Scientific Research Publishing

Open Journal of Applied Sciences, 2026, 16(6), 2196-2228
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Diagnostic Reliability of Medical Test Results as Routinely Reported to Physicians and Patients by Clinical Laboratories

Mark P. Silverman[1,2]

[1]G. A. Jarvis Professor of Physics, Emer., Trinity College, Hartford, USA
[2]Tall Pines Research, Simsbury, USA
Email: mark.silverman@trincoll.edu, jwmgibbs@gmail.com

How to cite this paper: Silverman, M.P. (2026) Diagnostic Reliability of Medical Test Results as Routinely Reported to Physicians and Patients by Clinical Laboratories. *Open Journal of Applied Sciences*, **16**, 2196-2228.
https://doi.org/10.4236/ojapps.2026.166125

Received: May 1, 2026
Accepted: June 21, 2026
Published: June 24, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Diagnostic clinical labs are frequently requisitioned by healthcare providers to perform tests of medical risk factors of their patients. The result of each test is ordinarily reported as a "point estimate", *i.e.* a single numerical value—the "mean" of presumably one measurement—with upper and lower limits of a reference range, but with *no* accompanying information regarding the uncertainty of the tests, the number of trials, or the correlation of the risk factors. On the basis of such incomplete information, a physician must then gauge whether a patient is or is not at risk for the illnesses or conditions associated with the measured factor. In this article rigorous methods drawn from statistical physics (Principle of Maximum Entropy) are employed to derive the *least biased*, *most probable* predictive distribution of medical test results consistent with the information reported by clinical labs, as described above. Among other things, one can predict the probability that a repeat test result obtained from the same sample falls within or outside the reference range. It is shown theoretically and by specific examples of risk factors in the standard Lipid and Metabolic Panels how unreliable would be diagnostic inferences based on these routine clinical laboratory reports. Suggestions are made to rectify this unnecessary deficiency, which can lead to serious misdiagnoses.

Keywords

Medical Risk Factors, Diagnostic Reliability, Diagnostic Clinical Labs, Test Uncertainty, Point Estimate, Information Entropy, Principle of Maximum Entropy, Maximum Entropy Probability, Lipid Panel, Lipoproteins, HDL Cholesterol, LDL Cholesterol, Metabolic Panel, Coefficient of Variation

"*The most misleading assumptions are the ones you don't even know you're making*."

Douglas Adams (1990)[1]

1. Introduction: A Critical Omission

The practice of medicine is both an art and a science. The art lies in the intuition and experience of the practitioner. The science lies in numbers, the results of careful measurement and analysis. These numbers are most often the results of tests of medical risk factors performed by clinical diagnostic laboratories at the requisition of physicians, both primary care and specialist. According to the U.S. Centers for Disease Control and Prevention (CDC), about 14 *billion* laboratory tests are performed each year in the United States by certified diagnostic laboratories [1].

For example, among the most common types of laboratory tests are a patient's 1) Lipid Panel (which identifies and quantifies the types of serum cholesterol, a risk factor for cardiovascular disease), 2) Metabolic Panel (which includes measurements of glucose, creatinine, and albumin, risk factors for diabetes and kidney disease), and 3) Complete Blood Count (which measures the concentrations of red cells, white cells, platelets, hemoglobin, and other constituents of blood which are risk factors for anemia, infections, and blood disorders).

The measured substances (referred to as analytes) in each panel are risk factors, *i.e.* biomolecular products whose excess or deficiency relative to some reference may signify a risk for some pathological (or at least anomalous) condition. In principle, the intent of each clinical test is to enable the requisitioning physician to decide whether a patient is currently at risk for the condition and/or to project whether such a condition is likely to arise in the future. Having this information, the physician can then propose therapeutic and/or preventative measures.

In the US, the outcome of each test is ordinarily reported as a point estimate—*i.e.* a single numerical value, the "mean" of *one* measurement—together with upper and lower limits of a reference range, but with *no* accompanying metric of uncertainty associated with the measurement. On the basis of such limited information, a physician must then gauge whether a patient is or is not at risk for the illnesses or conditions being screened. Although, whenever possible, a competent physician will take into consideration a patient's health history, test trends, and symptoms (if any), this additional information is not always available, especially if the patient has not been examined previously. Thus the physician's decision made during an exam, which in the US can be as short as 15 minutes, often amounts to "no risk" if a test value falls within or on the boundaries of the reference interval; "risk" if outside the boundaries.

