addressed them.

The requirement that the measure distribution be independent of prevalence ratio could be questioned in certain limits. In the extreme case, one can argue that, for a pair of diseases with very different prevalence $n_A >> n_B$, an unbiased measure should never assign maximal association even when the co-occurrence is the highest possible, $n_{AB} = n_B$. The basis for this is the observation that perfect association implies the pair co-occur in every case (aside from errors from misclassification or finiteness of study time scale), and hence divergent prevalence would not arise in the first place. A reasonable counter-argument would point to the potential scenario of B being an invariable cause of A. Whatever the consensus may be on this issue, most realistic situations contain very few or no such cases, and in large cohorts the relations in the generic pair follow $n_{AB} << \min\{n_A, n_B\}$ where one would expect measure distribution to be independent of prevalence ratio.

The results reported here on the nature of bias of standard measures and the performance of the new measure (together with the selection of γ) are valid for quantifying associations between diseases in a cohort. In a different context, the choice of variables used to test potential bias itself may not be appropriate.

For example, if we consider a sample where each data points is binary vector of different aircraft operating (or not) at a particular airport, then it should not be surprising that aircraft operating from only a few airports (i.e.,low prevalence) are strongly correlated (because they are likely to be of larger capacity and connect major destinations only) while those with higher prevalence are not. Hence, the choice of partitioning the pairs in the sample should be informed by what we desire the distribution of associations to be independent of.

	A	A	
B	p	q	$n_B = p + q$
B	r	s	$n_{\bar{B}} = r + s$
	$n_A = p + r$	$n_{\bar{A}} = q + s$	$N = p + q + r + s$

Table 1: Contingency table for a pair of diseases in a cohort

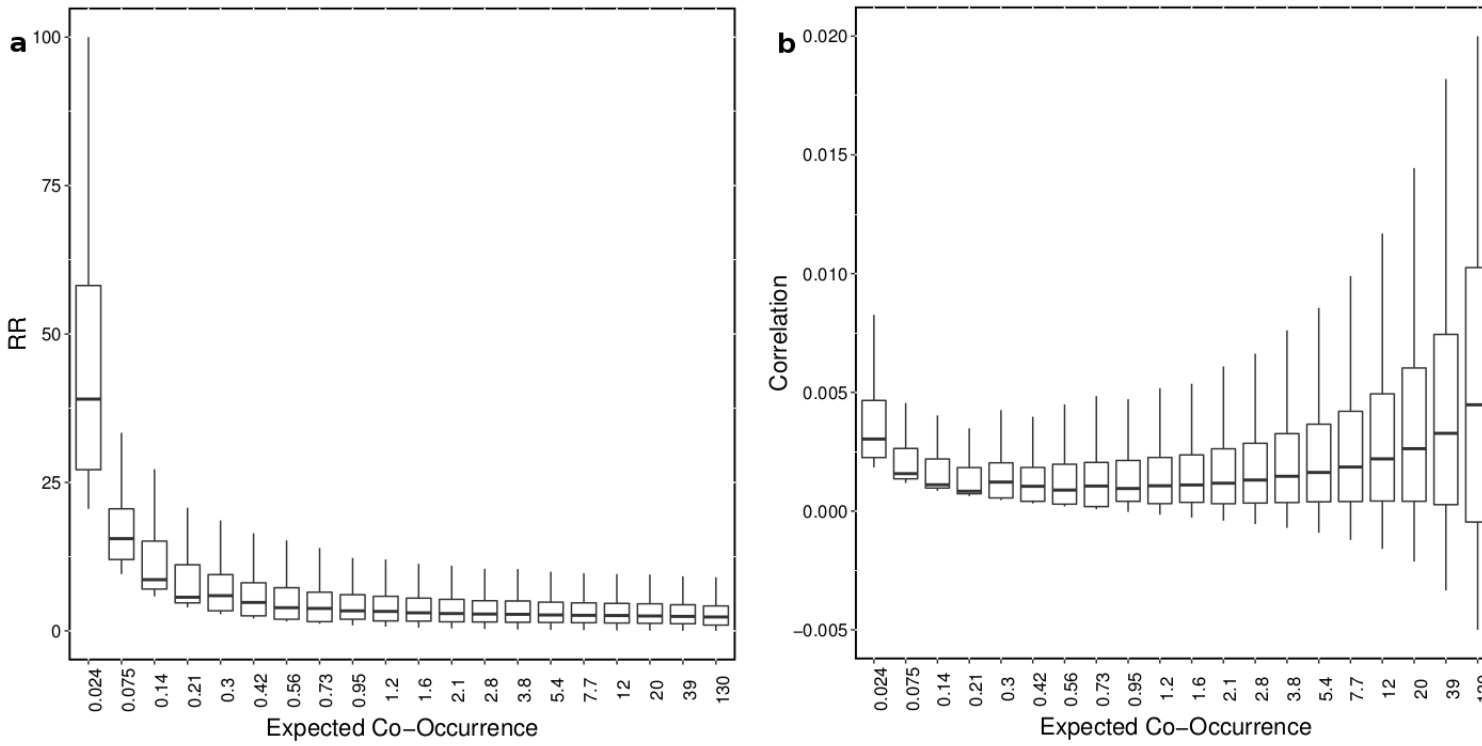

Figure 1: (a)Variation of RR as function of the expected co-occurrence shown as a box plot for all the pairs that fill within the particular window of expected co-occurrence. (b) The same analysis repeated for ϕ correlation.

