Relative Entropy as a Measure of Diagnostic Information

William A. Benish, MD, MS

Abstract

Relative entropy is a concept within information theory that provides a measure of the distance between two probability distributions. The author proposes that the amount of information gained by performing a diagnostic test can be quantified by calculating the relative entropy between the posttest and pretest probability distributions. This statistic, in essence, quantifies the degree to which the results of a diagnostic test are likely to reduce our surprise upon ultimately learning a patient’s diagnosis. A previously proposed measure of diagnostic information that is also based on information theory (pretest entropy minus posttest entropy) has been criticized as failing, in some cases, to agree with our intuitive concept of diagnostic information. The proposed formula passes the tests used to challenge this previous measure.
Diamond and coworkers observed in 1981 that information theory "has been little used in clinical medicine..." \(^1\) The situation has not changed significantly since that time. Language to quantify information has found its way into most niches of modern society, but phrases like "bits of diagnostic information" and "bits for the buck" are never heard on hospital wards.

The seminal work in information theory was published by Shannon in 1948.\(^2\)-\(^3\) He was primarily interested in the communication related issues of optimal data compression and data transmission rate. Information theory has subsequently been applied to a spectrum of disciplines including economics, computer science, physics, statistics, psychology, and biology.\(^4\)-\(^6\)

I believe that information theory has failed to establish a foothold in clinical medicine because an appropriate mathematical model of diagnostic information has yet to be introduced. The present report aspires to 1) provide an introduction to information theory and 2) demonstrate that "relative entropy" is an appropriate measure of diagnostic information.

### Surprisals and Entropy

The less likely a diagnosis, the more surprised we are when that diagnosis proves to be the case. The "surprisal" is a concept within information theory that provides a means of quantifying our surprise. The surprisal, \(u\), corresponding to an event with probability \(p\) is given by

\[
u = - \log_2 p
\]  

\(^{(1)}\) (reference 7). Although the choice of the base of the logarithm is arbitrary, \(\log\) to the base two gives a result expressed in bits. The surprisal corresponding to a diagnosis with probability 0.5 is one bit. If one diagnosis is half as likely as another diagnosis, then the
surprisal corresponding to the less likely diagnosis is one bit more than the surprisal corresponding to the more likely diagnosis. Moreover, the surprisal associated with the joint occurrence of two independent events is the sum of the surprisals associated with the two individual events. Since probabilities range from zero to one, and the log of numbers within this range are less than or equal to zero, surprisals are always greater than or equal to zero. If a diagnosis is certain (p = 1), then the surprisal is zero. If a diagnosis is not possible in a given situation (p = 0), then the surprisal is undefined--there is no number large enough to quantify our surprise at the occurrence of an impossible event.

Consider a patient with a clinical presentation that leads to a differential diagnosis consisting of n mutually exclusive possibilities: d_1, d_2, ..., d_n. Let p_i represent the probability that d_i is the correct diagnosis. Our surprise upon learning that d_i is the correct diagnosis is quantified by u_i. Before the diagnosis is known, our best estimate of the amount we will be surprised is

\[ \sum_{i=1}^{n} p_i u_i. \]  

(2)

This is the expected value of the function u. The expected value of a function is a weighted sum of all possible values the function can assume, with each weight equal to the probability of the corresponding value. Expression 2 is the expected surprisal and is equivalent to a fundamental concept in information theory, the entropy, H, of a random variable.\(^3\) By replacing u_i in expression 2 with its definition, stated by equation 1, we have

\[ H = - \sum_{i=1}^{n} p_i \log_2 p_i. \]  

(3)

H provides a measure of our uncertainty about a diagnosis.\(^\dagger\) For any given number of diagnostic possibilities, H achieves its maximum when the diagnoses are equally likely. H equals zero, on the other hand, when the diagnosis is known.

\[^\dagger\text{Although some of the references cited in this report apply the term “uncertainty” to H, our convention will be to refer to H as the “entropy” of the random variable.}\]
Relative Entropy