As a nuclear and medical physicist, the author finds the above-described widespread practice of communicating medical test results to be seriously deficient [2].

[1]Adams, D. and Carwardine, M (2018), *Last Chance to* See, (Ballantine Books, New York) 95; originally published in Great Britain (1990) by William Heinemann Ltd.

First, as a fundamental part of science education, especially in the physical sciences, one is frequently reminded that no empirically acquired number is reliable, or even meaningful, without a measure of its uncertainty, such as variance, or standard deviation of the mean, or confidence interval (CI), or coefficient of variation (CV). Second, and more specific to the context of diagnostic medicine, without a measure of uncertainty a physician cannot reliably judge whether a test result—and therefore a patient's risk—is *actually* within or outside reference bounds. Nor can a physician test consistency by ascertaining whether a repetition of the test on the same sample would yield a result statistically equivalent to the first result.

It is to be emphasized that certified diagnostic laboratories are undoubtedly aware of the statistical uncertainties of the tests they conduct. They would not be certified otherwise. There is, in fact, an international standards guide [3] produced by the *International Bureau of Weights and Measures* (BIPM) and the *Comité international des Poids et Mesures* (CIPM) that provides rules for the expression of measurement uncertainty and describes how different sources of uncertainties are arrived at. Moreover, there are numerous published references, such as [4]-[6], for informing diagnostic laboratory personnel of correct procedures for measuring and calculating uncertainties in laboratory medicine. So *why*, one might ask, is an estimate of uncertainty not provided together with the reported point estimate of each test? The simple, straightforward answer given to the author by a colleague who directed the clinical laboratory of a large regional hospital is this: Neither physicians nor patients would know what to do with it.

If such an assessment is generally valid, then clearly there is a serious educational deficiency in the training of physicians (not to mention the level of education of the general public). To rectify this deficiency medical school curricula would ultimately need to include basic courses in probability and statistics, or at least require that applicants for admission have taken such courses as undergraduates. Regrettably, educational reforms are usually slow in coming and, until such time, one can expect (at least in the US and probably elsewhere), that clinical laboratories will continue to report point estimates of test results in the same incomplete way described above.

In view of these circumstances, the objective of this paper is to examine the critically important practical problem of diagnostic inference under conditions of incomplete information. In particular, this paper addresses the fundamental issues of

1) how to predict the uncertainty—and therefore reliability—of a medical test result when such information as intrinsic test variability is not provided, and, as a corollary result;

2) how to ascertain whether the information provided is insufficient to permit a meaningful diagnostic interpretation of the reported outcome of a medical test.

Solutions to the problems set forth above, obtained within a comprehensive predictive statistical framework, are accomplished by means of the Principle of

Maximum Entropy (PME). The PME is an inferential procedure first employed in physics for deriving the formal structure of equilibrium statistical mechanics (ESM) [7]. This method has been employed subsequently to solve numerous problems in other domains including recent issues relating to health and medicine, such as the infectivity of SARS-Cov2 (COVID) [8], the distribution of human height and weight [9], and the diagnostic potential of body mass index (BMI) in the identification and treatment of obesity [10] [11], as well as critical social issues such as the detection of cheating or plagiarism [12].

This paper is organized as follows. Section 2 provides a synopsis of the Principle of Maximum Entropy and its application to the problem of statistical inference. Section 3 applies the PME to the simplest case—currently the most likely prevailing one—in which a clinical diagnostic lab reports only a point estimate of each risk factor whether or not it is correlated with other risk factors. Neither the variance, nor the covariance, nor the number of test trials on a given sample is reported. Section 4 examines the hypothetical case—to ascertain impact on diagnostic inference from more informative reporting—in which the lab provides the *exact* number N of test measurements (*i.e.* trials) of a risk factor, together with the point estimate of the outcome. Section 5 considers the less informative case in which only the *mean* number of trials is provided with the point estimate of the risk factor. Section 6 examines the case in which two risk factors are correlated, and the lab reports a point estimate of each factor together with some measure of their correlation. The specific example chosen for illustration is the important case of high density lipoprotein cholesterol (HDL-C) and low density lipoprotein cholesterol (LDL-C), two of the most widely tested analytes, given the global prevalence of cardiovascular disease. Section 7 summarizes and elaborates major points of the paper and the author's recommendations for reporting medical risk factors more informatively.

