International Journal of Approximate Reasoning 85 (2017) 159–177

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Dealing with under-reported variables: An information theoretic solution ☆

Konstantinos Sechidis [a], Matthew Sperrin [b], Emily S. Petherick [c], Mikel Luján [a], Gavin Brown [a,*]

[a] *School of Computer Science, University of Manchester, M13 9PL, UK*
[b] *Centre for Health Informatics, Institute of Population Health, University of Manchester, M13 9GB, UK*
[c] *School of Sport, Exercise & Health Sciences, Loughborough University, LE11 3TU, UK*

ARTICLE INFO

Article history:
Received 30 November 2016
Received in revised form 21 March 2017
Accepted 3 April 2017
Available online 7 April 2017

Keywords:
Under-reporting
Misclassification bias
Missing data
Mutual information

ABSTRACT

Under-reporting occurs in survey data when there is a reason for participants to give a false negative response to a question, e.g. maternal smoking in epidemiological studies. Failing to correct this misreporting introduces biases and it may lead to misinformed decision making. Our work provides methods of correcting for this bias, by reinterpreting it as a missing data problem, and particularly learning from positive and unlabelled data. Focusing on information theoretic approaches we have three key contributions: (1) we provide a method to perform valid independence tests with known power by incorporating prior knowledge over misreporting; (2) we derive corrections for point/interval estimates of the mutual information that capture both relevance and redundancy; and finally, (3) we derive different ways for ranking under-reported risk factors. Furthermore, we show how to use our results in real-world problems and machine learning tasks.

1. Introduction

Under-reporting usually occurs in survey data, and it refers to respondents that under-report the answer to a question, for example due to a perceived social stigma [35]. A famous example is maternal smoking during pregnancy, which is a key risk factor for adverse offspring outcomes including preterm birth and low birth weight (LBW). Like many health behaviours, accurate measurement of smoking habits can be difficult and expensive during pregnancy. For that reason, many studies use *self-reported* data, e.g. Wright et al. [37]. Given that most smokers know their habit to be harmful, both to themselves and their unborn child, there are strong motivations for women to under-report or deny their smoking status [10]. As such, the frequency of smokers in a sample is expected to be significantly lower than would be expected according to expert knowledge, derived for example from blood test result. Gorber et al. [15] presented a comprehensive analysis of the literature and compared the prevalence estimates of smoking based on self-reported data against the prevalence estimates based on directly measured smoking biomarkers. According to this analysis, self-reported smoking is generally under-reported in such a way that the true smoking figures may be underestimated by up to 47%.

☆ This paper is part of the Virtual special issue on the Eighth International Conference on Probabilistic Graphical Models, Edited by Giorgio Corani, Alessandro Antonucci, Cassio De Campos.

* Corresponding author.
E-mail address: Gavin.Brown@manchester.ac.uk (G. Brown).

http://dx.doi.org/10.1016/j.ijar.2017.04.002

Estimating the association between an *under-reported* (UR) variable and another will be biased in a manner that is specific to the degree/pattern of UR. Thus, any policy decisions made on the basis of such a biased result will be questionable. For example, government policies on tobacco control, e.g. [1], maybe ill-formed if they do not take into account UR. Maternal smoking and alcohol consumption are our focus for this paper, but there are many important health applications where *corrections* for UR are needed – e.g. HIV prevalence [38].

One method to correct UR bias is to spend time and resources to manually identify individuals that are likely to have misreported, and ignore/correct their testimony, i.e. identify smokers by performing cotinine blood tests. Unfortunately, in many applications it is impossible to completely correct the misreported cases. For example, Gorber et al. [15] present some possible flaws of the biochemical markers that identify smoking. As an alternative to this, authors in medical statistics treat it as a problem of *misclassification bias*, and combine data with a prior belief of the pattern of misclassification. They use this prior knowledge to derive corrected estimators for the log-odds ratio [7,12], and the relative-risk [27], or to suggest ways for performing tests of independence [5].

These solutions suffer from a number of weaknesses, which are addressed in this paper. For testing independence, they do not control both types of error (false positive/false negative). For estimating effect sizes, the suggested solutions are naturally only applied to estimate the correlation between a binary UR variable and a binary target variable – multi-class target variables with more than two categories are handled via a one-vs-one or one-vs-all strategy. Furthermore, ranking of variables in relation to a target – a common need in feature selection and other machine learning tasks – is not straightforward. Our strategy to overcome these limitations is to provide an information theoretic insight for the under-reporting problem, by disentangling three intimately related activities: *testing*, *estimation*, and *ranking* of features. Our main goal is to derive corrected estimators for the *mutual information* (MI), a measure of effect size widely used in machine learning applications with several interesting properties [4].

To achieve our goal we reinterpret the challenge not as dealing with misclassification and biased data, but as a problem of *learning from missing data*, and particularly learning from positive and unlabelled data [13]. By this interpretation, we present solutions using a graphical representation called *missingness* or *m*-graph [23], which is a tool to naturally incorporate a prior belief over the misreporting at the population (or appropriate sub-demographic) level. Furthermore, with our work, we show how to correct MI for under-reporting by examining independence properties observable via the *m*-graph representation.

In this paper, we present the following novel contributions[1] in relation to UR variables:

- Testing: **Section 3** suggests a way for testing independence between an UR feature and the class. Using our test and a derived correction factor, we can control both false positive/negative errors.
- Estimation: **Section 4** derives corrected estimators of MI terms that occur in UR scenarios. Section 4.1 presents estimators that capture the *relevance* between an UR variable and an arbitrary categorical variable, while Section 4.2 presents an estimator that captures the *redundancy* between two UR variables. Furthermore, we provide interval estimates where possible.

Section 5 presents two different methods for information theoretic feature ranking in the presence of UR variables, by using our suggested estimators. Section 5.1 presents Corrected-MIM, a univariate approach that provides rankings and captures only the relevance, while Section 5.2 presents Corrected-mRMR, a multivariate method that captures both the relevance and redundancy between the features. All the above contributions are novel, with the exceptions of Sections 4.1 and 5.1, which have been published in a conference paper [33]. Furthermore, we provide further experimental results in two applications. Firstly, **Section 6** derives rankings of risk factors related to low birth weight infants using a case study of 13,776 births in northern England, where we demonstrate some significant false conclusions that might be drawn when ranking variables without the correction factors. Finally, **Section 7** presents a machine learning application where we derive rankings of features when training/test distributions differ.

2. Background material

To the best of our knowledge, our work is the first that tackles the problem of estimating MI in under-reporting scenarios. In classic statistics there are some works that estimate other types of effect sizes (i.e. odds/risk ratios, limited to binary data) and we review them in Section 2.1. Section 2.2 shows how the under-reported can be phrased as a missing data problem. Finally, Sections 2.3, 2.4 and 2.5 give the background on testing, estimation and ranking using information theoretic measures.

2.1. Under-reporting as a misclassification bias problem

We assume that we have two random variables X and Y, representing a scenario where X is likely to be UR. In this case, we cannot observe the true value of X, but instead receive observations from a proxy variable $\widetilde{X}$. In the notation below we use lower case letters $(y, x, \widetilde{x})$ to denote a realisation from these variables. In our example of smoking during pregnancy,

[1] The software related to this paper will be available at github: https://github.com/sechidis.

$y \in \{0, 1\}$, is a binary indicator of LBW, $x \in \{0, 1\}$ is whether the mother smoked during pregnancy (1 for smoking and 0 for not smoking), and $\widetilde{x} \in \{0, 1\}$ is whether the mother *reported* that she smoked in pregnancy (1 for reported smoking and 0 for not smoking). While in our running example Y is binary, the techniques presented in this work are also applied to categorical data with more than two levels $|\mathcal{Y}| > 2$.

A classical solution to the under-reporting problem is to consider it as a *misclassification bias* [17]. Following Greenland [17] terminology, for an under-reported variable, the *specificity* is $p(\widetilde{x} = 0|x = 0) = 1$, while the *sensitivity* is $p(\widetilde{x} = 1|x = 1) < 1$. Here, the specificity is the probability that a non-smoker would tell the truth (equal to 1 in this setting) and the sensitivity is the probability that a smoker would tell the truth (in our setting strictly < 1, if it is equal to 1 the variable is not UR). As presented, this is the simplest scenario – referred to as *non-differential* – that is, the probabilities do not vary with respect to Y. The more complex case is when the sensitivity depends on Y, that is $p(\widetilde{x} = 1|x = 1, y)$, known as *differential misclassification* [17]. In this work, we will focus on the non-differential UR scenario, and leave the differential as a future work, outlined in Section 8. The non-differential assumption is reasonable to cohort studies (such as the Born in Bradford project [37] presented in Section 6), since as Greenland [16] states: "... studies that collect exposure data before the outcome occurs (such as most cohort studies) provide settings for reasonable employment of the non-differential assumption."

Estimating the strength of association between variables, using this misclassification approach, is a well explored challenge in epidemiology. For example, Chu et al. [7] derive corrected estimators for the log-odds ratio, while Rahardja and Young [27] did it for the relative-risk. To derive these corrections, knowledge of the specificities/sensitivities, or in other words knowledge of the misclassification rates, is needed. This can be derived in different ways, such as *validation studies* or domain *prior knowledge*. A different way of estimating these effect sizes is to use a model to impute the values of the possibly misclassified examples, for example Edwards et al. [12] present a way of using multiple imputations to estimate log-odds ratios. With our work we derive corrections for the MI, by incorporating simple forms of prior knowledge.

A further challenge other than estimation is to conduct a valid independence test. Testing independence under misclassification is a very old problem: Mote and Anderson [24] showed that the usual X^2-test of independence is valid when the misclassification is non-differential, but statistical power (i.e. true positive rate) is reduced. We found only one study [5], which suggested a correction factor that captures the amount of power loss in the X^2-test. In that work a binomial difference-of-proportions test is analysed and used to suggest properties of the X^2-test, using as an argument the equivalence of these two tests under the null hypothesis. With our work we derive a correction factor, which estimates the power loss more accurately than the correction factor suggested by Bross [5] for the same quality of prior knowledge.

2.2. Under-reporting as a missing data problem

A different way to phrase the under-reporting problem is by connecting it with the equivalent problem from the missing data literature. The first step is to consider the under-reporting bias as a *positive and unlabelled* (PU) problem [13]. That is, a semi-supervised binary classification problem where we have a set of positive examples and a separate set of unlabelled examples, which can be either positive or negative. The positive examples can be seen as the reported "smoking" cases ($\widetilde{x} = 1$), while the unlabelled can be seen as the reported "non-smoking" cases ($\widetilde{x} = 0$).

Furthermore, from the missing data literature we borrow a graphical representation which will help us to make apparent the assumptions behind the under-reporting mechanism. Mohan et al. [23] introduced a formalism for graphical modelling in the presence of missing data, known as *missingness graphs* or *m*-graphs. While in the literature of misclassification bias there is a different graphical representation [17], our modification of the *m*-graphs provides more useful information, by capturing both the data generation model and the causal mechanisms responsible for the misclassification process.

Fig. 1(a) shows the simplest case of *non-differential* UR.[2] A solid node indicates a fully observed variable, whilst dashed nodes represent unobserved variables. Associated with every unobservable variable X there are two additional nodes: firstly M_X, which controls whether a value from X is correctly reported ($m_x = 1$) or not ($m_x = 0$), and secondly, the proxy variable $\widetilde{X}$ which is fully observed. The major difference between missingness graphs used by Mohan et al. [23] and those here is that the mechanism M_X is not observed, and for that reason we must incorporate prior knowledge over the sensitivity $p(m_x = 1|x = 1)$ and specificity $p(m_x = 1|x = 0)$. The *m*-graph representation allows us to read off independence properties such as: $Y \perp\!\!\!\perp M_X|x = 1$ – which corresponds to the *selected completely-at-random* assumption in the positive and unlabelled literature [13]. Fig. 1(b) shows a more complex situation where we have two UR variables.

The current paper shows (in its simplest case) how to recover the value $I(X; Y)$ from $I(\widetilde{X}; Y)$ by deriving a correction based on prior belief over the mechanism M_X. And how to test the independence between X and Y, by using the observed variables $\widetilde{X}$ and Y.

2.3. Background on testing independence

The most usual way to decide independence between categorical variables is through the X^2-test, calculated from sample data, $\{(x^i, y^i)\}_{i=1}^N$. We note that this is related to the *squared-loss mutual information* (SMI) [31,34] as so:

[2] If we want to give a similar graphical representation for the differential under-reporting, a further arc should connect the variable Y with the variable that controls the under-reporting mechanism M_X.