Diagnostic testing can be conceptualized as identifying a characteristic of the patient that distinguishes him or her as belonging to a subset of the original population. If the diagnostic test is informative, the probabilities of the various diagnoses will differ between the original population and this subset. For a set of $n$ mutually exclusive diagnostic possibilities: $d_1, d_2, \ldots, d_n$, we let $a_i$ represent the pretest probability that a patient has $d_i$, and $b_i$ represent the posttest probability that a patient has $d_i$. In other words, $a_i$ is the probability that a patient randomly selected from the original population has $d_i$, and $b_i$ is the probability that a patient randomly selected from the subset of patients with “positive” test results has $d_i$.

Consider a patient with a positive test result. For an observer that is aware of this result, his expected surprise upon ultimately learning the patient’s diagnosis is quantified by

$$- \sum_{i=1}^{n} b_i \log_2 b_i$$ (4)

(since $b_i$ is the probability that the patient has $d_i$, and $- \log_2 b_i$ quantifies the surprise this observer will experience if he learns that the patient has $d_i$). For an observer that is not aware of the fact that the patient belongs to this particular subset of patients, his expected surprise upon learning the patient’s diagnosis is quantified by

$$- \sum_{i=1}^{n} b_i \log_2 a_i$$ (5)

(since $b_i$ is the probability that the patient has $d_i$, and $- \log_2 a_i$ quantifies the surprise this naive observer will experience if he learns that the patient has $d_i$).

The relative entropy between the posttest and pretest probability distributions, $D(b||a)$, is the reduction in the expected surprisal that is attributable to learning the test results. It is equal to expression 5 minus expression 4. This calculation gives
Example

Consider a population of patients with arthritis in which the differential diagnosis and associated probabilities are as shown in table 1. The posttest probabilities apply to the subset of patients who have a “positive” test result. For a patient selected at random from the original population, our expected surprise upon learning the diagnosis is quantified as

\[ (0.25)(- \log_2 0.25) + (0.5)(- \log_2 0.5) + (0.125)(- \log_2 0.125) + (0.125)(- \log_2 0.125) \]

\[ = 1.75 \text{ bits} \] (equation 3). For a patient found to have a positive test result, our expected surprise upon learning the diagnosis is quantified as 1.0 bit (expression 4). If we did not know the test results, our expected surprise upon learning this same patient’s diagnosis is quantified as 2.5 bits (expression 5), reflecting the fact that we will be more surprised to learn that the patient has gout or pseudogout than will an observer that was aware of the test results.

Given a patient who belongs to the subset of patients with positive test results, the relative entropy, \( D(b||a) \), is equal to the expected surprisal associated with not knowing the test results, minus the expected surprisal associated with knowing the test results: 2.5 bits - 1.0 bits = 1.5 bits. Knowledge of the test results in this example reduces our expected surprise upon learning the patient's diagnosis by an average of 1.5 bits.

Relative Entropy and Coding Theory

Although the approach of this report has been to define entropy and relative entropy in terms of surprisals, these concepts are more commonly related to the number of bits of information required to describe the outcome of a random variable. An
efficient code will use relatively short messages to describe relatively common events.\(^8\)

Table 2 presents efficient codes corresponding to the pretest and posttest probability distributions in the above example. The expected length of a message written in the code corresponding to the pretest probability distribution and indicating the diagnosis of a patient randomly selected from this population is:

\[
(2 \text{ bits})(0.25) + (1 \text{ bit})(0.5) + (3 \text{ bits})(0.125) + (3 \text{ bits})(0.125) = 1.75 \text{ bits},
\]

where each term in the sum is equal to the number of bits of code describing the outcome, multiplied by the probability of that outcome. The expected length of this message is equivalent to the entropy of the pretest probability distribution, \(H_a\). Shannon demonstrated that the minimum average length of a message reporting the results of a random event is the entropy associated with the event.\(^8\)

The other expected surprisals calculated in the above example can also be translated into the language of code length. In particular, observe that only \(H_b = 1 \text{ bit}\) of information is required to communicate the diagnosis of a patient with positive test results using the code based on the posttest distribution, but that \(H_b + D(b||a) = 1 + 1.5 = 2.5 \text{ bits}\) of information are required, on average, to communicate this message using the code based on the pretest distribution.