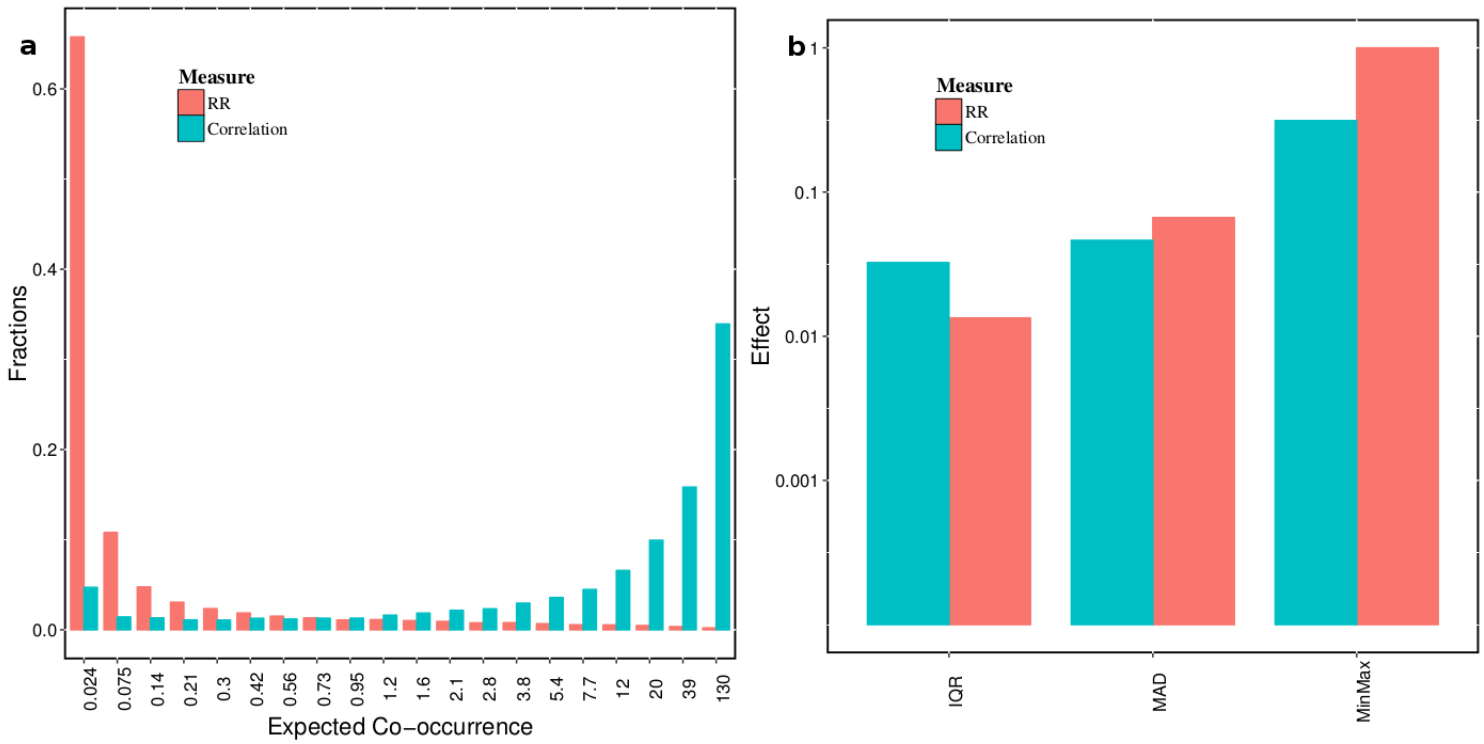

Figure 2: (a) Fraction of pairs above a certain fixed threshold (corresponding to 5% selection among all pairs) within each partition of the bias variable (expected co-occurrences) for RR and ϕ correlation measures. (b) Characterizing the variation of fractions across the different partitions using three indices. Here and in the following figures, MinMax has been diminished by a factor of 100.

