Information graphs for binary predictors

Hughes, G; McRoberts, N; Burnett, FJ

Published in:
Phytopathology

DOI:
10.1094/PHYTO-02-14-0044-R

First published: 01/01/2015

Document Version
Publisher's PDF, also known as Version of record

Information Graphs for Binary Predictors

G. Hughes, N. McRoberts, and F. J. Burnett

First and third authors: Crop and Soil Systems Research Group, SRUC, The King's Buildings, West Mains Road, Edinburgh EH9 3JG, UK; and second author: Plant Pathology Department, University of California, Davis, CA 95616-8751.
Accepted for publication 24 June 2014.

ABSTRACT

Hughes, G., McRoberts, N., and Burnett, F. J. 2015. Information graphs for binary predictors. Phytopathology 105:9-17.

Binary predictors are used in a wide range of crop protection decision-making applications. Such predictors provide a simple analytical apparatus for the formulation of evidence related to risk factors, for use in the process of Bayesian updating of probabilities of crop disease. For diagrammatic interpretation of diagnostic probabilities, the receiver operating characteristic is available. Here, we view binary predictors from the perspective of diagnostic information. After a brief introduction to the basic information theoretic concepts of entropy and expected mutual information, we use an example data set to provide diagrammatic interpretations of expected mutual information, relative entropy, information inaccuracy, information updating, and specific information. Our information graphs also illustrate correspondences between diagnostic information and diagnostic probabilities.

Additional keywords: diagnosis, disease management, entropy, information theory.

Information theory, as formulated by Shannon and Weaver (30), is the mathematical basis for the study of communication through noisy channels. Fortunately, Shannon's analysis does not impose upon us too strict an interpretation of what constitutes a channel, thus allowing information theory to be applied in disciplines beyond those directly concerned with the technical problems of data transmission. Here we apply some information theoretic analysis to the description of a class of predictors widely used in crop protection decision making. To be specific, we discuss binary predictors, typically derived by selection of an appropriate operational threshold on a receiver operating characteristic (ROC) curve (e.g., 1,6,10,12,15,24,25,33,34).

We consider a scenario in which crops either require crop protection measures or do not. A *predictor* (*test* and *forecaster* are synonymous) is any kind of diagnostic apparatus by means of which we may classify crops in terms of their requirement for protection measures. We predict this requirement (rather than measure it directly) because the objective is to use crop protection measures at a sufficiently early stage to prevent disease later developing to an economically significant level. With a binary predictor, we have two outcome categories; one indicative of the need for protection, the other indicative of no need for protection. We can think of such a (binary) predictor as a (binary) channel because it conveys a *message* (*prediction*, *test outcome*, *forecast* are synonymous) to the decision maker. We describe the channel as noisy because, typically, predictors are imperfect. This means that for some crops that actually require protection measures, a message may be conveyed that there is no need for protection; and for some crops that actually do not require protection measures, a message may be conveyed that there is a need for protection. In a crop protection context, these two kinds of noise are often characterized as, respectively, the false negative proportion and the false positive proportion (e.g., 35). Thus, as was Shannon, we are concerned with accuracy of information transfer from sender (the predictor) to receiver (the decision maker).

Information graphs were used to describe predictors in the context of clinical diagnosis at least as early as studies by Diamond et al. (11) and Somoza and Mossman (31), but Benish (5) is the paradigmatic account. Benish (5) used information graphs as a diagrammatic method of representing and comparing the properties of diagnostic tests. Here, we describe a number of information graphs for binary predictors, building both on Benish's work (5) and on more recent epidemiological applications of information theory in a phytopathological context (16,18).

The article is set out as follows. An introduction to the basic information theoretic concepts of entropy and expected mutual information is provided, followed by an outline of an example data set taken from the phytopathological literature. These introductory sections also set out some essential notation. Information graphs for relative entropy, information inaccuracy, information updating and specific information are then presented both analytically and numerically via calculations based on the example data set. Reference is made to the correspondence between diagnostic information and diagnostic probabilities. A discussion describes relationships between the information quantities illustrated in the information graphs, and briefly sets this material in the wider context of the application of predictors in diagnostic decision making.

THEORY AND APPROACHES

Information theoretic background. The generic probability of an event E_x is $\Pr(E_x)$, $0 \le \Pr(E_x) \le 1$. Then $h(\Pr(E_x)) = \ln(1/\Pr(E_x)) = -\ln(\Pr(E_x))$ is the information content of a message that conveys, without error, that the event E_x has occurred (thus, in the present context, such a message would constitute a perfect prediction). If $\Pr(E_x)$ is small, the information content of this message is large, and vice versa. We will use natural logarithms throughout, in which case the *nit* is the unit of information (32). Note that Benish (5) uses logarithms with base 2, in which case the *bit* is the unit of information. In order to convert from nits to bits, the information quantity as specified in nits is divided by ln(2).

Corresponding author: G. Hughes; E-mail address: gareth.hughes@sruc.ac.uk

http://dx.doi.org/10.1094/PHYTO-02-14-0044-R

We think of the event E_x as one of the possible messages that may be received by a decision maker. At the outset, we know the number of possible messages, but not which one will be received. Here, we are restricting our attention here to binary predictors, so generically there are two possible messages, E_1 and E_2, with probabilities $\Pr(E_1)$ and $\Pr(E_2) = 1 - \Pr(E_1)$. We cannot calculate the information content of a message until the message is received, because the message 'E_x occurred' may refer to either E_1 or E_2. However, we can calculate expected information content, referred to as the (Shannon) *entropy*, before the message is received, as follows. Since the message 'E_x occurred' is received with probability $\Pr(E_x)$, entropy, denoted $H(E) = -\sum_x \Pr(E_x) \cdot \ln(\Pr(E_x))$. We take $0 \cdot \ln(0) = 0$ and note that $H(E) \geq 0$. This quantity is the weighted average of the information contents of the possible messages. If $\Pr(E_x) = 1$ for either of the possible messages, $H(E) = 0$; i.e., we expect nothing from a message if we are already certain of the actual outcome. For a binary predictor, entropy has its maximum value ($H(E) = \ln(2)$ nits) when $\Pr(E_1) = \Pr(E_2)$; i.e., a message that tells us what actually happened has a larger expected information content when both outcomes are equally probable than when one outcome is more probable than the other. In the present context, entropy $H(E)$ can be thought of as characterizing either information or uncertainty (28). Thus, the situation at the outset is that either one of two events, E_1 or E_2, will occur, with corresponding probabilities $\Pr(E_1)$ and $\Pr(E_2)$. Entropy characterizes the amount of information obtained, on average, from use of a perfect predictor. Alternatively, entropy characterizes the extent of our uncertainty before use of the predictor.

A *prediction-realization table* is a table in which the rows refer to the predictions (categories of test outcome), the columns to the realizations (categories of actual status). For a binary predictor, the actual status of a crop is described in one of two categories (D_j, $j = 1,2$). D_1 denotes that the actual status is of a requirement for treatment (referred to as a *case*), D_2 denotes that the actual status is of no requirement for treatment (referred to as a *control*). The rows of the table comprise two categories (T_i, $i = 1,2$). T_1 denotes a prediction of requirement for treatment, T_2 denotes a prediction of no requirement for treatment. Theil (32) uses prediction-realization tables to refer both to the cross-tabulated frequencies of predictions and realizations, and to the estimated probabilities obtained by normalization of those frequencies. For a phytopathological example showing cross-tabulated frequencies, see Figure 2 in Capote et al. (8).