2. Principle of Maximum Entropy (PME)

In scientific usage, the term "entropy" conveys two basically different concepts. In the first sense, having roots in thermodynamics, entropy is a state function on par with energy in fundamental physical importance; it gauges quantitatively how efficiently a system converts heat to work. In the second sense, having roots in probability and communication theory, entropy quantifies system organization; specifically, it is a measure of the number of ways W (referred to as the multiplicity) by which an observable macroscopic state of a system can be realized by its constituent micro-states. It is this statistical second sense—also referred to as Shannon entropy or information entropy—that is pertinent to this paper. The two senses of entropy can be shown to be numerically equivalent for physical systems in thermodynamic equilibrium.

2.1. The Maximum Entropy (MaxEnt) Distribution—Formalism

Given a random variable (think medical risk factor) X whose possible values

x_i $(i=1,\cdots,n)$ are realized with respective probabilities p_i, the statistical entropy is defined by the expression

$$S(\boldsymbol{p}) = -\sum_{i=1}^{n} p_i \ln(p_i) \tag{1}$$

in which $\boldsymbol{p} \equiv (p_1,\ldots,p_n)$ stands for the complete set of probabilities. For simplicity, the outcomes and probabilities in this section are treated as discrete quantities. In cases for which the variable is continuous, the formalism replaces a discrete probability function p_x with a probability density function (PDF) $p(x)$ and the summation over outcomes in Equation (1) with an integral as follows

$$S(\boldsymbol{p}) = -\int p(x)\ln\big(p(x)\big)\mathrm{d}x\,. \tag{2}$$

Actually, extension of the theory to continuous variables entails redefining information entropy by incorporating a statistical measure function $m(x)$ to preserve invariance under a transformation of coordinates [13]. This modification does not affect the outcomes of the analyses in this paper and therefore is not explicitly dealt with in the cases to follow. A statistical problem requiring the measure function that is worked out in detail can be found in Ref. [12].

The principle of maximum entropy (PME) asserts that maximizing the information entropy of a random variable subject to constraints posed by known information, usually in the form of expectation values, leads to the *least biased* statistical distribution of that variable. The term "least biased" means that the maximum entropy (MaxEnt) distribution makes use only of known prior information. Other input data that can bias the distribution, such as might be drawn from model assumptions, is not employed. As a consequence, the MaxEnt distribution does not generate properties of variables or relations among variables not consistent with the given information. For example, if the given information says nothing about the correlation of two variables, then the resulting MaxEnt distribution will lead to no correlation of those variables. This is not necessarily the case in other methods of statistical inference.

The MaxEnt distribution is also the *most probable* of any distribution constrained by the same given information. This characteristic follows from the Boltzmann-Planck relation [14],

$$W \propto \mathrm{e}^{NS} \tag{3}$$

expressing an exponential dependence of the multiplicity W on the statistical entropy S and number of trials N. Maximization of entropy S therefore leads to a maximum multiplicity W. The more ways an observable system can be realized, the higher is its probability of occurrence. In cases of application with large number of trials, the resulting MaxEnt distribution can be many orders of magnitude more probable than any other statistical distribution constrained by the same prior information.

Mathematically, the PME leads to a variational equation for the set of probabilities $\boldsymbol{p}$. For example, in the absence of all prior information except for the prob-

ability completeness relation

$$\sum_{i=1}^{n} p_i = 1, \tag{4}$$

maximization of expression (1) subject to constraint (4) leads to a system of equal probabilities

$$p_i = 1/n, \tag{5}$$

in accordance with the Principle of Insufficient Reason [15] (also referred to as the Principle of Indifference).

