ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/6951188

Clinical and Pharmacogenomic Data Mining: 1. Generalized Theory of Expected Information and Application to the Development of Tools

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · JUNE 2003

Impact Factor: 4.25 · DOI: 10.1021/pr025587q · Source: PubMed

CITATIONS
12

READS
37

1 AUTHOR:

Barry Robson
St. Matthews University

274 PUBLICATIONS **7,873** CITATIONS

SEE PROFILE

Available from: Barry Robson
Retrieved on: 02 February 2016

research articles Journal of proteome research

Clinical and Pharmacogenomic Data Mining: 1. Generalized Theory of Expected Information and Application to the Development of Tools

Barry Robson

T. J. Watson Research Center, 1101 Kitchwan Road, Yorktown Heights, New York 10598

Received December 4, 2002; Revised Manuscript Received February 20, 2003

New scientific problems, arising from the human genome project, are challenging the classical means of using statistics. Yet quantified knowledge in the form of rules and rule strengths based on real relationships in data, as opposed to expert opinion, is urgently required for researcher and physician decision support. The problem is that with many parameters, the space to be analyzed is highly dimensional. That is, the combinations of data to examine are subject to a combinatorial explosion as the number of possible events (entries, items, sub-records) (*a*),(*b*),(*c*),... per record (*a*,*b*,*c*,..) increases, and hence much of the space is sparsely populated. These combinatorial considerations are particularly problematic for identifying those associations called "Unicorn Events" which occur significantly less than expected to the extent that they are never seen to be counted. To cope with the combinatorial explosion, a novel numerical "book keeping" approach is taken to generate information terms relating to the combinatorial subsets of events (*a*,*b*,*c*,..), and, most importantly, the ζ (Zeta) function is employed. The incomplete Zeta function $\zeta(s,n)$ with $s = 1$, in which frequencies of occurrence such as $n = n(a,b,c,...)$ determine the range of summation n, is argued to be the natural choice of information function. It emerges from Bayesian integration, taken over the distribution of possible values of information measures for sparse and ample data alike. Expected mutual information $I(a;b;c)$ in nats (i.e., natural units analogous to bits but based on the natural logarithm), such as is available to the observer, is measured as e.g., the difference $\zeta(s,o(a,b,c..)) - \zeta(s,e(a,b,c..))$ where $o(a,b,c,..)$ and $e(a,b,c,..)$ are, or relate to, the observed and expected frequencies of occurrence, respectively. For real values of $s > 1$ the qualitative impact of strongly (positively or negatively) ranked data is preserved despite several numerical approximations. As real s increases, and the output of the information functions converge into three values +1, 0, and −1 nats representing a trinary logic system. For quantitative data, a useful ad hoc method, to report σ-normalized covariations in an analogous manner to mutual information for significance comparison purposes, is demonstrated. Finally, the potential ability to make use of mutual information in a complex biomedical study, and to include Bayesian prior information derived from statistical, tabular, anecdotal, and expert opinion is briefly illustrated.

Keywords: data mining • association • covariance • negative association • proteomics • pharmacogenomics • patient record • information theory • expected information • zeta function

1. Introduction

1.1 Current Needs in Biological and Medical Data Analysis. There is an information explosion in biomedicine due to genomics, proteomics and the digitization of patient data.[1] Classical statistics,[2−5] as most widely taught and used, does not by itself prepare us for the complexity of data in the post-genomic era. Improved tools for analysis of medical and biomolecular information are required to discover the relationships between patient history, environment, lifetsyle, genomics, and disease for a variety of purposes. These include the desire to identify new drug targets, avoid adverse drug reactions in clinical trials and after physician visits, and to generate the rules and rule strengths for physician, patient, and researcher decision support systems. As patient records are becoming increasingly digitized, it is also evident that analysis of archives of these is of enormous value even in the absence of genomic data., for example to detect epidemiological features, trends and unexpected relationships between disease, and to maintain vigilance for natural and malicious biothreat.

The problem is not just one of trawling through great complexity to find simple solutions. The solutions themselves, hiding in the combined medical and genomic data, are sometimes complex. Although prior to the "genomic era" there were already more than 8500 human genetic disorders known to clinical geneticists and which may relate to one or very few genes, we now know that in many important disease cases, such as cardiovascular disease, polymophisms in not just one gene but many determine a disease of primarily genetic origin, and that patient history and lifetsyle have an important role introducing still more parameters into the problem space.

10.1021/pr025587q CCC: $25.00 © 2003 American Chemical Society

Published on Web 05/08/2003

1.2 Information Theory Approach. Following the pioneering work of Shannon[6] on the theory of the transmission of information, several workers were stimulated to apply a broader information theory to complex problems combining statistical[7,8] and decision theory[9] approaches. In particular, Fano[10] promoted the notion of "mutual information" as more closely related to knowledge. Hence, the program used in the present study is named FANO. In one of the earliest efforts in bioinformatics,[11] we used information measures weighted by classical methods (chi-square) to handle extremely sparse levels of protein structural data. It was soon appreciated that methods based on the work of Bayes[12] might be a better approach to such problems. Though use of Baysian methods was rare, especially in biology and chemistry, at that time (1970−1974), the University of Cambridge favored the approach[13] and it began to gain popularity.[14] Hence we applied the Bayes method to information theory,[16−17] to provide a self-consistent theory of expected information.[17] Though the Bayesian method has grown into a mainstay of bioinformatics, a Bayesian-linked information theory has been largely confined to one branch, protein science.[21−22] In large part, this has been due to the difficulties in "book-keeping" with regard to handling a much greater variety of terms arising from a much larger number of combinations of statistical events, so that a given widely used algorithm such as the GOR method[20] actually represented a fixed, specific expansion and specific neglect of certain terms. This is a restriction which is addressed here, allowing the method to be further generalized.

These previous methods, and especially the Theory of Expected Information,[17,20] form the basis of the present method. However, the method is generalized by use of the Zeta Function,[23] of the encoding of information terms and combinatorics of data using primes,[24] and the global optimization of multivaraiate coefficients.[25] The strengthened number−theoretic basis that links many of these aspects also provides a road map for future development by facilitating a coherent theory of data and data mining.

1.3 Association and Covariance. The essence of the Fano information measure is whether items such as (*a*) (*b*) (*c*), ... occur together as (*a*,*b*,*c*,...) more, or less, than expected. The measure quantifies the statistical degree of association of "events", "qualities", "items" or "entries" *a*,*b*,*c*,... The concept can be applied to pure qualitative data, or quantitive data by pooling numeric data into classes. In addition, when data is numeric, covariance can also be explicitly treated in a manner which allows comparison with association. Covariance differs by relating concerted trends in coherent data such a weight, alcohol consumption, and blood triglyceride concentrations, such that when the value of one goes up, the value of the other tends to go up too (in which case covariance is positive) or one goes down (in which case covariance is negative). The relationship between association and covariance is that in the former there are only two possible values of difference between items: they either match or they do not. In current versions, complex functional (including nonlinear) relations, that may need to be processed to show clear covariance between data, are converted by a user-programmable plug-in ("convert.dat") file.

1.4 Combinatorial Explosion. Even a quite short and relatively simple record may imply thousands or even trillions of rules to take full account of the information contained in it. In a spreadsheet, application of statistics is trivial if we know what combinations of item or column to address. But to address discovery, we must explore all combinations without prejudice. Moreover, we cannot even assume that there will be just one solution to do with just group of interrelated columns. Rather there could be several such groups, each with correlation statistics which are "eigensolutions", interrelating internally but not with each other. If there are n potential correlating columns of spreadsheet data drawn from N, then the number of possible combinations to analyze statistically is $N!/[(N - n)!n!]$, and if we do not know n in advance, the number of eigenvalues representing solutions is

$$\Sigma_{n=2,...N}\, N!/[(N - n)!n!]$$

Worse still, the solutions can go as a power law if the data is not of full spreadsheet form, and the records contain data items that may occur more than once; this creates a situation analogous to statistical sampling with replacement. Whereas, in principle, the matrix calculus can be applied to multivariance, in practice, such methods are not sufficiently robust to be practical. One is thus driven to consider either exploring all of the possible relatively small combinations of covariances ("bottom-up"), or using an optimization approach to the problem in which coefficients of different possible contributing associations are optimized to fit trial solutions to the actual data and at least obtain "bundled" multivariance measures (say of each column with all the rest ("top-down"). FANO combines both tactics in passes, to help "nail down" the eigensolutions.

1.5 Rectangularity, Lists, Sets and Collections. The classification of types of data can be made in terms of rectangular data (tables, spreadsheets, multidimensional arrays), lists (or sequences, including DNA and protein sequences), sets (which lack order), and (countable) collections (in which, unlike true sets, members can occur several times). The present approach treats the rightmost of these (collections) for two reasons. The first is that the data type becomes more general from left to right, so that whatever level we chose to handle, we can ultimately find a way to handle the types to the left. The second is that the properties of collections map very nicely to several useful concepts in number theory (Appendix 1), that provide us with insight, a road map to the generalized treatment of data, means of book keeping internal calculations, and even new data-analytic tools.

Much recent data provided for pharmacogenomics study has been in a sanitized columnar, spreadsheet form. Nonrectangularity arises when the data is not all in any particular order. Real spreadsheet data is also not so different. True rectangularity is often corrupted in that data items are missing (unknown or undetermined data), and conversely, in the rarer sense that the same metadata values (such as blood−glucose concentration, or cellular calcium level) can apply to several columns of a spreadsheet. In practice, there may incompleteness of metadata (column headings such as "ID" might be missing), and even in some cases the inability to distinguish data from metadata a priori. For example, in one of our studies the entries exemplified by "Alcohol in blood", "over legal driving limit", "under legal driving limit" could only be identified as metadata and two metadata values respectively after comparison of several records. A key concept to deal with this *and* to escape from rectangularity is to consider instead *qualification* by metadata. By this process, collections, sets and lists can all be addressed by the same algorithms. An example of qualification is Age:=42, where "Age" is metadata, and 42 is data. Rectangularity even without explicit metadata implies metadata such

as "column 1", "column 2", etc. Departure from rectangularity implies escape from the need to have metadata, though it is still permissible. Qualification also opens the opportunity to represent hierarchic relationships. i.e., to manage and deduce from formal ontologies relationships such as animal:=vertebrate:=primate:=human, where there is multiple metadata (qualification). Also branches can be represented, viz: animals:=(vertebrates:=mammals, invertebrates:=worms). In the present approach, internal working will take the rightmost individual item (human) as the data for analysis, and the rest (animal:=vertebrate:=primate:=) as metadata, but methods have also been explored in which the "split" between metdata and data is made at each ":=" separator, bringing exploration of ontological structure and degrees of association into correspondence.

1.6 Positive vs Negative Associations. A simple example of a positive association is that birthdays and gifts very commonly occur together even though statistically, based on their individual probabilities, they should not do so. Positive associations are natural to treat as they occur for example as commonly associated words or phrases in text. Negative associations can be more subtle. In text analysis, one can readily see that addressing associations of all possible combinations of word, to find which occurred less than expected, is a challenging and, on the face of it in most application areas, rather pointless pursuit. However, negative associations are among the most important associations in modern medicine. A life is pursued which is negatively correlated with disease. A chemical substance, for example, that negatively associates with disease, is a potential drug. A genomic constitution which negatively correlates with certain diseases confers protection and allows a more daring or broader lifestyle in certain specific areas: this is a position increasingly of interest to the tobacco industry.

Especially in real open records, very negative associations are very difficult to handle. An event which does occur at least once is said to be "existentially qualified". If it is not, the relevant combination of events may never be entered into the virtual computer tables, such as hash arrays, as items worthy of study: i.e., the key or reference is never created. To discover if it is interesting that an event does not occur, such entries need to be created as potentially interesting, and comparison must be made with the expected occurrence of a particularly huge combinatorial explosion of potential associations: many animals are observed which can be qualified as "white", "shy", "horned", or "horses", but white, horned horses, i.e., unicorns, have not been reproducibly observed (hence, "the unicorn event"). Also, despite the fact that pregnancies and males are common features of medical records, they never occur together in the same patient. This is biologically significant and, in the absence of prior knowledge, worth reporting. This touches on several matters raised by the philosopher Karl Popper, and, since the whole issue is that *we are interested in what we do not expect*, some associations to explore may seem bizarre and even the constituent events may not be considered. For example: cancer might be avoided by wearing an orange hat while playing the mandolin upside down in Belgium, but how would we ever know? Common sense can be brought to bear, but only to a point. Attempts have been made to scan this vast space assuming similar properties of "taxonomic siblings", i.e., entities with the same taxonomic heritage. Evidently, however, purchase of a mad dog is not the same as purchase of a lap dog, despite their taxonomic relationship. A more potent Artificial Intelligence is required and until that is available, brute force handling of the combinatorial explosions of potentially interesting concurrences is important, and internal bookkeeping methods for them, like those used here, are important.

2. Theory

The FANO program approach attempts to extend the breadth of Fano's theory to a Theory of Data by using a number−theoretic basis. Some basic principles are tabulated in Appendix 1, showing their one-to-one correspondence with statements in number theory as well as providing a convenient synopsis.

2.1 Fano's Mutual Information. The information Theory due to Shannon and others related to statistical mechanical entropy and was largely concerned with calculation of the simple or self-information such as $I(a) = -\log[P(a)]$ for one event (a) or $I(a,b) = -\log[P(a,b)]$ for two events (a,b), etc. Fano (1964) showed the importance of considering mutual information between e.g., (a) and (b), viz.

$$\boldsymbol{I(a;b) = \log [P(a,b)/(P(a)P(b))] = [P(a|b)/(P(a)] = [P(b|a)/(P(b))]} \quad (1)$$

Fano considered this as a better basis for considering the knowledge content of information, since it represents "how much event a has to say about event b, and vice versa". Such events can for example be items, or entries, on a record, and the relative frequency of occurrence of events a and b over several records leads to estimates explicitly or implicitly of the P, and hence, the mutual information. Equation 1 describes both positive ($I(a;b) > 0$) and negative ($I(a,b) < 0$) associations of conjoint ("concurrent", "complex") events. Information against an event is considered as negative information for an event.[17] Fano's notion can be extended simply to multiple events such as (a,b,c,...), so defining $I(a; b; c; ...)$ compared with the observations that the single ("simple", "disjoint", "individual") items ("entries", "events") (a),(b),(c),...come together on a chance basis. To do this, the observed frequency of occurrence (number of observations) $o(a,b,c)$ is counted, the individual frequencies such as $o(a) = \Sigma_b\, o(a,b) = \Sigma_c\, o(a,b,c)$ calculated, and the information measures are determined from functions with these frequencies as parameters.