In a normalized prediction-realization table (Table 1), the bottom row of the table contains the distribution $\Pr(D_j)$, from which we may calculate $H(D)$ using

$$H(D) = -\sum_j \Pr(D_j) \cdot \ln(\Pr(D_j)) \quad (1)$$

The right column of the table contains the distribution $\Pr(T_i)$, from which we may calculate $H(T)$ using

$$H(T) = -\sum_i \Pr(T_i) \cdot \ln(\Pr(T_i)) \quad (2)$$

The body of the table contains the joint probability distribution $\Pr(D_j \cap T_i)$ from which we may calculate the joint entropy $H(D,T)$ using

$$H(D,T) = -\sum_i \sum_j \Pr(D_j \cap T_i) \cdot \ln(\Pr(D_j \cap T_i)) \quad (3)$$

where $\Pr(D_j \cap T_i) = \Pr(T_i|D_j) \cdot \Pr(D_j) = \Pr(D_j|T_i) \cdot \Pr(T_i)$ (Bayes' theorem). If the test outcome were independent of actual status (which is not what we want), we would have $H(D,T) = H(D) + H(T)$. When the joint entropy is smaller than the sum of the individual entropies, this indicates association between test outcome and actual status (which is what we want). Then $H(D,T) = H(D) + H(T) - I_M(D,T)$; where the *expected mutual information*, denoted $I_M(D,T)$, is a measure of the association. We can write

$$I_M(D,T) = \sum_i \sum_j \Pr(D_j \cap T_i) \cdot \ln\left(\frac{\Pr(D_j \cap T_i)}{\Pr(D_j) \cdot \Pr(T_i)}\right) \quad (4)$$

and $I_M(D,T) \geq 0$, with equality only if D and T are independent. The term

$$\ln\left(\frac{\Pr(D_j \cap T_i)}{\Pr(D_j) \cdot \Pr(T_i)}\right)$$

is referred to as the *pointwise mutual information*.

An example data set. We will use data from a study of Fusarium head blight (FHB) of winter wheat in North America (21). These data were presented in the context of a discussion of decision making in epidemiology, including the derivation of a binary predictor and its use in making Bayesian probability revisions. In particular, we refer to the data denoted Scenario A in Madden (21). In view of the comprehensive description of the data and their use in a decision-making context in Madden (21), we repeat as little as possible of that material here and encourage interested readers to consult the original source. The exception is that we will summarize Madden's (21) account of the assessment of predictor accuracy in order to make clear how this relates to the analysis described in the present study.

Our starting point is a normalized prediction-realization table for the FHB data (Table 2). From Table 2 we may calculate conditional probabilities representing the properties of the binary predictor as follows: $\Pr(T_1|D_1)$ (true positive proportion, sensitivity, *TPP*) = 0.30/0.36 = 0.833, $\Pr(T_2|D_2)$ (true negative proportion, specificity, *TNP*) = 0.54/0.64 = 0.844, and $\Pr(T_2|D_1)$ (false negative proportion, *FNP*) = 1 – *TPP* = 0.167, $\Pr(T_1|D_2)$ (false positive proportion, *FPP*) = 1 – *TNP* = 0.156. Sensitivity and specificity represent two kinds of accuracy, respectively, for cases and controls. Sensitivity and specificity are independent of the proportions of cases and controls in a data set and can therefore be viewed as properties of a test (16). The sensitivity and specificity values noted above are consequent on the choice of a particular threshold value to distinguish predicted cases from predicted controls on the basis of the FHB risk algorithm. The particular threshold adopted here represents a balanced predictor at which both sensitivity and specificity are reasonably high (see Figure 3 in reference 21).

Here, for the present analysis, we refer to $\Pr(D_1)$ as the prior probability of an epidemic and from Table 2 take $\Pr(D_1) = 0.36$

TABLE 1. The prediction-realization table for a predictor with two categories of actual status D_j, $j = 1,2$; and two categories of predicted status T_i, $i = 1,2$[a]

	Realization, D_j		
Prediction, T_i	D_1	D_2	Row sums
T_1	$\Pr(T_1 \cap D_1)$	$\Pr(T_1 \cap D_2)$	$\Pr(T_1)$
T_2	$\Pr(T_2 \cap D_1)$	$\Pr(T_2 \cap D_2)$	$\Pr(T_2)$
Column sums	$\Pr(D_1)$	$\Pr(D_2)$	1

[a] In the body of the table are the joint probabilities denoted $\Pr(T_i \cap D_j)$.

TABLE 2. The normalized prediction-realization table for Fusarium head blight scenario A from Madden (21), based on 50 location-year observations[a]

	Realization, D_j		
Prediction, T_i	D_1	D_2	Row sums
T_1	0.30	0.10	0.40
T_2	0.06	0.54	0.60
Column sums	0.36	0.64	1.00

[a] This describes a predictor with two categories of actual status D_j, $j = 1,2$; and two categories of predicted status T_i, $i = 1,2$. The joint probabilities $\Pr(T_i \cap D_j)$ are shown in the body of the table.

(so $h(\Pr(D_1)) = -\ln(\Pr(D_1)) = 1.022$ nits) and $\Pr(D_2) = 1 - \Pr(D_1) = 0.64$ (so $h(\Pr(D_2)) = -\ln(\Pr(D_2)) = 0.446$ nits). Thus, in this case the prior probability has been calculated from the data, as in Madden (21). The posterior probabilities are (via Bayes' theorem, see reference 21) $\Pr(D_1|T_1) = 0.75$, $\Pr(D_2|T_1) = 0.25$, $\Pr(D_2|T_2) = 0.90$, and $\Pr(D_1|T_2) = 0.10$. Then we calculate the information quantities $H(D) = 0.653$ nits (equation 1), $H(T) = 0.673$ nits (equation 2), $H(D,T) = 1.093$ nits (equation 3) and $I_M(D,T) = 0.233$ nits (equation 4), and note that $I_M(D,T) = H(D) + H(T) - H(D,T)$.

Introduction to information graphs. As described by Benish (5), information graphs are graphical plots of information quantities expressed as a function of prior probability $\Pr(D_1)$. Thus, an information graph for entropy $H(D)$ is provided directly by equation 1 (Fig. 1). The information graph for entropy is symmetrical about $\Pr(D_1) = 0.5$. Expected mutual information $I_M(D,T)$ can be written as a function of prior probability $\Pr(D_1)$ in a number of different formats, for example:

$$I_M(D,T) = H(TPP \cdot \Pr(D_1) + FPP \cdot \Pr(D_2)) - H(TPP) \cdot \Pr(D_1) - H(FPP) \cdot \Pr(D_2) \quad (5)$$

(equation 1 in ref. 2; see also equation 7 in ref. 5; equation 20 in ref. 16) (Fig. 1). The information graph for expected mutual information may be symmetrical about $\Pr(D_1) = 0.5$, but is not necessarily so (Fig. 1; see also Fig. 2 in ref. 5, Fig. 5 in ref. 16).

Our interpretation of the information graphs shown in Figure 1 is that at the use of an imperfect binary predictor T at any specified prior probability $\Pr(D_1)$ on average reduces $H(D)$ (uncertainty about D) by $I_M(D,T)$ (expected mutual information). Thus, we can interpret the vertical difference between $H(D)$ and $I_M(D,T)$ at any specified prior probability $\Pr(D_1)$ as the average remaining uncertainty about D after use of T. This latter quantity is the conditional entropy $H(D|T)$:

$$H(D|T) = -\sum_i \Pr(T_i) \sum_j \Pr(D_j|T_i) \cdot \ln(\Pr(D_j|T_i)) = \sum_i \Pr(T_i) H(D|T_i) \quad (6)$$

For the example data set, we calculate $H(D|T) = 0.420$ nits at $\Pr(D_1) = 0.36$ and note $I_M(D,T) = H(D) - H(D|T)$.

Now, consider a predictor such that D and T are identical, so that use of the predictor T accounts for all the uncertainty in D. Then $H(D|T) = H(D|D)$ and $I_M(D,T) = H(D) - H(D|D) = H(D)$. Thus, the upper limit of the expected mutual information $I_M(D,T)$ is the entropy $H(D)$, and $I_M(D,T) = H(D)$ would characterize a perfect predictor. Also, we have $H(D) - H(D|T) = I_M(D,T) \geq 0$, so $H(D|T) \leq H(D)$ with equality only if T and D are independent. Thus, *on average*, if T and D are not independent, use of a predictor T will decrease uncertainty in D.