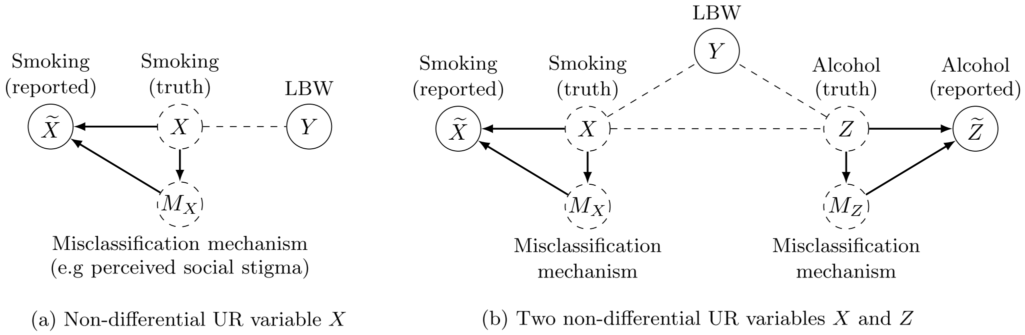

Smoking (reported)
Smoking (truth)
LBW
$\widetilde{X}$
X
Y
M_X
Misclassification mechanism (e.g perceived social stigma)
(a) Non-differential UR variable X
LBW
Y
Smoking (reported)
Smoking (truth)
Alcohol (truth)
Alcohol (reported)
$\widetilde{X}$
X
Z
$\widetilde{Z}$
M_X
M_Z
Misclassification mechanism
Misclassification mechanism
(b) Two non-differential UR variables X and Z

Fig. 1. A graphical representation for under-reporting. (a) Non-differential, where low birth weight (LBW) Y is assumed to be associated with smoking X, so we want to know the strength of association $I(X;Y)$ on this arc. However, X is *under-reported*, so the true value is unobservable, and instead we have a proxy $\widetilde{X}$, determined by X and the misclassification mechanism M_X. (b) Two non-differential correlated under-reported variables X and W. In this case we want to know the strength of $I(X;Y)$ and $I(Z;Y)$, but also we are interested in the strength of $I(X;Z)$. The dashed lines indicate that there may or may not be a correlation between the variables.

$$\chi^2 = N \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{\left(\hat{p}(x,y) - \hat{p}(x)\hat{p}(y)\right)^2}{\hat{p}(x)\hat{p}(y)} = 2N\widehat{I}_2(X;Y),$$

where $\widehat{I}_2(X;Y)$ is the maximum likelihood estimate of the squared-loss mutual information. Under the null hypothesis that X and Y are statistically independent, the X^2-statistic is asymptotically χ^2-distributed, with $\nu = (|\mathcal{X}|-1)(|\mathcal{Y}|-1)$ degrees of freedom [2].

To decide independence between X and Y, for a given sample of data, we calculate the statistic and check whether it exceeds the critical value defined by $F^{-1}(1-\alpha)$, where α is the user-specified significance level of the test and F^{-1} is the inverse cumulative distribution function of the χ^2-distribution with ν degrees of freedom. If the critical value is not exceeded, we *fail* to reject the null hypothesis of independence. The user specified significance level defines the probability of *type I error* (α), which is the probability that the test will falsely reject the null hypothesis. To calculate the probability that the test will falsely reject the null hypothesis, or probability of *type II error* ($\beta = 1 - \text{power}$), we should perform a power analysis [8]. To do this we need to know the sampling distribution of the test statistic under the alternative hypothesis. The X^2-statistic asymptotically follows a non-central χ^2 distribution under the alternative hypothesis, with the same degrees of freedom ν and with non-centrality parameter $\lambda_{\chi^2(X;Y)} = 2NI_2(X;Y)$ [2, Section 6.6.4]. Thus, the probability of a type II error depends also on the sample size N and on the population value of the SMI $I_2(X;Y)$.

Given this context, a very important tool of a priori power analysis is *sample size determination* [8]. In this prospective procedure we specify the significance level of the test (e.g. $\alpha = 0.05$), the desired power (e.g. power $= 0.99$ or the probability of false negative to be 0.01) and the population value of the desired effect size described in terms of $I_2(X;Y)$ – from this we determine the minimum number of examples required to detect that effect with the given probabilities of error. Section 3 presents how we can use the above methodology when we have UR features.

2.4. Background on estimating mutual information

In practical applications we want to explore relationships between random variables. Just giving a yes/no answer through a hypothesis test may not be of much interest, and estimating the size of the effect gives more useful information, for example, how strongly smoking is correlated with low birth weight. In machine learning one of the main ways of measuring the strength of this association is by estimating Shannon's mutual information (MI) [4]. The maximum likelihood (ML) estimate of the MI is:

$$\widehat{I}(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \widehat{p}(x,y) \ln \frac{\widehat{p}(x,y)}{\widehat{p}(x)\widehat{p}(y)}. \tag{1}$$

Firstly, it is a non-negative quantity which takes its minimum zero value when the random variables are independent. Furthermore, it can be associated with both upper and lower bounds on the Bayes error [14,19]. Brown et al. [6] present an extensive discussion of this in the context of feature selection, including various heuristics which provide approximations for high dimensional data, resulting in a unifying theoretical framework derived from a simple probabilistic model.

Together with point estimates, it is a good practice to give an *interval* estimate, a range of possible values that the mutual information can take. Asymptotic distribution theory has a set of tools to derive the sample distribution of the ML-MI estimator and the following theorem presents this known result [4].

Theorem 1 *(ML-MI estimator, asymptotic distribution). For the estimator $\widehat{I}(X;Y)$ it holds that: $\sqrt{n}\left(\widehat{I}(X;Y) - I(X;Y)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{MI}\right)$, where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. The standard error of the estimator is:*

$$SE\left[\widehat{I}(X;Y)\right]=\frac{\sigma_{MI}}{\sqrt{n}}=\frac{1}{\sqrt{n}}\left(\sum_{x\in\mathcal{X},y\in\mathcal{Y}}p(x,y)\left(\ln\frac{p(x,y)}{p(x)p(y)}\right)^2-I(X;Y)^2\right)^{\frac{1}{2}}. \tag{2}$$

Proof sketch: This result can be proved by using delta methods [2].

While the asymptotic variance here depends on the population values $p(x,y)$, in practice, for interval estimation we replace them by their sample values $\widehat{p}(x,y)$. This standard procedure [2, Section 3.1.7] is followed for all the sampling distributions that we present in this work. Section 4 presents how we can estimate mutual information in UR scenarios.

2.5. Background on information theoretic feature ranking and selection

In most real world problems we have more than one feature, i.e. we observe a sample dataset $\{\mathbf{x}^i, y^i\}_{i=1}^N$, where the feature vector $\mathbf{x}=[x_1...x_d]$ is a d-dimensional realisation of the joint random variable $\mathbf{X}=X_1...X_d$. With a slight abuse of notation, in the rest of this section, we interchange the symbol for a set of variables and for their joint random variable.

In this scenario, it is also useful to order the features according to their relationship with the target variable, a procedure known as *feature ranking*. Feature rankings provide very useful information, and applications of this principle range from model selection to decision tree construction. There are two main categories of feature rankings [29]: univariate methods, which consider only the individual relevance of each feature and multivariate methods, which also take into account dependencies between features.

In information theoretic feature selection, firstly we rank the features according to a score measure and then select the ones that contain most of the useful information (i.e. higher score). By ranking the features with respect to their mutual information with the target variable, we derive a ranking that takes into account only the *relevance* with the target. Choosing the features according to this ranking corresponds to the univariate *Mutual Information Maximization* (MIM) criterion [21]; where the score of each feature X_k is given by:

$$J_{MIM}(X_k)=\widehat{I}(X_k;Y).$$

This approach does not take into account the *redundancy* between the features. More advanced multivariate techniques take into account both relevance and redundancy, *without* having to compute very high dimensional distributions. For example, a popular multivariate criterion is the minimal Redundancy Maximal Relevance (mRMR), which ranks the features according to the score [25]:

$$J_{mRMR}(X_k)=\widehat{I}(X_k;Y)-\frac{1}{|\mathbf{X}_\theta|}\sum_{X_j\in\mathbf{X}_\theta}\widehat{I}(X_k;X_j),$$

where $\mathbf{X}_\theta$ is the set of the features already selected. The first term of the *RHS* captures the relevance and the second the redundancy. Section 5 derives extensions of MIM and mRMR that handle UR features.

3. Testing independence in under-reported scenarios

To answer the question of whether two variables are independent or not we need a hypothesis testing procedure, where we can have a control over the two probabilities of error: false positive (Type I error) and false negative (Type II error). Testing independence in under-reported scenarios is not straightforward and this section explores the dynamics of this test.

In non-differential under-reporting, it is *valid* to test independence by using the under-reported variable $\widetilde{X}$ [24,16]. This can be easily proved since we have: $X \perp\!\!\!\perp Y \Leftrightarrow \widetilde{X} \perp\!\!\!\perp Y$. This can be also read directly from the m-graph in Fig. 1(a): when there is no direct arc between X and Y there is no path that connects $\widetilde{X}$ with Y and vice versa. Quantifying the loss of power in the χ^2-test is more challenging. Now we will show how to derive a correction factor that captures this loss effectively.

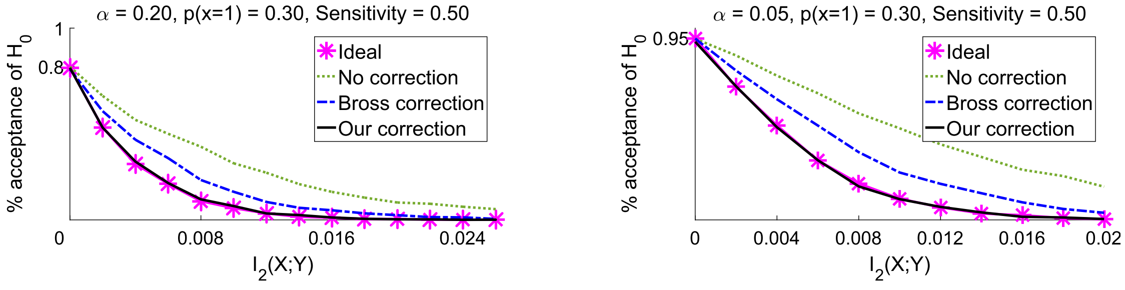
We first need to write the non-centrality parameter of the under-reported test $\chi^2(\widetilde{X};Y)$ in terms of the same parameter of the unobservable test $\chi^2(X;Y)$.

Theorem 2 *(UR test of independence). In the non-differential under-reported scenario the non-centrality parameter of the ideal test,* $\lambda_{\chi^2(\widetilde{X};Y)}=2NI_2(\widetilde{X};Y)$*, and the non-centrality parameter of the unobserved test,* $\lambda_{\chi^2(X;Y)}=2NI_2(X;Y)$*:* $\lambda_{\chi^2(\widetilde{X};Y)}=\kappa\lambda_{\chi^2(X;Y)}$*, have the following relationship:*

$$\kappa=\frac{1-p(x=1)}{p(x=1)}\frac{p(\widetilde{x}=1)}{1-p(\widetilde{x}=1)}. \tag{3}$$

Proof can be found in Appendix A.1.

The proof builds upon the fact that the following relationship holds between the population values of the MI terms: $I_2(\widetilde{X};Y)=\kappa I_2(X;Y)$. Since $\kappa<1$, the under-reported test will be always less powerful than the unobserved correctly



(a) $\alpha = 0.20$, $|\mathcal{Y}| = 2$ and $N = 400$

(b) $\alpha = 0.05$, $|\mathcal{Y}| = 2$ and $N = 500$

Fig. 2. Comparing the under-reported tests of independence and verify experimentally correctness of our correction factor, where *Ideal* means the unobservable test $\chi^2(X;Y)$ with N examples, *No correction* means the under-reported test $\chi^2(\widetilde{X};Y)$ with N examples, *Bross correction* means the under-reported test $\chi^2(\widetilde{X};Y)$ with the corrected effective sample size N/κ', while *Our correction* means the under-reported test $\chi^2(\widetilde{X};Y)$ with the corrected effective sample size N/κ.

reported test. Furthermore, by having a correction factor, we can quantify the amount of the power loss by incorporating knowledge over $p(x=1)$.

As a result, a χ^2-test between X and Y with N examples will have the same power as a χ^2-test between the under-reported $\widetilde{X}$ and Y with N/κ examples, referred to as *effective sample size*. Bross [5] suggested a different approach to derive correction factor in non-differential misclassification, using as a starting point the equivalence, under the null hypothesis, between χ^2-test and the difference-of-proportions-test. This correction factor under our notation in the UR scenario is written as (the analytical derivation can be found in Appendix A.2):

$$\kappa' = p(\widetilde{x}=1)(1-p(\widetilde{x}=1))/\left(p(x=1)(1-p(x=1))\right). \tag{4}$$

By comparing the equations for the two correction factors, (3) and (4), we can derive the following proposition.