More generally, when additional information is given in the form of known expectation values $F_j \quad (j = 1, \cdots, m)$ of a set of functions $f_j(x_i)$

$$F_j \equiv \langle f_j \rangle = \sum_{i=1}^{n} p_i f_j(x_i), \tag{6}$$

the variational function to be maximized takes the form

$$H(\boldsymbol{p}) = S(\boldsymbol{p}) - \lambda_0 \left(\sum_{i=1}^{n} p_i - 1 \right) - \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{n} p_i f_j(x_i) - F_j \right) \tag{7}$$

in which the set of functions $\lambda_k \quad (k = 0, 1, \cdots, m)$ are Lagrange multipliers to be determined from the given information. The symbol *H*, widely used in communication theory, is actually meant to be an upper case Greek eta, which stands for "Entropy".

Variation of Equation (7)

$$\delta H(\boldsymbol{p}) = 0, \tag{8}$$

with respect to each p_i leads to the solution

$$p_i = \frac{\exp\left(-\sum_{j=1}^{m} \lambda_j f_j(x_i) \right)}{Z(\boldsymbol{\lambda})} \tag{9}$$

in which the partition function $Z(\boldsymbol{\lambda})$

$$Z(\boldsymbol{\lambda}) = Z(\lambda_1, \cdots, \lambda_m) \equiv \sum_{i=1}^{n} \exp\left(-\sum_{j=1}^{m} \lambda_j f_j(x_i) \right) \tag{10}$$

has replaced the multiplier λ_0 after utilization of the completeness relation (4).

Knowledge of the partition function $Z(\boldsymbol{\lambda})$ of a system permits one to calculate all the statistical information that can be known or inferred about that system. In particular,

- $Z(\boldsymbol{\lambda})$ is a normalization factor ensuring that the completeness relation Equation (4) is satisfied.
- First derivatives of $Z(\boldsymbol{\lambda})$ with respect to each multiplier λ_j yield expressions

$$-\partial \ln(Z) / \partial \lambda_j = \langle f_j \rangle = F_j \tag{11}$$

for the given expectation values from which the set of Lagrange multipliers $\boldsymbol{\lambda} \equiv (\lambda_1, \cdots, \lambda_m)$ can be determined.

- Homogeneous second derivatives of $Z(\boldsymbol{\lambda})$ give rise to expressions

$$\partial^2 \ln(Z)/\partial\lambda_j^2 = \langle f_j^2 \rangle - \langle f_j \rangle^2 \equiv \Delta^2 F_j \tag{12}$$

for the (initially not given) variances of the functions $f_j(x_i)$.

- Mixed second derivatives of $Z(\boldsymbol{\lambda})$ yield 1st order correlation functions

$$\partial^2 \ln(Z)/\partial\lambda_j \partial\lambda_k = \langle (f_j - F_j)(f_k - F_k) \rangle = \langle f_j f_k \rangle - F_j F_k \equiv C_{jk} \tag{13}$$

which reduce to Equation (12) for $j = k$.

Clearly, higher-order mixed derivatives of $Z(\boldsymbol{\lambda})$ can generate higher correlation functions, such as worked out in Ref. [9] for a pair of correlated variables (human height and weight).

Further details regarding the foregoing methodology, including derivations of relations, can be found in Reference [9] and the collected papers of E. T. Jaynes [13].

2.2. The Maximum Entropy (MaxEnt) Distribution—Reliability

In summary, implementation of the PME to a particular system of random variables yields a mathematical expression, similar to that of Equation (9), for the least biased, most probable statistical distribution of that system consistent with known prior information. The expression permits one to predict initially unknown uncertainties and correlations of functions of the variables in terms of the given information, including the simplest case (relevant to this paper) where $f(x_i) \equiv x_i$.

The preceding characteristics notwithstanding, one might inquire as to the *reliability* of the predictions drawn from the MaxEnt distribution in any given case. Perhaps the most spectacularly successful case is that of equilibrium statistical mechanics (ESM), the domain of physics for which the PME was first proposed. Given only the mean system energy E and number N of particles (which plays the same role as the number of trials N in this paper), ESM can accurately account for the thermal properties of a vast array of systems ranging from the scale of elementary particles to that of stars, galaxies, and beyond. What makes this possible is the astronomical number of particles in the physical system. One mol (*i.e.* gram-molecular-mass) of a gas, for example, contains an Avogadro's number (~6×10^{23}) of particles. Since the variance $\Delta^2 E$ of mean energy (or $\Delta^2 N$ of mean particle number) is inversely proportional to the number of particles, the ratio $\sqrt{\Delta^2 E}/E$ is so small (generally on the order of ~10^{-12}) that no further improvement in predictability is achieved by explicitly including the variances of energy and particle number in the analysis.