RESULTS

An information graph for relative entropy. Our starting point for the analysis leading to equation 1 for entropy was to consider the information content of a message that constitutes a perfect prediction. For relative entropy we consider instead the information content of a message that transforms a set of prior probabilities into a corresponding set of posterior probabilities. As previously, the prior probabilities of actual status categories D_1 and D_2 are denoted $\Pr(D_1)$ and $\Pr(D_2)$, respectively. A message T_i is received which serves to transform these prior probabilities into the posterior probabilities $\Pr(D_j|T_i)$, with $\sum_j \Pr(D_j|T_i) = 1$, $\Pr(D_j|T_i) \geq 0$, $j = 1,2$. The information content of this message with respect to actual status D_j is

$$\text{information content of } T_i = \ln\left(\frac{\Pr(D_j|T_i)}{\Pr(D_j)}\right) \quad (7)$$

Note that this includes the message that constitutes a perfect prediction (i.e., $\Pr(D_j|T_i) = 1$) as a special case. Then the expected information content of the message T_i is the weighted average of the information contents, the weights being the posterior probabilities $\Pr(D_j|T_i)$. This is referred to here as *relative entropy*, denoted $I(T_i)$:

$$I(T_i) = \sum_j \Pr(D_j|T_i) \cdot \ln\left(\frac{\Pr(D_j|T_i)}{\Pr(D_j)}\right) \quad (8)$$

Relative entropy, also widely known as the Kullback-Leibler divergence (19), can be thought of as a measure of the divergence between two probability distributions (9); in this case between the posterior probability distribution (taken as the comparison distribution) and the prior distribution (taken as the reference distribution). For a binary predictor we can calculate two versions of equation 8, respectively for outcomes T_i ($i = 1,2$). For the present example, based on Table 2, we obtain $I(T_1) = 0.315$ nits, $I(T_2) = 0.179$ nits (equation 8). Generally, $I(T_i) \geq 0$ with equality only if $\Pr(D_j|T_i) = \Pr(D_j)$ for all j; thus, the expected information content of a message which leaves the prior probabilities unchanged is equal to zero, which is reasonable. We can interpret relative entropy as a measure of diagnostic information (4,5): $I(T_i)$ is a measure of the amount of information provided by outcome T_i on average over both categories of actual status, cases D_1 and controls D_2. The expected value of relative entropy $I(T_i)$ over both outcomes of a binary predictor is expected mutual information:

$$I_M(D,T) = \sum_i \Pr(T_i) \cdot I(T_i) \quad (9)$$

Figure 2 shows an information graph for relative entropy, based on the example set. To begin, we consider entropy from the point of view of the prior probability $\Pr(D_1)$; the corresponding entropy curve is denoted $H(D)$ (equation 1). We draw the tangent to the entropy curve at prior probability $\Pr(D_1) = 0.36$ (slope = 0.575, intercept = 0.446). Then, we consider entropy from the point of

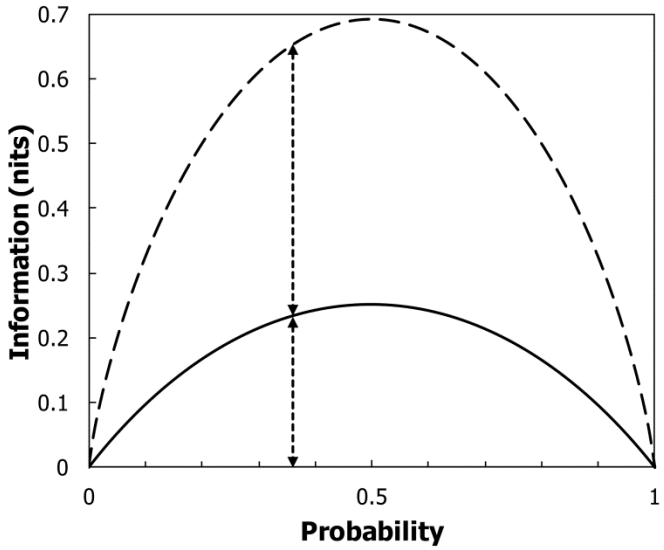

Fig. 1. Information graph for entropy and expected mutual information. The long-dashed line shows the Shannon entropy curve: information axis $H(D)$, probability axis $\Pr(D_1)$ (equation 1). The solid line shows the expected mutual information curve: information axis $I_M(D,T)$, probability axis $\Pr(D_1)$ (equation 5 with $TPP = 0.833$, $FPP = 0.156$). The short-dashed line between the expected mutual information curve and the horizontal axis shows $I_M(D,T) = 0.233$ nits at prior probability $\Pr(D_1) = 0.36$. The short-dashed line between the entropy curve and the expected mutual information curve shows $H(D|T) = 0.420$ nits at prior probability $\Pr(D_1) = 0.36$. Values refer to calculations based on the example data set (Table 2).

view of the posterior probabilities $\Pr(D_j|T_i)$; the corresponding conditional entropy curve is denoted $H(D|T_i)$:

$$H(D|T_i) = -\sum_j \Pr(D_j|T_i) \cdot \ln(\Pr(D_j|T_i)) \quad (10)$$

The two entropy curves are of course identical; only our point of view has changed, from prior to posterior probability. Now, at posterior probability $\Pr(D_1|T_1) = 0.75$, the vertical distance between the tangent and the conditional entropy curve $H(D|T_i)$ is $I(T_1) = 0.315$ nits. The corresponding vertical distance between the conditional entropy curve $H(D|T_i)$ and the horizontal axis is $H(D|T_1) = 0.562$ nits. At posterior probability $\Pr(D_1|T_2) = 0.10$, the vertical distance between the tangent and the conditional entropy curve $H(D|T_i)$ is $I(T_2) = 0.179$ nits. The corresponding vertical distance between the conditional entropy curve $H(D|T_i)$ and the horizontal axis is $H(D|T_2) = 0.325$ nits. Points on the tangent line thus represent the quantity $I(T_i) + H(D|T_i)$, referred to as the *cross-entropy* $H_C(D|T_i, D)$ (23). From equations 8 and 10 we have

$$H_C(D|T_i, D) = I(T_i) + H(D|T_i) = \sum_j \Pr(D_j|T_i) \cdot \ln\left(\frac{1}{\Pr(D_j)}\right) \quad (11)$$

An information graph for information inaccuracy. An imperfect binary predictor provides us with outcomes T_1 and T_2. T_1 is a good prediction if the actual status subsequently turns out to be D_1, but not if the actual status turns out to be D_2. T_2 is a good prediction if the actual status subsequently turns out to be D_2, but not if the actual status turns out to be D_1. Theil (35) discusses this as information inaccuracy, in the context of forecast evaluation. For the example data set, consider a T_1 prediction. If the actual status is subsequently revealed to be D_1 (i.e., the prediction was correct), we can calculate the divergence:

$$1 \cdot \ln\left(\frac{1}{\Pr(D_1|T_1)}\right) + 0 \cdot \ln\left(\frac{0}{\Pr(D_2|T_1)}\right) = \ln\left(\frac{1}{0.75}\right) = 0.288 \text{ nits}$$

Alternatively, if the actual status is subsequently revealed to be D_2 (i.e., the prediction was incorrect), we can calculate the divergence:

$$0 \cdot \ln\left(\frac{0}{\Pr(D_1|T_1)}\right) + 1 \cdot \ln\left(\frac{1}{\Pr(D_2|T_1)}\right) = \ln\left(\frac{1}{0.25}\right) = 1.386 \text{ nits}$$

Similarly, for a T_2 prediction, if the actual status is subsequently revealed to be D_2 (i.e., the prediction was correct), we can calculate the divergence:

$$1 \cdot \ln\left(\frac{1}{\Pr(D_2|T_2)}\right) + 0 \cdot \ln\left(\frac{0}{\Pr(D_1|T_2)}\right) = \ln\left(\frac{1}{0.9}\right) = 0.105 \text{ nits}$$