Proposition 1 *(Comparing effective sample sizes). Deriving effective sample size by using correction factor κ results to more powerful test than using κ'.*

The proof of this proposition is straightforward. In the UR scenarios it holds $p(x=1) > p(\widetilde{x}=1) \Leftrightarrow \kappa < \kappa'$. Thus the relationship between the effect sample sizes is: $N/\kappa > N/\kappa'$, which means that using our correction factor κ results to higher sample size, and as a result more powerful test.

To verify experimentally our theoretical results we generate synthetic random variables X and Y with different degrees of dependency and we explore the false positive/negative rates through the graphs presented in Fig. 2. In the x-axis we have different effect sizes in terms of the squared loss mutual information between X and Y, while in the y-axis we have the acceptance rate of the null hypothesis H_0 (over 2,000 repeats). The y-intercept represents *1 – False Positive Rate*, and should be close to $1-\alpha$ in order for the tests to be valid, while elsewhere the plots indicate the *False Negative Rate*.

Fig. 2 verifies experimentally the correctness of our correction factor κ (verification of Theorem 2) and it shows the superiority against κ' (verification of Proposition 1). Having a known correction factor is very useful for power analysis activities, such as sample size determination [30]. Furthermore, it shows that testing using $\widetilde{X}$ instead of X is a valid approach, since all lines have the same intercept at $1-\alpha$, and thus the tests have the same false positive rate.

Section 6 shows how our results on UR test of independence can be useful in analysing a clinical dataset.

4. Estimating mutual information in under-reported scenarios

In many applications, just giving an informed answer through a hypothesis test may not be of much interest, while estimating the size of the effect gives more useful information. In this section we will present different ways to estimate mutual information, despite under-reporting.

The *ideal* method to completely correct UR is to spend resources to identify the individuals that have misreported, and correct their testimony. For example, it could be done by performing cotinine blood tests to all women that reported non-smoking ($\widetilde{x}=0$). This approach is expensive, and still it may not be possible to identify the individuals that have misreported [15]. On the other hand, the simplest way to estimate mutual information in under-reported scenarios is to follow a *naive* approach and just use the observed data. Unfortunately, this estimator, $\widehat{I}(\widetilde{X};Y)$, is asymptotically biased for estimating $I(X;Y)$. This can be easily proved, since under the model of Fig. 1(a) the following strict inequality holds[3]: $I(\widetilde{X};Y) < I(X;Y)$.

[3] We can prove this result by using Jensen's inequality, and the fact that in non-differential under-reporting the following strict inequality holds: $p(\widetilde{x}=1) < p(x=1)$.

Another way to estimate mutual information is by trying to "predict" the real values of the misclassified examples using some prediction model. Then, impute new values for these examples, and finally, estimate MI using the imputed data. This is similar to solving the missing data problem by *imputation* [3]. In our running example this means imputing the actual values of the women who reported not smoking ($\widetilde{x}=0$). To do so we need to build a model to derive the Bayesian posterior distribution[4] $p(x=1|y,\widetilde{x}=0)$, and we use this model to impute the values for the examples with $\widetilde{x}=0$. Then, we can use these imputed values to derive point and interval estimates of the MI using the expressions presented in Section 2.4. One limitation of single-imputation is that estimating standard error using conventional methods – such as eq. (2) – does not take into account the fact that some of the data were imputed [28]. One solution to this problem is to perform multiple-imputations and use improved ways of estimating the standard errors, such as Rubin's rule presented in [3, Chapter 5]. Multiple-imputation has some limitations; for example, it is computationally expensive, while, in the case of estimating[5] MI, there are no guarantees that the confidence intervals derived by Rubin's rule will have the coverage defined by the nominal (user specified) level. For more details on the strengths and weaknesses of multiple-imputation we refer to Rubin [28].

In the next section we present a corrected estimator for the mutual information that takes into account the under-reporting and overcomes the above limitations: (1) it is consistent, unlike the naive approach, (2) it produces valid interval estimates, unlike the simple-imputation, and (3) it is computational-efficient, unlike multiple-imputations.

4.1. Correcting for under-reporting the mutual information that captures relevance

To estimate mutual information in the under-reported scenario, we need to come up with a way to estimate marginal and joint/conditional probabilities, despite the restrictions of the problem. While we can estimate the marginal $p(y)$ from all data, the conditionals are more challenging. For example, the conditional $p(y|x=1)$ is inaccessible, as we do not have access to the full set of the examples with $x=1$, i.e. we do not know the identities of all smokers, but only those that self-reported it ($\widetilde{x}=1$). Because of the event based independence assumption $Y \perp\!\!\!\perp M_X|x=1$ it holds that $p(y|x=1)=p(y|x=1,m_x=1) \Leftrightarrow$

$$p(y|x=1)=p(y|\widetilde{x}=1). \tag{5}$$

To find the other conditional $p(y|x=0)$ we use a simple trick first introduced by Denis et al. [9] in the context of positive and unlabelled data. By using (5) we can write the marginal as $p(y)=p(y|\widetilde{x}=1)p(x=1)+p(y|x=0)p(x=0)$ and solving for $p(y|x=0)$:

$$p(y|x=0)=\frac{p(y)-p(y|\widetilde{x}=1)p(x=1)}{1-p(x=1)}. \tag{6}$$

Finally, since we do not have access to the marginal distribution $p(x=1)$, and since it cannot be estimated without modelling assumptions, we incorporate prior knowledge[6] as a parameter γ_x, provided by a user's belief over the true prevalence $p(x=1)$. Incorporating prior knowledge over the true prevalence is a widely used approach in the positive and unlabelled literature [30].

By assuming perfect knowledge over the prevalence $\gamma_x=p(x=1)$ and using only the observed variables Y and $\widetilde{X}$ we can estimate $I(X;Y)$ using the following corrected estimator.

Definition 1 *(Corrected ML-MI estimator).* The corrected estimator of the MI between an UR variable and the target is given by:

$$\widehat{I}_{\gamma_x}(\widetilde{X};Y)=\sum_{y\in\mathcal{Y}}\left(\gamma_x\,\widehat{p}(y|\widetilde{x}=1)\ln\frac{\widehat{p}(y|\widetilde{x}=1)}{\widehat{p}(y)}+(\widehat{p}(y)-\gamma_x\widehat{p}(y|\widetilde{x}=1))\ln\frac{\widehat{p}(y)-\gamma_x\widehat{p}(y|\widetilde{x}=1)}{\widehat{p}(y)\,(1-\gamma_x)}\right). \tag{7}$$

The following Lemma proves the consistency of the estimator.

Lemma 1 *(Corrected ML-MI estimator, consistency). When we have perfect prior knowledge over the prior probabilities, i.e. $\gamma_x=p(x=1)$, the suggested estimator in Definition 1 is consistent, since it holds: $I_{\gamma_x}(\widetilde{X};Y)=I(X;Y)$.*

[4] It is worth pointing out that imputation usually assumes missing at random (MAR), which is not the case in UR – where we have missing not at random (MNAR). To calculate this posterior, we also need prior knowledge for $p(x=1)$.

[5] Rubin's rule assumes normality, while ML-MI estimator can be severely non-normal. For example, in small effect sizes non-central χ^2-distribution provides a better fit – for more details see [31, Section 2.1.1].

[6] Using prior knowledge over $p(x=1)$ is equivalent to using prior knowledge over the sensitivity, an approach that is followed to correct the misclassification bias of epidemiological effect sizes [17]. This can be shown by the fact that in the non-differential under-reporting it holds that: Sensitivity $=p(m_x=1|x=1)=\frac{p(m_x=1,x=1)}{p(x=1)}=\frac{p(\widetilde{x}=1)}{p(x=1)}$, and the $p(\widetilde{x}=1)$ can be estimated by the observed data.

To prove that the estimator is consistent is straightforward, since when we have perfect prior knowledge $\gamma_x = p(x=1)$, by using (5) and (6) it holds that $I_{\gamma_x}(\widetilde{X};Y) = I(X;Y)$. Only proving the consistency of the corrected estimator is not useful, and we need to capture also the variance that it has in the finite sample size. We do so through the following theorem.

Theorem 3 *(Corrected ML-MI estimator, asymptotic distribution). For the estimator $\widehat{I}_{\gamma_x}(\widetilde{X};Y)$ it holds that: $\sqrt{n}\big(\widehat{I}_{\gamma_x}(\widetilde{X};Y) - I(X;Y)\big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{MI_{\gamma_x}}\right)$, when we have perfect prior knowledge $\gamma_x = p(x=1)$. The standard error is:*

$$SE\left[\widehat{I}_{\gamma_x}(\widetilde{X};Y)\right] = \frac{\sigma_{MI_{\gamma_x}}}{\sqrt{n}} = \frac{1}{\sqrt{n}}\left(\sum_{\widetilde{x}\in\widetilde{\mathcal{X}},y\in\mathcal{Y}}\left(p(\widetilde{x},y)\phi^2_{\widetilde{x},y}\right) - \left(\sum_{\widetilde{x}\in\widetilde{\mathcal{X}},y\in\mathcal{Y}}\left(p(\widetilde{x},y)\phi_{\widetilde{x},y}\right)\right)^2\right)^{\frac{1}{2}}, \tag{8}$$

$\phi_{\widetilde{x}=0,y} = \ln\frac{p(y)-\gamma_x p(y|\widetilde{x}=1)}{p(y)}$, $\phi_{\widetilde{x}=1,y} = \phi_{\widetilde{x}=0,y} + \frac{\gamma_x}{p(\widetilde{x}=1)}\sum_{y'\in\mathcal{Y}}\left(p(y'|\widetilde{x}=1) - \delta_{yy'}\right)\ln\frac{p(y')-\gamma_x p(y'|\widetilde{x}=1)}{\gamma_x p(y'|\widetilde{x}=1)}$.

Proof can be found in Appendix A.3.

4.1.1. Experiments with synthetic data and perfect prior knowledge

As a “sanity check” for our theoretical results we generated synthetic random variables X and Y with different degrees of dependency. To create the data, firstly we generate the values of X, by taking N samples from a Bernoulli distribution with parameter $p(x=1)$. Then, we randomly choose the parameters $p(y|x)$ that guarantee the desired degree of dependency, expressed in terms of $I(X;Y)$, and we use these parameters to sample the values of Y. To create the under-reported variable $\widetilde{X}$ we sample with Sensitivity $= p(\widetilde{x}=1|x=1)$ the examples with $x=1$. We estimate mutual information using five different methods:

- **Ideal:** using the unobservable estimator $\widehat{I}(X;Y)$ and eq. (2) for standard error.
- **No correction:** using the under-reported estimator $\widehat{I}(\widetilde{X};Y)$ and eq. (2) for standard error.
- **Single imputation:** using a model to impute possible misclassified data and then estimate MI and standard error by eq. (2).
- **Multiple imputations:** using a model to impute multiple times and then average MI across the imputed datasets and using Rubin's rule [3] for standard error. To decide this number, we used the White et al. [36] guideline that the number of imputations should be approximately 100 times the fraction of missing information. In under-reporting this can be phrased as using 100 × (1 − Sensitivity) imputations.
- **Our correction:** using our corrected estimator $\widehat{I}_{\gamma_x}(\widetilde{X};Y)$ presented in eq. (7) and using the results of Theorem 3 for standard error.

For the imputation-based approaches, we imputed the potentially misclassified examples through the following posterior, which can be naturally derived by the model of Fig. 1(a):

$$p(x=1|y,\widetilde{x}=0) = \frac{p(y,\widetilde{x}=0|x=1)\gamma_x}{p(y,\widetilde{x}=0)} = \frac{\left(p(y|x=1) - p(y,\widetilde{x}=1|x=1)\right)\gamma_x}{p(y,\widetilde{x}=0)} = \frac{p(y|\widetilde{x}=1)\left(\gamma_x - p(\widetilde{x}=1)\right)}{p(y,\widetilde{x}=0)}.$$

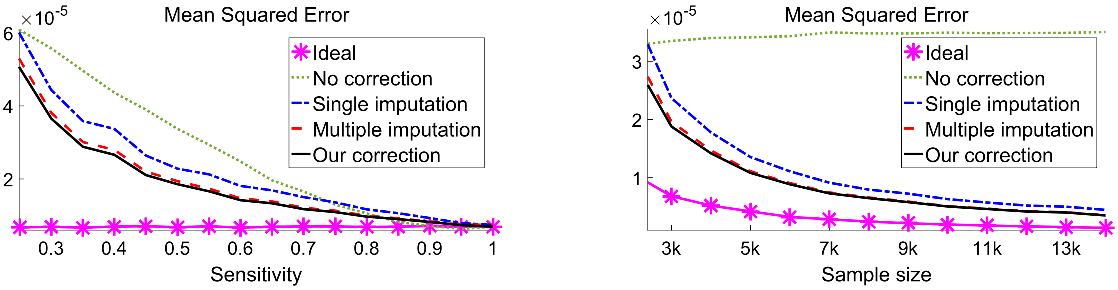
As we mentioned, we use perfect prior knowledge over γ_x, while the rest of the parameters are estimated through ML from the observed data. To get a fair comparison between the last three methods, we used the same modelling assumptions and γ_x is assumed to be known and equal to $p(x=1)$.