Here, we define the multiple mutual information as

$$\boldsymbol{I(a;b;c;d;...) = \log [P(a,b,c,d..)/(P(a)P(b)P(c)P(d)...)]} \quad (2)$$

This form is chosen because it is fairly general. Other forms such as $I(a,b;c,d)$, the mutual information between (a,b) and (c,d) can be estimated. For example, $I(a,b; c,d) = I(a;b;c;d) - I(a;b) - I(c;d)$. As in eq 1 vertical bar can be introduced within the probability functions I and P, e.g., $I(a; b \mid z)$ reads as "conditional upon". Estimates of such information measures should be "coherent" so that terms are both additive and, ideally, comparable. "Additive" means that, for example, information functions such as $I(a;b,c,d,...)$ can be expanded into a series of additive positive or negative terms.[17] "Comparable" means that pairs, triplets, etc. can be ranked together (as in the present paper). Neglect of any of the terms (e.g., terms for which data may not be available) is permissible and represents the model for, or estimate of, $I(a;b,c,d)$.

2.2 Expected Information. If the amount of data obtained were always infinite, then probability and information measures could be obtained directly from the frequencies of concurrence and Fano measures could be obtained directly via log of the probabilities as for $I(a;b)$ above. In practice, the required estimate according to Robson[17] is in the form of the expectation

E [$I(a;b;c;d;...)$ | D] of $I(a;b;c;d;...)$ Given data D. The integration of E[$I(a;b;c;d;...)$ | D] chosen by Robson[17] is described in more detail in section 2.6: based on that, expected values of mutual information measures $I(a;b;c;...)$ are expressed as linear combinations of the Zeta function

$$I(a;b;c;...) = \zeta[s,f_o(a,b,c,...)]- \zeta[s,f_e(a,b,c,...)]$$

$$f_o(a,b,c,...) = o(a,b,c,...) + h_o(a,b,c,..) - A$$

$$f_e(a,b,c,...) = e(a,b,c,...) + h_e(a,b,c,..) - A \quad (3)$$

The arguments $f_o(a,b,c,...)]$ and $f_e(a,b,c,...)$ relate to the frequencies of occurrence, i.e., the numbers of obserbations of, or the abundance of, events $(a,b,c,...)$. The function $o(a,b,c,..)$ is the actual observed frequency of occurrence of concurrent (conjoint) event (a,b,c) and $e(a,b,c,..)$ is the corresponding expected value as defined below. Parameters h_o and h_e respectively, expressed in the dimensions of frequency of occurrence, represent some prior opinion about the frequencies.[17] "Constant" A equals 1 in the present study and this choice is discussed below. The term $o(a,b,c)$ was the number of times that a, b, and c are seen to occur together, e.g., in a record. We could neglect h and A terms so $f_o(a,b,c) = o(a,b,c)$ for zero or more of concurrent events $a,b,c,...$ The corresponding expected frequency in this case is that in the chi-square sense, which can be represented comprehensively as

$$e(a,b,c,...) = N_{tot} (o(a)/N_{tot}) (o(b)/N_{tot}) (o(c)/N_{tot})... \quad (4)$$

where $N_{tot} = o(a) + o(b) + ...$

The need for estimation of the I and hence eq 3 represents the fact that probabilities and information cannot be directly known. The information measured is that available to the observer about information in a system.[17] The unusual nature of the integration and its result at that time (1970−1974) is primarily due to the way in which classical probabilities P and Bayesian probabilities Pr are intertwined. Yet this choice is reasonable. We cannot be sure that such a "bias in nature" as classical probability P always actually exists "behind the data", but we can chose to hold a *belief* in the existence of such biases P, and in *degrees of belief* in their values as, e.g., Pr[$P(x = 1, y)|D(x,y)$], conditional on data $D(x,y)$ that we see. That includes holding degrees of belief about functions of P, such as Pr[$I(x = 1:2;y)|D(x,y)$. Then, the estimate of the information $I(x = 1:2;y) = I(x = 1;y) - I(x = 2;y)$ is interpreted as the expectation of the information *which is accessible to the mind of the observer* as a consequence of those beliefs, and is

$$E[I(x=1:2;y)] = \int I(x=1:2;y)\, Pr[P(x = 1:2,y)|D(x,y)] \, dP(x = 1:2;y) \quad (5)$$

According to Robson,[17] the proper form of integration for terms log P and log$(1 - P)$ used building up information expressions where $n[\ ]$ was a frequency of observation analogous to $o(\ )$

$$\Gamma(n[1,y] + n[2,y] - 1)/\Gamma(n[1,y] - 1)\Gamma(n[2,y] - 1) \int \log(P/(1 - P)\, P^{n(1,y)-1}(1 - P)^{n(2,y)-1}\, dP$$

$$=\zeta(s=1, n[1,y] - 1) - \zeta(s = 1, n[2,y] - 1) \quad (6)$$

where $\zeta(s = 1, n[1,y])$ can be considered as relating to the estimate of the component log(P) and $\zeta(s = 1, n[2,y])$ as relating to the estimate of the component log$(1 - P)$. Robson[17] did not describe the result in terms of the Zeta function, but only in terms of the Euler harmonic series which corresponds to the Zeta function for $s = 1$. However, the *ad hoc* but demonstrated generalization to other values of s is of interest in the present work as the behavior of the function has useful properties. Note in eqs 5 and 6 that $I(x = 1:2; y) = I(x = 1;y) - I(x = 2; y)$, where the latter components are the true Fano mutual information measure for events $x = 1$ and y. The $I(x = 1;y)$ rather than $I(x = 1:2; y)$ are used in the present study because, for open records with large numbers of states to which new states are constantly added, the complement $x = 2$ to $x = 1$ (or "not x" as opposed to "x") evolves with the record archive, and original Fano measure seems more appropriate.

2.3 Prior and Absolutely Prior Decision Distributions. Influences of prior belief or knowledge are most readily discussed in terms of underlying probabilities P, the density function Pr of which is identified with the Bayes degree of belief. The above terms h above can mathematically represent any judicious opinion about prior data, expressed in easy to understand terms of frequency of occurrence. There is also an issue of the "absolutely prior" distribution, prior to any observation or opinion. This is reflected by the frequent appearance of "−1" in eq 6, which leads by the integration to the choice of "constant" $A = 1$ for eq 3. This subtraction of 1 in using eq 3 is most generally of little importance because obviously $\zeta[s,n - A) - \zeta(s,n - A] \rightarrow \zeta(s,n) - \zeta(s,n)$, for small A as $n \rightarrow \infty$. The argument below favors this "−1" even though it means that when there are no other h, o, or e terms, the distribution of eq 6 is "improper" i.e., cannot be integrated using real numbers. When data is not sparse throughout, the distinction really only matters as a practical computational issue here, for large s when eq 3 converges to +1,0,−1, i.e., to trinary logic (see below). Subtracting 1 (i.e., using $A = 1$ in eq 3 is sometimes referred to as the DAPD (Dirichlet Absolutely Prior Probability Density) choice, and is sometimes considered as inelegant. In contrast the ZAPD choice when 1 is not subtracted seems, at first consideration., more natural, and in the last resort we are free to believe what we like, a priori. However, the need for "−1" in the generality of the result for eq 5 as represented by eq 6 rests on a requirement by Wilks[4] and others about the distributions of relevant underlying probability distributions. The condition for their requirements to be met is that the likelihood is rendered as multinomial β-distributed, i.e., of form $P(1,y)^{n(1,y)}P(2,y)^{n(2,y)}$ with parameters n which are derived from the $D(x,y)$, and that this is achieved by the inclusion of the effect of the Dirichelet absolutely prior density with components P^{-1}, i.e., $P^{-1}\,(1 - P)^{-1}$ in the binomial case. This guarantees that binomial, multinomial and marginal probability densities Pr all follow the β-distribution and are hence coherent (consistent). In effect, we are then entitled to add and subtract information contributions in certain ways, in a consistent manner irrespective of the amounts of data[14]. For example, $I(x = 1:2; y)$ can then be estimated as $\zeta(s,n[1,y] - 1) - \zeta(s,n[2,y] - 1) - \zeta(s,n[1] - 1) + \zeta(s,n[2] - 1)$.

2.4 s-Generalization of the Information Measures based on the Zeta Function. The Zeta function appearing in eq 3 has been expressed in a variety of forms. For the general case of s for all real and imaginary values, the incomplete (partially summed) Riemann's fully extended Zeta Funtion would be required. Development of this requires the techniques of complex analysis and the technique analytic continuation.[23] The "partially extended" Riemann Zeta Function below can also be regarded as an approximation; in any event it is only valid

for $0 < s <= 2$, $x >= |t|/\pi$ with order s^2

$$\Delta\zeta(s,n) = -[\, n^{(1-s)} \cos(t \ln(n))\,(s-1)/((s-1)^2 + t^2) - n^{(1-s)} \sin(t \ln(n))\, t/((s-1)^2 + t^2)$$

$$- i\, n^{(1-s)} \sin(t \ln(n))\,(s-1)/((s-1)^2 + t^2) - i\, n^{(1-s)} \cos(t \ln(n))\, t/((s-1)^2 + t^2) \quad (7)$$

Here, the imaginary component is made explicit for computation convenience. The "delta form" $\Delta\zeta(s,n) = \zeta(s,n) - \zeta(s)$ avoids a term $\zeta(s)$ on the right-hand side of eq 5, corresponding to the case when n is approaching infinity. Though Riemann considered his fuller form the "correct" choice of continuation and this is so from the perspective of the field of complex analysis, the Zeta function can be considered as branching at $s < 1$. The simple application of the Euler form defined for $s = 1$

$$\zeta(s,n) = 1 + 1/2^s + 1/3^s + \ldots + 1/(n-1)^s + 1/(n)^s \quad (8)$$

can be also considered by definition as the "Euler branch" for $s < 1$.

Note that

$$\zeta(s,n) - \zeta(s,m) = 1/(m-1)^s + 1/(n)^s + \ldots\ldots + 1/(n)^s \quad (9)$$

and that$\zeta(s,n) > \zeta(s,m)$ if $n > m$. Though $\zeta(s,n) - \zeta(s,m)$ decreases as $m + n$ increases, it is easy to show that we should expect insensitivity to $s > 1$ of the rank order of the expected information for concurrent events. Similarly one expects a degree of stability (internal consistency of rank) for $s < 1$, when the Euler form, eq 9, is the one extended to that region. Departures are due to approximations in sampling and approximations (interpolations and extrapolations from the Zeta function table) and in the treatment of absolutely prior density.

There are some noteworthy matters that relate to scaling. As noted above, the resulting measures are "nats" or "natural units" analogous to "bits" or "binary units". The use of natural log units is retained because it arises in a natural way from the above and in the theory[17] and is analogous to the treatment of entropy and free energy. We recall that amounts of data, say $n[1,y]$, govern the "degree of completeness" of the incomplete Zeta function: *if n is sufficiently large,* $\zeta(s = 1, n[1,y])$ *can be replaced by the natural logarithm of n plus the Euler−Mascheroni constant.* The Euler−Mascheroni constant is 0.577 215 664 9... This cancels in normal applications in our case and the above therefore simply means that $\zeta(s = 1, n[1,y]) - \zeta(s = 1, n[2,y])$ may be evaluated as the natural logarithm of ratio $\log_e (n/m)$ when n and m are sufficiently large

$$\mathbf{Lt}_{n \to \infty},\ \zeta(s = 1,n) - \zeta(s = 1,m) \rightarrow \log_e (n/m) \quad (10)$$

Here, the log is specifically $\log_e$. When n and m are between 10 and 20 or larger, they produce reasonable approximations of $\log(n/m)$. At $s = 1$, the point of the branch, the Zeta function returns infinity for infinite n. With the choice of $s = 2$ and increasing n, both eqs 7 and 8 converge to $\pi^2/6$ and, for the case $s = 4$, $\pi^4/90$. At $s = 1$, the point of the branch, the function returns infinity for infinite n. For $s < 1$ and infinite n, the Riemann branch returns e.g., $\zeta(s = 0,\ n = \infty) = -1/2$ and $\zeta(= -1,\ n = \infty) = -1/12$. If we consider the Euler form as a valid branch below $s < 1$, then it returns values based on a more linear dependence on n, and notably $\zeta(s = 0, n) = n$. In contrast to using the log limit of eq 10 for all cases, the use of the Zeta function for all cases handles data down to zero levels, and the position could be taken that this spans the gap between qualitative and quantitative research. A value close to zero nats can arise either because there is no departure from expectation, or little or no data: it is the information available to the researcher which is implied in the theory. This approach allows methods in which the search space can be pruned because sampling, which would lead to close-to-zero measures, can be predicted in advance. In any event, filters are usually applied to prune out "low" measures, i.e., measures implying information close to zero, as being of less interest.

2.5 Prime Representations of Records, Sub-Records, and Concurrent (Conjoint) Events. Euclid's "Elements of Geometry" observed that every positive integer can be written in one and only one way as a product of larger and larger primes, this representation being its "prime decomposition", and hence any factorization into components is also unique ("unique factorization theorem"). Implementation of this represents "Goedelization" of lists of any kind, including representations of statements in mathematical logic, as primes. Typically, any list S with items $(i,j,k,l,...)$ is represented uniquely by the products $2^i\, 3^j\, 5^k\, 7^l$... However, use here is different because we wish to treat records more generally as collections, not lists, and the role of the powers of the primes is more simply to represent the abundance of the items coded by primes, except that with zero occurrence (implying the term $P^0 = 1$ for item coded by prime P) need not be explicitly included in the product. Thus each unique item is assigned a prime code, and it is raised to the power representing the abundance in that record. If $a,b,...$etc are each represented by primes, then an event such as $(a,b,c...)$ represented as the product $a \times b \times c \times ...$ uniquely implies the content (a), (b), (c) by the unique factorization theorem. For example, items (a,b,b,d) are a record, and $a = 2$, $b = 3$, $d = 7$) then $2 \times 3 \times 3 \times 7 = 126$ uniquely identifies that record. Providing that we assign any metadata to each entry, viz:- Age:=42, it adequately qualifies the entry and Age:=42 becomes the item to which a prime number is assigned. We may shuffle the entries per record without loss of information. When the data item is numeric, the large number of potential items can be reduced by pooling number into ranges, e.g., Age: >42 and Age:<50. (In the program FANO, the symbol := is the universal "metadata" sign indicating qualification by metadata to the left, so that the forms :=> and :=< are used for greater than or equal to, and less than, respectively).