If the actual status is subsequently revealed to be D_1 (i.e., the prediction was incorrect), we can calculate the divergence:

$$0 \cdot \ln\left(\frac{0}{\Pr(D_2|T_2)}\right) + 1 \cdot \ln\left(\frac{1}{\Pr(D_1|T_2)}\right) = \ln\left(\frac{1}{0.1}\right) = 2.303 \text{ nits}$$

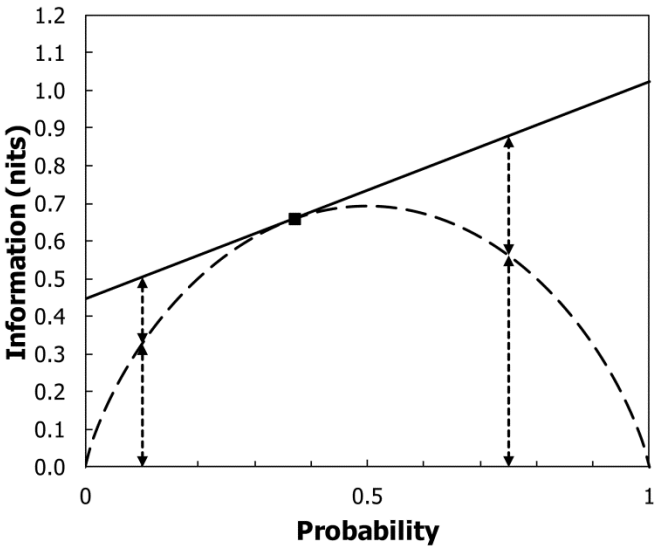

Fig. 2. Information graph for relative entropy. The basis is a generic graph which shows information quantities on the vertical axis and probabilities on the horizontal axis. We can depict different information quantities on the vertical axis by adopting different reference probabilities as our point of view on the horizontal axis. (i) Reference point of view: prior probability $\Pr(D_1)$. The long-dashed line is interpreted as the entropy curve: information axis $H(D)$, probability axis $\Pr(D_1)$ (equation 1). The solid line shows the tangent to the entropy curve at prior probability $\Pr(D_1) = 0.36$ (point marked ■). The tangent line (slope = $-\ln(0.36/(1-0.36)) = 0.575$, intercept = 0.446) depicts cross-entropy (equation 11). (ii) Reference point of view: posterior probability $\Pr(D_1|T_i)$. The long-dashed line is interpreted as the conditional entropy curve: information axis $H(D|T_i)$, probability axis $\Pr(D_1|T_i)$ (equation 10). At posterior probability $\Pr(D_1|T_1) = 0.75$, the short-dashed line between the tangent and the conditional entropy curve shows relative entropy $I(T_1) = 0.315$ nits; the short-dashed line between the conditional entropy curve and the horizontal axis shows $H(D|T_1) = 0.562$ nits. At posterior probability $\Pr(D_1|T_2) = 0.10$, the short-dashed line between the tangent and the conditional entropy curve shows relative entropy $I(T_2) = 0.179$ nits; the short-dashed line between the conditional entropy curve and the horizontal axis shows $H(D|T_2) = 0.325$ nits. Calculations are based on the example data set (Table 2).

Figure 3 shows an information graph for information inaccuracy, based on the example set. To begin, we consider entropy from the point of view of the posterior probabilities $\Pr(D_j|T_i)$; the corresponding conditional entropy curve is denoted $H(D|T_i)$ (equation 10). We draw one tangent to the conditional entropy curve at posterior probability $\Pr(D_1|T_1) = 0.75$ (with slope = –1.099, intercept = 1.386) and another at posterior probability $\Pr(D_1|T_2) = 0.10$ (with slope = 2.197, intercept = 0.105) (Fig. 3). To read Figure 3, observe the following: the tangent at $\Pr(D_1|T_1) = 0.75$ (complement $\Pr(D_2|T_1) = 0.25$) characterizes T_1 predictions; the tangent at $\Pr(D_1|T_2) = 0.10$ (complement $\Pr(D_2|T_2) = 0.90$) characterizes T_2 predictions; the right vertical axis refers to cases (actual D_1 subjects); the left vertical axis refers to controls (actual D_2 subjects). Thus, a correct T_1 prediction is characterized by intersection of the tangent at $\Pr(D_1|T_1) = 0.75$ on the right vertical axis, which is 0.288 nits. An incorrect T_1 prediction is characterized by the intersection of the tangent at $\Pr(D_1|T_1) = 0.75$ (i.e., $\Pr(D_2|T_1) = 0.25$) on the left vertical axis, which is 1.386 nits. A correct T_2 prediction is characterized by the intersection of the tangent at $\Pr(D_1|T_2) = 0.10$ (i.e., $\Pr(D_2|T_2) = 0.90$) on the left vertical axis, which is 0.105 nits. An incorrect T_2 prediction is characterized by the intersection of the tangent at $\Pr(D_1|T_2) = 0.10$ on the right vertical axis, which is 2.303 nits. We observe that a T_1 prediction has a low information inaccuracy when the actual status is D_1 and a high information inaccuracy when the actual status is D_2. Similarly, a T_2 prediction has a low information inaccuracy when the actual status is D_2 and a high information inaccuracy when the actual status is D_1. Note that for a binary predictor (i = 1,2; j = 1,2), the terms $-\ln(\Pr(D_j|T_i))$ (for a correct prediction, i.e., $i = j$) and $-\ln(\Pr(D_j|T_i))$ (for an incorrect prediction, i.e., $i \neq j$) arise as a scoring rule (e.g., 20), often invoked without any detailed reference to the analysis outlined above.

An information graph for information updating. Madden (21) discusses in detail the use of Bayes' theorem in the context of crop protection decision making. In essence, Bayes' theorem allows us to update the prior (pre-test) probability of an epidemic

(or of the need for a control intervention) to a corresponding posterior (post-test) probability, after taking account of evidence in the form of data related to risk factors. As discussed by Madden (21), a predictor is in effect a device that combines data related to risk factors in such a way as to allow Bayes' theorem to be applied for the purpose of probability updating. Here, we characterize the information contents corresponding to prior and posterior probabilities. Write equation 7 as follows: *information content of* $T_i = -\ln(\Pr(D_j)) - (-\ln(\Pr(D_j|T_i))$. Based on the example data set, for the priors $\Pr(D_j)$ we have: $-\ln(\Pr(D_1)) = -\ln(0.36) = 1.022$ nits and $-\ln(\Pr(D_2)) = -\ln(0.64) = 0.446$ nits. For the posteriors $\Pr(D_j|T_i)$ we have: $-\ln(\Pr(D_1|T_1)) = -\ln(0.75) = 0.288$ nits and $-\ln(\Pr(D_1|T_2)) = -\ln(0.10) = 2.303$ nits; $-\ln(\Pr(D_2|T_1)) = -\ln(0.25) = 1.386$ nits and $-\ln(\Pr(D_2|T_2)) = -\ln(0.90) = 0.105$ nits.