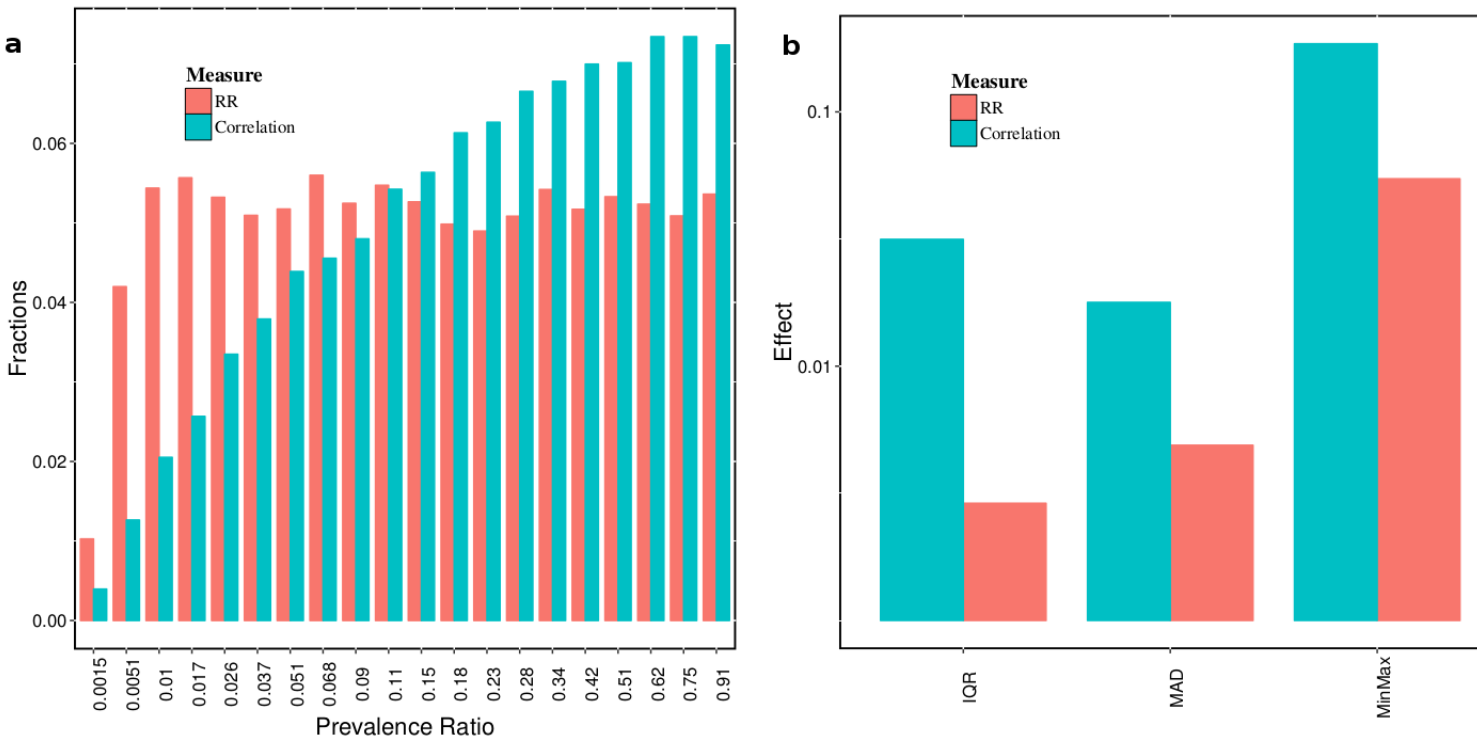

Figure 3: Fraction of pairs above a certain fixed threshold (corresponding to 5% selection among all pairs) within each partition of the bias variable (prevalence ratio, $\frac{n_<}{n_>}$, where $n_<(n_>)$ is the prevalence of the less (more) frequent disease) for RR and ϕ correlation measures. (b) Characterizing the variation of fractions across the different partitions using three indices. MinMax has been diminished by a factor of 100.