Note that any two or more records coded by the same product of primes is considered as having the same content. We can select any subset of digits from two representations. Done to sufficient precision, it can at least be stated that two records and so forth cannot be identical in content if they are different at a specified precision level. The use of logs of the primes coding for items is also a useful quantity[24]. Finally, if there is always a unique patient identifier and hence a unique prime, no two whole records are equivalent in content.

2.6 Generation of Information Terms (All Possible Collections of Concurrent Events). For present practical purposes, the most important of the prime number considerations relates to how information measures can be expanded as a series of terms, in which the "event" parameters are combinations or, more generally, collections of events or items from records. Items might include Ag:=42 or "nonsmoker". The problem of how to expand all these terms, is also the question of how a record can be broken down into contributions.

Congruence issues in number theory map to a useful method for generating all possible sub-concurrent events (sub records) or a record. From $(a,b,c,d,e,f,g,...)$ the different simple and

conjoint events such as (*a*), (*a*,*b*), (*a*,*b*,*d*), (*b*,*d*,*g*,*p*) ...are generated as follows. Let *N* be the product of primes encoding event (*a*,*b*,*c*,*d*,*e*,*f*,*g*,...).Any integer *k*, which can be divided into the above product *N*, without leaving a remainder generates a valid simple or conjoint event, i.e., $k|N$ or $k \equiv k \bmod(N)$. By repeated division by $k = 1,2,3,4,\ldots N - 1$ all valid simple and conjoint events are generated correctly once. This corresponds to those divisors counted in the number theoretic function $\tau(N)$ (Appendix 1). This again assumes that the data type is of the *collection* type used in the underlying theory, i.e., a set-like collection in which, however, items can reappear in the same record. The above division process is not as time-consuming as full factorization, because we require to keep sub-products such as (*a*,*b*,*d*) from (*a*,*b*,*c*,*d*,*e*), not to further decompose them. Nonetheless, because of the relative slowness of division over other mathematical operations this approach is used to pre-generate code which is "hard wired" into the program, and division is performed only for more complex cases if justified by the data available.

In some instances, it is computationally more efficient to form the product, and in others, to avoid division or factorization. To handle the multiple approaches, objects consisting of strings of primes are defined, e.g., "*2*2*5*11*", and functions are defined, which enable the dynamic transition indicated by ⇌ between string and product representations

"2*2*5*11*" ⇌ 4*5*11 ⇌ "2*2*55*" ⇌ "*2 × 10*11* ⇌ "*220*"

Important functions acting on such strings include the function which interprets any and all of the above equivalent forms as their content in English, e.g., (Gender:=male, Age:=>50). Also, all events such as (*a*,*b*,*c*..) when first seen create and increment a hash array entry count{"2*2*5*11"}. Normally, the function ord() ensures that the primes in the string are returned in ascending order, viz: count{ord("2*2*5*11")}. For certain purposes defining rings, i.e., closed, end-connected strings, is of value.

2.7 Normalization of Expected Frequencies. The use of expected frequencies, exemplified by eq 3, is the choice of Robson.[17] Apart from the effects of any prior belief, the question of normalization resides in the second Zeta term and specifically within the formulation of expected frequencies. Expected frequencies may be computed as in the chi-square test frequencies e.g., as

$$e(a;b) = o(a)\, o(b)/[N_{\text{tot}}] \tag{11}$$

An expected frequency is in general a real rather than integer number. In practice, the value of a decimal argument is adequately obtained by linear interpolation between the results for the integer values. An expected frequency is a kind of probability estimate which is not absolutely normalized, and in fact $e(a;b) = p(a)\, p(b)\, N_{\text{tot}}$ for adequate data. In eq 11, the term $[o(a) + o(b)] = N_{\text{tot}}$ is the total amount of data which may be also expressed as $N_{\text{tot}} = \Sigma_a\, o(a) = \Sigma_b\, o(b) = \Sigma_a\Sigma_b\, o(a,b)$. It is not the only choice. The considered volume of the sampling space, i.e., the size of the playing field in which we believe or chose to believe that we are playing, is a a key concept in the choice of N_{tot} and is a matter of conditioning the space, which is essentially dividing it up by metadata assignment. For concurrence of independent nonoverlapping items or events, such as distinct species of butterfly which we catch together in a sample from the jungle, it is simply $o(a) + o(b) + o(c)$. Some uses may require the expected frequency conditional on another event or events, say *z*, in which case *z* appears in all frequency terms

$$e(x = 1;y|z) = o(x{=}1,z)\, o(y,z)/[o(x = 1,z) + o(x{=}2,z)] \tag{12}$$

and conversely on *y*

$$e(x = 1;z|y) = o(x = 1,y)\, o(y,z)/[o(x = 1,y) + o(x = 2,y)] \tag{13}$$

Note for completeness examples of the specially defined cases which arguably arise naturally from forms (expected frequency equals probability of relevant event times an observed frequency)

$$e(x = 1) = o(x = 1)\, N_{\text{tot}}/N_{\text{tot}} = o(x = 1) \tag{14}$$

$$e(x = 1|\, y) = o(x = 1,y)\, [o(x = 1,y) + o(x = 2,y)]/[o(x = 1,y) + o(x = 2,y)] = o(x = 1,y) \tag{15}$$

$$e(x = 1|\, y,z) = o(x = 1,y,z)\, [o(x = 1,y,z) + o(x = 2,y,z)]\, /\, [o(x = 1,y,z) + o(x = 2,y,z)] = o(x = 1,y,z) \tag{16}$$

If $x = 1$ and $x = 2$ are not mutually exclusive but overlapping sets, then some other value of N_{tot} must be invoked. The advantage of the information-theoretic approach is that the choice of normalization is approximately additive for given *n*-plets such as triplets (*a*,*b*,*c*). For reasonable levels of data the mutual information converges to $\log(n(a,b,c,\ldots) - \log(o(a)) - \log(o(b)) - \log(o(c)) - \ldots. + (T - 1)\log(N_{\text{tot}}'/N_{\text{tot}}'')$, where *T* is the number of items and terms (*a*),(*b*),(*c*),..... Thus, the baseline shift can be identified with a "correction" factor $(T - 1)\log(N_{\text{tot}}'/N_{\text{tot}}'')$, when N_{tot}'' is the alternative choice of total frequency for normalization compared with original choice N_{tot}'.

2.7.1 Metadata Duplicated in Columns of Rectangular Arrays. Usually for rectangular data, as in spreadsheets, a specific class of item can only occur in a particular column, and classic approach is to determine the expectation of two items concurring in a column from the average density (expected density) per column. Internally, in the Fano program, the analogous but more general practice is that the expected total frequency of occurrence of all items is based on the sum over all those with the same metadata, whatever the initial format. This also allows that there can be more than one columns with the same metadata and columns can be of unequal length. Multiple entries of value per row with same metadata are automaticaly treated as further observations per row which add to statistical power and they are included in covariance.

2.7.2 Unknown Data. Conversely, entries may be missing from some records. An issue is whether "unknown" data are present because of their ambiguous character of unknowns, which could be considered as "hiding" real values. For example, if there is one observation on blood pressure showing it to be high, and 99 nonobservations, the use of the "more sophisticated" counting (which carefully ignores the unknowns) leads to a density of one observation divided by a total of one observation, for the whole set or records, and hence 100% chance of occurring in a record. This is despite the fact that the unknown observations, if made, would have revealed that all the rest were low blood pressure and hence the expected frequency of occurrence of high blood pressure per record is a low one. In FANO, an option (**ignore unknown**=) is available so that all data items matching the parameter are treated as

Table 1. Equivalent Non-metadata and Metadata Forms

	boy	girl	tall	short
boy	15	0	10	5
girl	0	19	12	7
tall	10	12	22	0
short	5	7	0	12

	height:=tall	height:=short
gender:=boy	10	5
gender:=girl	12	7

unknowns, say, the word "unknown", as a flag. Also, an empty entry such as two successive commas in a .csv file also automatically implies the unknown value as above.

2.7.3 Resolving Metadata Assignment. Where the distinction of data and metadata is unclear, as in text analysis, FANO helps identify the relationships by a first pass in which associations are analyzed. The probability that a male and female will turn up together at a dinner party is a different consideration from a male and a tall person, and the improbability that a person is both male and female at the same time is a metadata issue. Consider the data for the properties of persons in a 4 × 4 table (Table 1) of boy−girl−tall−short vs boy−girl−tall−short.

Identities such as boy vs boy should either disallowed and assigned zero, or set so that (boy,boy) = (boy). We assume the latter; the self-case here is equally well distinguished as the same as the sum for the remaining cells in the rows or columns. The FANO approach basically takes the position that all items are potential AND events until proven that they are never seen in association. Whence they might be mutually exclusive OR (XOR) events and should be considered for data under the same metadata. For example, by inspection, boy and girl with short, and also short with tall, are "unicorn events" (see above) and the table can be reduced to the 2 × 2 table (Table 1) boy−girl vs tall−short.

In the real world the distinction between the AND and the XOR situation is not always so clear-cut, and an appeal may be made to areas of symmetry theory and quantum mechanics for useful tools. These include the relative weightings of quantum mechanical state functions, the cosine of the angle a graph by which two orthogonal dimensions of description such as "*a*-ness" AND "*b*-ness" pivots and moves toward "metagonality". For example, in the above tabular boy−girl−tall−short example, a "break" occurred more fully between the table dimensions for boy−girl and for tall/short to reduce a sparse 4 × 4 table to a 2 × 2.

2.8 Covariance Methods. When data is numerical such that comparison is possible, FANO automatically calculates not only associations for data pooled above and below the mean (see Sections 4.1 and 4.4), but also σ-normalized correlations, i.e., covariances, and compares them. It does so by an ad hoc treatment of covariances intended to allow them to be co-ranked as intuitively as possible with associations.

Covariance is usually considered as a relationship between *columns* of numeric data, though in fact within Fano it is not identified by column but the numerically qualified by the metadata (Column title, such as "Age"), which may or may not correspond to a physical column. Hence, if a tabular format is presented to FANO, then it can have several columns with the same metadata heading, or data may be missing. Then, measures of covariance are converted to effective frequency-of-occurrence terms, thereby so-called "fuzzy associations" analogous to associations in presentation. The effect is as analogous as possible to doing association on the same data. i.e., calculating the average value $\langle v \rangle$ of every corresponding item with that metadata, and then creating a state of pooled data which corresponds to being equal to or greater than, and then less than, the cut point.

Covariance typically implies pairs of columns. Multivariance (between more than two classes of metadata at a time) is done by generalizing the classical treatment for the two column (a,b) case to three or more columns $(a,b,c,...)$

$$\mathrm{cov}(a,b,c,..) = \Sigma\,[(a - \langle a\rangle)(b - \langle b\rangle)(c - \langle c\rangle)....]/(\sigma(a)\,\sigma(b)\,\sigma(c)...) \quad (17)$$

where $\langle a \rangle$ is the mean value of a and so on, and sigma σ is the variance. For further discussion below, note

$$s(a) = \mathrm{abs}[(a - \langle a\rangle)/(\sigma(a)] \quad (18)$$

and similarly for $b,c,.....$ Here, $\mathrm{cov}(a,b,..) = \Sigma\,[s(a)\,s(b)\,s(c)...]$ is not in general a correct alternative to eq 17 but provides model terms which can be use to build exploratory covariance-like functions to fit the data by optimization (see Program and Methods).

Effective fuzzy frequencies n' might be calculated by resolving *effective* components such as

$$n' = N'\,\langle v\rangle_{\mathrm{age}>=50}/[\langle v\rangle_{\mathrm{age}>=50} + \langle v\rangle_{\mathrm{age}<50}] \quad (19)$$

where N' is the total frequency of occurrence of items seen with that metadata. In practice, this is done by taking

$$n' = N' \times (1+ \mathrm{cov})/2 \quad (20)$$

Covariance between columns can be euqivalently considered as comprising n' data items which covary, and $N - n'$ with zero covariance (as opposed to $N - n'$ with negative covariance). The required effect is achieved if the "pseudoexpected" "fuzzy frequencies" e' are therefore calculated as

$$e' = N' - n' \quad (21)$$

Other tempting values for e' lead to a bound on the information value which does not represent a reasonable comparison with similar events explored by associations.

3. Program and Methods

3.1 Program and its Control. A more detailed account on a command-by-command basis is available as Supporting Information, and will also be reported elsewhere. The simplest overall statement of the overall process (relating input and output), is that it takes records of input in forms such as $[a,b,c,d,...]$ and converts them collectively to data summary records that each describe occurrence of complex events, such as "*a* with *d* with *e*", found in the input records. The input records are (typically) each either a row of spreadsheet, or a list of items on a file record (file line). The output complex events are the pairs, triplets etc. of items (entries) that are found among the input records (not necessarily side by side, but within the same records). These output records include also numbers representing the information with which these occur, compared with what would be expected for the occurrence in input records on a chance basis. This information is positive or negative according to whether it is more or less than expected. Moreover, although input order of records is ir-

relevant, the output records are in ranked order from most positive information down to most negative. When associations contain numeric values and metadata is invoked, covariances are calculated as well as associations, but these are co-ranked along with the associations in terms of "fuzzy associations". In contrast multivariate covariances, discussed in Section 3.4 below, are treated separately by a novel and experimental method, as a separate sequence of output records.