Figure 4 shows an information graph for information updating, based on the example set. In Figure 4, Bayesian probability updating is characterized on the horizontal axis, while the corresponding information updating is characterized on the vertical axis. To begin, we consider information from the point of view of the prior probability $\Pr(D_j)$; the corresponding information curve is denoted $h(\Pr(D_j))$. Then, we consider information from the point of view of the posterior probabilities $\Pr(D_j|T_i)$; the corresponding information curve is denoted $h(\Pr(D_j|T_i))$. The two information curves are of course identical; only our point of view has changed, from prior to posterior probability. In Figure 4A, the prior probability is $\Pr(D_1) = 0.36$ and the posterior probability may be $\Pr(D_1|T_1) = 0.75$ or $\Pr(D_1|T_2) = 0.10$. The information contents resulting from information updating are then, respectively:

$-\ln(\Pr(D_1)) - (-\ln(\Pr(D_1|T_1)) = 1.022 - 0.288 = 0.734$ nits

$-\ln(\Pr(D_1)) - (-\ln(\Pr(D_1|T_2)) = 1.022 - 2.303 = -1.281$ nits

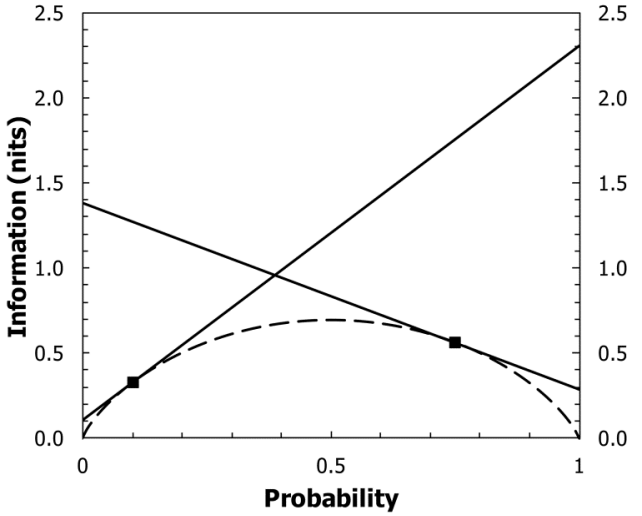

Fig. 3. Information graph for information inaccuracy. The basis is a generic graph which shows information quantities on the vertical axis and probabilities on the horizontal axis. We can depict different information quantities on the vertical axis by adopting different reference probabilities as our point of view on the horizontal axis. (i) Reference point of view: posterior probability $\Pr(D_1|T_i)$. The long-dashed line is interpreted as the conditional entropy curve: information axis $H(D|T_i)$, probability axis $\Pr(D_1|T_i)$ (equation 10). The solid lines show the tangents to the conditional entropy curve at posterior probabilities $\Pr(D_1|T_1) = 0.75$ and $\Pr(D_1|T_2) = 0.10$ (points marked ■). (ii) At the vertical axis where the (horizontal) probability axis = 1, actual status = D_1; at the vertical axis where the (horizontal) probability axis = 0, actual status = D_2. The tangent at posterior probability $\Pr(D_1|T_1) = 0.75$ (slope = $-\ln(0.75/(1-0.75)) = -1.099$, intercept = 1.386) intersects the vertical axis where the probability axis = 1 (for a correct T_1 outcome) at $-\ln(0.75) = 0.288$ nits; and intersects the vertical axis where the probability axis = 0 (for an incorrect T_1 outcome) at $-\ln(0.25) = 1.386$ nits. The tangent at posterior probability $\Pr(D_1|T_2) = 0.10$ (slope = $-\ln(0.10/(1-0.10)) = 2.197$, intercept = 0.105) intersects the vertical axis where the probability axis = 1 (for an incorrect T_2 outcome) at $-\ln(0.10) = 2.303$ nits; and intersects the vertical axis where the probability axis = 0 (for a correct T_2 outcome) at $-\ln(0.90) = 0.105$ nits. The information quantities calculated represent information inaccuracy; lower values correspond to more accurate test outcomes. Calculations are based on the example data set (Table 2).

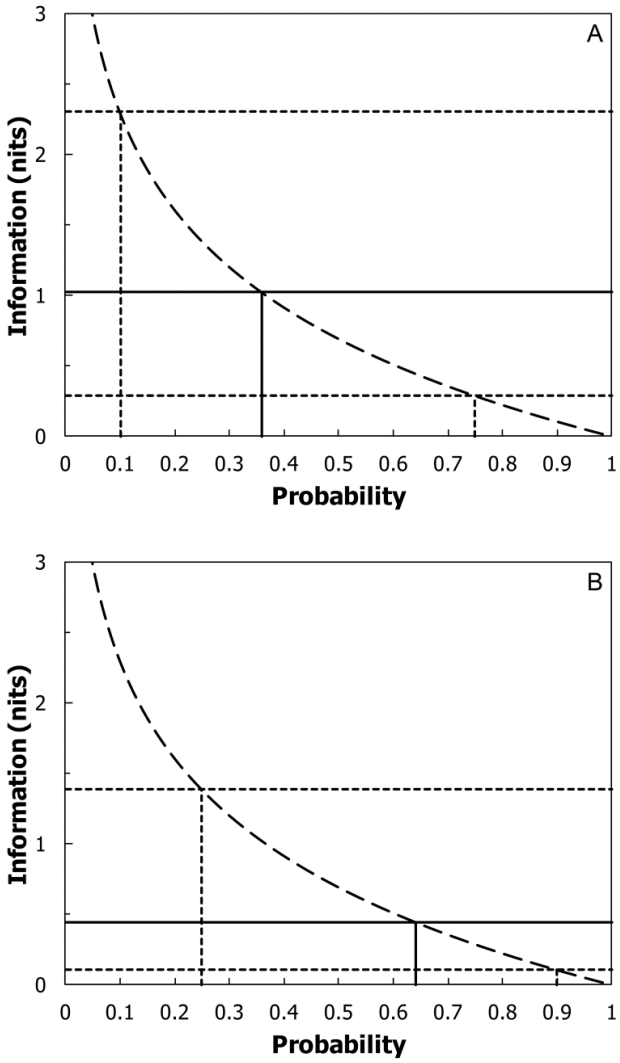

Fig. 4. Information graphs for information updating. The basis is a generic graph which shows information quantities on the vertical axis and probabilities on the horizontal axis. We can depict different information quantities on the vertical axis by adopting different reference probabilities as our point of view on the horizontal axis. In both parts, the long-dashed line shows the generic information curve: information axis $-\ln(\Pr(E_x))$, probability axis $\Pr(E_x)$; $\Pr(E_x)$ and the corresponding information quantity may refer to prior or posterior probability. Values refer to calculations based on the example data set (Table 2). **A,** The probability axis refers to prior probability $\Pr(D_1)$ and posterior probabilities $\Pr(D_1|T_1)$ and $\Pr(D_1|T_2)$. The solid vertical line at prior probability $\Pr(D_1) = 0.36$ intersects the information curve at $-\ln(\Pr(D_1)) = 1.022$ nits. The short-dashed vertical line at posterior probability $\Pr(D_1|T_1) = 0.75$ intersects the information curve at $-\ln(\Pr(D_1|T_1)) = 0.288$ nits, and then $-\ln(\Pr(D_1)) - (-\ln(\Pr(D_1|T_1))) = 1.022 - 0.288 = 0.734$ nits. The short-dashed vertical line at posterior probability $\Pr(D_1|T_2) = 0.10$ intersects the information curve at $-\ln(\Pr(D_1|T_2)) = 2.303$ nits, and then $-\ln(\Pr(D_1)) - (-\ln(\Pr(D_1|T_2))) = 1.022 - 2.303 = -1.281$ nits. **B,** The probability axis refers to prior probability $\Pr(D_2)$ and posterior probabilities $\Pr(D_2|T_2)$ and $\Pr(D_2|T_1)$. The solid vertical line at prior probability $\Pr(D_2) = 0.64$ intersects the information curve at $-\ln(\Pr(D_2)) = 0.446$ nits. The short-dashed vertical line at posterior probability $\Pr(D_2|T_2) = 0.90$ intersects the information curve at $-\ln(\Pr(D_2|T_2)) = 0.105$ nits, and then $-\ln(\Pr(D_2)) - (-\ln(\Pr(D_2|T_2))) = 0.446 - 0.105 = 0.341$ nits. The short-dashed vertical line at posterior probability $\Pr(D_2|T_1) = 0.25$ intersects the information curve at $-\ln(\Pr(D_2|T_1)) = 1.386$ nits, and then $-\ln(\Pr(D_2)) - (-\ln(\Pr(D_2|T_1))) = 0.446 - 1.386 = -0.940$ nits.