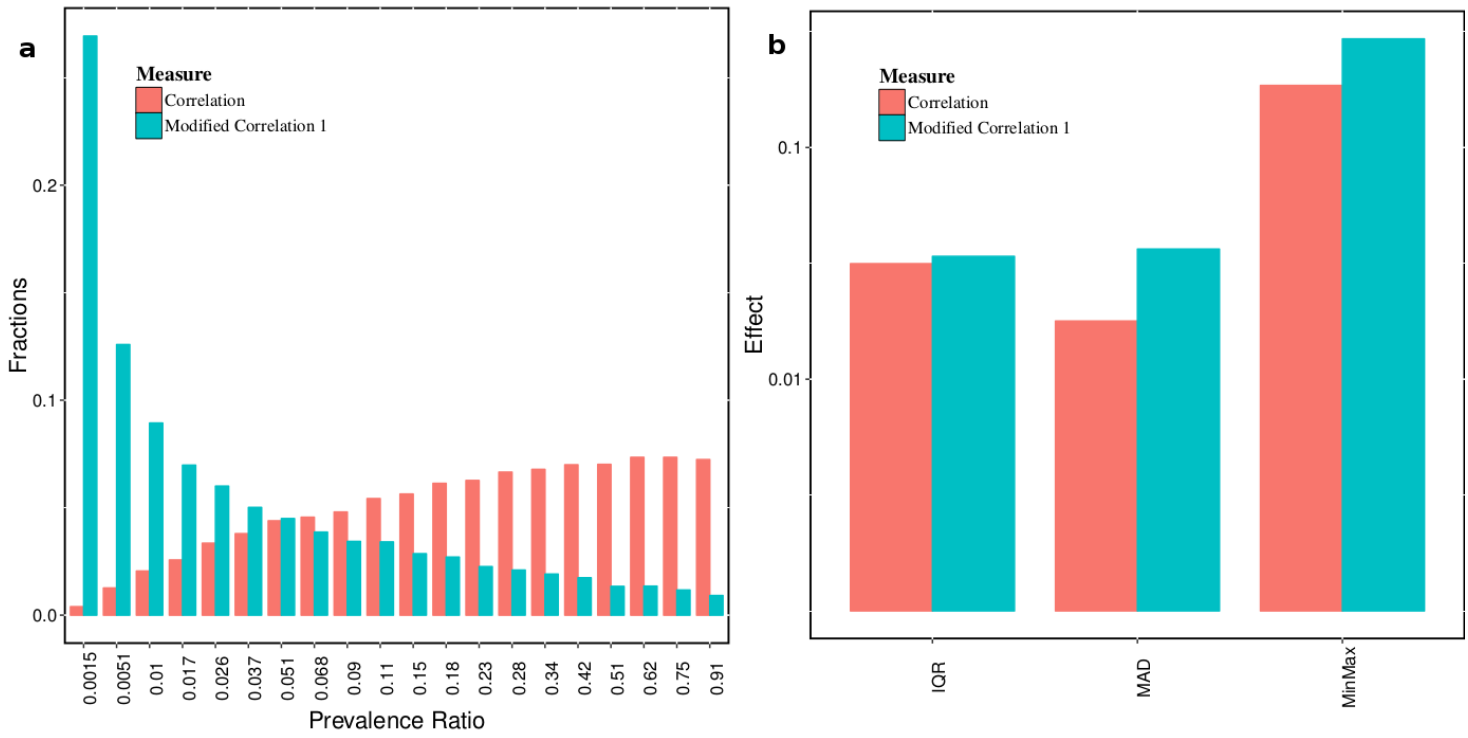

Figure 4: (a) Variation of fractions with prevalence ratios for the original correlation ϕ and its modification ϕ^M. We find that the ϕ^M has its own bias inflating the significant fractions for smaller prevalence ratios, which is the opposite of the original measure. (b) The variation of fractions characterized by the three indices. MinMax has been diminished by a factor of 100.

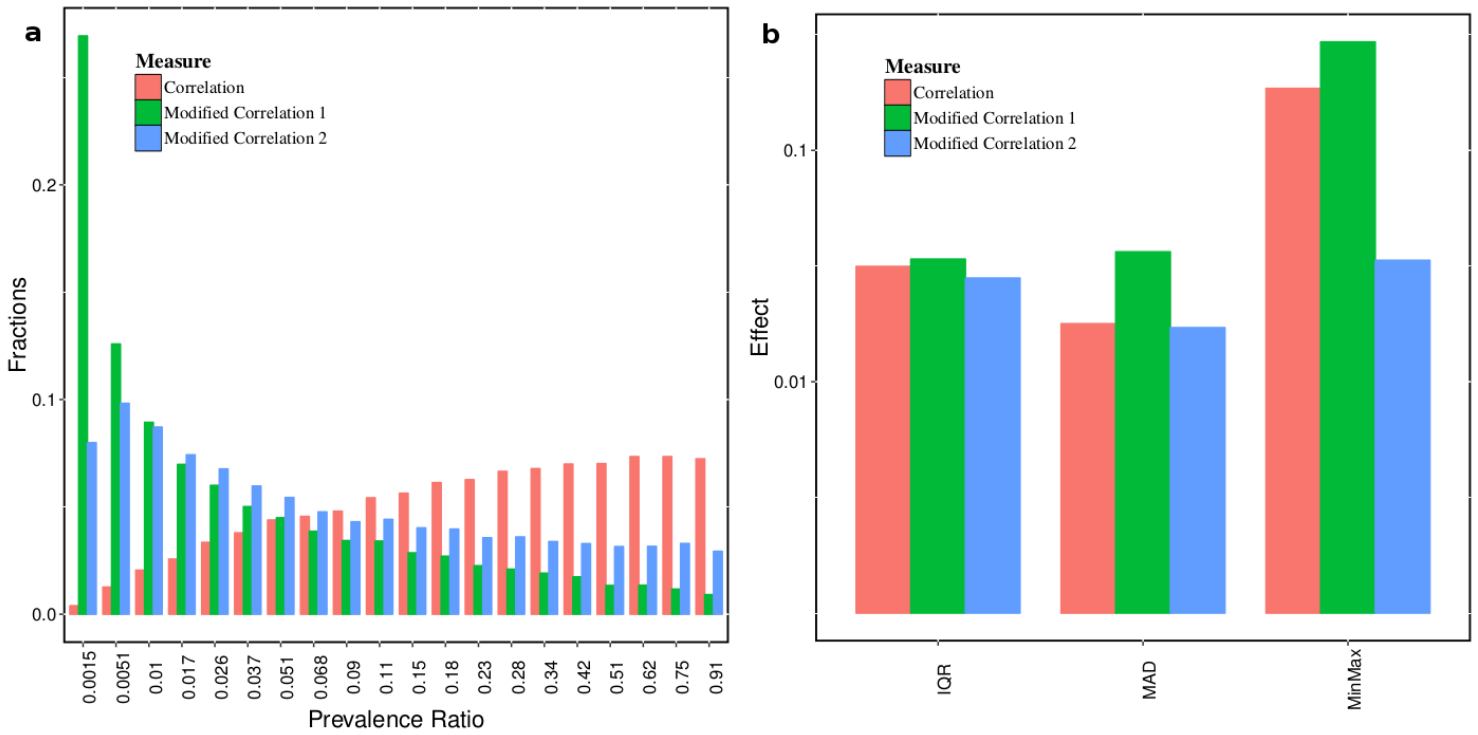

Figure 5: (a)Variation of fractions with prevalence ratios across ϕ, ϕ^M and ϕ^{M2}. (b) The same is characterized by the three indices. MinMax has been diminished by a factor of 100.