The MaxEnt distribution, however, does not require an astronomical number of samples for its success. A recent example of its spectacular reliability in a matter relating to health and medicine was demonstrated in the case of human weight and height [9]. As part of an investigation of the application of the body mass

index (BMI) to identify and classify obesity [10], the author has shown that human height and weight are distributed empirically according to a joint lognormal distribution to such a degree of exactness, that one might suspect there is an underlying biophysical mechanism. Although there may well be such an intrinsic mechanism (that question is yet unanswered), use of the PME led to a MaxEnt distribution *identical* to the empirical one. Theoretical predictions agreed within statistical uncertainties with all moments and correlation functions testable with a large anthropometric data base [16] comprising several thousands of individuals in the male and female cohorts.

With regard to the incomplete reporting of medical risk factors by clinical laboratories, the MaxEnt distribution could in principle enable physicians to predict the variability of test results and thereby give more informed guidance to patients. However, the number of test measurements N of a specified risk factor is far fewer than the sample size in the two preceding examples. Ordinarily one presumes $N=1$ because this information is not routinely included in the medical report.

In the following sections of this paper the MaxEnt distribution will be derived for several practical cases distinguished by prior information regarding (a) the number of trials (*i.e.* tests per variable) and (b) the independence of the variables. The MaxEnt distribution itself will reveal how sharp are its predictions in any particular case. And if the predicted statistics should turn out to be highly uncertain, that is *not* a failure of the PME. *It is an indication that the paucity of information provided by the clinical laboratory is insufficient to permit a reliable estimate of a patient's risk*. More generally, as emphasized by Jaynes: "...the principle is most useful in just those cases where the empirical distribution fails to agree with the one predicted by maximum entropy." [17].

A final point to note: The examples to follow use test data from an unidentified healthy patient. The predictive distributions to be derived from the principle of maximum entropy are general in that they depend solely on the *kind* of prior information provided; e.g. a single expectation value, or a pair of correlated expectation values, etc. Given a certain set of information, the MaxEnt procedure leads to a predictive distribution of fixed functional form containing unknown Lagrange multipliers. When the specific *numerical values* of the prior information are taken into account to solve for the Lagrange multipliers, the predictive distribution becomes applicable to an individual, rather than a statistical population.

3. Case 1: Point Estimate of an Uncorrelated Risk Factor

Consider a non-negative risk factor X, continuous over the interval $[0,\infty]$, for which the only reported information is the point estimate x_0 of what is presumed to be a single trial, although that datum is not provided. Also included are the lower and upper bounds, s_L and s_U, of the reference range.

3.1. The Maximum Entropy Distribution

Implementation of the procedure in Section 2.1 with function $f(x)\equiv x$ and ex-

pectation value $F \equiv \langle X \rangle = x_0$ leads to an exponential PDF

$$p(x) = \exp(-\lambda x)/Z(\lambda) \tag{14}$$

with partition function

$$Z(\lambda) = \int_0^\infty \exp(-\lambda x)\mathrm{d}x = \lambda^{-1} \tag{15}$$

and expectation

$$-\partial \ln(Z)/\partial\lambda = \lambda^{-1} = \langle X \rangle = x_0 . \tag{16}$$

Thus, from Equations (14)-(16) the MaxEnt probability density takes the forms

$$p(x) = \lambda \exp(-\lambda x) = x_0^{-1} \exp(-x/x_0) \tag{17}$$

in which the first is for a general exponential, and the second presents the PDF explicitly in terms of the prior information x_0.

Following the procedure of Section 2.1, it is now possible to predict the variance of X

$$\mathrm{var}_X \equiv \sigma_X^2 = \left.\frac{\partial^2 \ln Z(\lambda)}{\partial \lambda^2}\right|_{\lambda = x_0^{-1}} = x_0^2 \tag{18}$$

from which follows the standard deviation of a single trial

$$\sigma_X = x_0 . \tag{19}$$

It is a well-known characteristic of the exponential distribution that the standard deviation of the variable equals the mean [18]. This property has profound implications for the diagnostic utility of risk factor test results reported by clinical laboratories.