3.2 General Features Facilitating Program Development. FANO was developed in such a way as to facilitate experimentation and exploration of various forms of implementation, by considering two distinct aspects.

(1) Parsing. This is controlled by commands on the "command file". These parsing functions include handling of lists (a) by specifying separators such as commas, whitespaces or tabulation markers for items, for physical file line, and for record, (b) for DNA or protein or other sequences from which consecutive or overlapping strings of specified length, and (c) for extraction of metadata and data from xml and even whole patient records in xml though the latter study is still in progress.

(2) Conversion. Commands for converting data item by item (once parsed) reside on an optional "convert file". Conversions include scaling of data (e.g., to reflect choice of units), conversion such as nonmoker or male to 0 and smoker or female to 1, yes to 1, no to −1 and 'do not know' to 0, and conversion of dates and times to standard forms.

3.3 Association Sampling. Any record such as (*a*,*b*,*c*,*d*,*e*,*f*..) contains a large number of simpler conjoint (concurrent) events (*a*,*b*) (*a*,*c*,*d*), (*c*,*d*,*e*), (*a*,*c*,*d*,*f*) ("sub-records") which must be independently counted. For all but the smaller records, generation of all possible conjoint events more complex than pairs would be prohibitive in memory. It may be noted that it there are K kinds of item, perhaps hundreds, then going from doublets to triplets and generally from J-plets to $J + 1$-plets requires approximately an increase in the number of records by a factor of K to obtain equal significance. In FANO, pairs for association and pairs and optionally triplets for covariance analysis are treated exactly, without random sampling. For more complex events, the order of items in records are randomized after assignment of metadata, and split into sections of length L using the command **maximum number of items per record**= to set L. While large L is expensive, decreasing the size L increasingly loses varieties of conjoint event which cannot be recognized because they lie in separate record sample sections. Optionally, a technique known as "extreme zone sampling" cuts into effect when $L > 6$. Its aim is to catch the most positive *and* negative associations. All the triplets, quadruplets, ... L-plets which contain one, two, or three of the three "most interesting" items are sampled, and the rest are neglected. "Most interesting" is interpreted in two opposing senses: those items which are the most common, and those items which are the most rare, throughout the whole data (collection of records). An optional monitoring mechanism identifies and deletes those combinations of items which are inappropriately sampled more than once.

3.4 Covariance and Relevance Sampling. As for association, covariance is applied explicitly to covariance between numeric items in all pairs of sets of metadata (typically, columns). Optionally, triplet covariances can be treated exhaustively or by a sampling procedure. Higher multiplets such as 100-plets are treated separately and in statistical summary, and results are really measures of "potentially relevant variation".The detection of weakly variant columns is as follows. From eq 18, one may define the generator of a "covariance-like" function

$$\Phi = \Sigma_{\text{records}} \Pi_{\text{a}}[(s(a)/M_{\text{a}}]^{c(a)} \qquad (22)$$

Here, Σ_{records} indicates summation over all records (i.e., rows of a spreadsheet). M_{a} is either 1 or the mean value of $s(a)$ (see eq 18), depending on method option, for metadata "a". $S(a)/M_{\text{a}}$ is positive for both positive and negative covariances, and 0 for no covariance, and more than 1 for deviations from the mean values which are more than average deviations. Equation 22 is minimized as a function of the parameters $c(a)$,.... in such a way that when condition $S(a)/M_{\text{a}} > 1$ applies then $c(a)$ will decrease the value of the Φ function. Thus, the method picks out metadata whose data has "outlying" values with strong conjoint trends or anti-trends (in general, "interactions") with data of one or more other metadata.

True multi-covariance methods are also being explored. In general, optimization methods are preferred, but fitting models to sampled data and evaluating multivariate coefficients (and the kind of optimization considered above) cannot use optimizers which assume smooth surfaces and a single minimum.

The minimization method used must be of the type which can handle rough function surfaces with multiple minima. A simplex-based minimization method, "Globex" used here is based on that of Robson and Platt.[25] A "simplex" is a collection of $p + 1$ points in a parameter space of p dimensions. To assign coefficients to a problem of p parameters involves generating $p + 1$ conformations and repeatedly reflecting the point of highest function value through the centroid of the remaining p points, combined with operations of expansion and contraction for efficiency and convergence. Because the surface defined by eq 22 in this context is not only rough but discontinuous, and for another reason described below, in this case, it is the average value of the points so far which is the actual function value used. However, much of the activity resides in a higher level cycle involving use of a record of good configurations of coefficients, called the "Globex Stack". After a satisfactory convergence with the lowest function value point as closest to the centroid, the variables for that configuration of coefficients are stored on the "Globex Stack". The contents of up to p configurations of lowest function value and one random configuration are used to define the $p + 1$ points of a simplex for a new cycle in which fresh averaging commences (as if this were the first run). The other reason for using an average is that it promotes fast convergence. Normally, this might be seen as prohibiting discovery of the global minimum, but the present problem is not quite of the same class as, say, finding the statistical mechanical least free energy of a physical system. Rather, a number of eigensolutions can be found in sperate runs, each of which can correspond to clusters of intercorrelating metadata ("columns"). However, given sufficient cycles to provide some possibility of a global solution, the result reported will reflect the extent to which items under the same metadata may interact with other items, and move quickly, those insufficiently variant to be interesting.

3.5 FANO Signing of Multivariance. Consider the matrix **M** representing the two records each as rows, and which implies (symbol:=) covariance −1

$$\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix} := -1 \qquad (23)$$

That is, the matrix represents an archive of two records (+ − and − +) making up two columns (horizontal, + − and − +). Then, the sum of the products of the two columns is $(+1 \times -1) + (-1 \times 1) = -1 - 1 = -2$ which normalizes to −1. However, the analogous case for two records making up three columns is

$$\frac{-1+1-1}{+1-1+1} := 0 \tag{24}$$

In this case $(-1 \times +1 \times -1) + (+1 \times -1 \times +1) = +1-1 = 0$. Here, the statistic correctly describes that the trends in data in the columns are varying against each other, effectively "neutralizing" the overall covariation, but hides the overall correlation. In the present study, mutivariance is treated such the final information measure is positive if all columns have data with a trend in the same direction, and negative if at least one column has a trend going in the opposite direction (this is, of course, independent of where the columns are physically located with respect to each other in the spreadsheet). Also, in one more specific option of the FANO program, the multivariance of order M (e.g., between M columns) is treated as the sum of absolute covariances of every implied $M(M-1)/2$ constituent pairs, which are added. That is, if columns 1,2,3 are being tested together, then the covariances for columns 1−2, 2−3, and 1−3 are added. Then the above "mixing residue", i.e., the absolute value of a true multivariance (eq 17) for all M interactions, is added as a corrective "cross term" up to and including item c. The sum of $1 + M(M-1)/2$ terms is then divided by $1 + M(M-1)/2$ to produce the final measure.

3.6 Example Input Data. Sparse data like Table 2 occur often in clinical research practice in the present time of transition to digital records. "Microscalibility", i.e., robustly handling very little data, is as important as (macro-) scalability. Each line is a patient record. Although these data are somewhat contrived, the phenotypic features are real record extracts from information from a number of sources (see, for example, refs 27 and 28), and, judiciously used, such "contrivance" is actually a useful research tool too (see Discussion). The SNPs reflect preliminary findings which may emerge as haplotype-disease relationships. For example, there is a possible ideotype and likely haplotype for reduced levels of glutathione or other antioxidant components, and factors associated with high oxidant levels associate with pancreatitis.[28] A challenge would be to attempt by visual inspection to say which events or entries such as "alcoholism", "snp b", or other events occurring three, four of more items at a time, are most associated. A second challenge would be to spot any which are strongly avoiding each other, particularly identify a genetic polymorphism ("snp") which provides a clinically interesting protection. The answers are given in Results.

Table 2. Example Qualitative Input Data[a]

alcoholism, hepatic dysfunction, snp a, snp c, snp d, snp e, snp p, snp q
alcoholism, hepatic dysfunction, snp a, snp c, snp d, snp e, snp r
alcoholism, hepatic dysfunction, pancreatitis, snp a, snp d, snp e, snp p, snp q
alcoholism, hepatic dysfunction, pancreatitis, snp a, snp e, snp p, snp q
alcoholism, hepatic dysfunction, pancreatitis, snp a, snp c, snp d, snp e, snp p, snp q
alcoholism, hepatic dysfunction, snp a, snp c, snp e, snp p, snp q
alcoholism, hepatic dysfunction, pancreatitis, spider bite
snp a,snp b, hepatitis c, pancreatitis
hepatic dysfunction, alcoholism, snp a, pancreatitis
cancer,snp a,snp d,snp b, snp c, snp e, snp r, snp f, snp g, snp h, snp i, snp j
scorpion bite, pancreatitis, snp q, hepatitis A
cancer,snp d,snp b, snp r, snp c, snp h
hepatic dysfunction, spider bite, hepatitis c
hepatic dysfunction, snp c, snp d, snp f, snp g, snp p, snp q, hepatitis c
hepatic dysfunction, snp c, snp f, snp g, snp p, snp q, hepatitis B
hepatic dysfunction, snp c, snp d, snp f, snp g, snp p, snp q
cancer,snp a,snp b, snp c, snp r, spider bite, hepatitis B
cancer,snp a, snp d,snp b, snp r, hepatitis c
scorpion bite, snp e, snp p, pancreatitis
cancer,snp a,snp d,snp b, snp c, snp e, snp r, hepatitis B
schizophrenia, snp a,snp b,snp c,snp d, snp e
schizophrenia, snp a,snp b,snp c,snp e
schizophrenia, snp a,snp b,snp c
snp a, snp b, scorpion bite, scorpion bite, pancreatitis
snp b, pancreatitis, scorpion bite
hepatic dysfunction, alcoholism, pancreatitis
scorpion bite, snp d, hepatic dysfunction, cancer
hepatic dysfunction, snp p
hepatic dysfunction, snp a, snp c, snp d, snp p
hepatic dysfunction, snp a, snp c, snp d, snp x
alcoholism, snp a
hepatic dysfunction, alcoholism
alcoholism, snp c, hepatic dysfunction
alcoholism, schizophrenia
schizophrenia, scorpion bite
heart attack, snp a, snp b, snp p
heart attack, scorpion bite, snp b, snp p, pancreatitis
heart attack, hepatitis A, snp b
heart attack, snp p, snp a
snp b, snp c, spider bite
stress, snp a, snp b, snp c,snp d,snp e
stress, hepatitis A, snp b, snp c
stress, schizophrenia, snp a, snp e
stress, heart attack, spider bite
stress, heart attack, snp p, pancreatitis
stress, heart attack, snp a, snp p
stress, snp a, snp p
stress, scorpion bite, snp p, snp b, pancreatitis
stress, snp c, snp p
schizophrenia, snp b, snp e
schizophrenia, snp a,snp b,snp c
pancreatitis, schizophrenia, scorpion bite
schizophrenia, snp a,snp b,snp c
snp a, snp e
snp b,snp c,snp d
snp b,snp c,snp d, pancreatitis
snp p,snp q
snp p,snp q, scorpion bite, pancreatitis
snp p,snp q, scorpion bite, scorpion bite, pancreatitis
snp a,snp b,snp c
cancer, scorpion bite, snp d, pancreatitis, alcoholism, hepatic dysfunction, schizophrenia
hepatic dysfunction, alcoholism, snp a,snp c
snp a, snp c
hepatic dysfunction, alcoholism, snp b, snp a, snp c
hepatic dysfunction, hepatitis A
hepatic dysfunction, alcoholism, snp a

[a] One patient record abstract per line.

4. Results

4.1 Result Forms for Simple Test Cases. First, a highly structure rectangular form is tested in which it is easier to see the basic principles. It is helpful to think of the analysis as a complex matrix transformation from an input matrix to an output vector which is an estimate of the value and rank order of features worthy of attention, except that the absolute value (very positive or negative) refelects the level of interest. Note

that columns A with C correlate, and A with B, and B with C, anticorrelate.

a,	b,	c
0,	1,	0
0,	1,	0
1,	0,	1
1,	0,	1
0,	1,	0
0,	1,	0
1,	0,	1
1,	0,	1

These generate qualified (i.e., metadata qualifies data via the string ":=") associations. True associations are such as a:=>0 indicating that the pooled set in which the value of the item is equal to or greater than the mean, and a:=< in which it is less than the mean. Because the data is symmetric around the mean in this case, there are four data items in each set. FANO also generates "fuzzy associations" which are really covariance results expressed in terms of effective frequencies of occurrence of events. These have the qualification ":=av_". The average given is extra information only; there is no pooling and relationships all known, i.e., 8, values are used to estimate the covariance.

The following output is obtained for the output "vector". The other principle output, the weighting coefficients from the optimization of eq 22, were 100% for metadata A, B, and C showing that each had a stronger correlation (positive or negative) with at least one of the other two.

```
2.59=%=a:=av_0.5 c:=av_0.5
1.83===c:=<0.5 a:=<0.5 b:=>0.5
1.83===b:=<0.5 a:=>0.5 c:=>0.5
0.83===b:=<0.5 a:=>0.5
0.83===b:=<0.5 c:=>0.5
0.83===a:=<0.5 b:=>0.5
0.83===a:=>0.5 c:=>0.5
0.83===c:=<0.5 b:=>0.5
0.83===c:=<0.5 a:=<0.5
-1=Z=b:=<0.5 a:=<0.5
-1=Z=c:=<0.5 a:=>0.5
-1=Z=a:=>0.5 b:=>0.5
-1=Z=a:=<0.5 c:=>0.5
-1=Z=c:=>0.5 b:=>0.5
-1=Z=b:=<0.5 c:=<0.5
-2.45=%=a:=av_0.5 b:=av_0.5 c:=av_0.5
-2.59=%=a:=av_0.5 b:=av_0.5
-2.59=%=b:=av_0.5 c:=av_0.5
```

The following may be noted. There is a significant expansion of output even from this relatively simple data. Normally, limits might be set to filter output between −2 and +2 nats, for example. "Fuzzy" associations designated "=%=" as opposed to true associations "= = =" are really processed covariances in comparable numerical representation to associations. The information content in covariance is +2.59 nats for all A with C data and −2.59 with all A with B data and B with C data, indicating negative covariance. A, B, and C are also reported to covary together "as a threesome" but with weaker negative covariance −2.45 nats, reflecting the fact that two are positive correlations, but one is negative. This is because the correction at the end of section 3.5 is used here. Associations for ("unicorn") events that are not observed but which would be expected to be observed on the basis of the occurrence of component events, are indicated by starting with "=Z=". The ">" or "<" for any metastate will be the complementary (opposite) between the associated columns a and c, and the same for the inversely covarying columns a with b and b with c, and "=Z=" events always have negative information. Associations imply estimates of e.g., $\log[P(X,Y)/(P(X)\,P(Y))]$ which is not in general the same as $\log[P(\sim X,\sim Y)/(P(\sim X)\;P(\sim Y))]$, where $\sim X$ implies not X and $\sim Y$ implies not Y. Associations are however symmetrical here in this contrived data. To reduce output, not all equivalent descriptions are reported. Covariances such as would be a 4-fold redundancy of output and

```
+ve= a:=>av  b:=>av
+ve= a:=<av  b:=<av
-ve= a:=>av  b:=<av
-ve= a:=<av  b:=<av
```

redundancy is limited to two.