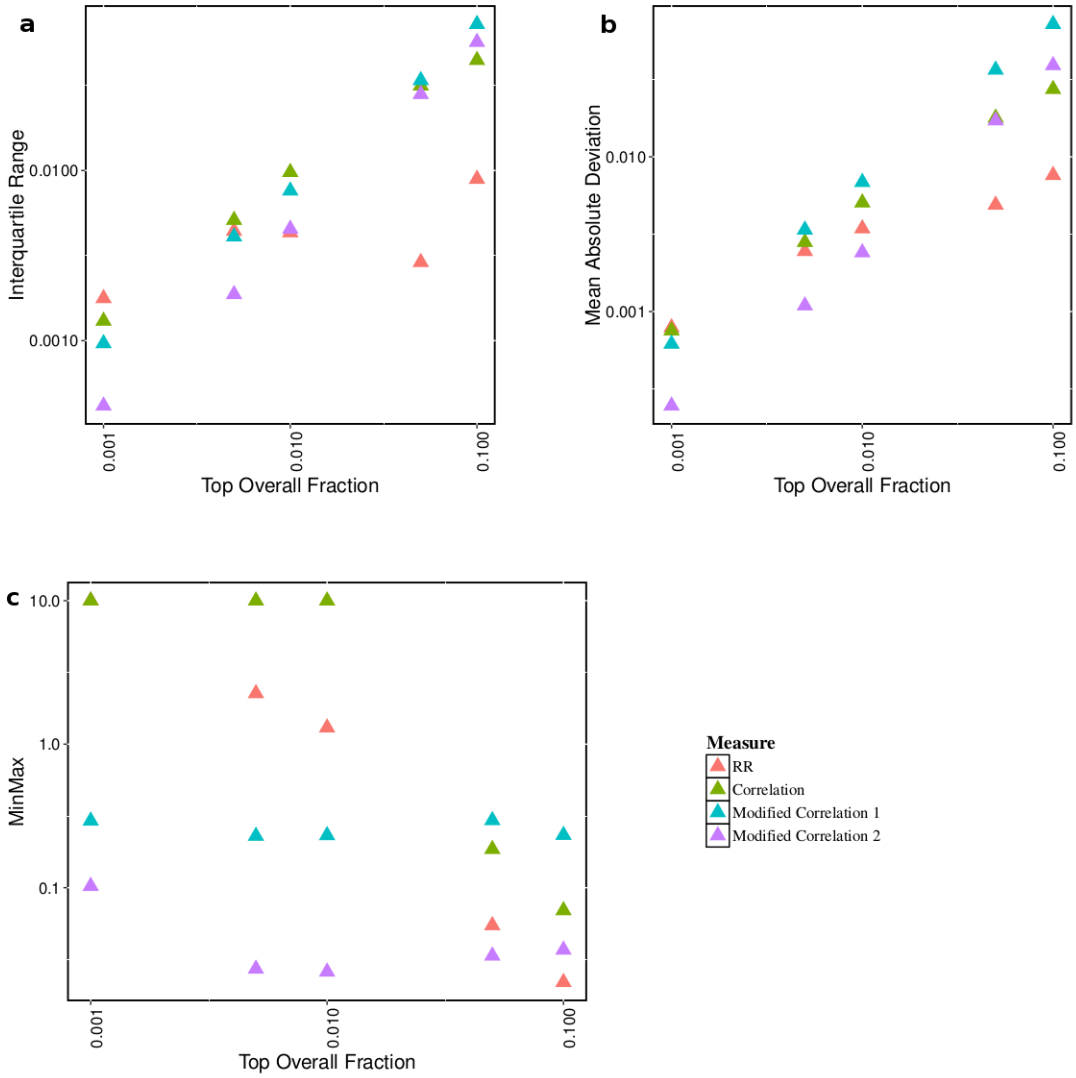

Figure 6: The three indices (a) Interquartile Range (b) Mean Absolute Deviation and (c)MinMax of the fractions obtained for the bias with respect to the prevalence ratio, for different overall (full data) fractions (x-axis) and for all the measures. MinMax has been diminished by a factor of 100.