In Figure 4B, the prior probability is $\Pr(D_2) = 0.64$ and the posterior probability may be $\Pr(D_2|T_1) = 0.25$ or $\Pr(D_2|T_2) = 0.90$. The information contents resulting from information updating are then, respectively:

$$-\ln(\Pr(D_2)) - (-\ln(\Pr(D_2|T_1)) = 0.446 - 1.386 = -0.940 \text{ nits}$$

$$-\ln(\Pr(D_2)) - (-\ln(\Pr(D_2|T_2)) = 0.446 - 0.105 = 0.341 \text{ nits}$$

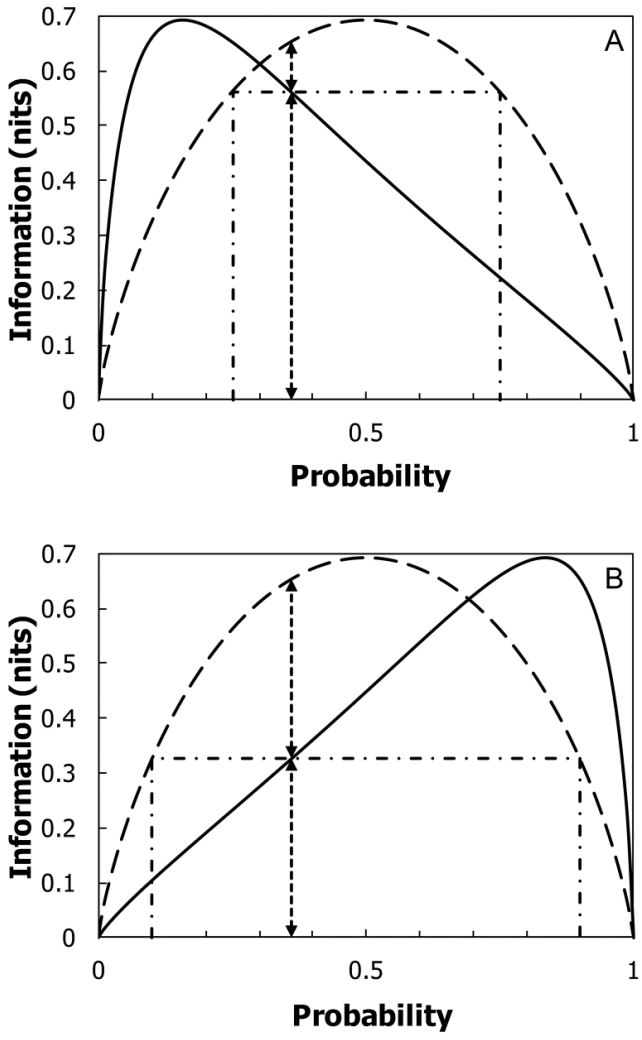

Fig. 5. Information graphs for specific information. In both parts, the long-dashed line shows the entropy curve: information axis $H(D)$, probability axis $\Pr(D_1)$ (equation 1); and the solid line shows the conditional entropy curve: information axis $H(D|T_i)$ (equation 10), probability axis $\Pr(D_1)$. All values refer to calculations based on the example data set (Table 2). **A,** Test outcome T_1. At prior probability $\Pr(D_1) = 0.36$, the short-dashed line between the entropy curve $H(D)$ and the conditional entropy curve $H(D|T_1)$ shows specific information $I_S(T_1) = 0.091$ nits, the short-dashed line between the conditional entropy curve and the horizontal axis shows $H(D|T_1) = 0.562$ nits (equation 12). The horizontal dash-dot line through the point of intersection of the conditional entropy curve $H(D|T_1)$ and prior probability $\Pr(D_1) = 0.36$ extends in each direction to intersect the entropy curve $H(D)$. The Bayesian posterior probabilities $\Pr(D_1|T_1) = 0.75$ (upper intersection) and $\Pr(D_2|T_1) = 0.25$ (lower intersection) are obtained as the projections onto the probability axis of these points of intersection. **B,** Test outcome T_2. At prior probability $\Pr(D_1) = 0.36$, the short-dashed line between the entropy curve $H(D)$ and the conditional entropy curve $H(D|T_2)$ shows specific information $I_S(T_2) = 0.328$ nits, the short-dashed line between the conditional entropy curve and the horizontal axis shows $H(D|T_2) = 0.325$ nits (equation 12). The horizontal dash-dot line through the point of intersection of the conditional entropy curve $H(D|T_2)$ and prior probability $\Pr(D_1) = 0.36$ extends in each direction to intersect the entropy curve $H(D)$. The Bayesian posterior probabilities $\Pr(D_2|T_2) = 0.90$ (upper intersection) and $\Pr(D_1|T_2) = 0.10$ (lower intersection) are obtained as the projections onto the probability axis of these points of intersection.

For a binary predictor calibrated as in the current FHB example, information content is positive if the conditional probability given the prediction exceeds the prior probability, zero if the two probabilities are equal, and is negative if the conditional probability given the prediction is smaller than the prior. Thus, correct predictions yield positive information contents and incorrect predictions yield negative information contents.

An information graph for specific information. From equation 1 and equation 10 we can calculate *specific information* for test outcome T_i, denoted $I_S(T_i)$ as

$$I_S(T_i) = H(D) - H(D|T_i) \tag{12}$$

As is the case with relative entropy $I(T_i)$ (equation 8), specific information $I_S(T_i)$ can be interpreted as a measure of diagnostic information provided by outcome T_i on average over both categories of actual status, cases D_1 and controls D_2. And again as with relative entropy (equation 9), the expected value of specific information over both outcomes of a binary predictor is expected mutual information:

$$I_M(D,T) = \sum_i \Pr(T_i) I_S(T_i) \tag{13}$$

(16). However, a distinction between relative entropy and specific information readily becomes apparent from their respective information graphs.

Figure 5 shows an information graph for specific information, based on the example set. To illustrate specific information, we plot entropy $H(D)$ (equation 1) and conditional entropy $H(D|T_i)$ (equation 10), both with probability $\Pr(D_1)$ on the horizontal axis. For outcome T_1, the information graph (Fig. 5A) shows specific information as $H(D) - H(D|T_1) = I_S(T_1)$; for outcome T_2 (Fig. 5B), $H(D) - H(D|T_2) = I_S(T_2)$. On the basis of the example data set, specific information $I_S(T_i)$ is the vertical difference between $H(D)$ and $H(D|T_i)$ at $\Pr(D_1) = 0.36$ (Fig. 5). Here, we calculate $I_S(T_1) = 0.091$ nits, $I_S(T_2) = 0.328$ nits. In this case both $I_S(T_1)$ and $I_S(T_2) > 0$, but in general $I_S(T_i)$ may be positive (when $H(D) > H(D|T_i)$, in which case uncertainty has decreased upon receipt of T_i), or negative (when $H(D) < H(D|T_i)$, in which case uncertainty has increased upon receipt of T_i). Note that although specific information for a particular outcome may be negative, *average* specific information (i.e., expected mutual information, equation 13) is ≥ 0.

To view Bayesian probability revisions for the example data set on Figure 5A and B, we first draw a horizontal line through the point of intersection of the $H(D|T_i)$ curve and the vertical line $\Pr(D_1) = 0.36$. In Figure 5A, this horizontal line is $H(D|T_1) = 0.562$ nits; in Figure 5B the corresponding line is $H(D|T_2) = 0.325$ nits. On each graph, the horizontal line is extended as far as the $H(D)$ curve in each direction. Then in Figure 5A, the Bayesian posterior probabilities for a T_1 outcome, $\Pr(D_1|T_1) = 0.75$ (upper intersection) and $\Pr(D_2|T_1) = 0.25$ (lower intersection), are obtained as the projections onto the probability axis of the points of intersection of the horizontal line $H(D|T_1) = 0.562$ nits and the $H(D)$ curve. Similarly, in Figure 5B, the Bayesian posteriors for a T_2 outcome, $\Pr(D_2|T_2) = 0.90$ (upper intersection) and $\Pr(D_1|T_2) = 0.10$ (lower intersection), are obtained as the projections onto the probability axis of the points of intersection of the horizontal line $H(D|T_2) = 0.325$ nits and the $H(D)$ curve.