Fig. 3 compares the five methods in terms of their mean squared error. The three methods that take into account the under-reporting (single/multiple imputation and our corrected estimator) outperform the naive estimator, which is not consistent. As the sample size/sensitivity increases, all of these three approaches tend to behave in a similar way to the ideal estimator. Our corrected estimator outperforms the imputation-based approaches, especially in small sample sizes and small levels of sensitivity – which are the most challenging situations. Interestingly, our method clearly outperforms methods with the same complexity (no correction and simple imputation).

Fig. 4 verifies that the suggested standard error in Theorem 3 is correct, and that our method is a valid way to derive interval estimates, similar to those derived using the ideal estimator. In this figure we estimate the proportion of times that the 90% confidence intervals, derived by using different standard errors for the different methods, contain the true value of the mutual information $I(X;Y)$. Since the estimated coverage probability for the ideal and our proposed method are at the nominal (user specified) level of 90%, we can conclude that only these methods produce accurate interval estimates.

4.1.2. Experiments with synthetic data and uncertain prior knowledge

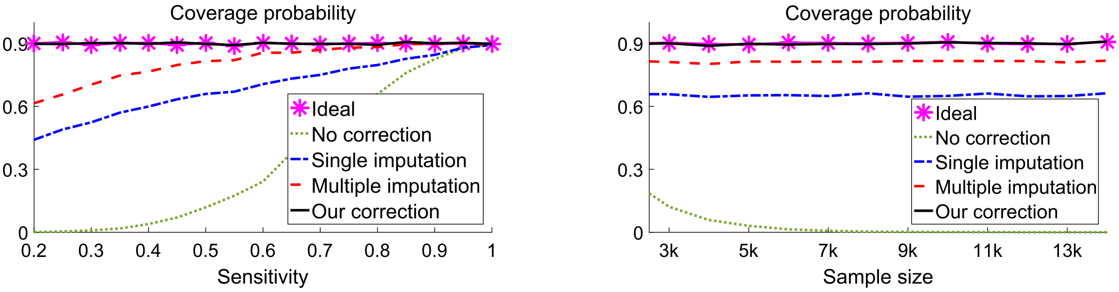
Perfect prior knowledge, i.e. $\gamma_x = p(x=1)$, will not always be available. Therefore it is important to explore ways to deal with uncertain knowledge and examine the behaviour with incorrect priors – results are presented below for an artificial scenario where we can exert control over the “quality” of prior knowledge.



(a) Different levels of UR with $N = 3\mathrm{k}$

(b) Different sample sizes with Sensitivity $= 0.50$

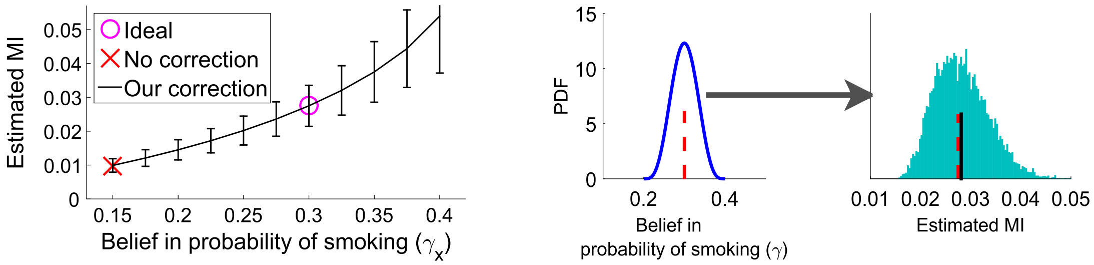
Fig. 3. Comparison in terms of mean (over 5,000 repetitions) squared error. In each repetition we set $I(X;Y) = 0.01$ and we randomly choose: $|\mathcal{Y}| \in \{2\text{–}5\}$ and $p(x = 1) \in \{0.1\text{–}0.5\}$.



(a) Different levels of UR with $N = 5\mathrm{k}$

(b) Different sample sizes with Sensitivity $= 0.50$

Fig. 4. Comparing in terms of coverage. We set the nominal level to be 0.90 (90% confidence intervals) and we observe the proportion (over 5,000 repetitions) of the times that suggested intervals contain the true value of the mutual information.



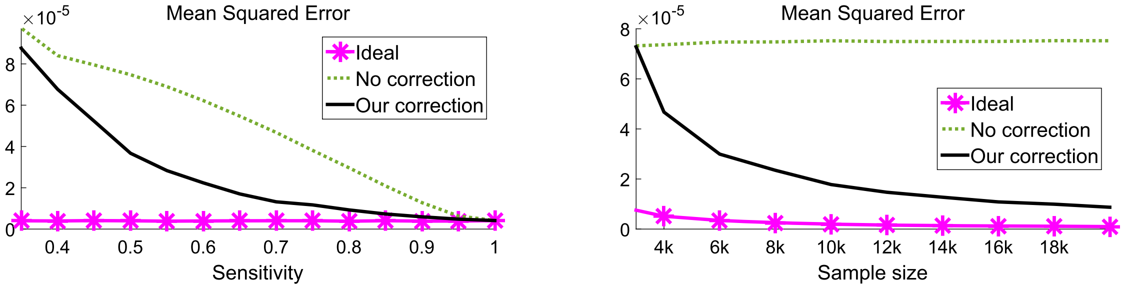
(a) Sensitivity analysis.

(b) Simulation based analysis.

Fig. 5. Different ways to handle uncertain prior knowledge. (a) Sensitivity analysis. (b) Simulation based analysis: [left] the user's prior belief over γ_x, [right] the resultant uncertainty in the estimated mutual information through our correction. The dashed line shows the true (but unknown) value, while the solid line the average over the simulations.

Let us assume that non-smoker births are drawn from a normal distribution with $\mu = 3500$ g and $\sigma = 500$ g, while weight of smoker births are drawn from a normal distribution with $\mu = 3000$ g and $\sigma = 500$ g. Birthweight was considered to be "low", $x = 1$, if the weight was < 2500 g [37]. We assume that in a cohort of $N = 5000$ pregnant mothers, 30% are smokers, so $p(x = 1) = 0.3$. However, only half of the mothers on average would admit to this, so $p(\widetilde{x} = 1) = 0.15$. In a typical draw from this simulation, the mutual information is estimated with an under-reported variable. However, after using our corrected estimator and by incorporating the prior knowledge that the X variable is non-differential under-reported, the estimated mutual information increases by a factor of three (Fig. 5(a)).

One way to handle uncertain prior knowledge is by performing a *sensitivity analysis* as Fig. 5(a) shows. To do so we plot the interval estimates for the corrected MI, calculated by eq. (8), for different values of our belief over the probability of smoking (γ_x). As we observe the point estimate for $\gamma_x = p(x = 1) = 0.30$ (perfect knowledge) is the same with the true (ideal) value of the MI. A different way to handle uncertainty is through a *simulation based analysis*, where we represent uncertainty over γ_x as a probability distribution, sample from this distribution many times, and estimate the corrected MI for each value. For example, in Fig. 5(b) we model γ_x as a generalised Beta distribution (bounded between a minimum and



(a) Different levels of UR with $N = 5\text{k}$

(b) Different sample sizes with Sensitivity $= 0.50$

Fig. 6. Comparison in terms of mean (over 5,000 repetitions) squared error. In each repetition we set $I(X;Z) = 0.01$ and we randomly choose: $p(x=1)$, $p(z=1) \in \{0.1\text{–}0.5\}$. For illustration proposes we assumed the same sensitivity for the two UR variables.

a maximum value) and we explore the resultant uncertainty in the estimate of the corrected mutual information. As we observe, the true value of the MI is very close to the average over the simulations.

4.2. Correcting for under-reporting the mutual information that captures redundancy

Using the results of the previous section, we can measure redundancy terms when only one of the features is under-reported, i.e. $I(\widetilde{X};Z)$, but we cannot measure terms when both of the features are under-reported, i.e. $I(\widetilde{X};\widetilde{Z})$ in Fig. 1(b). To do so we will use the following estimator.

Definition 2 *(Corrected ML-MI estimator between two UR variables).* The corrected estimator of the MI between two UR variables is given by:

$$\begin{aligned}
\widehat{I}_{\gamma_x,\gamma_z}(\widetilde{X};\widetilde{Z}) &= \frac{\gamma_x\gamma_z\widehat{p}(\widetilde{x}=1,\widetilde{z}=1)}{\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)} \ln \frac{\widehat{p}(\widetilde{x}=1,\widetilde{z}=1)}{\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)} \\
&+ \frac{\gamma_z\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)-\gamma_x\gamma_z\widehat{p}(\widetilde{x}=1,\widetilde{z}=1)}{\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)} \ln \frac{\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)-\gamma_x\widehat{p}(\widetilde{x}=1,\widetilde{z}=1)}{(1-\gamma_x)\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)} \\
&+ \frac{\gamma_x\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)-\gamma_x\gamma_z\widehat{p}(\widetilde{x}=1,\widetilde{z}=1)}{\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)} \ln \frac{\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)-\gamma_z\widehat{p}(\widetilde{x}=1,\widetilde{z}=1)}{(1-\gamma_z)\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)} \\
&+ \frac{(1-\gamma_x-\gamma_z)\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)+\gamma_x\gamma_z\widehat{p}(\widetilde{x}=1,\widetilde{z}=1)}{\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)} \ln \frac{(1-\gamma_x-\gamma_z)\widehat{p}(\widetilde{x}=1)\widehat{p}(\widetilde{z}=1)+\gamma_x\gamma_z\widehat{p}(\widetilde{x}=1,\widetilde{z}=1)}{(1-\gamma_z)(1-\gamma_x)p(\widetilde{x}=1)p(\widetilde{z}=1)}.
\end{aligned} \tag{9}$$

The following lemma proves that with perfect prior knowledge this estimator is consistent.

Lemma 2 *(Corrected ML-MI estimator between two UR variables, consistency). When we have perfect prior knowledge over the prior probabilities, i.e. $\gamma_x = p(x=1)$ and $\gamma_z = p(z=1)$, the suggested estimator in Definition 2 is consistent, since it holds: $I_{\gamma_x,\gamma_z}(\widetilde{X};\widetilde{Z}) = I(X;Z)$.*

Proof can be found in Appendix A.4.

As a "sanity check" for our theoretical results we generated synthetic random variables X and Z with different degrees of dependency. To create the under-reported variables $\widetilde{X}$ and $\widetilde{Z}$ we sample with $\text{Sensitivity}_x = p(\widetilde{x}=1|x=1)$ the examples with $x=1$ and with $\text{Sensitivity}_z = p(\widetilde{z}=1|z=1)$ the examples with $z=1$. We estimate mutual information using three different methods:

- **Ideal:** using the unobservable estimator $\widehat{I}(X;Z)$.
- **No correction:** using the under-reported estimator $\widehat{I}(\widetilde{X};\widetilde{Z})$.
- **Our correction:** using our corrected estimator $\widehat{I}_{\gamma_x,\gamma_z}(\widetilde{X};\widetilde{Z})$ presented in eq. (9).

We will not compare against imputation based methods, since, in this scenario, we do not have any correct reported variable to build the imputation model.

Fig. 6 compares the three methods in terms of their mean squared error. As the sample size/sensitivity increases, our suggested method tends to behave in a similar way to the ideal estimator. Our corrected estimator clearly outperforms the naive method (no correction).

5. Ranking features in under-reported scenarios

By using our theoretical results on estimating mutual information terms in under-reported scenarios, we can derive two different algorithms for producing feature rankings. Before that we will introduce some extra notation. We can assume that the feature vector $\mathbf{X}$ consists of two parts: $\mathbf{X} = \{\mathbf{X}_{cr}, \widetilde{\mathbf{X}}_{ur}\}$. Where $\mathbf{X}_{cr}$ contains features that are correctly reported, while $\widetilde{\mathbf{X}}_{ur}$ the ones that are under-reported.