**Discussion**

The relative entropy between the posttest and pretest probability distributions is a quantitative representation of the amount of information gained by performing a diagnostic test. It is not a measure of the absolute amount of information that a test provides; information that is already known or that has no bearing on disease probability is not included in this index. The results of using the relative entropy formula to determine the amount of information gained by diagnostic testing in several clinical situations is shown in table 3.
A diagnostic test should be understood as any event that potentially provides diagnostic information. From this perspective, reviewing a patient's medical record or asking a patient to describe his symptoms is no less a diagnostic test than a serum potassium determination or an electrocardiogram.

The merits of an information theory approach to quantifying diagnostic information has been the subject of debate. The assumption made in these discussions is that diagnostic information is properly quantified by subtracting posttest entropy from pretest entropy. We will refer to this statistic as $\Delta H$. Using the same symbols that were used to define $D(b||a)$,

$$\Delta H = - \sum_{i=1}^{n} a_i \log_2 a_i + \sum_{i=1}^{n} b_i \log_2 b_i. \quad (7)$$

Asch et al examined this definition of diagnostic information and concluded that it "fails to capture reasonable intuitions about the quantity of information provided by diagnostic tests." The three examples Asch used to critique this statistic help illustrate the reasonableness of relative entropy as a measure of diagnostic information. In all three examples, the differential diagnosis is limited to two possible disease states.

Example 1. Let the pretest probability of two diseases be 0.1 and 0.9 and the corresponding posttest probabilities be just the opposite, 0.9 and 0.1. Although a great deal is learned by performing this diagnostic test, the posttest entropy equals the pretest entropy and, hence, $\Delta H$ equals zero. The relative entropy formula, on the other hand, provides an intuitively satisfying result--2.54 bits of information.

Example 2. Let the pretest probability of two diseases be 0.1 and 0.9 and the corresponding posttest probabilities be 0.91 and 0.09. $\Delta H$ is calculated to be 0.033 bits. This is contrasted with an alternative diagnostic test in which the respective posttest probabilities are 0.95 and 0.05. In the case of this alternative test, $\Delta H$ is found to be 0.183 bits. Asch et al correctly point out that, "although the second test clearly provides more information, it is hard to accept that it provides more than five times as much
information." The corresponding values of $D(b|a)$ for the two tests are 2.60 bits and 2.88 bits--results that are not "hard to accept."

Example 3. Let the pretest probability of two diseases be 0.1 and 0.9 and the corresponding posttest probabilities be 0.91 and 0.09. This scenario was described in example 2. We now contrast this with a different scenario in which the two pretest probabilities are identical, 0.5, and the posttest probabilities are the same as in the first scenario, 0.91 and 0.09. Intuition tells us that more information is gained in the first situation, where the pretest and posttest probabilities are more disparate, than in the second situation. $\Delta H$, however, is found to be smaller in the first scenario, 0.33 bits, than in the second scenario, 0.56 bits. Conversely, the respective values of $D(b|a)$ are 2.60 bits and 0.56 bits.

Another problem with $\Delta H$ as a measure of diagnostic information is that if posttest entropy happens to be larger than pretest entropy, information is calculated to be a negative number. Relative entropy, on the other hand, is always greater than or equal to zero.

Relative entropy is also known, in honor of the originators of the formula, as the Kullback Leibler distance.\textsuperscript{5,13} It is a measure of the distance between two probability distributions. More precisely, it is a directed divergence.\textsuperscript{14} Unlike the conventional concept of distance, in which the distance from point X to point Y is equal to the distance from point Y to point X, it is generally not the case that $D(b||a)$ equals $D(a||b)$. A final example illustrates this point.