The following abstract is from the beginning of 209 contrived test data records with qualitative, textual entries. It was generated by combinatorial considerations with a combination missing (brown eyed males with rheumatoid heart valve type). This again illustrates detection of "unicorn events".

```
CRS_Number,Gender,Valve_Type,Eye_Color
1,M,normal,brown
2,M,normal,blue
3,F,normal,brown
4,F,normal,blue
5,M,normal,brown
6,M,rheumatic,blue
7,F,rheumatic,brown
8,F,rheumatic,blue
9,M,normal,brown
10,M,normal,blue
```

The full output is as follows

```
0.49===Valve_Type:=normal Gender:=m Eye_Color:=brown
0.32===Gender:=f Valve_Type:=rheumatic Eye_Color:=brown
0.32===Eye_Color:=blue Valve_Type:=rheumatic Gender:=m
0.23===Valve_Type:=normal Gender:=m
0.23===Valve_Type:=normal Eye_Color:=brown
0.20===Gender:=f Eye_Color:=brown
0.20===Eye_Color:=blue Gender:=m
0.11===Eye_Color:=blue Valve_Type:=rheumatic
0.11===Gender:=f Valve_Type:=rheumatic
0.01===Gender:=f Valve_Type:=normal Eye_Color:=brown
0.01===Eye_Color:=blue Valve_Type:=normal Gender:=m
-0.05===Gender:=f Eye_Color:=blue Valve_Type:=rheumatic
-0.17===Gender:=f Eye_Color:=blue
-0.19===Gender:=f Valve_Type:=normal
-0.19===Eye_Color:=blue Valve_Type:=normal
-0.20===Valve_Type:=rheumatic Gender:=m
-0.20===Valve_Type:=rheumatic Eye_Color:=brown
-0.37===Gender:=f Eye_Color:=blue Valve_Type:=normal
-0.41===Gender:=m Eye_Color:=brown
-3.56=Z=Valve_Type:=rheumatic Gender:=m Eye_Color:=brown
```

4.2 Diagnosis: Record Incidences. The FANO program can also report all incidents (i.e., the record, and hence patient) for which the reported instances were observed. This is of importance for preventative diagnosis and alerting patients at risk. To facilitate compliance of systems with HIPAA and other privacy regulations these can be removed and demographic references are replaced by indirect references to records which irreversibly encrypted keys which nonetheless identify a unique unknown individual for statistical purposes.

The following is a fuller output (but again excluding XML tags) from a simple but real study of clinical record extracts from some 420 records relating diseases to diseases and to polymorphisms of the mitochondrial genome (although the security mode was switched off to display the incidences, the patient IDs are arbitrary and untraceable). The number of

Table 3. Simple Real Case Study Illustrating Diagnostic Benefits of Associations between Diseases and with (mitochondrial) Genomic Features

3.56002===myopathy leber_hereditary_optic_neuropathy [saw 30 expected 1] (coded +2+29+)
INCIDENTS: 29 112 207 208 211 212 213 214 215 216 217 218 219 220 225 232 235 237 238 242 244 249 249 250 251 254 255 261 264 294

3.25156===dystonia stroke-like_episodes [saw 15 expected 0] (coded +103+107+)
INCIDENTS: 206 240 243 247 248 252 256 265 274 276 278 285 287 290 296

3.25156===stroke-like_episodes ldyt_leber's_hereditary_optic_neuropathy [saw 15 expected 0] (coded +107+113+)
INCIDENTS: 206 240 243 247 248 252 256 265 274 276 278 285 287 290 296

3.25156===lactic_acidosis encephalomyopathy ldyt_leber's_hereditary_optic_neuropathy [saw 15 expected 0] (coded +101+109+113+) INCIDENTS: 206 240 243 247 248 252 252 256 265 276 278 285 287 290 296

3.25156===stroke-like_episodes encephalomyopathy ldyt_leber's_hereditary_optic_neuropathy [saw 15 expected 0] (coded +107+109+113+) INCIDENTS: 206 240 243 247 248 252 256 265 274 276 278 285 287 290 296

2.85301=== site x4 a-g mtnd4 [saw 11 expected 1] (coded +11+41+) INCIDENTS: 126 128 130 131 136 140 142 144 147 243 246

2.71785===myopathy mttk trna_lys [saw 9 expected 0] (coded +2+227+241+) INCIDENTS: 328 329 330 331 333 333 334 334 334

2.71785===deafness diabetes_mellitus [saw 9 expected 0] (coded +43+223+) INCIDENTS: 209 270 271 279 289 292 327 341 346

2.67693===myopathy mttk trna_leu [saw 11 expected 1] (coded +2+59+) INCIDENTS: 280 283 284 286 286 288 294 295 297 345 345

1.5000 ===myopathy mttk trna_lys glycoside-induced_deafness [saw 3 expected 0] (coded +2+227+241+383+)
INCIDENTS: 334 334 334

1.00390===myopathy site x1 c-t [saw 5 expected 2] (coded +2+19+) INCIDENTS: 9 65 153 213 255

-1.5872===myopathy site x4 a-g [saw 1 expected 3] (coded +2+2+11+) INCIDENTS: 249

-1.8360===myopathy site x20 g-a [saw 1 expected 4] (coded +2+2+3+) INCIDENTS: 257

incidents of the concurrence of the events (e.g., myopathy and leber_hereditary_optic_neuropathy) actually seen and expected are reported, and the numbers after "INCIDENTS" relate to the patient (in this case, arbitrarily numbered here for privacy). As before, the number at the beginning of each line is the information content in nats. The disease states or other associated items are separated by blanks and words relating to the same item are joined by an underscore. The corresponding codes as prime numbers is given in association with the word "coded" (Table 3).

4.3 Solution to data of Table 2. The reader should attempt to identify by eye the most positively and most negatively associated events in the data of Table 2. The question is not unreasonable: the human brain is moderately good at making such decisions for small amounts of data, presumably by rapid elimination of certain possibilities. However, from the combinatorial perspective even this simple case is challenging, as shown by Table 4.

In Table 5, results as real medical mechanisms are, as ever, putative. Both hepatic and pancreatic dysnfunction do strongly correlate with alcohol. Hepatic protection is apparently conferred by a genetic feature, snp b. Snp_D seems to associate with hepatic dysfunction and snp p with pancreatitis. These diseases show some tendency to occur together, with alcohol as a common environmental factor. However a surprising feature might be rather stronger association of pancreatitis with scorpion bite than with alcohol, as discussed below. Snps a, c, d, and e do occur together and so may up the feature of a haplotype, but snp_d seems to increase the chances of hepatic dysfunction, whereas b confers some protection.

4.4 Effect of Varying *s*. To choose the branch of the Zeta function and the value of s is to choose a particular knowledge model. For example the choice $s = 0$ in the Euler (non-Riemann) branch yields $1 + 1/2^s + 1/3^s + 1/n^s + \ldots = n$ and brings the measure into alignment with direct counting process, which is to say that the information measure is mostly simply the difference between the actual and expected frequency (assuming for simplicity the ZAPD choice of not subtracting one from the frequencies). At the very least, this property at $s = 0$ is a useful device for validating the program, and to assess the effects of approximations used for sampling, interpolation, extrapolation and limiting values of the Zeta function as calculated. These include the impact of the Dirichlet absolutely prior density on low values, as implemented in Fano, and the assumption of linear interpolation of the returned value of the Zeta function for arguments less than 2 and from fractional arguments.

For real $s < 1$, the order is typically mostly preserved but we would not expect it to be preserved exactly. For the case of $s =$

Table 4. Example Screen Output

```
SUMMARY: Information for conjoint vs. random events.
297 counts distributed over 26 events, maximum record length 26 events, and
 maximum possible sample chunk size per record was set as 6.
Multiplets sampled from the highest & lowest of each 6 items/record sample,
Skew seen in test of random number generator = 49.9968% vs ideal 50%.
_____________________________________________________________
Potential complexity of problem:-
Number of potential conjoint events for maximum 26/record:-
Each event can appear in a record only once :67108863
One event can appear in a record x 2        :134217726
Two events can appear in a record x 3       :268435426
Number of potential conjoint events from 26 events to calculate expectations:-
Each event appears only once                :67108863
If events can each appear x 2               :4.5035996273705e+015
If events can each appear x 3               :3.02231454903657e+023
If events can each appear x 4               :2.02824096036517e+031
_____________________________________________________________
These were pruned as follows:-
60 events were > 1 nat or <-1 nat and therefore selected
(including the null state for test purposes).
955 types of combinatorial events (plus 1 catchall) were generated.
955 types  survived to information calculation.
3164 combinations were generated to test zero occurrence strong negative associations.
955 were processed non-arithmetically from lists
0 were recovered numerically from small codes <10000000000000
0 were recovered numerically from big codes >10000000000000
1786 of these were from nonzero events > 1 nat,
Start time Wed Jul 10 18:12:42 2002
*Stop time Wed Jul 10 18:12:52 2002
```

0, the differences between observed frequencies, say, A and expected frequencies, say B, of events are, of order $A-B$ and for $s = 1$, at least for large data, they are on a $\log A - \log B$ scale. For this reason, in the case below, $s = 1$ is here loosely called the "nonlogarithmic case" and that for $s = 1$ and above is colloquially called "logarithmic case". Discussion of the relations between the cases such as $0 < s < 1$ and their utility have been omitted for brevity and will be discussed elsewhere. For real $s > 1$, then with the exception of approximations including sampling for triplets and higher, rank order will tend to be preserved (see eq 6) irrespective of real s subject to sampling and other approximations noted above. This may be seen in comparison of Tables 4 (case $s = 1$) and 6 (case, $s = 2$). For even powers the function converges to the "ceiling" $2^{s-2}\pi^2|B_s|/s!$ and, for example, the choice of $s = 2$ sets $\pi^2/6$ and $s = 4$ sets $\pi^4/90$. In Table 7 the rank order found in Table 5 is given in brackets, and for negative values omitted because this depends on the filter value (i.e., what is omitted round zero) and the above ceiling values. Largest deviations from previous order are due to sampling of triplets and higher associations when the number of records is small.

Increasing $s > 1$ has interesting properties. In Table 8, the effect of very large s is to classify the data simply into $+1$, 0 or -1 nats according to whether the observed frequency of occurrence respectively exceeds, exactly equals, or is exceed by, the expected frequency. This is interesting because it maps the method to trinary logic. It arises because the progressive fall in the ceilings compress the calculated range of Zeta function values to $+1$, and the information (from the difference between two Zeta functions) to $-1...+1$, as s increases. For most purposes, this is reduction to a $+1/-1$ binary logic. Though there would be few exact matches giving zero (which could be enhanced by considering only integral values of expected frequencies) a value of 0 is also a consequence of very sparse data in away determined by the prior probability density, e.g., the Dirichlet absolutely prior density (DAPD) choice.

As illustated by Table 9, trinary logic is a not achieved by large negative values of s. These give similar results to the $s = 0$ case. A large value of negative s ("Euler branch") analogous to the large positive value above cannot be computed because of precision, but -1 down to ca. -10 yields results similar to $s = 0$. The dominance of the concurrence of snp_a and snp_c is particularly notable.

4.5 Analysis of Quantitative Data in Practice. The examples in the present paper concentrate mainly on qualitative associations: in some respects these, and especially very negative associations, of qualitative data are among the most challenging analyses. However, FANO is being extensively involved in analysis of biomedical and other data in numerical and tabular form. For brevity, accounts will be reported elsewhere and information is available on request. Two separate examples are retained here for illustrative purposes. Because the metadata are assigned to qualify the data, e.g., AGE_1:=29 from Table 10, and because then the order is randomized by FANO to ensure no bias, the principle is as for associations in examples above, except that pooling of values can occur and also covariance can be analyzed. Pooling gives two items with metadata WEIGHT_10, for example, which might be WEIGHT_10: =>190 (equal to or great than the mean in that archive of records) and WEIGHT_10:=<190 (by FANO convention, less than the mean in that archive of records). Another option, a three-way division into "normal" within the standard deviation, and below and above the limits of standard deviation, appears of particular value for relating natural or "ideoptypic" variations to haplotypes, and may be of great value in establishing routine normal ranges for patients not based on some arbitrary universal human but rather on their individual molecular ethnicity and variations.