5.1. Rankings that capture only relevance

Using our findings from Section 4.1 we suggest **Corrected-MIM**, which is an extension of MIM (Section 2.5), and it is suitable when we have UR features. The score of each feature X_k is estimated by:

$$J'_{MIM}(X_k) = \widehat{I}'(X_k;Y),$$

where

$$\widehat{I}'(X_k;Y) = \begin{cases} \widehat{I}(X_k;Y) & \text{when } X_k \in \mathbf{X}_{cr}, \text{ estimate MI using eq. (1)} \\ \widehat{I}_{\gamma_{x_k}}(\widetilde{X}_k;Y) & \text{when } X_k \in \widetilde{\mathbf{X}}_{ur}, \text{ estimate MI using eq. (7)} \end{cases} \tag{10}$$

5.2. Rankings that capture both relevance and redundancy

To derive feature rankings that capture *both* relevance and redundancy, we need to combine our findings from Sections 4.1 and 4.2, and the mRMR paradigm presented in Section 2.5. Our suggested method, **Corrected-mRMR**, ranks the features according to the score:

$$J'_{mRMR}(X_k) = \widehat{I}'(X_k;Y) - \frac{1}{|\mathbf{X}_\theta|}\sum_{X_j \in \mathbf{X}_\theta} \widehat{I}''(X_k;X_j),$$

where $\mathbf{X}_\theta$ is the set of the features already selected, the term that captures relevance I' is given by eq. (10) and redundancy I'' by:

$$\widehat{I}''(X_k;X_j) = \begin{cases} \widehat{I}(X_k;X_j) & \text{when } X_k \in \mathbf{X}_{cr} \text{ and } X_j \in \mathbf{X}_{cr}, \text{ estimate MI using eq. (1)} \\ \widehat{I}_{\gamma_{x_k}}(\widetilde{X}_k;X_j) & \text{when } X_k \in \widetilde{\mathbf{X}}_{ur} \text{ and } X_j \in \mathbf{X}_{cr}, \text{ estimate MI using eq. (7)} \\ \widehat{I}_{\gamma_{x_j}}(X_k;\widetilde{X}_j) & \text{when } X_k \in \mathbf{X}_{cr} \text{ and } X_j \in \widetilde{\mathbf{X}}_{ur}, \text{ estimate MI using eq. (7)} \\ \widehat{I}_{\gamma_{x_k},\gamma_{x_j}}(\widetilde{X}_k;\widetilde{X}_j) & \text{when } X_k \in \mathbf{X}_{cr} \text{ and } X_j \in \widetilde{\mathbf{X}}_{ur}, \text{ estimate MI using eq. (9)} \end{cases}$$

In the following two sections we present two applications of our theoretical results.

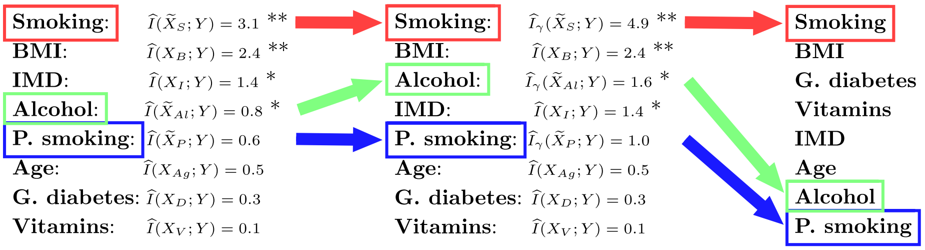
6. Application in ranking risk factors for low birth weight infants

In this section we present a real-world application of our results — ranking the risks factors that may lead to adverse birth outcomes derived from a large real-world dataset.

To describe the usefulness of our theoretical findings we will use data from a prospective birth cohort, the Born in Bradford (BiB) study. BiB is a longitudinal multi-ethnic birth cohort study aiming to examine the impact of environmental, psychological and genetic factors on maternal and child health and well-being [37]. Bradford is a city in northern England with high levels of socio-economic deprivation and ethnic diversity. The full BiB cohort recruited 12,453 women with 13,776 pregnancies between 2007 and 2010, and the cohort is broadly characteristic of the city's maternal population in terms of age, deprivation and ethnicity. Ethics approval for the study was granted by Bradford Research Ethics Committee (Ref. 07/H1302/112). In our analyses we focus on term births only, and we excluded ethnic groups (such as Pakistani mothers) that are much less likely to smoke and drink alcohol than the rest of the cohort [37]. As a result, the number of suitable pregnancies reduced to 5,457.

We show how to rank several risk factors according to their association with LBW. The risk factors that we focus on are the correctly-reported *categorical* variables: ethnicity X_E (3 levels), age X_{Ag} (3 levels), Body Mass Index (BMI) X_B (4 levels), index of multiple deprivation X_I (5 levels), gestational diabetes X_G (binary), taken vitamins X_V (binary), and the following binary UR variables: any smoking $\widetilde{X}_S$, passive smoking $\widetilde{X}_P$ and alcohol $\widetilde{X}_{Al}$ consumption during pregnancy. Let us assume that 1/3 of the overall women under-report these three variables, and we assume non-differential UR.

In Fig. 7(a) we observe the MIM ranking by using the MI of the observed covariates and the target variable (LBW). Then, to correct the three UR variables, we use prior knowledge and our corrected estimators presented in Sect. 4, and we derive the Corrected-MIM ranking of Fig. 7(b). Finally, to take into account the correlations between the risk-factors we derive the Corrected-mRMR ranking of Fig. 7(c). From these three rankings we can arrive to the following interesting conclusions:


Smoking: $\widehat{I}(\widetilde{X}_S;Y)=3.1$ **
BMI: $\widehat{I}(X_B;Y)=2.4$ **
IMD: $\widehat{I}(X_I;Y)=1.4$ *
Alcohol: $\widehat{I}(\widetilde{X}_{Al};Y)=0.8$ *
P. smoking: $\widehat{I}(\widetilde{X}_P;Y)=0.6$
Age: $\widehat{I}(X_{Ag};Y)=0.5$
G. diabetes: $\widehat{I}(X_D;Y)=0.3$
Vitamins: $\widehat{I}(X_V;Y)=0.1$
Smoking: $\widehat{I}_\gamma(\widetilde{X}_S;Y)=4.9$ **
BMI: $\widehat{I}(X_B;Y)=2.4$ **
Alcohol: $\widehat{I}_\gamma(\widetilde{X}_{Al};Y)=1.6$ *
IMD: $\widehat{I}(X_I;Y)=1.4$ *
P. smoking: $\widehat{I}_\gamma(\widetilde{X}_P;Y)=1.0$
Age: $\widehat{I}(X_{Ag};Y)=0.5$
G. diabetes: $\widehat{I}(X_D;Y)=0.3$
Vitamins: $\widehat{I}(X_V;Y)=0.1$
Smoking
BMI
G. diabetes
Vitamins
IMD
Age
Alcohol
P. smoking

(a) MIM (b) Corrected-MIM (b) Corrected-mRMR

Fig. 7. Variable ranking by their association with the LBW. (a) Ranked by MIM, uncorrected. (b) Ranked by Corrected-MIM. (c) Ranked by Corrected-mRMR. Units are milli-nats. The single star * means the null hypothesis (independence between the reported covariate and LBW) is rejected at $\alpha = 0.01$, while double stars ** at $\alpha = 0.001$. Failure to reject the null does not imply insignificance as the test may not have sufficient power, which is likely the case in an under-reported test as we showed in Section 3.

Table 1
Datasets used in the feature rankings experiments.

Dataset	# examples	# features	$\lvert\mathcal{Y}\rvert$
Chess	3196	38	2
Congress	435	48	2
LUCAP0	2000	143	2
LUCAS0	2000	11	2
Mushroom	4062	112	2
Splice	3175	240	3

1. Smoking is the most important risk for LBW even without taking into account the UR (Fig. 7(a)).
2. By correcting for UR, alcohol gains one position in the ranking (Fig. 7(b)). Another interesting observation is that the test of independence between passive smoking and LBW does not reject the null hypothesis. Failure to reject the null does not necessarily imply insignificance, as the test may not have sufficient power to detect an actual effect, which is likely the case in an under-reported test[7] as we showed in Section 3.
3. By using our method for deriving corrected multi-variate rankings (Corrected-mRMR), we see that alcohol and passive smoking become the least important factors (Fig. 7(c)). This result matches with the known correlation between (passive) smoking, BMI and alcohol [11,22]. By using our Corrected-mRMR methodology, we take into account the redundancy between these factors. Thus, by conditioning over smoking, both alcohol and passive smoking become less important.

The differences between the three rankings illustrate the importance of having techniques that are able to produce estimates that correct under-reporting and take into account the correlation between the risk factors. We have demonstrated that failure to correct for potential under-reporting of exposure will lead to biased estimates of the ranking of the relative effect between variables and outcomes. To appropriately validate our statistical findings access to the true values of X_S, X_P and X_{Al} will be necessary, i.e. through a blood test. Unfortunately, this information was not available. For that reason, in the following section we present the merits of our analysis in a machine learning application using datasets for which we have access to the ground truth.

7. Application in feature ranking in under-reported scenarios

In this section we present a machine learning application of our findings: producing feature rankings when the training/test distributions differ, and particularly, when the features are non-differentially under-reported.[8] We used four categorical UCI[9] datasets (Splice, Chess, Mushroom and Congress) and two artificial lung cancer[10] datasets (LUCAS0 and LUCAP0) from the Causation and Prediction Challenge [18]. Splice is a 3-class classification problem, while the rest are binary. Categorical attributes are expanded into several binary attributes. Table 1 shows the characteristics of each dataset. We assumed that the features of these datasets are correctly reported, and to generate UR datasets we randomly under-reported the original features.

[7] In https://github.com/sechidis we provide Matlab scripts with code and illustrative examples of the power-loss that is caused by under-reporting. The inputs can be adjusted in the script, to demonstrate how to achieve a given false positive/negative rate tradeoff for a given data scenario chosen by the user.

[8] This scenario can be seen as an event-level version of covariate shift [26], since, because of eq. (5), we have $p_{test}(y|x=1) = p_{train}(y|x=1)$ but not for $x=0$.

[9] Available in the UCI Machine Learning Repository http://archive.ics.uci.edu/ml/.

[10] Available in http://www.causality.inf.ethz.ch/challenge.php.

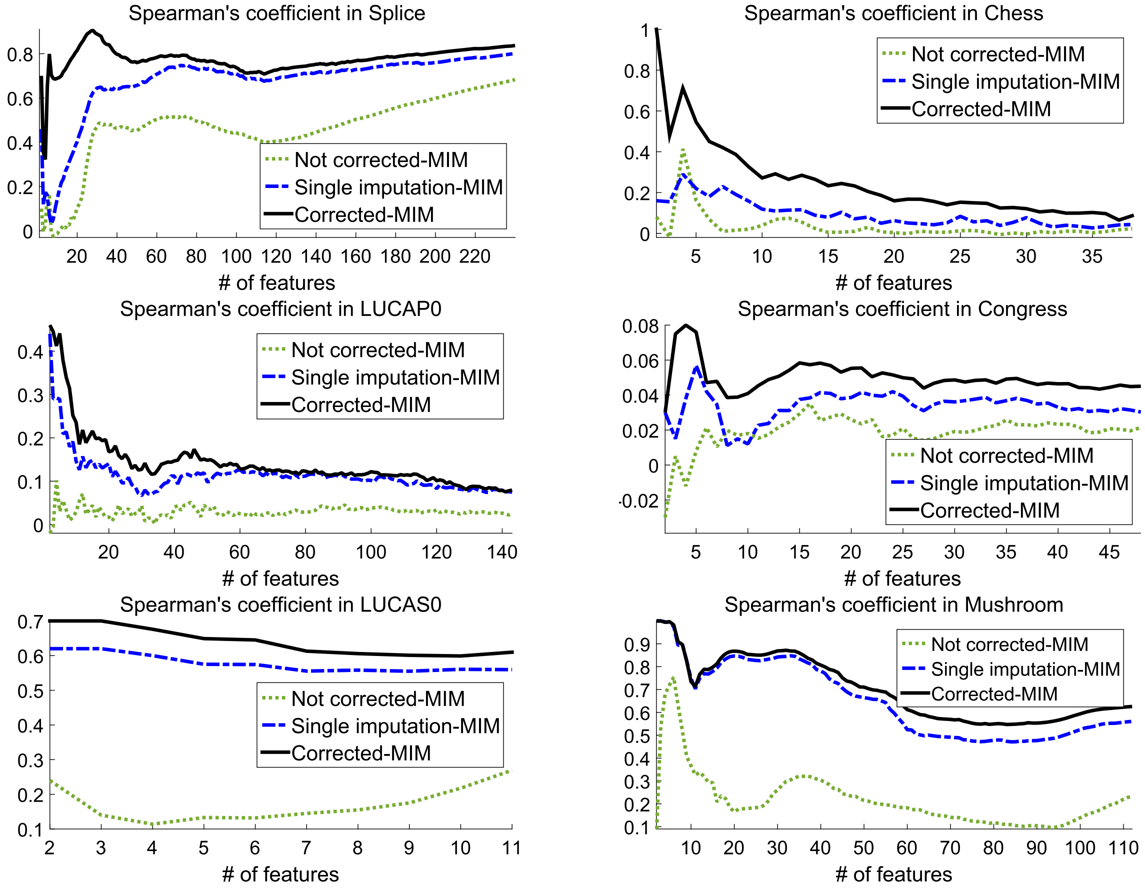


Fig. 8. Quality of under-reported MIM rankings: Average Spearman's ρ correlation coefficient between the features returned by Ideal-MIM method and the features returned by MIM rankings derived by the three under-reported methods (Not corrected-MIM, Single imputation-MIM and our suggested method, Corrected-MIM). The average is calculated over 100 under-reported versions of the original datasets. To generate these versions we bootstrap the original data and then we under-report the features with sensitivities chosen randomly in the range [0.50–1].