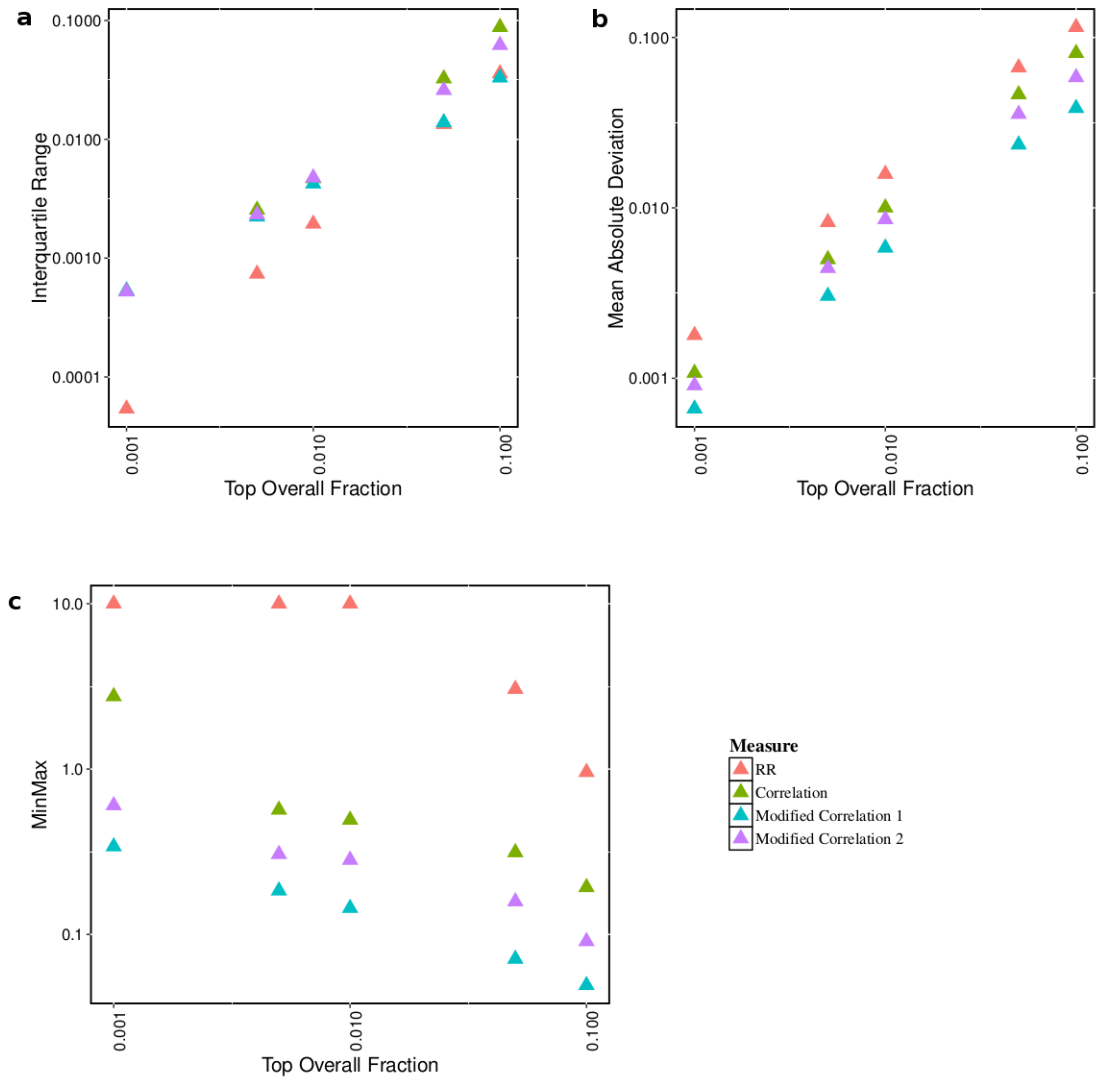

Figure 7: The three indices (a) Interquartile Range (b) Mean Absolute Deviation and (c)MinMax of the fractions obtained for the bias with respect to the expected co-occurrences, for different overall (full data) fractions (x-axis) and for all the measures. MinMax has been diminished by a factor of 100.

Appendices

A Effect of parameter γ on measure bias

We consider how the measure properties change as we modify the value of $\gamma = \beta/\alpha$ in Eq. 5.19 (in main text). As our primary focus is the bias of the measure with respect to the prevalence ratio, we consider how the the measure changes with γ for a given prevalence ratio. First, observe that when the prevalence of the two diseases are equal, the γ factor in the measure expression vanishes. Following the discussion in the last paragraph of the Methods section, in this limit of equal prevalence the new measure is equal the original ϕ and the modified ϕ^M upto a factor of n/N, which is assumed to negligible as in the rest of the analysis.

In the limit $\gamma = 1$, we have an intermediate measure,

$$\phi^I = \phi_{\gamma=1} = \left(\frac{n_1 + n_2}{4n_1n_2/N}\right)\left(\frac{n_{12}}{N} - \frac{n_1n_2}{N^2}\right)$$

and Fig. 8a shows how the fractions selected by this measure vary when partitioning the bias variable (prevalence ratio). Similar to ϕ^M, the fractions obtained with ϕ^I increases with decrease in prevalence ratio, although it is relatively more balanced and this fact is confirmed by the three indices in Fig. 8b.

The bias pattern suggests that we should examine the relative variation across the partitions of two measures from the new family corresponding to γ and γ'. To make this more precise, consider a hypothetical set of disease pairs $L_c = \{(n_{12} = c(k), n_1, n_2 = n_1k, N)|0 \leq k \leq 1\}$ where $n_1 < N$ is a constant and each element l_k of the set is identified by the prevalence ratio $n_2/n_1 = k$. The set is parametrized by a continuous function $c : [0, 1] \longrightarrow \mathbb{N}$ such that $c(k) < n_1/k$.

For a given c and k, the following quantity

$$\frac{\phi_\gamma(l_k)/\phi_\gamma(l_1)}{\phi_{\gamma'}(l_k)/\phi_{\gamma'}(l_1)} = \frac{1 - \gamma'\left(\frac{1-k}{1+k}\right)^2}{1 - \gamma\left(\frac{1-k}{1+k}\right)^2} \tag{20}$$

reflects how the ratio of the measure at prevalence ratio k and unity, depends on γ. The monotonic decrease of the ratio with γ together with the observation that $\phi^I = \phi_{\gamma=1}$ systematically increases with decrease in k implies that we should choose $\gamma < 1$ in order to stabilize the measure.

To probe the dependence of the bias on γ, it is important to determine how the ratio decreases with γ. For a given k, what should γ be for the ratio in Eq. (20) to decrease by a factor m compared to its value for $\gamma = 1$.