3.2. Application to Metabolic and Lipid Panels

As determined in Section 3.1, the least biased, most probable predictive distribution is exponential in form leading to a standard deviation also of magnitude x_0. The reference range $[s_L, s_U]$ is ordinarily thought of as the region within which a patient's test result is "normal". This inference can be misleading in several ways. The term "normal" evokes an image of the widely encountered Gaussian distribution, referred to as the "normal" distribution with implication that the test result suggests "healthy" or "low risk" because it falls within 1 or 2 standard deviations about the mean where most healthy people's test result would fall. But this is not necessarily the case.

In a Gaussian distribution, which is symmetric about the mean, the probability of a result falling within a range of $\pm\sigma_X$ about the mean x_0 is 68.3%. However, in the case of an exponential distribution, which is a monotonically decreasing function, the corresponding range is $[0, 2x_0]$ with integrated probability 86.5%. This suggests that the outcome of a repeat trial could fall almost anywhere within the *full* range, and therefore well outside the *reference* range.

To examine this property quantitatively, one can calculate from PDF (17) the

probability P_{out} that a point estimate falls *outside* the bounds of the reference range

$$P_{\text{out}} = 1 - \int_{s_L}^{s_U} p(x)\,\mathrm{d}x = 1 + \mathrm{e}^{-s_U/x_0} - \mathrm{e}^{-s_L/x_0}. \tag{20}$$

Table 1 summarizes the values of P_{out} for selected risk factors from two frequently requisitioned screening panels, Metabolic and Lipid, of a patient's medical record. For each risk factor, the point estimate is within the reference range. However, the probability P_{out}, calculated from Equation (20), is well over 50% that a repeated measurement from the same sample would fall *outside* the reference range in all cases except one. For the one exception no lower limit was posted for the reference range.

In regard to the content of Table 1 it is to be noted that the reference range of a risk factor is not an absolute quantity. Clinical labs, health service providers, and medical organizations may cite different ranges for the same risk factor [1]. The Lipid Panel provides an important timely example. Popular health and medical media have long conveyed to the public that high density lipoprotein cholesterol (HDL-C) is the "good" cholesterol and low density lipoprotein cholesterol (LDL-C) is the "bad" cholesterol. While this belief is largely valid, the implication that there is no upper threshold of harm for HDL-C or lower threshold of harm for LDL-C is not. Although still controversial issues, research has established a more nuanced situation whereby a too high HDL-C is associated with a variety of conditions, including elevated risk of dementia [19]-[22], and a too low LDL-C (hypocholesterolemia) is likewise associated with a variety of adverse conditions including cancer, pancreatitis, hyperthyroidism, and kidney failure [23] [24]. Therefore, HDL-C and LDL-C values in Table 1 include reference ranges with upper and lower limits.

Table 1. Probability that a trial falls outside the reference range.

Panel	Risk Factor	Point Estimate	Units	Reference Range	*P* (out) (%)
METABOLIC	Glucose	87	mg/dL	65 - 99	84.7
	Creatinine	0.79	mg/dL	0.7 - 1.28	78.6
	Urea (BUN)	18	mg/dL	7 - 25	57.1
	Sodium	141	mmol/L	135 - 146	97.1
LIPID	Total-C	165	mg/dL	116 - 200	80.2
	HDL-C	70	mg/dL	40 - 90	71.2
	LDL-C	81	mg/dL	40 - 100	68.1
	Triglycerides	60	mg/dL	<150	8.2

Another point to emphasize is that the test results in Table 1 would generally be regarded as indicative of a nominally healthy patient. Few primary care physi-

cians or even specialists would be concerned (if they thought about it at all) that the results of a repeat trial from the same sample had a high probability of falling *outside* the reference range. This possibility would be regarded as a purely hypothetical, rather than practical, matter, since there *is no* second measurement provided by the clinical lab. Indeed, the fact that Equations (17) and (20) predict such results might be construed as a failure of the MaxEnt distribution, and one would wonder what the "true" probability distribution is.