7.1. Deriving MIM ranking

In this section, we will compare the rankings derived by using different MIM methods. Fig. 8 compares the similarities between the MIM rankings derived by UR methods and the ideal ranking – the MIM ranking that we would have if we had access to the actual values of the features. To check the similarity between the rankings we use Spearman's ρ correlation coefficient [20]. The range of values that this coefficient takes is $[-1, 1]$, where 1 means that the two rankings are identical, 0 means that there is no correlation between them, while -1 means that the two rankings are inverse. We compare the three UR methods with the same complexity: no correction, simple imputation and our correction (for a fair comparison we used perfect prior knowledge for the last two approaches). Fig. 8 shows that our suggested approach, **Corrected-MIM**, outperforms the other approaches in all the settings.

Now we will compare the different methods in terms of their misclassification error. As a classification model we will use a k-nearest neighbour ($k = 3$) classifier, which makes few assumptions about the data and treats all features equally, a desirable property when we compare different feature selection/ranking approaches [6]. Fig. 9 compares the different UR approaches in terms of their misclassification error. In the Splice and Chess datasets our approach outperforms the others and achieves similar performance as using the ideal estimator. In the rest of the datasets, all methods have similar performance, and always our method performs similarly with the ideal.

7.2. Deriving mRMR ranking

Now we will compare the rankings derived by using different mRMR methods. Fig. 10 compares the similarities between the mRMR rankings derived by UR methods and the ideal ranking. Our suggested approach, **Corrected-mRMR**, outperforms the naive not-corrected approach in all the settings. Fig. 11 compares the different approaches in terms of their misclassification error. In the Mushroom and Chess dataset our approach outperforms the naive method and achieves similar performance as using the ideal estimator. In the rest of the datasets, all methods have similar performance, and always our method performs similarly with the ideal.

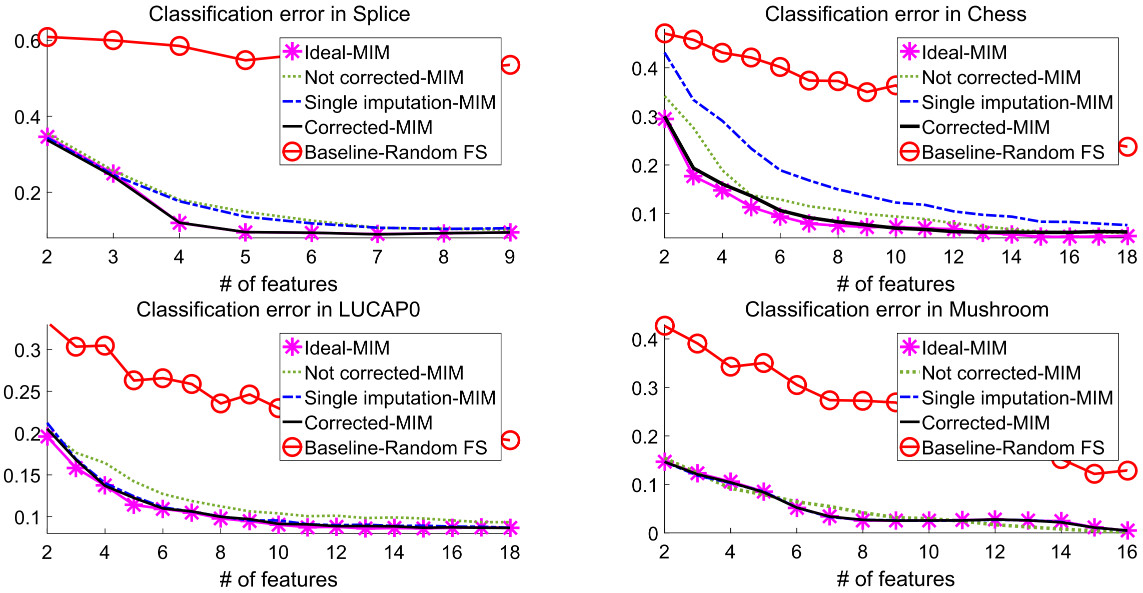


Fig. 9. Misclassification error using top-*k* features from different MIM rankings: Average testing error over 30 random splits of the data into 50% training and 50% testing. For each training dataset we generate UR datasets, by randomly non-differential under-reporting ten of the features with sensitivities chosen in the range [0.50–1]. The results of the LUCAS0 and Congress datasets show similar trend with the Mushroom dataset, thus we omit them for brevity. The baseline random feature selection method helps in determining the complexity of the task.

8. Conclusions and future work

In this work we have provided an information theoretic solution to the problem of under-reported variables. Initially, by reinterpreting under-reporting as a missing data problem, we presented how we can use the tool of *missingness graphs* [23] for providing graphical representations of the different under-reported scenarios. Then, by using these representations, we explored valid ways to test independence, and we derived a correction factor that quantifies the power loss of the χ^2-test. Furthermore, by incorporating simple prior knowledge, we derived ways for estimating mutual information quantities that capture both relevance and redundancy. Our suggested estimators are computationally efficient, while they have similar error with more complex imputation based approaches. Additionally, for the mutual information that captures the relevancy, we derived confidence intervals that achieve the ideal coverage. Using our suggested estimators we proposed two methods for feature ranking: Corrected-MIM, which captures only relevance, and Corrected-mRMR, which captures both relevance and redundancy. Our theoretical results are supported through experiments with synthetic data. Finally, we showed how we can use our findings in a real-world health care application (ranking the risk factors that may lead to low birth weight) and in a machine learning application (feature ranking when training/testing distributions differ).

In many practical applications we have misreporting mechanisms that are correlated, for example by having a latent variable which is a parent of both missingness mechanisms M_X and M_X in Fig. 1(b). One limitation of our work is that it assumes independent misreporting mechanisms. Thus an interesting future direction is to explore ways of estimating redundancy terms without making any independence assumption. Furthermore, providing ways for testing, estimation and ranking in *differential* under-reporting (i.e. when there is a direct arc between the missingness mechanism M_X and the variable Y in Fig. 1(a)) seems challenging. Lastly, finding ways to consistently estimate *conditional* mutual information terms will provide us with algorithms for structure learning or Markov blanket discovery in UR scenarios [32]. In our earlier work [33] we suggested a way for correcting conditional mutual information when we condition on correctly reported variables, while a promising future direction is to derive a corrected estimator, when we condition on under-reported variables.

We believe our results are highly applicable in a wide variety of machine learning applications, when we face the problem of under-reporting. Estimating mutual information, testing independence, ranking sets of features according to their relevance/redundancy, learning Bayesian network structures and sample size determination for experimental design are some – but not all – of the possible applications.

Acknowledgements

This study was partly supported by the University of Manchester's Health eResearch Centre (HeRC) funded by the Medical Research Council (MRC) Grant MR/K006665/1. Konstantinos Sechidis, Mikel Luján and Gavin Brown were supported by the Engineering and Physical Sciences Research Council through the Centre for Doctoral Training grant [EP/I028099/1] and the Anyscale Apps project grant [EP/L000725/1]. Born in Bradford is only possible because of the enthusiasm and commitment

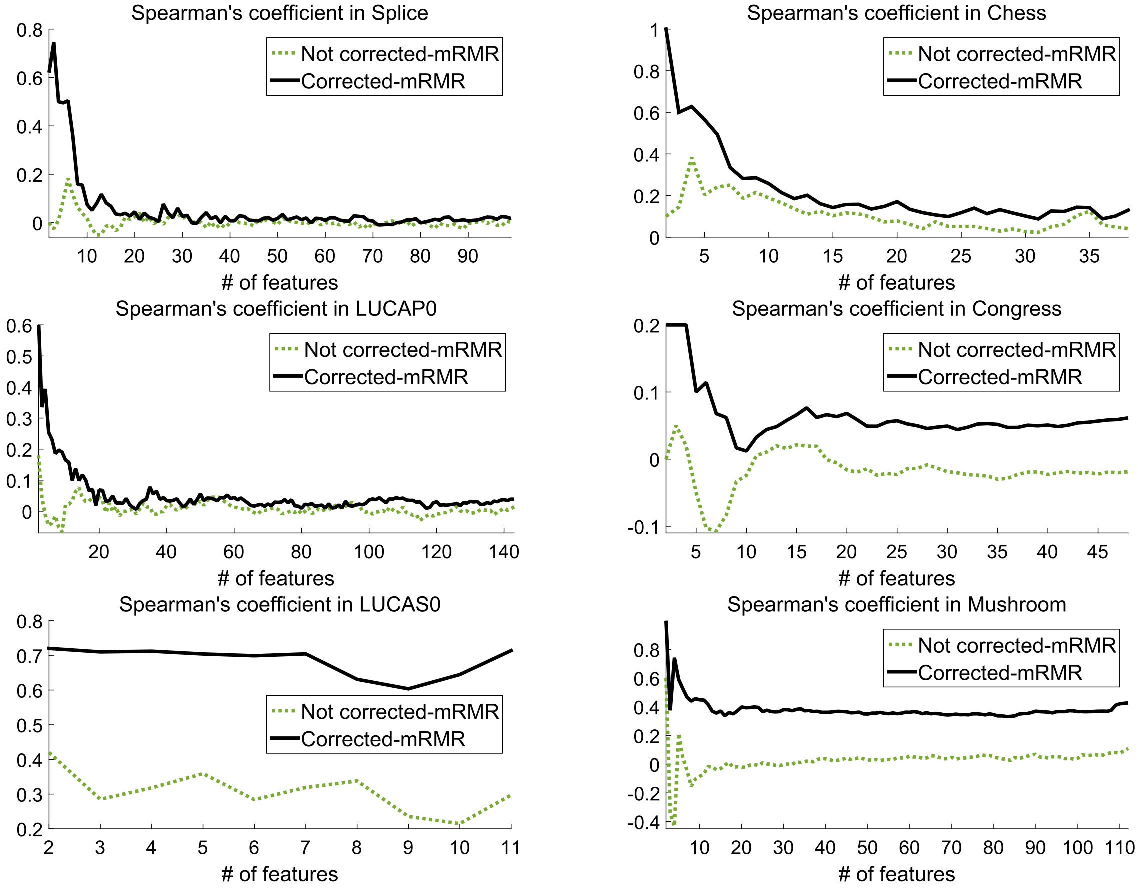


Fig. 10. Quality of under-reported mRMR rankings: Average Spearman's ρ correlation coefficient between the Ideal-mRMR ranking and the mRMR rankings derived by two under-reported methods (Not corrected-mRMR and our suggested method Corrected-mRMR). The average is calculated over 100 under-reported versions of the original datasets. To generate these versions we bootstrap the original data and then we under-report the features with sensitivities chosen randomly in the range [0.50–1].

of the Children and Parents in BiB. We are grateful to all the participants, practitioners and researchers who have made Born in Bradford happen.

Data access statement: All research data supporting this publication are directly available within this publication, apart from the Born in Bradford dataset, which is obtained upon request and subject to licence restrictions. Due to the potentially identifiable nature of this dataset coming from a small geographical area we are unable to deposit it in the public domain. Full details of how these data were obtained are available in [37], while further details on the application procedure can be found on the Born in Bradford website (http://www.borninbradford.nhs.uk/research-scientific/how-to-request-access-to-raw-bib-data/).

Appendix A. Proofs of the theorems

A.1. Proof of Theorem 2

The non-centrality parameter of the under-reported χ^2 test is equal to $\lambda_{\chi^2(\widetilde{X};Y)} = 2NI_2(\widetilde{X};Y)$, while the non-centrality parameter of the unobserved test is equal to $\lambda_{\chi^2(X;Y)} = 2NI_2(X;Y)$. In order to derive a relationship between the parameters we should derive a relationship between the two squared loss mutual information terms.