Table 5. Output Ranking of Significant Conjoint Events $s = 1$. Example Extract of Highest and Lowest Ranked Items circa −1<information>2.2 nats[a]

```
3.25===hepatic dysfunction alcoholism
<fano:assn events="hepatic_dysfunction
alcoholism"information="3.25" saw="15"
expected="1.37" coded="+7+19+"
incidents=" 0 1 2 3 4 5 6 8 25 31 32 60
61 63 65"/>.
```

```
3.25===hepatic_dysfunction alcoholism
3.18===snp_c snp_d
3.10===snp_a snp_e
2.92===snp_a alcoholism
2.92===snp_a snp_d
2.92===snp_p snp_q
2.82===snp_a snp_c snp_b
2.82===hepatic_dysfunction snp_d
2.82===hepatic_dysfunction snp_p
2.82===pancreatitis scorpion_bite
2.71===snp_p pancreatitis
2.71===snp_c snp_e
2.59===hepatic_dysfunction snp_q
2.59===snp_d snp_e
2.59===snp_b snp_d
2.45===pancreatitis alcoholism
2.45===snp_c alcoholism
2.45===hepatic_dysfunction pancreatitis
2.28===snp_a schizophrenia
2.28===snp_c snp_b snp_d
2.28===snp_b schizophrenia
2.28===snp_b snp_e
2.28===snp_b pancreatitis
2.28===snp_d cancer
:
:
:
-1.01===snp_b hepatic_dysfunction
-2.18===scorpion_bite x 2
```

[a] The full XML-compliant form is illustrated by the above.

Table 6. Output Ranking of Significant Conjoint Events $s = 0$ ("Euler branch" range of s), Extract[a]

```
15.77===snp_a snp_c <fano:assn
events="snp_a snp_c"
information="15.77" saw="19"
expected="3.22" coded="+2+3+"
incidents=" 0 1 4 5 9 16 19 20 21
22 28 29 40 50 52 59 61 62 63"/>
```

```
15.77===snp_a snp_c
14===hepatic_dysfunction alcoholism
13.55===snp_c snp_b
13===snp_c snp_d
12.22===snp_a snp_b
12===snp_a snp_e
10===snp_p snp_q
10===snp_a alcoholism
10===snp_a snp_d
9.65===snp_c hepatic_dysfunction
9.33===snp_a hepatic_dysfunction
9===pancreatitis scorpion_bite
9===snp_a snp_c snp_b
9===hepatic_dysfunction snp_d
9===hepatic_dysfunction snp_p
8===snp_p pancreatitis
8===snp_c snp_e
7.55===snp_a snp_p
7===snp_d snp_e
7===hepatic_dysfunction snp_q
7===snp_b snp_d
6===snp_c alcoholism
6===snp_a hepatic_dysfunction alcoholism
6===pancreatitis alcoholism
:
:
:
-1.02===snp_b hepatic_dysfunction
-12===scorpion_bite x 2
```

[a] The full XML-compliant form is illustrated by the above.

5. General Results and Conclusions

This paper sought to establish the basis of a broader general theory of Expected Information and hence of data mining, and to give some illustrative applications. A coherent theory seems possible.

A deep implication of the choice of the Zeta with $s = 1$ parameter is that this form arises from the simplest information model concerning the amount of information in a system which is available to the observer. It is also applicable to the matter of holding of strong prior degrees of belief, which merely adds numbers to the frequencies of observation which are the parameters of the Zeta functions.[17] A further interesting consequence of this study is the observation that the higher s values are also relevant.

A typical question relates to the concept of significance. In the perspective of the FANO program at least, nothing in the world is assured, there are only issues of the amounts of available information. In practice, without strong priors, it is usually sufficient to appreciate that measure $\zeta[1,f_o(a,b,c,...)] - \zeta[1,f_e(a,b,c,...)]$ approaches $\log[f_o(a,b,c,...)] - \zeta[2,f_e(a,b,c,...)]$ to better than 90% when there 10−15 observations (observed plus expected) and the ratio of observed to expected frequency is about 2:1. Arguably, notions of significance in classical statistics are illusory because they themselves are based on arbitrary thresholds, e.g., the probability of 0.9 or 0.95 of observing an event by chance. Of course, if one considers $I(a;b;c;..]$ as a random variable, one can evaluate the variance of the information $E\{ I(a;b;c;d;...) - E[I(a;b;c;d;...) \mid D]^2 \mid D\}$, and also use it as a means of normalization, e.g., $I(a;b;c;d;...) \mid D]/ \sqrt{E\{ I(a;b;c;d;...) - E[I(a;b;c;d;...) \mid D]^2 \mid D\}}$. Similarly, the founders of application of covariance methods arbitrarily recommended threshold as to significance, e.g., a positive covariance is customarily considered significant if greater than 0.2 and a negative one is significant if less than −0.2.

Of course, in reporting to other groups, as in the scientific literature, one can agree to speak of associations as being significant at a certain level, even if this is arbitrary. If we ignore any modification of frequencies prior to entering those frequencies to the Zeta function, 1 nat emerges as the information in the first observation for, and minus one nat as the information against, a hypothesis. Given adequate data, events which have an association of 1 nat occur e times (approximately 2.7 times) more than expected. By calculation of events which occur at least 10 times and comparing $E\{I(a;b;c;d;...) \mid D\}$ with

Table 7. Output Ranking of Significant Conjoint Events $s = 2$, Extract with the Rank Position in Brackets that was Found for the Case $s = 1$

```
1.57===hepatic_dysfunction alcoholism (1)
1.57===snp_c snp_d (2)
1.56===snp_a snp_e (3)
1.54===snp_a alcoholism (4)
1.54===snp_a snp_d  (5
1.54===snp_p snp_q (6)
1.53===hepatic_dysfunction snp_d (8)
1.53===hepatic_dysfunction snp_p (9)
1.53===pancreatitis scorpion_bite (10)
1.52===snp_a snp_c snp_b (7)
1.52===snp_p pancreatitis (11)
1.52===snp_c snp_e (12)
1.51===hepatic_dysfunction snp_q (13)
1.51===snp_d snp_e (14)
1.51===snp_b snp_d (15)
1.49===pancreatitis alcoholism (16)
1.49===snp_c alcoholism (17)
1.49===snp_a snp_c hepatic_dysfunction (32)
1.49===hepatic_dysfunction pancreatitis (18)
:
:
:
-1.00===snp_b hepatic_dysfunction
```

Table 8. Conversion to trinary logic (-1,0,+1) by large positive values of s. Output ranking of significant conjoint events s=+1000000, extract

```
1.00===hepatic_dysfunction alcoholism
1.00===snp_c snp_d
:
:
:
:
-1.00===snp_b hepatic_dysfunction
```

Table 9. Effect of Extreme Negative Values of *s*. Output Ranking of Significant Conjoint Events $s = -10$, Extract

```
7792505395211.19===snp_a snp_c
1106532664885.17===snp_c snp_b
529882277575===hepatic_dysfunction alcoholism:
:
:
-23.87===snp_b hepatic_dysfunction
-240627622598===scorpion_bite x 2
```

because any increase in expected frequency subtracts from this value. For covariance, it follows from the properties of the Zeta Function[23] that, whenever s is an even integer, division of the information measure by $2^{s-2}\pi^2|B_s|/s!$ gives an indication comparable to that of covariance on a range $-1...0...+1$. Use of $s = 2$ is hinted at by analogy to second moment variance, in which case $\pi^2/6$ (ca. $\sqrt{e}$) is the divisor for normalization. Then, also assuming the recommended covariance limit of 0.2 in covariance, 1 nat remains a reasonable safe limit given adequate data. For routine use both for associations and covariances, the following simple guidelines are recommended:

3 nats "Definitely worthy of further investigation". Strong effect.

2 nats "Well worthy of further investigation" if observed and expected frequencies $\geqslant 10$.

1 nat "Marginally worthy of further investigation" if observed and expected frequencies $\geqslant 10$.

These considerations are affected by utility and cost, which can vary in circumstances and can be included within the theoretical framework.[17] In medicine, emergency services and military defense there is also the cost of deferment of the decision to act. In an emergency, a physician cannot always take the option of going out and collecting more data for statistical support. The optimal decision must be made however weak the evidence, because, across the sample of many such decisions for many patients, the patients will then benefit. The golden rule, expressed colloquially, is "*When someone's head is on the block, go in the direction for improvement for which the available appropriate Fano mutual information is maximal*". Unfortunately, whether FANO or any other codes can as yet attain the quality justify such a golden rule is unclear, and rather than assume liability, such methods should presume only to serve as decision companions.

6. Discussion

6.1 Illustration of the Method in the Exploration of the Etiology of Immunologicaly Based Diseases. By way of putting discussion in context, we consider investigation of the etiologies of pancreatitis. It is well-known that autoimmune disease can arise from similarity between self-epitopes and those of infecting pathogens, but that there is considerable delay between infection onset of disease and a strong dependence on the haplotype of the host,[30] which makes both medical detection of the association, and a guarantee of etiology, difficult. The patient records were not all real records but plausible forms constructed from collated medical data over many patients. By "contrived", however, is meant something positive, namely, the capability to introduce "prior degrees of belief" or "additional degrees of belief. It can shown equivalent to assigning values to the additive prior data parameters h associated with eq 3, which have consistent form with frequencies of observations, i.e., counts of record contents. Hence, various data sources ranging from unstructured data through to tabular summaries can be combined. The resulting Table 12 could of course be further refined by experts. This is no worse, and arguably much better than, than the Expert System approach of purely assigning expert opinion alone.[29]

One example from the above Tables 4−7 was the association of pancreatitis with scorpion bite, which, until the mechanism is elucidated, might be biologically unexpected. On search of the literature, it turns out that this is experimentally reproduc-

$\sqrt{}$ $E\{ I(a;b;c;d;...) - E[I(a;b;c;d;...) \mid D]^2 \mid D\}$, a choice of 1 nat as a threshold limit also seems reasonable in practice. Thus, as a FANO guideline, $|I| = |\zeta[1,f_o(a,b,c,...)] - \zeta[1,f_e(a,b,c,...)]| > 1$ is recommended. Any higher value such as two nats (a ratio af 7.4 times what is expected) simply adds stronger and stronger support. However, with the DAPD choice of subtracting 1 from frequencies of observation, to reach 3 nats when the expected frequency is zero would require 10 observations, the above-mentioned "critical" number. There is no way to achieve such a positive value with less than 10 observations

Table 10. Example Metadata Plus First Record of 2862 Records, 254 Items a Record of Realistic Collated Metadata and One Record Reflecting Clinical Phenotype of a Patient at Different Ages As Partitioned into Age Groups

```
ID,AGE_1,AGE_2,AGE_3,AGE_4,AGE_5,AGE_6,AGE_7,AGE_8,AGE_9,AGE_10,AGE_11,AGE_12,AGE_13,A
GE_14,AGE_15,AGE_16,AGE_17,AGE_18,AGE_19,AGE_20,AGE_21,CHOL_1,CHOL_2,CHOL_3,CHOL_4,CHO
L_5,CHOL_6,CHOL_7,CHOL_8,CHOL_9,CHOL_10,CHOL_11,CHOL_12,CHOL_13,CHOL_14,CHOL_15,CHOL_1
6,CHOL_17,CHOL_18,CHOL_19,CHOL_20,CHOL_21,CPD_1,CPD_2,CPD_3,CPD_4,CPD_5,CPD_6,CPD_7,CP
D_8,CPD_9,CPD_10,CPD_11,CPD_12,CPD_13,CPD_14,CPD_15,CPD_16,CPD_17,CPD_18,CPD_19,CPD_20
,CPD_21,DRINK_1,DRINK_2,DRINK_3,DRINK_4,DRINK_5,DRINK_6,DRINK_7,DRINK_8,DRINK_9,DRINK_
10,DRINK_11,DRINK_12,DRINK_13,DRINK_14,DRINK_15,DRINK_16,DRINK_17,DRINK_18,DRINK_19,DR
INK_20,DRINK_21,EXAMDTH,GLUC_1,GLUC_2,GLUC_3,GLUC_4,GLUC_5,GLUC_6,GLUC_7,GLUC_8,GLUC_9
,GLUC_10,GLUC_11,GLUC_12,GLUC_13,GLUC_14,GLUC_15,GLUC_16,GLUC_17,GLUC_18,GLUC_19,GLUC_
20,GLUC_21,HBP_1,HBP_2,HBP_3,HBP_4,HBP_5,HBP_6,HBP_7,HBP_8,HBP_9,HBP_10,HBP_11,HBP_12,
HBP_13,HBP_14,HBP_15,HBP_16,HBP_17,HBP_18,HBP_19,HBP_20,HBP_21,HDL_1,HDL_2,HDL_3,HDL_4
,HDL_5,HDL_6,HDL_7,HDL_8,HDL_9,HDL_10,HDL_11,HDL_12,HDL_13,HDL_14,HDL_15,HDL_16,HDL_17
,HDL_18,HDL_19,HDL_20,HDL_21,HGT_1,HGT_2,HGT_3,HGT_4,HGT_5,HGT_6,HGT_7,HGT_8,HGT_9,HGT
_10,HGT_11,HGT_12,HGT_13,HGT_14,HGT_15,HGT_16,HGT_17,HGT_18,HGT_19,HGT_20,HGT_21,HRX_1
,HRX_2,HRX_3,HRX_4,HRX_5,HRX_6,HRX_7,HRX_8,HRX_9,HRX_10,HRX_11,HRX_12,HRX_13,HRX_14,HR
X_15,HRX_16,HRX_17,HRX_18,HRX_19,HRX_20,HRX_21,SBP_1,SBP_2,SBP_3,SBP_4,SBP_5,SBP_6,SBP
_7,SBP_8,SBP_9,SBP_10,SBP_11,SBP_12,SBP_13,SBP_14,SBP_15,SBP_16,SBP_17,SBP_18,SBP_19,S
BP_20,SBP_21,TRIG_1,TRIG_2,TRIG_3,TRIG_4,TRIG_5,TRIG_6,TRIG_7,TRIG_8,TRIG_9,TRIG_10,TR
IG_11,TRIG_12,TRIG_13,TRIG_14,TRIG_15,TRIG_16,TRIG_17,TRIG_18,TRIG_19,TRIG_20,TRIG_21,
WGT_1,WGT_2,WGT_3,WGT_4,WGT_5,WGT_6,WGT_7,WGT_8,WGT_9,WGT_10,WGT_11,WGT_12,WGT_13,WGT_
14,WGT_15,WGT_16,WGT_17,WGT_18,WGT_19,WGT_20,WGT_21
4,29,31,33,35,37,39,41,43,45,47,49,51,53,55,57,-9,61,-9,65,-9,-
9,195,193,188,188,190,181,207,192,190,195,194,-9,208,213,205,-9,-9,-9,-9,-9,-9,20,20,-
9,20,20,-9,20,20,20,20,20,20,20,20,-9,20,-9,20,-9,-9,-9,0,-9,-9,-9,-9,0,-9,-9,-9,-
9,0,0,0,0,-9,0,-9,-9,-9,-9,0,103,102,96,102,-9,102,-9,100,103,93,-9,105,97,101,102,-
9,103,-9,98,-9,-9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,-9,0,-9,0,-9,-9,-9,-9,-9,-9,-9,-9,-
9,-9,-9,34,-9,-9,-9,35,-9,-9,-9,-9,-9,-9,65,-9,-9,-9,66,-9,-9,-9,-9,66,-9,-
9,65,65,65,-9,66,-9,66,-9,-9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,-9,0,-9,0,-9,-
9,113,117,115,114,121,119,114,119,119,124,117,124,126,126,131,-9,128,-9,132,-9,-9,-9,-
9,-9,-9,-9,-9,-9,-9,-9,-9,118,-9,-9,-9,-9,-9,-9,-9,-9,-9,-
9,146,145,145,145,154,140,150,153,147,152,147,143,150,145,143,-9,146,-9,151,-9,-9
```