It is important to understand that there *is no* "true" distribution. Probability is a measure of the *knowledge* one has of a system, not a characterization of the system, itself. The form taken by the MaxEnt probability function depends on what prior information is available. And the more information one has, the more reliably one can predict a system's statistical behavior. This apparent subjectivity does not detract from the concept of probability as an objectively rigorous part of mathematics. When the PME is implemented correctly, two analysts with the same prior information about a system should be able to arrive at equivalent predictions of the system's statistical properties.

In summary, the MaxEnt distribution does not guarantee agreement with an empirical distribution. What it has been proven to provide is the least biased, most probable predictive probability distribution consistent with the information available. Accordingly, as revealed by the content of **Table 1**, more information needs to be provided than a single point estimate of each risk factor if a clinical lab report is to be diagnostically useful to a physician.

4. Case 2: Point Estimate of Risk Factor and Number of Trials

Consider next a non-negative, uncorrelated risk factor X for which the information from a clinical lab comprises a point estimate $\overline{x}$ of the mean of N independent measurements as expressed in the relation below

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \,. \tag{21}$$

The individual trial outcomes x_n $(n = 1, \cdots, N)$ are not provided. Ordinarily, N is not provided either, but it is of utility to answer the question of how knowledge of N affects the prediction of risk.

4.1. Use of the Characteristic Function (CF)

In Case 1, only a point estimate x_0 was reported. In the absence of knowledge of the number of trials, the MaxEnt distribution was found to be of exponential form with mean and standard deviation both equal to x_0 and a Lagrange multiplier $\lambda = x_0^{-1}$. Now, given the number of trials N, how does the point estimate x_0 of indeterminate trial number relate to the mean $\overline{x}$?

Expressed as a relation connecting the expectation values of random variables $\left(\overline{X}, X_n\right)$, Equation (21) becomes

$$\overline{x} = \left\langle \overline{X} \right\rangle = \frac{1}{N} \sum_{n=1}^{N} \left\langle X_n \right\rangle = \frac{N x_0}{N} = x_0 \tag{22}$$

in which the expectation $\left\langle X_n \right\rangle$ of each variable X_n $\left(n = 1, 2, \cdots, N\right)$ is calculated by the integral

$$\left\langle X_n \right\rangle = \int_0^\infty x\, p(x) \mathrm{d}x = x_0 \tag{23}$$

where $p(x)$ is the MaxEnt PDF Equation (17). One finds from Equations (22) and (23) that the expectation of the sample mean equals the expectation of a single trial, or $\overline{x} = x_0$.

When deriving the MaxEnt distribution, one ordinarily takes a given expectation value to be an exact quantity. Or, if this value is uncertain, then one supplements the information in the functional equation (7) with a known variance (or the equivalent). In the present case, the uncertainty in x_0 (and therefore $\overline{x}$) is not provided by the clinical lab. However, having been informed from Equation (21) that $\overline{x}$ is a sample mean of N trials, it is now possible to derive the PDF that governs its distribution. An expedient way to do this is to make use of the characteristic function (CF) [25] of the variable X

$$h_X(t) \equiv \left\langle \mathrm{e}^{iXt} \right\rangle = \int_0^\infty \mathrm{e}^{ixt} p(x) \mathrm{d}x = \left(1 - \frac{it}{\lambda}\right)^{-1}. \tag{24}$$

Equation (24) shows that the CF and PDF are Fourier transforms of one another.

From Equation (24) one obtains the CF of the variable $\overline{X}$ by applying the expectation operation to the function $\exp\left(i\overline{X}t\right)$ as follows [25]

$$h_{\overline{X}}(t) \equiv \left\langle \mathrm{e}^{i\overline{X}t} \right\rangle = h_X\left(N^{-1}t\right)^N = \left(1 - \frac{it}{N\lambda}\right)^{-N}. \tag{25}$$

The PDF of $\overline{X}$ is then the inverse Fourier transform of $h_{\overline{X}}(t)$

$$p_{\overline{X}}(\overline{x}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-i\overline{x}t} h_{\overline{X}}(t) \mathrm{d}t = \frac{N\lambda\left(N\lambda\overline{x}\right)^{N-1} \mathrm{e}^{-N\lambda\overline{x}}}{\Gamma(N)} \tag{26}$$

where $\Gamma(N)$ is the gamma function

$$\Gamma(N) \equiv \int_0^\infty t^{N-1} \mathrm{e}^{-t} \mathrm{d}t \tag{27}$$

equal to $(N-1)!$ for integer N. Evaluation of the integral in Equation (26) is accomplished by contour integration around the N^{th}-order pole at $-iN\lambda$ in the lower half complex plane.