We will start by re-expressing $I_2(X;Y)$ as follows

$$I_2(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{(p(x,y)-p(x)p(y))^2}{p(x)p(y)}$$

$$I_2(X;Y) = \sum_{y\in\mathcal{Y}} \frac{(p(y,x=1)-p(y)p(x=1))^2}{p(y)p(x=1)} + \sum_{y\in\mathcal{Y}} \frac{(p(y,x=0)-p(y)p(x=0))^2}{p(y)p(x=0)}$$

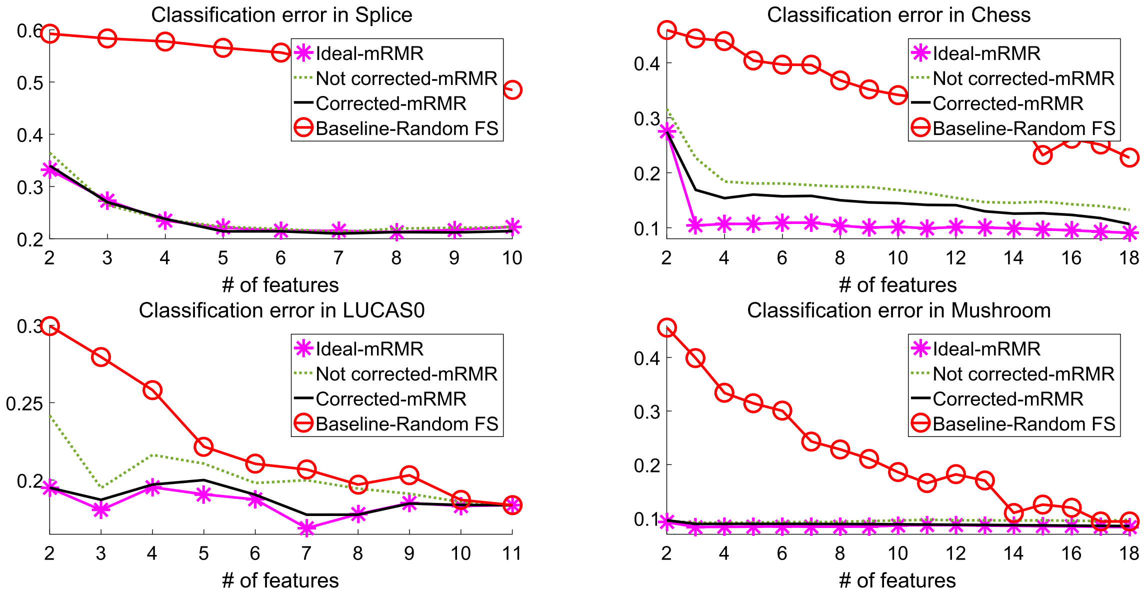


Fig. 11. Misclassification error using top-*k* features from different mRMR rankings: Average testing error over 30 random splits of the data into 50% training and 50% testing. For each training dataset we generate UR datasets, by randomly non-differential under-reporting ten of the features with sensitivities chosen in the range [0.50–1]. Because the results of the LUCAP0 and Congress datasets show similar trend with Splice dataset, we omit them for brevity. The baseline random feature selection method helps in determining the complexity of the task.

$$
\begin{aligned}
I_2(X;Y) &= \sum_{y\in\mathcal{Y}} \frac{(p(y,x=1)-p(y)p(x=1))^2}{p(y)p(x=1)} + \sum_{y\in\mathcal{Y}} \frac{(p(y)-p(y,x=1)-p(y)(1-p(x=1)))^2}{p(y)(1-p(x=1))} \\
I_2(X;Y) &= \sum_{y\in\mathcal{Y}} \frac{(p(y,x=1)-p(y)p(x=1))^2}{p(y)p(x=1)} + \sum_{y\in\mathcal{Y}} \frac{(p(y,x=1)-p(y)p(x=1))^2}{p(y)(1-p(x=1))} \\
I_2(X;Y) &= p(x=1) \sum_{y\in\mathcal{Y}} \frac{(p(y|x=1)-p(y))^2}{p(y)} + \frac{p(x=1)^2}{1-p(x=1)} \sum_{y\in\mathcal{Y}} \frac{(p(y|x=1)-p(y))^2}{p(y)} \\
I_2(X;Y) &= \frac{p(x=1)}{1-p(x=1)} \sum_{y\in\mathcal{Y}} \frac{(p(y|x=1)-p(y))^2}{p(y)}. \qquad \text{(A.1)}
\end{aligned}
$$

Following exactly the same procedure for the $I_2(\widetilde{X};Y)$ we get

$$I_2(\widetilde{X};Y) = \frac{p(\widetilde{x}=1)}{1-p(\widetilde{x}=1)} \sum_{y\in\mathcal{Y}} \frac{(p(y|\widetilde{x}=1)-p(y))^2}{p(y)} \qquad \text{(A.2)}$$

Under the non-differential assumption because of the event based independence assumption $Y \perp\!\!\!\perp M_X|x=1$ it holds that $p(y|x=1) = p(y|x=1, m_x=1) \Leftrightarrow p(y|x=1) = p(y|\widetilde{x}=1)$, so from (A.1) and (A.2) we derive that:

$$I_2(\widetilde{X};Y) = \frac{1-p(x=1)}{p(x=1)} \frac{p(\widetilde{x}=1)}{1-p(\widetilde{x}=1)} I_2(X;Y).$$

And by multiplying both sides with $2N$ we can derive the relationship between the non-centrality parameters:

$$\lambda_{\chi^2(\widetilde{X};Y)} = \frac{1-p(x=1)}{p(x=1)} \frac{p(\widetilde{x}=1)}{1-p(\widetilde{x}=1)} \lambda_{\chi^2(X;Y)},$$

with

$$\kappa = \frac{1-p(x=1)}{p(x=1)} \frac{p(\widetilde{x}=1)}{1-p(\widetilde{x}=1)}. \qquad \square$$

A.2. Derivation of the correction suggested by Bross [5]

Bross [5, p. 484] suggests that in order to calculate the power of the χ^2-test we should use an effective sample size of $1/\kappa'$ times the actual sample size, where the correction factor κ' is given in eq. (1.02) of [5]. Bross derived this correction factor by starting from the equivalence under the null hypothesis between the X^2-test and the different-of-proportions-test. Using our notation, this correction factor can be written as:

$$\kappa' = (1 - p(\widetilde{x}=1|x=0) - p(\widetilde{x}=0|x=1))^2 + \frac{p(\widetilde{x}=0|x=1)\,(1-p(\widetilde{x}=0|x=1))}{p(x=0)} + \frac{p(\widetilde{x}=1|x=0)\,(1-p(\widetilde{x}=1|x=0))}{p(x=1)} \tag{A.3}$$

In our UR setting we have $p(\widetilde{x}=0|x=0)=1 \Leftrightarrow p(\widetilde{x}=1|x=0)=0$. By substituting $p(\widetilde{x}=1|x=0)=0$ in eq. (A.3) we get:

$$\begin{aligned}\kappa' &= (1-p(\widetilde{x}=0|x=1))^2 + \frac{p(\widetilde{x}=0|x=1)\,(1-p(\widetilde{x}=0|x=1))}{p(x=0)} \Leftrightarrow \\ \kappa' &= p(\widetilde{x}=1|x=1)^2 + \frac{(1-p(\widetilde{x}=1|x=1))\,p(\widetilde{x}=1|x=1)}{1-p(x=1)}\end{aligned} \tag{A.4}$$

The conditional probability can be written as: $p(\widetilde{x}=1|x=1) = \frac{p(\widetilde{x}=1,x=1)}{p(x=1)}$. Because of the UR constraint, details in Section 2.1, whenever an example has $\widetilde{x}=1$ then it also holds that $x=1$. This means that: $p(\widetilde{x}=1|x=1)=p(\widetilde{x}=1)$, and as a result the conditional probability takes the following form: $p(\widetilde{x}=1|x=1)=\frac{p(\widetilde{x}=1)}{p(x=1)}$. By substituting this expression in eq. (A.4) we get:

$$\begin{aligned}\kappa' =& \frac{p(\widetilde{x}=1)^2}{p(x=1)^2} + \frac{(p(x=1)-p(\widetilde{x}=1))\,p(\widetilde{x}=1)}{p(x=1)^2(1-p(x=1))} \Leftrightarrow \kappa' = \frac{p(\widetilde{x}=1)}{p(x=1)^2}\left(p(\widetilde{x}=1) + \frac{p(x=1)-p(\widetilde{x}=1)}{1-p(x=1)}\right) \Leftrightarrow \\ \kappa' =& \frac{p(\widetilde{x}=1)}{p(x=1)^2}\left(\frac{p(\widetilde{x}=1)-p(\widetilde{x}=1)p(x=1)+p(x=1)-p(\widetilde{x}=1)}{1-p(x=1)}\right) \Leftrightarrow \\ \kappa' =& \frac{p(\widetilde{x}=1)}{p(x=1)^2}\left(\frac{p(x=1)\,(1-p(\widetilde{x}=1))}{1-p(x=1)}\right) \Leftrightarrow \\ \kappa' =& \frac{p(\widetilde{x}=1)(1-p(\widetilde{x}=1))}{p(x=1)(1-p(x=1))}.\end{aligned}$$

The last expression for κ' is the one presented in Section 3. □

A.3. Proof of Theorem 3

To derive the asymptotic distribution of $\widehat{I}_{\gamma_x}(\widetilde{X};Y)$ we will use the delta method [2, Section 16.1.4], which we formally present in the following lemma.

Lemma 3 *(Delta method). Suppose that cell counts $\boldsymbol{n}=\{n_{x,y}\}$ have a multinomial distribution with cell probabilities $\boldsymbol{p}=\{p(x,y)\}$, $\forall\ x\in\mathcal{X},\ \ y\in\mathcal{Y}$. Let $N=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} n_{x,y}$, and let $\hat{\boldsymbol{p}}$ denote the sample proportions: $\hat{p}(x,y)=n_{x,y}/N$. Let $g(\boldsymbol{p})\in\mathbb{R}$ be a differentiable function, and let $\phi_{x,y}=\frac{\partial g}{\partial p(x,y)}(\boldsymbol{p})$, $\forall\ x\in\mathcal{X},\ \ y\in\mathcal{Y}$. Assume that at least one $\phi_{x,y}$ is nonzero then the distribution $\sqrt{N}\left[g(\hat{\boldsymbol{p}})-g(\boldsymbol{p})\right]$ converges to the normal distribution $\mathcal{N}\left(0,\sigma^2\right)$ when $N\to\infty$, where $\sigma^2=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\phi_{x,y}^2 - \left(\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\phi_{x,y}\right)^2$.*

The expression of the corrected estimator is:

$$\widehat{I}_{\gamma_x}(\widetilde{X};Y) = \sum_{y\in\mathcal{Y}}\left(\gamma_x\,\widehat{p}(y|\widetilde{x}=1)\ln\frac{\widehat{p}(y|\widetilde{x}=1)}{\widehat{p}(y)} + (\widehat{p}(y)-\gamma_x\widehat{p}(y|\widetilde{x}=1))\ln\frac{\widehat{p}(y)-\gamma_x\widehat{p}(y|\widetilde{x}=1)}{\widehat{p}(y)\,(1-\gamma_x)}\right).$$

Since in the expression of $\widehat{I}_{\gamma_x}(\widetilde{X};Y)$ we have the maximum likelihood estimates for the probabilities $p(y)$, $p(y|\widetilde{x}=1)$ the first step is to calculate the partial derivatives of these quantities with respect to the parameters of this model $p(y,\widetilde{x}=1)$ and $p(y,\widetilde{x}=0)$:

$$\begin{aligned}\frac{\partial p(y')}{\partial p(y,\widetilde{x}=1)} &= \delta_{yy'}, & \frac{\partial p(y')}{\partial p(y,\widetilde{x}=0)} &= \delta_{yy'}, \\ \frac{\partial p(y'|\widetilde{x}=1)}{\partial p(y,\widetilde{x}=1)} &= \frac{\delta yy' - p(y'|\widetilde{x}=1)}{p(\widetilde{x}=1)}, & \frac{\partial p(y'|\widetilde{x}=1)}{\partial p(y,\widetilde{x}=0)} &= 0,\end{aligned}$$

where $\delta_{yy'}$ is the Kronecker delta, which takes the value of 1 if $y=y'$ and 0 otherwise. By using the above partial derivatives we have the following results:

$$\phi_{y,\widetilde{x}=1} = \frac{\partial I_{\gamma_x}(\widetilde{X};Y)}{\partial p(y,\widetilde{x}=1)} = \ln \frac{p(y) - p(y|\widetilde{x}=1)\gamma_x}{p(y)} + \frac{\gamma_x}{p(\widetilde{x}=1)} \sum_{y'\in\mathcal{Y}} \left(p(y'|\widetilde{x}=1) - \delta_{yy'}\right) \ln \frac{p(y') - p(y'|\widetilde{x}=1)\gamma_x}{p(y'|\widetilde{x}=1)\gamma_x}$$

$$\phi_{y,\widetilde{x}=0} = \frac{\partial I_{\gamma_x}(\widetilde{X};Y)}{\partial p(y,\widetilde{x}=0)} = \ln \frac{p(y) - p(y|\widetilde{x}=1)\gamma_x}{p(y)}.$$