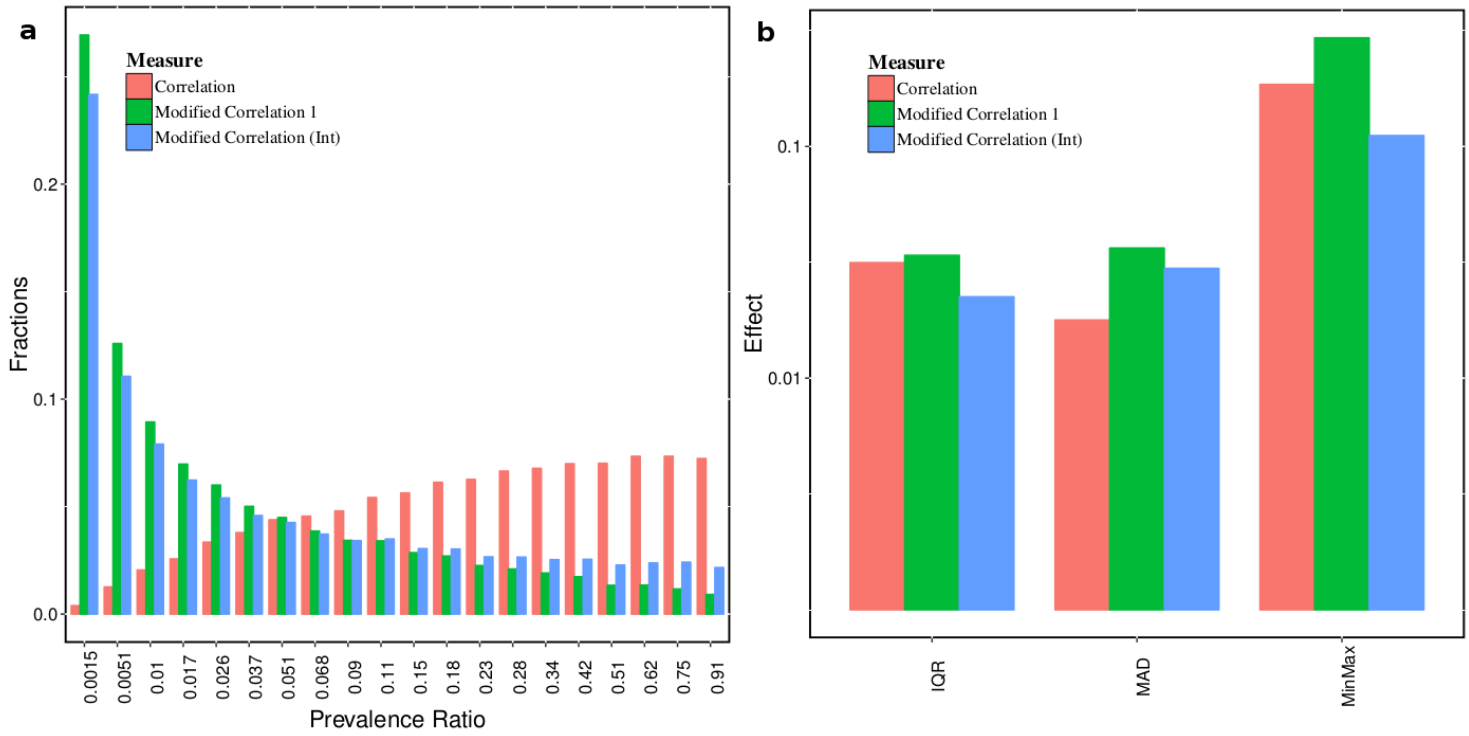

Figure 8: Fraction of pairs above a certain fixed threshold (corresponding to 5% selection among all pairs) within each partition of the bias variable (expected co-occurrences) for ϕ, ϕ^M and ϕ^I correlation measures. (b) Characterizing the variation of fractions across the different partitions using three indices. MinMax has been diminished by a factor of 100.

$$\frac{1-\left(\frac{1-k}{1+k}\right)^2}{1-\gamma\left(\frac{1-k}{1+k}\right)^2} = 1/m \tag{21}$$

$$\Longrightarrow$$

$$\frac{1-m\left(1-\left(\frac{1-k}{1+k}\right)^2\right)}{\left(\frac{1-k}{1+k}\right)^2} = \gamma \tag{22}$$

The plot of the dependence of γ on m and k is shown in Fig. 9. We find, as expected that, to attain the same reduction m, γ increases as we decrease the ratio $1/k$. This is because, for a fixed value of γ, m increases as we increase $1/k$.

More important, observe that to attain ratios of about $m = 2$, at $1/k = 300$, we need γ to deviate from unity only by a very small amount (0.015). The appropriate choice of γ is guided by the desired suppression of selected fractions at large $1/k$.

References

[1] N E Breslow and N E Day. Statistical methods in cancer research. Volume I - The analysis of case-control studies. *IARC Sci. Publ.*, (32):5–338, jan 1980.

[2] S Greenland and J M Robins. Conceptual problems in the definition and interpretation of attributable fractions. *Am. J. Epidemiol.*, 128(6):1185–97, dec 1988.

[3] César A Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.*, 5(4):e1000353, apr 2009.

[4] Anna Chmiel, Peter Klimek, and Stefan Thurner. Spreading of diseases through comorbidity networks across life and gender. *New J. Phys.*, 16(11):115013, nov 2014.

[5] Venkateshan Kannan, Fredrik Swartz, Narsis A. Kiani, Gilad Silberberg, and Giorgos Tsipras. Conditional Disease Development from Longitudinal Health Care Cohort Data using Layered Network Construction. *Sci. Rep.*, 6:1–27, may 2016.

[6] Shu Kay Ng, Libby Holden, and Jing Sun. Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics. *Stat. Med.*, 31(27):3393–405, nov 2012.