ible in animal models.[26] There is a very high association between scorpion bite and pancreatitis. Also, anecdotal evidence that a second bite in a few unlucky cases almost guaranteed a subsequent attack suggests that the mechanism may hint at an autoimmunization effect. This hints at induction of antibodies against a protein or proteins of the host which directly or indirectly regulate, or are associated with, pancreatic activity. To test this, one could look for correlation with the protein sequences of host and pathogen, infectious disease, and possibly patient haplotype. FANO can handle sections of biological sequences as items in a larger record represented by the DNA or protein. This allows *records in which genes of pathogens or venoms are brought together* with a disease state on the clinical record. There are, of course, several established sequence-based bioinformatics tools by which can perform such an investigation to track similar sections of sequence of ca. 10 contiguous amino acids, which might represent similar loops which could function as epitopes, so causing an immunological cross-reaction. However, we may "show off" FANO flexibility and the scope of the existing core algorithm. In sampling sequences, the ZAPD option is chosen and the FANO input control commands allow the sampling to be "staggered", considering characters 1....M as the first item, 2...$M + 1$ as the second, and so on. This identified homologous sections of about 8−15 residues between human proteins and those on venoms and pathogens to which the patients had been exposed, according to their records. These detailed results have been removed for the brevity of this paper but examples include rabies virus glycoprotein with insulin receptor, Papilloma virus E2 with insulin receptor, Poliovirus VP2 with acetylcholine receptor, Measles Virus P3 with corticotropin, HIV p24 protein Human IgG constant region, reptile venom peptides with Helodesmin, and helospectins etc, with VIP and the secretin group, and, notably, scorpion venom neurotoxin with secretin. In the case of the scorpion venom toxin, it is likely that antibodies are raised against secretin but in this case do not neutralize it, but rather enhance its effect by preventing clearance, leading to hyperstimulation of the pancreas. A similar clearance prevention phenomena was also noted in relation to autoimmunization against growth hormone. One might now perform the experiment of synthesizing these peptides and performing immunological studies. Hence, although the above are putative, the above constitutes a valid protocol for etiological investigation. In a more realistic but ambitious protocol, the overall "SHAMAN" research framework that includes FANO, does have such capabilities which can be mobilized to model discontinuous and conformational epitopic determinants (See ref 22 and Appendix thereof).

6.2 Other Issues Arising. Several secondary discussions as follows arose in this study which, for brevity, will now discussed in more detail elsewhere. [1] A comparison of FANO with other techniques shows that FANO is differentiated by the complexity of events, and especially highly negatively associated events, that it can detect, and its unified approach to qualified and numeric data. [2] "How universal is role of the Zeta function as an information measure?" Zeta functions relate to the basic notions of information in counting, and may generally be considered as "atomic" to a wide variety of information expressions. By "atomic" is meant that they can appear as recurrent terms in expansions or expressions for a variety of information theoretic forms, but not that they are indivisible because these can certainly be still further resolved further into constituent terms $1/n^s$ (see above). [3] It may help understanding to say that the present methods bring the Information Theoretic approach closer to a neural network approach, except that the learning process by which weights are assigned is

Table 11. Example Output from Another Real Clinical Study, Extract of XML File[a]

```
-2.73=Z=Lymphocyte_Count_(g/L):=<6.11 Eosinophil_Count_(g/L):=<0.25
Basophil_Count_(g/L):=<3.59  <fano:assn events="Lymphocyte_Count_(g/L):=<6.11
Eosinophil_Count_(g/L):=<0.25 Basophil_Count_(g/L):=<3.59 " information="-2.73"
saw="0" expected="9.12" coded="+11+13+17+" incidents=""/>
-2.73===Lymphocyte_Count_(g/L):=<6.11 Basophil_Count_(g/L):=<3.59
Monocyte_Count_(g/L):=<1.66 <fano:assn events="Lymphocyte_Count_(g/L):=<6.11
Basophil_Count_(g/L):=<3.59 Monocyte_Count_(g/L):=<1.66" information="-2.73" saw="1"
expected="9.12" coded="+11+17+19+" incidents=" 28"/>
-2.73=%=Potassium_(mmol/L):=av_4.21 Platelet_count_(g/L):=av_173.20 <fano:covn
events="Potassium_(mmol/L):=av_4.21 Platelet_count_(g/L):=av_173.20" information="-
2.73" saw="2.24" of="29" coded="+0+157+" incidents="all with numeric
Potassium_(mmol/L) Platelet_count_(g/L)"/>
-2.73===Lymphocyte_Count_(g/L):=<6.11 Eosinophil_Count_(g/L):=<0.25 CRS#:=>7881.69
<fano:assn events="Lymphocyte_Count_(g/L):=<6.11 Eosinophil_Count_(g/L):=<0.25
CRS#:=>7881.69" information="-2.73" saw="1" expected="9.12" coded="+11+13+127+"
incidents=" 17"/>
-2.73=Z=Lymphocyte_Count_(g/L):=<6.11 Eosinophil_Count_(g/L):=<0.25
Monocyte_Count_(g/L):=<1.66  <fano:assn events="Lymphocyte_Count_(g/L):=<6.11
Eosinophil_Count_(g/L):=<0.25 Monocyte_Count_(g/L):=<1.66 " information="-2.73"
saw="0" expected="9.12" coded="+11+13+19+" incidents=""/>
:
:
:
<fano:multivariance status="experimental">
<fano:covariance_Coefficient_optimization pass="1" function_value="2626.48802168178">
<fano:interest metatstate="Date_of_Examination" column="0" likely_value="3%"
optimized_omnivariate_value="3%" estimated_from_fuzzy="0%"/>
<fano:interest metatstate="Person_Responsible" column="1" likely_value="0%"
optimized_omnivariate_value="87%" estimated_from_fuzzy="0%"/>
<fano:interest metatstate="CRS#" column="2" likely_value="35%"
optimized_omnivariate_value="19%" estimated_from_fuzzy="100%"/>
<fano:interest metatstate="White_blood_cell_count_(g/L)" column="3"
likely_value="100%" optimized_omnivariate_value="100%" estimated_from_fuzzy="100%"/>
:
:
```

[a] The format of "equals" sign to the right of the first number, the information value, indicates type of measure. === positive or negative association, =Z= negative associations where concurrence was expected but not observed in sample, events =%= covariance. The extract at the foot is the statistical summary uses the method related to eq 4 to warn the use of high dimensional covariances which may have been missed, or weak covariances for which columns of data can be removed in a subsequent study (in the original file, symbols > and < are converted to > and < resepctively for xml compliance).

Table 12. Some Potential Etiological Correlations Obtained from Extended Preliminary Fano Study on Pancreatitis[a]

nats	etiology	nats	etiology
4	gallstones	1	mumps
3	scorpion bite	1	hepatitis b
3	malnutrition, restricted tropical diets	1	mumps
3	cystic fibrosis	1	HIV
2	alcohol	1	hyperglycaemia
2	oxidative stress due to selnium/cartene/vitamin	1	porphyria
	a, c or e deficiency	1	chronic renal failure
1	insecticides	1	inflamatory bowel disease
1	duodenal obstruction or infection	1	anorexia nervosa
1	hydrophobic agents (e.g., aliphatic/hydrogenated	1	essential fatty acid deficiency
	hydrcoarbos(aviation fuel) anesthetics,	1	Reye's syndrome
1	cardiovascular bypass surgery	1	hemolytics uraemic syndrome
1	diazinon	1	long-term antibiotics
1	fast	1	hypothermia
1	cardiopulmonary bypass	1	trans-lumbar aortography
1	serum triglycerides >20 mM/L	1	vasculitis
1	surfeit after fast	<0.5	pregnancy
1	post-renal bypass/transplant	0...−2	unknown causes
1	trauma		

[a] These are placed in rank order but data is sparse in many cases. "Unknown causes"(last row of table) was treated as an event.

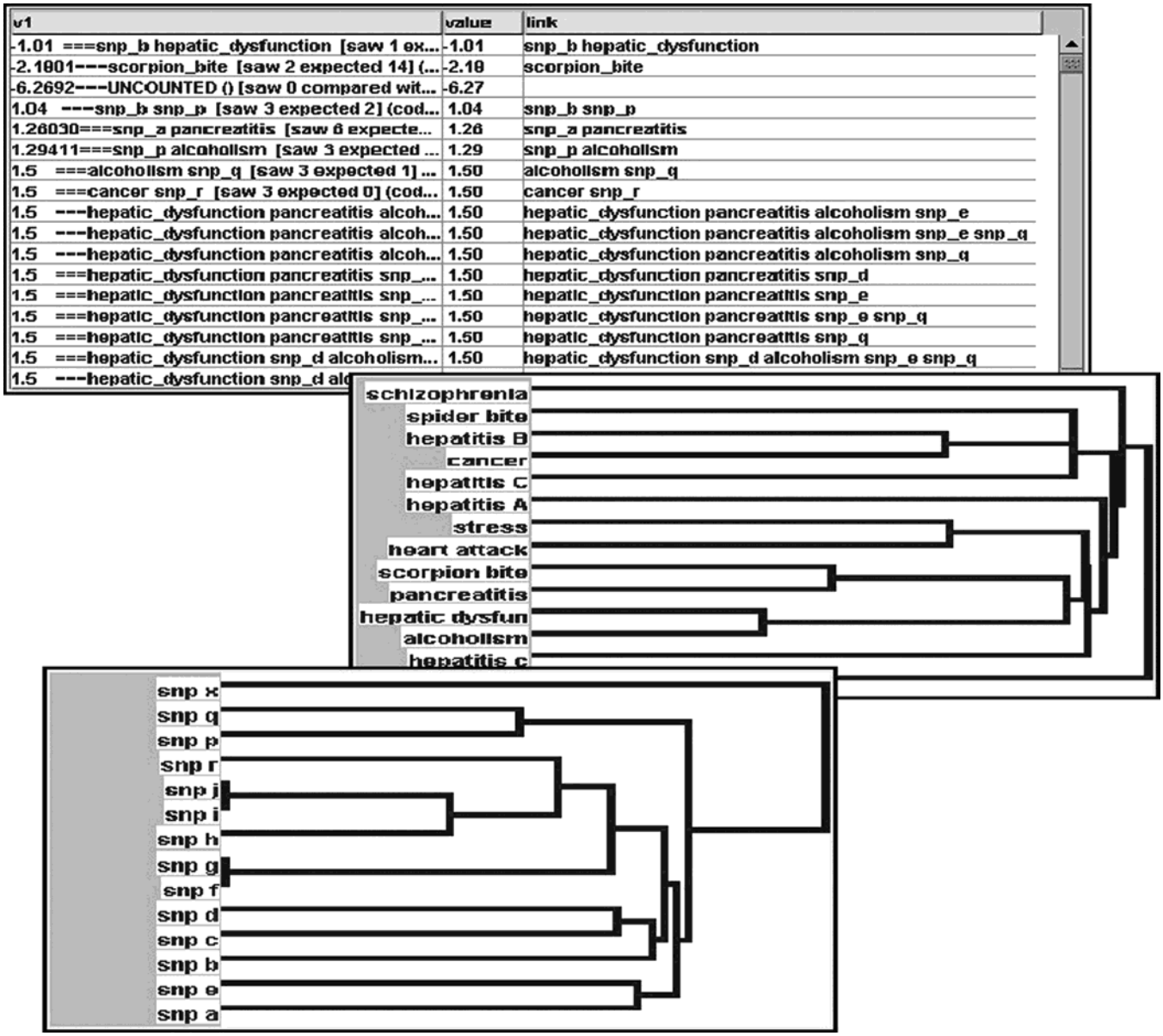

Figure 1. Output from an example Fano run. Above screen display with dendrogram indicating disease groupings. Below, SNPs dendrogram in haplotype research. Data similar to that of Table 2.

directly, by source data analysis, rather than by feedback based on output performance. Importantly, these weights as distributed by the "learning process" are much more readily understood and applicable to generate rules and rule strengths for a potential rule-based decision-support system. [4] FANO is installed with customer-collaborators. Accounts of the configurations and subsequent various manipulation tools are available at this time from IBM Research. An example graphic output of the FANO ranking using PRIMA (see acknowledgments) is shown in Figure 1.

Acknowledgment. I am grateful to the referees and also to Chid Apte and Emmanuel Yashchin of IBM Research, for critical reading of the manuscript, and to my manager William Pulleyblank, a mathematician who made many insightful comments in presentations of the theory. I am also grateful to Richard Mushlin, my close collaborator who collated, preprocessed, and provided much pharmacogenomic data, as well as Professors Bruce McManus, Kosta Steliou, and others for relevant real data. I thank also OK Baek, Mike Koranda, and Ronald Martin of IBM Global Services and Life Sciences for management of a pilot installation project making extensive use of FANO, and especially their colleague Erik Voldal for provision of clinical data, for substantial testing and suggestions, and for providing many plug-ins for graphical display of Fano output. I am similarly grateful to members of the SHAMAN medical informatics project team. The figure in this paper were generated using the PRIMA system developed from the IBM OPAL vizualization package by the team of Bernice Rogowitz at IBM Hawthorne specifically for the SHAMAN project. This project was supported as part of the SHAMAN project by IBM Life Sciences and the Computational Biology Center of IBM Research in New York.