So by using delta method the asymptotic variance of the estimator equals

$$\sigma^2_{MI_{\gamma_x}} = \sum_{\widetilde{x}\in\widetilde{\mathcal{X}}} \sum_{y\in\mathcal{Y}} p(y,x)\phi^2_{y,x} - \left(\sum_{\widetilde{x}\in\widetilde{\mathcal{X}}} \sum_{y\in\mathcal{Y}} p(y,x)\phi_{y,x}\right)^2$$

$$\sigma^2_{MI_{\gamma_x}} = \sum_{y\in\mathcal{Y}} \left(p(y,\widetilde{x}=1)\phi^2_{y,\widetilde{x}=1} + p(y,\widetilde{x}=0)\phi^2_{y,\widetilde{x}=0}\right) - \left(\sum_{y\in\mathcal{Y}} \left(p(y,\widetilde{x}=1)\phi_{y,\widetilde{x}=1} + p(y,\widetilde{x}=0)\phi_{y,\widetilde{x}=0}\right)\right)^2,$$

where $\phi_{y,\widetilde{x}=1}$ and $\phi_{y,\widetilde{x}=0}$ are calculated earlier and are functions of γ_x. Furthermore, when $\gamma_x = p(x=1)$ it holds that $I_{\gamma_x}(\widetilde{X};Y) = I(X;Y)$ and so the estimator $\widehat{I}_{\gamma_x}(\widetilde{X};Y)$ is asymptotically normally distributed around $I(X;Y)$. □

A.4. Proof of Lemma 2

Using Lemma 1 we can re-write the mutual information $I(X;Z)$ as:

$$\begin{aligned} I(X;Z) &= I_{\gamma_x}(\widetilde{X};Z) \\ &= \sum_{z\in\mathcal{Z}} \left(\gamma_x p(z|\widetilde{x}=1) \ln \frac{p(z|\widetilde{x}=1)}{p(z)} + (p(z) - \gamma_x p(z|\widetilde{x}=1)) \ln \frac{p(z) - \gamma_x p(z|\widetilde{x}=1)}{p(z)\,(1-\gamma_x)}\right) \\ &= \gamma_x p(z=1|\widetilde{x}=1) \ln \frac{p(z=1|\widetilde{x}=1)}{\gamma_z} + (\gamma_z - \gamma_x p(z=1|\widetilde{x}=1)) \ln \frac{\gamma_z - \gamma_x p(z=1|\widetilde{x}=1)}{\gamma_z\,(1-\gamma_x)} \\ &+ \gamma_x p(z=0|\widetilde{x}=1) \ln \frac{p(z=0|\widetilde{x}=1)}{1-\gamma_z} + (1 - \gamma_z - \gamma_x p(z=0|\widetilde{x}=1)) \ln \frac{1 - \gamma_z - \gamma_x p(z=0|\widetilde{x}=1)}{(1-\gamma_z)(1-\gamma_x)} \end{aligned}$$

When both the random variables are non-differentially under-reported it holds

$$p(z=1|\widetilde{x}=1) = \frac{p(\widetilde{x}=1|z=1)p(z=1)}{p(\widetilde{x}=1)} = p(z=1)\frac{p(\widetilde{x}=1|\widetilde{z}=1)}{p(\widetilde{x}=1)} = \gamma_z \frac{p(\widetilde{x}=1,\widetilde{z}=1)}{p(\widetilde{x}=1)p(\widetilde{z}=1)}$$

and

$$p(z=0|\widetilde{x}=1) = 1 - p(z=1|\widetilde{x}=1) = \frac{p(\widetilde{x}=1)p(\widetilde{z}=1) - \gamma_z p(\widetilde{x}=1,\widetilde{z}=1)}{p(\widetilde{x}=1)p(\widetilde{z}=1)}$$

By substituting we get:

$$\begin{aligned} I(X;Z) &= \frac{\gamma_x\gamma_z p(\widetilde{x}=1,\widetilde{z}=1)}{p(\widetilde{x}=1)p(\widetilde{z}=1)} \ln \frac{p(\widetilde{x}=1,\widetilde{z}=1)}{p(\widetilde{x}=1)p(\widetilde{z}=1)} \\ &+ \frac{\gamma_z p(\widetilde{x}=1)p(\widetilde{z}=1) - \gamma_x\gamma_z p(\widetilde{x}=1,\widetilde{z}=1)}{p(\widetilde{x}=1)p(\widetilde{z}=1)} \ln \frac{p(\widetilde{x}=1)p(\widetilde{z}=1) - \gamma_x p(\widetilde{x}=1,\widetilde{z}=1)}{(1-\gamma_x)p(\widetilde{x}=1)p(\widetilde{z}=1)} \\ &+ \frac{\gamma_x p(\widetilde{x}=1)p(\widetilde{z}=1) - \gamma_x\gamma_z p(\widetilde{x}=1,\widetilde{z}=1)}{p(\widetilde{x}=1)p(\widetilde{z}=1)} \ln \frac{p(\widetilde{x}=1)p(\widetilde{z}=1) - \gamma_z p(\widetilde{x}=1,\widetilde{z}=1)}{(1-\gamma_z)p(\widetilde{x}=1)p(\widetilde{z}=1)} \\ &+ \frac{(1-\gamma_x-\gamma_z)p(\widetilde{x}=1)p(\widetilde{z}=1) + \gamma_x\gamma_z p(\widetilde{x}=1,\widetilde{z}=1)}{p(\widetilde{x}=1)p(\widetilde{z}=1)} \ln \frac{(1-\gamma_x-\gamma_z)p(\widetilde{x}=1)p(\widetilde{z}=1) + \gamma_x\gamma_z p(\widetilde{x}=1,\widetilde{z}=1)}{(1-\gamma_z)(1-\gamma_x)p(\widetilde{x}=1)p(\widetilde{z}=1)} \\ &= I_{\gamma_x,\gamma_z}(\widetilde{X};\widetilde{Z}) \qquad \square \end{aligned}$$

Appendix B. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.ijar.2017.04.002.

References

[1] E. Kathleen Adams, Sara Markowitz, Viji Kannan, Patricia M. Dietz, T. Van Tong, Ann M. Malarcher, Reducing prenatal smoking: the role of state policies, Am. J. Prev. Med. 43 (1) (2012) 34–40.

[2] A. Agresti, Categorical Data Analysis, 3rd edition, Wiley–Interscience, 2013.

[3] P.D. Allison, Missing Data, SAGE Publications, Inc., 2001.

[4] D.R. Brillinger, Some data analyses using mutual information, Braz. J. Probab. Stat. 18 (6) (2004) 163–183.

[5] Irwin Bross, Misclassification in 2 × 2 tables, Biometrics 10 (4) (1954) 478–486.

[6] G. Brown, A. Pocock, M. Zhao, M. Lujan, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (2012) 27–66.

[7] H. Chu, Z. Wang, S.R. Cole, S. Greenland, Sensitivity analysis of misclassification: a graphical and a Bayesian approach, Ann. Epidemiol. 16 (11) (2006) 834–841.

[8] Jacob Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd edition, Routledge Academic, 1988.

[9] F. Denis, A. Laurent, R. Gilleron, M. Tommasi, Text classification and co-training from PU examples, in: ICML Workshop: The Continuum from Labeled to Unlabeled Data, 2003.

[10] P. Dietz, D. Homa, L. England, K. Burley, V. Tong, S. Dube, J. Bernert, Estimates of nondisclosure of cigarette smoking among pregnant and nonpregnant women of reproductive age in the US, Am. J. Epidemiol. 173 (3) (2011) 355–359.

[11] Alexis E. Duncan, Julia D. Grant, Kathleen Keenan Bucholz, Pamela A.F. Madden, Andrew C. Heath, Relationship between body mass index, alcohol use, and alcohol misuse in a young adult female twin sample, J. Stud. Alcohol Drugs 70 (3) (2009) 458–466.

[12] J. Edwards, S. Cole, M. Troester, D. Richardson, Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data, Am. J. Epidemiol. 177 (9) (2013) 904–912.

[13] C. Elkan, K. Noto, Learning classifiers from only positive and unlabeled data, in: 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 213–220.

[14] Robert M. Fano, Transmission of Information: A Statistical Theory of Communications, MIT Press Classics, Massachusetts Institute of Technology Press, 1961.

[15] Sarah Connor Gorber, Sean Schofield-Hurwitz, Jill Hardt, Geneviève Levasseur, Mark Tremblay, The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status, Nicotine Tob. Res. 11 (1) (2009) 12–24.

[16] S. Greenland, Variance estimation for epidemiologic effect estimates under misclassification, Stat. Med. 7 (7) (1988) 745–757.

[17] S. Greenland, Sensitivity analysis and bias analysis, in: Handb. of Epidemiology, 2014, pp. 685–706, Ch. 19.

[18] Isabelle Guyon, Constantin F. Aliferis, Gregory F. Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, Alexander R. Statnikov, Design and analysis of the causation and prediction challenge, in: WCCI Causation and Prediction Challenge, 2008, pp. 1–33.

[19] Martin E. Hellman, Josef Raviv, Probability of error, equivocation, and the Chernoff bound, IEEE Trans. Inf. Theory 16 (4) (1970) 368–372.

[20] Alexandros Kalousis, Julien Prados, Melanie Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowl. Inf. Syst. 12 (1) (2007) 95–116.

[21] David D. Lewis, Feature selection and feature extraction for text categorization, in: Proceedings of the Workshop on Speech and Natural Language, Association for Computational Linguistics, 1992, pp. 212–217.

[22] Chunxiao Liao, Wenjing Gao, Weihua Cao, Jun Lv, Canqing Yu, Shengfeng Wang, Bin Zhou, Zengchang Pang, Liming Cong, Zhong Dong, et al., The association of cigarette smoking and alcohol drinking with body mass index: a cross-sectional, population-based study among Chinese adult male twins, BMC Public Health 16 (1) (2016) 1.

[23] K. Mohan, J. Pearl, J. Tian, Graphical models for inference with missing data, in: Advances in Neural Information Processing Systems (NIPS), vol. 26, 2013, pp. 1277–1285.

[24] Victor L. Mote, Richard L. Anderson, An investigation of the effect of misclassification on the properties of χ^2-tests in the analysis of categorical data, Biometrika (1965) 95–109.

[25] Hanchuan Peng, Fuhui Long, Chris Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[26] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence, Dataset Shift in Machine Learning, The MIT Press, 2009.

[27] D. Rahardja, D.M. Young, Confidence intervals for the risk ratio using double sampling with misclassified binomial data, J. Data Sci. 9 (4) (2011) 529–548.

[28] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, J. Wiley & Sons, 2004.

[29] Yvan Saeys, Iñaki Inza, Pedro Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[30] K. Sechidis, B. Calvo, G. Brown, Statistical hypothesis testing in positive unlabelled data, in: ECML/PKDD, Springer, Berlin, Heidelberg, 2014, pp. 66–81.

[31] Konstantinos Sechidis, Hypothesis Testing and Feature Selection in Semi-Supervised Data, PhD thesis, School of Computer Science, University of Manchester, UK, November 2015.

[32] Konstantinos Sechidis, Gavin Brown, Markov blanket discovery in positive-unlabelled and semi-supervised data, in: Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), Springer, 2015, pp. 351–366.

[33] Konstantinos Sechidis, Matt Sperrin, Emily Petherick, Gavin Brown, Markov blanket discovery in positive-unlabelled and semi-supervised data, in: International Conference on Probabilistic Graphical Models (PGM) 2016, J. Mach. Learn. Res. Workshop Conf. Proc. 52 (2016).

[34] Masashi Sugiyama, Machine learning with squared-loss mutual information, Entropy 15 (1) (2012) 80–112.

[35] Eva L.H. Tsui, Gabriel M. Leung, Pauline P.S. Woo, Sarah Choi, Su-Vui Lo, Under-reporting of inpatient services utilisation in household surveys – a population-based study in Hong Kong, BMC Health Serv. Res. 5 (1) (2005) 31.

[36] I.R. White, P. Royston, A.M. Wood, Multiple imputation using chained equations: issues and guidance for practice, Stat. Med. 30 (4) (2011) 377–399.

[37] J. Wright, N. Small, P. Raynor, D. Tuffnell, R. Bhopal, N. Cameron, L. Fairley, D.A. Lawlor, R. Parslow, E.S. Petherick, et al., Cohort profile: the born in Bradford multi-ethnic family cohort study, Int. J. Epidemiol. 42 (4) (2013) 978–991.

[38] P.L. Yudkin, E.H. Burger, D. Bradshaw, P. Groenewald, A.M. Ward, J. Volmink, Deaths caused by HIV disease under-reported in South Africa, AIDS 23 (12) (2009) 1600–1602.