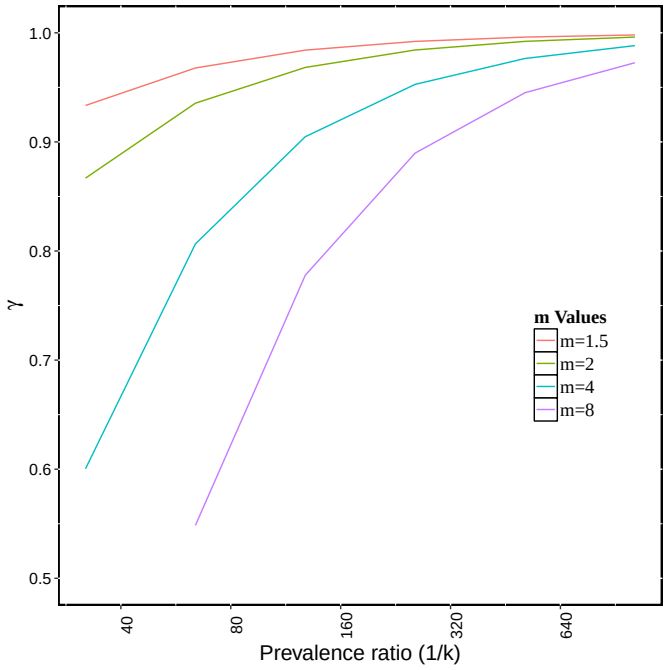

Figure 9: Dependence of γ on the ratio of the prevalences for a range of values of m.

[7] David A Grimes and Kenneth F Schulz. Bias and causal associations in observational research. *Lancet*, 359(9302):248–252, 2002.

[8] Jaideep Tan, Pang-Ning and Kumar, Vipin and Srivastava. Selecting the right interestingness measure for association patterns. In *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pages 32—-41, New York, New York, USA, 2002. ACM Press.

[9] Frederick Mosteller. Association and Estimation in Contingency Tables. *J. Am. Stat. Assoc.*, 63(321):1, mar 1968.

[10] Mark S Pearce, Jane A Salotti, Mark P Little, Kieran McHugh, Choonsik Lee, Kwang Pyo Kim, Nicola L Howe, Cecile M Ronckers, Preetha Rajaraman, Alan W Sir Craft, Louise Parker, and Amy Berrington de González. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *Lancet (London, England)*, 380(9840):499–505, aug 2012.

[11] Pamela H. Wolf, Jennifer H. Madans, Fanchon F. Finucane, Millicent Higgins, and Joel C. Kleinman. Reduction of cardiovascular diseaserelated mortality among postmenopausal women who use hormones: Evidence from a national cohort. *Am. J. Obstet. Gynecol.*, 164(2):489–494, feb 1991.

[12] Lena Kramer, Oliver Hirsch, Kathrin Schlössler, Susanne Träger, Erika Baum, and Norbert Donner-Banzhoff. Associations between demographic, disease related, and treatment pathway related variables and health related quality of life in primary care patients with coronary heart disease. *Health Qual. Life Outcomes*, 10:78, jan 2012.

[13] Qianghu Wang, Weisha Liu, Shangwei Ning, Jingrun Ye, Teng Huang, Yan Li, Peng Wang, Hongbo Shi, and Xia Li. Community of protein complexes impacts disease association. *Eur. J. Hum. Genet.*, 20(11):1162–7, nov 2012.

[14] Jun Zhang and Kai F. Yu. What's the Relative Risk? *JAMA*, 280(19):1690, nov 1998.

[15] Magdalena Szumilas. Explaining odds ratios. *J. Can. Acad. Child Adolesc. Psychiatry*, 19(3):227–9, aug 2010.

[16] G Udny Yule. Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134, 1903.

[17] Peter Cummings. The relative merits of risk ratios and odds ratios. *Arch. Pediatr. Adolesc. Med.*, 163(5):438–45, may 2009.

[18] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20(1):37–46, apr 1960.

[19] MG Kendall. A new measure of rank correlation. *Biometrika*, 30(1):81–93, jun 1938.

[20] Jake Olivier and Melanie L Bell. Effect sizes for 22 contingency tables. *PLoS One*, 8(3):e58777, jan 2013.

[21] Christopher J. Ferguson. An effect size primer: A guide for clinicians and researchers. *Prof. Psychol. Res. Pract.*, 40(5):532–538, 2009.

[22] Fiona Fidler, Geoff Cumming, Neil Thomason, Dominique Pannuzzo, Julian Smith, Penny Fyffe, Holly Edmonds, Claire Harrington, and Rachel Schmitt. Toward improved statistical reporting in the journal of consulting and clinical psychology. *J. Consult. Clin. Psychol.*, 73(1):136–43, feb 2005.

[23] Patricia Snyder and Stephen Lawson. Evaluating Results Using Corrected and Uncorrected Effect Size Estimates. *J. Exp. Educ.*, 61(4):334–349, 2014.

[24] Jacob Cohen. A power primer. *Psychol. Bull.*, 112(1):155–159, 1992.

[25] T Byrt, J Bishop, and J B Carlin. Bias, prevalence and kappa. *J. Clin. Epidemiol.*, 46(5):423–9, may 1993.

[26] E. C. Davenport and N. A. El-Sanhurry. Phi/Phimax: Review and Synthesis. *Educ. Psychol. Meas.*, 51(4):821–828, dec 1991.

[27] Edna Schechtman. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat - Which of these should we use? *Value Heal.*, 5(5):431–436, sep 2002.

[28] James A. Breaugh. Journal of Management. *J. Manage.*, 29(1):79–97, feb 2011.

[29] Scott B. Morris and Richard P. DeShon. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol. Methods*, 7(1):105–125, 2002.