Supporting Information Available: A more detailed account on a command-by-command basis is given for the program and control. This material is available free of charge via the Internet at http://pubs.acs.org.

Appendix 1.

Statements in number theory and their mapping to elements of a general theory of data and of data mining.

Table A1.1 Some Basic "Mappings" can be Summarized as Follows

number theoretic element	description	statement in data theory
P	A prime number, divisible only by itself and 1.	The "Goedel number" or code for a unique item (e.g., Weight > 200) in a record (e.g., A clinical record). If an entry item (e.g., male) has metadata (e.g., Gender) then the item is the entry qualified by the metadata (e.g., Gender:=male).
P^a	A prime number to the (a)th power	The "Goedel number" or code for a unique item seen (a) times in a record or subrecord such as a medical record.
$N = P_1^{a(1)} P_2^{a(2)} P_3^{a(3)} \ldots$	The natural number N represented as a product of powers of primes	"Goedel number" or code of a record (e.g., a medical record) with items represented by primes P_i and occurring $a(i)$ times.
UFT, $N = L\ M$	Unique Factorization Theorem: all products of primes factorize uniquely to those primes. Any product of primes N is equivalent to the arithmetic product of a prime or product of primes L with a prime or product of primes M.	Any two records or subrecords match, i.e., are identical by content, when their Goedel number is the same. Concatenation of records into one record is independent of item order (when metadata is treated as above).
$J\|N$	J is a divisor of N (divides into N without remainder)	J is the Goedel number of an item or subrecord (or "observation") within the record of Geodel number N. The subrecord coded by with Goedel number J matches that coded by Goedel number N by one or more items, e.g., (Caucasian, Gender:=Male, weight > 200,). All items and subrecords of a record are generated exhaustively and nonreduntantly by dividing the Goedel number of the record by 2,3,4,5,.... Without remainder (division by 1 generates the record itself).
K, $1 \leqslant K \leqslant N$	Any positive integer less than or equal to N, i.e., a "potential divisor" prior to the division test K/N.	K is the Goedel number of a "Potential Subrecord" or "Potential Observation" wrt. N, The potential observation may, or may not, be actual.
$F(N) => F(K), K\|N$	Defined number theoretic functions F of N are also defined number theoretic functions of K.	An actual subrecord or observation can be considered as and treated as a record
$N = \phi(N) + \tau(N) + \xi(N) - 1$	Statement in number theory on the relations between number theoretic functions (defined below)	Statement in Data Theory about the relations between potential subrecords (observations) according to whether and how potential entries are actually seen in the record of Goedel number N
$\tau(N)$	The number of divisors of N, including 1 (i.e., the number of divisions without remainder)	One plus the number of actual "subrecords" ("observations") in the record of Goedel number N

Table A1.1 (Continued)

number theoretic element	description	statement in data theory
$\phi(N)$	The number of real numbers K, $1 \leqslant K \leqslant N$, which are coprime to N	The number of potential subrecords (potential observations) for which no items actually occur in the record of Goedel number N
$\xi(N)$	The number of real numbers K, $1 \leqslant K \leqslant N$, which are not divisors of N and which have a greatest common divisor other than 1.	The number of potential subrecords (potential observations) that intersect, i.e., for which at least one item actually occurs in the record of Goedel number N. *Useful to find subrecords from which unions can be formed.*
$\omega(N)$	The number of distinct primes $P_1\ P_2\ P_3$ .... In N.	The number of types of item (of distinct items) in the record of Geodel number N, i.e., the number if items ignoring recurrences
$\mu(N)$	$(-1)^r$if r is the number of distinct primes $P_1\ P_2\ P_3$ ...., 1 if $N = 1$, 0 otherwise	-1 if the number of distinct items in a record of Goedel number N is odd, $+1$ if even.
$\sigma(N)$	The sum of the divisors of N, including 1 (i.e., the number of divisions without remainder)	One plus the sum of the Goedel numbers of all the actual "subrecords" ("observations") in the record of "Goedel number" N
$\sigma(N)/N < \log(N) + 1 <$ N. $\zeta(s = 2) = \mathrm{N}/\Sigma_{J=1..\infty}(\mu(J)/J^2) = \pi^2\mathrm{N}/6$	The arbitrarily close upper bound on $\sigma(N)$ (Dirichlet, 1838). ζ is defined below.	Specification of the maximum information content in record of Goedel number N
$1/\xi(s)$	The probability that any s arbitrarily selected real numbers at least one will be coprime to the others (proven for $s = 2$, Cesaro, 1881)	The probability that at least one of s arbitrarily selected numbers will be a Goedel code implying a record or subrecord (observation) unique in all items wrt the others.
$\xi(s = 1,n)$	The incomplete summation of a Zeta function with s:1, up to term n.	Some general information measure where n is the number of observations of a record or subrecord plus a value which represents the impact of prior belief. Shown to be the Bayes expected value of information given the likelihood, in the case $s = 1$ (Robson, 1974).
$\xi(s, n)$, n:=ZAPD, DAPD; $(\xi(s, n)$, ZAPD) $\leftrightarrow \xi(s, n)$, DAPD), $n \rightarrow \infty$	The Zeta function with n defined according to two different bases formulated such that the results converge mutually for large n. DAPD means subtract one from all arguments, ZAPD means leave as-is.	Two relations relating information measure arguments to number of observation of records or subrecords coded as N. ZAPD defines observed frequency $n = o(N)$ and implies a zero absolutely prior probability density, and $n = o(N) - 1$ implies use of Dirichlet's ("improper") choice of absolutely prior probability density.
$n = \xi(\mathrm{s} = 0, n)$ Euler branch	The value of the Euler Zeta Function (i.e., without Riemann's analytical continuations), at $s = 0$.	The simple counting of n observations with zero absolutely prior probability (ZAPD) or or $n + 1$ observations according to Dirichlet's absolutely prior (DAPD)
$\xi(s = 1,n) \rightarrow \log(n) + \gamma$, $n \rightarrow \infty$ $\gamma = 0.577\ 215\ 6...$	The limit in the incomplete summation of a Zeta function with $s = 1$ up to term n	The natural information in n (ZAPD) or $n + 1$ (DAPD) observations of a record pr subrecord
$\xi(s = 1,\mathrm{n}) - \xi(s = 1,m) \rightarrow \log(n/m)$, $n \gg 1$, $m \gg 1$	The difference between two Zeta functions for $s = 1$	The natural information in n (ZAPD) or $n + 1$ (DAPD) observations of a record or subrecord, relative to that in some other number m (ZAPD) or $m + 1$ (DAPD) of observations the same or some different other record or subrecord.
$\xi(s = 1,O(N)) - \xi(s = 1,E(N)) \rightarrow \log(O(N)/E(N))$, $O(N) \gg 1$, $E(N) \gg 1$	The difference between two Zeta functions for $s = 1$	A measure of the association between the individual items in the record or subrecord coded by the Godel number N, where $O(N)$ (ZAPD) or ($O(N + 1)$ DAPD) is the observed and $E(N)$ (ZAPD) or $(1 + E(N) + 1)$ (DAPD) is the expected frequency of occurrence of that record.
$\xi(s,O(N)) - \xi(s,E(N)) \rightarrow [-1,0,+1]$ $s \gg 1$, $O(N) > 1$, $E(N) > 1$	The limit of the difference between incomplete Zeta function for large s converges to -1 or 0 or $+1$.	The trinary logic measure (true, do not know, false) that a specified record or subrecord coded by Goedel number N will be observed more than expected.

Table A1.1 (Continued)

number theoretic element	description	statement in data theory
$\xi(s,O(N)) - \xi(s,E(N)) \rightarrow 0$, $s \gg 1$, $O(N) \leqslant 1$, $E(N) \leqslant 1$, $\mid O(N) - E(N) \mid = 0$, ZAPD $\xi(s,O(N)) - \xi(s,E(N)) \rightarrow 0$, $s \gg 1$, $O(N) \leqslant 1$, $E(N)$Ÿ1, $\mid O(N) - E(N)) \mid < 1$, DAPD	The zero limit of the difference between complete Zeta function for large s conditional on n defined on two different bases	The choice of absolutely prior probability density determines the outcome of zero information at low frequencies of observation.
$-\xi(s = 1,E(N)) \rightarrow -\log(E(N)) - \gamma$, $n \rightarrow \infty$, $\gamma = 0.5772156...$	The limit of the value of (−1).Zeta function for large s.	Some general negative information that a record or subrecord (observation) was expected to occur $E(N)$ times (ZAPD) or $E(N) + 1$ times (DAPD), but was not in fact observed.
$\xi(s,c.M) - \xi(s,(1 - c).M) \rightarrow \log(c/(1 - c))$, $M \gg 1$	The difference between complete Zeta functions for large s where the arguments are expressed as a ratio $(c/(1 - c)$.	Definition of the degree of "fuzzy association" between metadata with M (ZAPD) or $M + 1$ (DAPD) comparable (defined) numerical entries (e.g., columns). Here $c = C(\mathrm{cov} + 1)/2$, $-1 \leqslant c \leqslant +1$ for Covariance Cov of the numeric entries under those metadata.
$\xi(s = s + i.t, n)$	The incomplete Extended Riemann Zeta function	Some general information measure (with real and imaginary components) for vector logic (possible basis of a quantitative predicate calculus).

References

(1) Lander, S. E. Foreward to *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 2nd ed.; Baxevanis, D. A., Ouellette, B. F. F., Eds.; John Wiley & Sons: New York, 2001.
(2) Fischer, R. A. *Statistical Methods for Reserach Workers*; Oliver and Boyd: Edinburgh, 1941.
(3) Aitken, A. C. *Statistcial Mathematics*; Oliver and Boyd: Edinburgh, 1945 .
(4) Wilks, S. S. *Mathematical Statistics*; Wiley: New York, 1962; Section 7.7.
(5) Whittle, P. *Probability, Library of University Mathematics*; Penguine Books: London, 1970.
(6) Shannon, C.; Weaver *The Mathematical Theory of Communication*; University of Illinois Press, Champaign, 1949.
(7) Kullback, S. *Information Theory and Statistics*; Wiley: New York, 1959.
(8) Savage, L. J. *Recent Advances in Information and Decision Processes*; Macmillan: New York, 1962; pp 164−194.
(9) Goode, H. *Recent Developments in Information and Descision Processes*; Machol, Grey, Eds.; Macmillan: New York, 1962; pp 74−76.
(10) Fano, R *Transmission of Information*; Wiley: New York, 1961.
(11) Pain, R. H.; Robson B. Analysis of the code relating sequence to conformation in globular proteins. *Nature* **1970**, *227*, 62−63.
(12) Bayes, T. *Philos. Trans. R. Soc. London, Ser. B* **1763**, *53*, 370−148.
(13) Lindley, D. V. *An Introduction to Probability and Statistics from a Bayesian Viewpoint: Part 2, Inference*; Cambridge University Press: New York, 1965.
(14) Silvey, S. D. *Statistical Inference, Library of University Mathematics*; Penguine Books: London, 1969.
(15) Robson, B. Theoretical and Experimental Studies on Protein Folding. Ph.D. Thesis, University of Newcastle Upon Tyne, 1971.
(16) Robson, B.; Pain, R. H. Analysis of the code relating sequence to conformation in globular proteins: Possible implications for the mechanism of formation of helical regions. *J. Mol. Biol.* **1971**, *58*, 237−256.
(17) Robson, B. Analysis of the code relating sequence to conformation in globular proteins: Theory and application of expected information. *Biochem. J.* **1974**, *141*, 853−867. (First full development of Bayes Expected Information theory method).
(18) Robson, B.; Suzuki, E. Conformation properties of amino acid residues in globular proteins. *J. Mol. Biol.* **1976**, *107*, 327−356.
(19) Garnier, J.; Osguthorpe, D. J.; Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **1978**, *120*, 97−120. (The GOR paper—paper with the large mutiple citation award.)
(20) Garnier, J.; Robson, B. The GOR Method for Predicting Secondary Structure in Proteins. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum Publishing Corp.: New York, 1989; pp 417−465. Garnier, J.; Francoise, J.; Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **1996**, 266 (*Computer Methods for Macromolecular Sequence Analysis*; Dolittle, R. F., Ed.).
(21) Crampin, J.; Nicholson, B. H.; Robson, B. Protein folding and heterogeneity inside globular proteins. *Nature* **1978**, *272*, 558−560.
(22) Robson, B. Fold diagnosis: Studies in the assessment of folding quality for protein modeling and simulation when the experimental structure is unknown. *J. Proteome Res.* **2002**, in press.
(23) Edwards, H. M. *Riemann's Zeta Function*; Dover Publications: New York, 1974.
(24) Robson, B. Alignment-Free Method for Rapid Determination of Differences between a Test Data Set and Known Data Sets. U.S. Patent 6, 434, 488, 2002.
(25) Robson, B.; Platt, E. Refined models for computer calculations in protein engineering. Calculation and testing of atomic functions compatible with more efficient calculations. *B. Mol. Biol.* **1986**, *188*, 259−281.
(26) Williams, J. A.; Gallagher, S.; Sankaran, S. Scorpion toxin-induced amylase secretion in guinea pig pancreas: Evidence for a new neurotransmitter. *Proc. Soc. Exp. Biol. Med.* **1982**, *170*, 384−389.5
(27) Personal communication. Compiled from patient data with the help and information particularly of Dr. J. Braganza and colleagues, The University of Manchester, Machester Royal Infirmary United Kingdom, and from Mount Sinai Hospital, New York.
(28) Braganza, J. M., Ed. *Pathogenisis of Pancreatitis*; Manchester University Press: Manchester, 1991.
(29) Buchanan B. G.; Shortliffee, E. H. *Rule Base Expert Systems: The Mycin experiments of the Stanford Heuristic Programming Project*; Addison-Wesley: Reading, London, Amsterdam, Ontario, Sydney, 1984.
(30) Marsh, G. E., Parham, P.; Barber, L. A. *The HLA Facts Book*; Academic Press: San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo, 2000; pp 79−83, .

PR025587Q