Overview. As Madden (21) pointed out, Bayesian updating as applied in crop protection decision making is based on determination of the probability of a disease outbreak (or need for a control intervention) before (the prior probability) and after (the posterior probability) using the predictor. These probabilities have corresponding information contents which we have illustrated diagrammatically by means of information graphs. The main phytopathological application of these information graphs is in

characterizing and evaluating predictors that facilitate deployment of Bayesian updating in crop protection decision making.

Recall that, like Madden (21) we are concerned here with binary predictors. When characterizing and evaluating a predictor from an information theoretic perspective, we identify the prior probability of case status, the posterior probability of case status given the prediction, and the probability of case status after the actual status is made known; respectively $\Pr(D_1)$, $\Pr(D_1|T_i)$ ($i = 1$, predicted case; $i = 2$, predicted control), and either 1 (actual case) or 0 (actual control). Because the predictors in question are binary, we can always calculate prior, posterior and actual probabilities of control (D_2) status as complements of the corresponding probabilities of case status. Now, consider the amount of information that would be required to take us from the prior probability to the actual status: if we have an imperfect predictor this can only supply enough information to take us part of the way, from the prior to the posterior probability, thus leaving a deficit, the amount of information still required to take us from the posterior probability to the actual status. Figure 4 illustrates the information contents corresponding to Bayesian updating from prior to posterior probabilities given either a T_1 or a T_2 prediction, and Figure 3 illustrates the remaining information deficits, following Bayesian updating, between posterior probabilities and actual status. Figure 2 illustrates the expected values, calculated over both categories of actual status, both of the information contents corresponding to Bayesian updating and of the remaining information deficits. Figure 5 illustrates the difference between entropies calculated before and after Bayesian updating from prior to posterior probabilities.

DISCUSSION

The information graphs presented above provide diagrammatic interpretations of the statistical decomposition of diagnostic information resulting from the Bayesian probability revisions that we observe when using a binary predictor. All the graphs are calculated from the same phytopathological data used to characterize the true positive and true negative proportions and their complements for such a predictor. In the present context, these data are formatted as a prediction-realization table (Tables 1 and 2).

Figure 1 provides a diagrammatic interpretation of the partition of entropy $H(D)$ (equation 1) into expected mutual information (equation 5, which shows that expected mutual information is a function of the properties of the predictor *TPP* and *FPP*, and the prior probabilities) and expected information inaccuracy (equation 6): $H(D) = I_M(D,T) + H(D|T)$. We can think of entropy as the uncertainty prior to the use of a predictor, and thus quantified in a particular case by specification of the prior probability. Then at the specified prior probability, use of the predictor, on average over both actual status categories D_j and both test outcome categories T_i, reduces uncertainty by an amount equal to the expected mutual information. Expected information inaccuracy is then the average remaining uncertainty after use of the predictor. Since entropy as specified by equation 1 is the upper limit of expected mutual information as specified by equation 5, we can calculate a normalized version of expected mutual information as $(H(D) - H(D|T))/H(D)$ (3). Thus, in the case of the example data set about one-third of the uncertainty prior to the use of a predictor is, on average, explained by use of the predictor. For a discussion that allows consideration of this value in a wider epidemiological context, see section 2.4 of Hughes (16).

Benish (5) plotted information graphs for relative entropies $I(T_1)$ and $I(T_2)$ as functions of prior probability of case status. Figure 2 instead provides a diagrammatic interpretation of the partition of cross-entropy $H_C(D|T_i, D)$ (equation 11) into relative entropy $I(T_i)$ (equation 8) and expected information inaccuracy for a T_i prediction (equation 10). An alternative version of this relative entropy calculation appears in Hughes (17). We can obtain the information quantities illustrated in Figure 1 from those in Figure 2 by taking the expected values over T: expected mutual information is calculated from relative entropy via equation 9 and overall expected information inaccuracy is calculated from expected information inaccuracy for a T_i prediction via equation 6.

At this stage we refer back to our earlier introduction to the example data set and in particular the account of predictor accuracy, as characterized on the basis of sensitivity and specificity, for context. There, we were concerned with calibration of the predictor from a data set where the actual status (D_1, case; D_2, control) and predicted status (T_1, predicted case; T_2, predicted control) were known for all crops in the data set. Here, we are concerned with the accuracy (or inaccuracy) of predictions characterized on the basis of posterior probabilities. Figure 3 characterizes information inaccuracies in terms of a logarithmic scoring rule. Over a series of predictions for which actual status is subsequently identified, these information inaccuracies provide a basis for forecast evaluation (e.g., 29). Consider a situation where it is regarded as important that no-treatment decisions are restricted almost entirely to controls. To this end, a threshold low on the risk scale is adopted. Then sensitivity (*TPP*) is high and so *FNP* is low, so a high proportion of negative test outcomes would (correctly) be for controls, as required. In this situation, only a small proportion of cases would (incorrectly) not be treated. In terms of posterior probabilities, such a calibration corresponds to $\Pr(D_2|T_2) \approx 1$, $\Pr(D_1|T_2) \approx 0$ (see, e.g., Table 2 of ref. 7). An obvious matter for consideration in relation to the logarithmic scoring rule for binary predictors is that such a calibration sets a highly exacting standard for forecast evaluation. Suppose that for the purpose of evaluation, a series of predictions is made for which the actual status is subsequently identified. If the predictor is calibrated so that $\Pr(D_2|T_2) \approx 1$, $\Pr(D_1|T_2) \approx 0$, then observation of a D_1 subject subsequent to a T_2 test outcome will result in a very large value for information inaccuracy. Note that from the individual logarithmic scores characterized in Figure 3 we can calculate the expected information inaccuracy (i.e., the average remaining uncertainty) for a T_i prediction via equation 10. Thus, in the current example, for a T_1 prediction we have

$$H(D|T_1) = -(0.75 \cdot \ln(0.75) + 0.25 \cdot \ln(0.25)) = 0.562 \text{ nits}$$

and for a T_2 prediction we have

$$H(D|T_2) = -(0.10 \cdot \ln(0.10) + 0.90 \cdot \ln(0.90)) = 0.325 \text{ nits}$$

as in Figure 2.

Figure 4 characterizes the information contents of predictions arising from Bayesian probability updating. These are calculated directly via equation 7, and illustrated graphically in terms of the difference between the information content of the prior probability and that of the posterior probability. Relative entropy $I(T_i)$ is expected information content for a prediction T_i, so we can obtain the relative entropies (as in Fig. 2) from the information contents (as in Fig. 4) via equation 8. Thus, in the current example, for a T_1 prediction we have

$$I(T_1) = 0.75 \cdot \ln\left(\frac{0.75}{0.36}\right) + 0.25 \cdot \ln\left(\frac{0.25}{0.64}\right) = 0.315 \text{ nits}$$

and for a T_2 prediction we have

$$I(T_2) = 0.10 \cdot \ln\left(\frac{0.10}{0.36}\right) + 0.90 \cdot \ln\left(\frac{0.90}{0.64}\right) = 0.179 \text{ nits}$$

as in Figure 2.

We recall that information content

$$\ln\left(\frac{\Pr(D_j|T_i)}{\Pr(D_j)}\right)$$

can alternatively be written as the pointwise mutual information

$$\ln\left(\frac{\Pr(D_j \cap T_i)}{\Pr(D_j)\cdot\Pr(T_i)}\right)$$

in which case we can characterize information contents resulting from information updating directly from data in the form of a prediction-realization table and so calculate expected mutual information via equation 4.