Example 4. Let the pretest probabilities of two diseases both equal 0.5 and the posttest probabilities equal 1.0 and 0. The test provides one bit of information. This conclusion is reached regardless of whether one calculates $\Delta H$ or $D(b||a)$; the two formulas give the same result when there is maximal uncertainty prior to testing. Now consider what happens if we try to calculate information assuming that the pretest and posttest probabilities are reversed. $\Delta H = -1$ bits, a result that does not make sense. The
relative entropy calculation, on the other hand, is undefined on the basis of requiring
division by zero. This latter observation is consistent with the realization that if we are
certain of a diagnosis (which we never are), no finite amount of information can change
our minds.

Conclusion

Scientific progress is dependent upon innovations in measurement. It remains to
be seen whether the ability to quantify diagnostic information will contribute to medical
progress. The author hopes that he has helped stimulate an interest in the application of
information theory to diagnostic test evaluation and other areas of clinical medicine.

The author is grateful to Thomas D. Schneider, PhD for his mentorship and critical review of this report.

Table 1  A Hypothetical Example

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Pretest Probability</th>
<th>Posttest Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>gout</td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td>osteoarthritis</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>pseudogout</td>
<td>0.125</td>
<td>0.5</td>
</tr>
<tr>
<td>other possibilities</td>
<td>0.125</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2  Efficient Codes

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Pretest Population</th>
<th>Posttest Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>gout</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>osteoarthritis</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>pseudogout</td>
<td>110</td>
<td>1</td>
</tr>
<tr>
<td>other possibilities</td>
<td>111</td>
<td></td>
</tr>
</tbody>
</table>
Table 3 • Examples of the Information Gained by Performing Diagnostic Tests

<table>
<thead>
<tr>
<th>Testing Situation</th>
<th>Pretest Probability of Disease</th>
<th>Test Operating Characteristics: Sensitivity/Specificity</th>
<th>Test Result</th>
<th>Posttest Probability of Disease</th>
<th>Information Gained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer screening</td>
<td>0.01</td>
<td>0.75/0.94</td>
<td>positive</td>
<td>0.11</td>
<td>0.25 bits</td>
</tr>
<tr>
<td>with mammography</td>
<td></td>
<td></td>
<td>negative</td>
<td>0.003</td>
<td>0.006 bits</td>
</tr>
<tr>
<td>Mammography given</td>
<td>0.2</td>
<td>0.80/0.90</td>
<td>positive</td>
<td>0.67</td>
<td>0.74 bits</td>
</tr>
<tr>
<td>palpable breast mass</td>
<td></td>
<td></td>
<td>negative</td>
<td>0.05</td>
<td>0.13 bits</td>
</tr>
<tr>
<td>Screening for HIV</td>
<td>0.001</td>
<td>0.99/0.998</td>
<td>positive</td>
<td>0.33</td>
<td>2.4 bits</td>
</tr>
<tr>
<td>with antibody test</td>
<td></td>
<td></td>
<td>negative</td>
<td>0.00001</td>
<td>0.001 bits</td>
</tr>
<tr>
<td>Presence of tonsillar</td>
<td>0.1</td>
<td>0.45/0.84</td>
<td>positive</td>
<td>0.24</td>
<td>0.11 bits</td>
</tr>
<tr>
<td>exudate in diagnosing</td>
<td></td>
<td></td>
<td>negative</td>
<td>0.07</td>
<td>0.01 bits</td>
</tr>
<tr>
<td>group A streptococci</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon cancer screening</td>
<td>0.005</td>
<td>0.40/0.90</td>
<td>positive</td>
<td>0.02</td>
<td>0.02 bits</td>
</tr>
<tr>
<td>by fecal occult blood</td>
<td></td>
<td></td>
<td>negative</td>
<td>0.003</td>
<td>0.0005 bits</td>
</tr>
<tr>
<td>testing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*See reference 9 for details.