Figure 5 illustrates specific information $I_S(T_i)$ diagrammatically in terms of the difference between entropy $H(D)$ and expected information inaccuracy $H(D|T_i)$ (equation 12). We can obtain the information quantities illustrated in Figure 1 from those in Figure 5 by taking the expected values over T: expected mutual information is calculated from specific information via equation 13 and overall expected information inaccuracy is calculated from expected information inaccuracy for a T_i prediction via equation 6. While specific information and relative entropy both have expected mutual information as their expected value (equations 9 and 13, respectively), we see from their information graphs (Figs. 2 and 5, respectively) that relative entropy $I(T_i) \geq 0$ while specific information may take either positive values (when $H(D) > H(D|T_i)$) or negative values (when $H(D) < H(D|T_i)$). For comparative discussions of specific information and relative entropy, see Fano (13) (for an information theoretic point of view) and Hughes and McRoberts (18) (for an epidemiological point of view). One important distinction between specific information and relative entropy from an epidemiological point of view is that the former is additive while the latter is not. The consequence of this is that in order to correctly accumulate diagnostic information in a sequential diagnosis, we calculate specific information and not relative entropy (18). Relative entropy is interpretable as the amount of diagnostic information provided by an individual test in the diagnostic sequence (4), but these amounts are not cumulative.

The information graphs described here allow us to view binary predictors used in diagnostic decision making from the perspectives of both probability revision and entropy reduction. There are machine-learning approaches to classification that use algorithms based on entropy reduction (26,27). Quinlan (27) provides an example based on a phytopathological data set from Michalski (22), and more recently the C4.5 algorithm has been used as a statistical diagnostic to classify the temporal dynamics of plant disease epidemics (14). A diagrammatic approach to the information theoretic concepts deployed in such phytopathological applications can help to reduce the extent to which the analysis is just viewed as a 'black box' by end-users. And while in general the application of information theory to epidemiology of course transcends the analysis of binary decision problems, necessitating equation- or software-based analysis, the underlying concepts described in our information graphs remain applicable to an understanding of the diagnostic decision-making process.

ACKNOWLEDGMENTS

SRUC receives grant-in-aid from the Scottish Government.

LITERATURE CITED

1. Aegerter, B. J., Nuñez, J. J., and Davis, R. M. 2003. Environmental factors affecting rose downy mildew and development of a forecasting model for a nursery production system. Plant Dis. 87:732-738.
2. Amblard, P.-O., Michel, O. J. J., and Morfu, S. 2005. Revisiting the asymmetric binary channel: Joint noise-enhanced detection and information transmission through threshold devices. Pages 50-60 in: Proceedings of the SPIE, vol. 5845, Noise in Complex Systems and Stochastic Dynamics III. L. B. Kish, K. Lindenberg, and Z. Gingl, eds. The International Society for Optical Engineering. doi:10.1117/12.609436
3. Attneave, F. 1959. Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results. Holt, Rinehart and Winston, New York.
4. Benish, W. A. 1999. Relative entropy as a measure of diagnostic information. Med. Decis. Making 19:202-206.
5. Benish, W. A. 2002. The use of information graphs to evaluate and compare diagnostic tests. Method. Inform. Med. 41:114-118.
6. Bondalapati, K. D., Stein, J. M., Neate, S. M., Halley, S. H., Osborne, L. E., and Hollingsworth, C. R. 2012. Development of weather-based predictive models for Fusarium head blight and deoxynivalenol accumulation for spring malting barley. Plant Dis. 96:673-680.
7. Caffi, T., Rossi, V., Bugiani, R., Spanna, F., Flamini, A., Cossu, L., and Nigro, C. 2009. A model predicting primary infections of *Plasmopara viticola* in different grapevine-growing areas of Italy. J. Plant Pathol. 91:535-548.
8. Capote, N., Bertolini, E., Olmos, A., Vidal, E., Martínez, M. C., and Cambra, M. 2009. Direct sample preparation methods for the detection of *Plum pox virus* by real-time RT-PCR. Int. Microbiol. 12:1-6.
9. Cover, T. M., and Thomas, J. A. 2006. Elements of Information Theory, 2nd ed. Wiley-Interscience, Hoboken, NJ.
10. Dewdney, M. M., Biggs, A. R., and Turechek, W. W. 2007. A statistical comparison of the blossom blight forecasts of *MARYBLYT* and *Cougarblight* with receiver operating characteristic curve analysis. Phytopathology 97:1164-1176.
11. Diamond, G. A., Hirsch, M., Forrester, J. S., Staniloff, H. M., Vas, R., Halpern, S. W., and Swan, H. J. 1981. Application of information theory to clinical diagnostic testing: The electrocardiographic stress test. Circulation 63:915-921.
12. Elegbede, C. F., Pierrat, J. C., Aguayo, J., Husson, C., Halkett, F., and Marçais, B. 2010. A statistical model to detect asymptomatic infectious individuals with an application in the *Phytophthora alni*-induced alder decline. Phytopathology 100:1262-1269.
13. Fano, R. M. 1961. Transmission of Information: A Statistical Theory of Communications. The M.I.T. Press and John Wiley & Sons, Inc., New York.
14. Franke, J., Gebhardt, S., Menz, G., and Helfrich, H.-P. 2009. Geostatistical analysis of the spatiotemporal dynamics of powdery mildew and leaf rust in wheat. Phytopathology 99:974-984.
15. Gent, D. H., and Ocamb, C. M. 2009. Predicting infection risk of hop by *Pseudoperonspora humuli*. Phytopathology 99:1190-1198.
16. Hughes, G. 2012 Applications of Information Theory to Epidemiology. The American Phytopathological Society, St. Paul, MN.
17. Hughes, G. 2014. Information graphs for epidemiological applications of the Kullback-Leibler divergence. Method. Inform. Med. 53:IV-VI.
18. Hughes, G., and McRoberts, N. 2014. The structure of diagnostic information. Australas. Plant Pathol. 43:267-286.
19. Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. Ann. Math. Statist. 22:79-86.
20. Lindley, D. V. 1985. Making Decisions. 2nd ed. John Wiley & Sons, London, UK.
21. Madden, L. V. 2006. Botanical epidemiology: Some key advances and its continuing role in disease management. Eur. J. Plant Pathol. 115:3-23.
22. Michalski, R. S., Davis, J. H., Bisht, V. S., and Sinclair, J. B. 1983. A computer-based advisory system for diagnosing soybean diseases in Illinois. Plant Dis. 67:459-463.
23. Nielsen, F., and Nock, R. 2011. Entropies and cross-entropies of exponential families. Pages 3621-3624 in: Proceedings of the 17th International Conference on Image Processing, Hong Kong. Institute of Electrical and Electronics Engineers. doi:10.1109/ICIP.2010.5652054
24. Nita, M., Ellis, M. A., and Madden, L. V. 2008. Variation in disease incidence of Phomopsis cane and leaf spot of grape in commercial vineyards in Ohio. Plant Dis. 92:1053-1061.
25. Pethybridge, S. J., Gent, D. H., Esker, P. D., Turechek, W. W., Hay, F. S., and Nutter, F. W., Jr. 2009. Site-specific risk factors for ray blight in Tasmanian pyrethrum fields. Plant Dis. 93:229-237.
26. Quinlan, J. R. 1986. Induction of decision trees. Mach. Learn. 1:81-106.
27. Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.
28. Renyi, A. 1965. On the foundations of information theory. Int. Stat. Rev. 33:1-14.
29. Roulston, M. S., and Smith, L. A. 2002. Evaluating probabilistic forecasts using information theory. Mon. Weather Rev. 130:1653-1660.
30. Shannon, C. E., and Weaver, W. 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.
31. Somoza, E., and Mossman, D. 1992. Comparing and optimizing diagnostic tests: An information-theoretical approach. Med. Decis. Making 12:179-188.

32. Theil, H. 1967. Economics and Information Theory. North-Holland, Amsterdam, The Netherlands.
33. Turechek, W. W., Hartung, J. S., and McCallister, J. 2008. Development and optimization of a real-time detection assay for *Xanthomonas fragariae* in strawberry crown tissue with receiver operating characteristic curve analysis. Phytopathology 98:359-368.
34. Turechek, W. W., and Wilcox, W. F. 2005. Evaluating predictors of apple scab with receiver operating characteristic curve analysis. Phytopathology 95:679-691.
35. Yuen, J., Twengström, E., and Sigvald, R. 1996. Calibration and verification of risk algorithms using logistic regression. Eur. J. Plant Pathol. 102:847-854.