One sees from Equation (26) that the variable $\overline{X}$ representing the mean of N independent identically distributed variables is governed by a gamma distribution [18] with parameter $N\lambda$. Applying PDF Equation (26), one can calculate the first and second moments of $\overline{X}$

$$\left\langle \overline{X} \right\rangle = \int \overline{x} p_{\overline{X}}(\overline{x}) \mathrm{d}\overline{x} = \lambda^{-1} = x_0 \tag{28}$$

$$\left\langle \bar{X}^2 \right\rangle = \int \bar{x}^2 p_{\bar{X}}(\bar{x}) \mathrm{d}\bar{x} = \left(1 + N^{-1}\right) \lambda^{-2} = \left(1 + N^{-1}\right) x_0^2, \tag{29}$$

which validate the mean obtained in a different way in Equation (22) and now provide the best predictive estimate of the variance of $\bar{X}$

$$\Delta^2 \bar{x} = \left\langle \bar{X}^2 \right\rangle - \left\langle \bar{X} \right\rangle^2 = N^{-1} x_0^2. \tag{30}$$

From Equation (30), it follows that the predicted standard deviation of the reported point estimate x_0

$$\sigma_{\bar{X}} = \frac{x_0}{\sqrt{N}} \tag{31}$$

can be made arbitrarily small by increasing the number of independent measurements of the associated risk factor.

The number of measurements performed by a diagnostic lab, however, is at best likely to be relatively few, if not just one. Nevertheless, use of machine automation of multiple trials simultaneously from the same sample is a future possibility not to be excluded. To assess the impact of knowing the number of trials on the reduction in uncertainty, and therefore enhancement of reliability, of the MaxEnt predictions, two functions are of particular utility: the coefficient of variation CV

$$\mathrm{CV}(N) \equiv \frac{\text{standard deviation}}{\text{mean}} = \frac{\sigma_{\bar{X}}}{x_0} = \frac{1}{\sqrt{N}} \tag{32}$$

and the conditional probability $P_{\text{in}}(\bar{x}_0 \mid N)$ that the mean $\bar{x}_0$ of a repeat set of N trials

$$\begin{aligned} P_{\text{in}}(\bar{x}_0 \mid N) &\equiv P(s_U \geq \bar{x}_0 \geq s_L \mid N) = \int_{s_L}^{s_U} p_{\bar{X}}(\bar{x} \mid N) \mathrm{d}\bar{x} \\ &= \Gamma(N)^{-1} \left[\Gamma\left(N, \frac{N s_L}{x_0}\right) - \Gamma\left(N, \frac{N s_U}{x_0}\right) \right] \end{aligned} \tag{33}$$

lies *within* the reference range. The two-parameter function of the form $\Gamma(a, z)$ in Equation (33) is the incomplete gamma function

$$\Gamma(a, z) \equiv \int_z^\infty t^{a-1} \mathrm{e}^{-t} \mathrm{d}t. \tag{34}$$

4.2. Application to the Measurement of Serum Glucose

As an example of the information provided by knowledge of the number of trials, we consider again measurement of serum glucose—which is a risk factor for diabetes, among other conditions—as summarized in Table 1 for an indeterminate number of trials. Pertinent data in mg/dL are reproduced below for convenience:

Mean :	87
Lower Ref Limit :	65
Upper Ref Limit :	99

(35)

Use of the preceding data in Equations (32) and (33) lead respectively to the blue traces in Figure 1 and Figure 2. The figures actually display results for Case

1, Case 2, and Case 3 (to be taken up in Section 5). In viewing the figures, note that the phrase "mean number" in the abscissa label is to be interpreted as "no given number" for Case 1, "exact number" for Case 2, and "mean number" for Case 3.

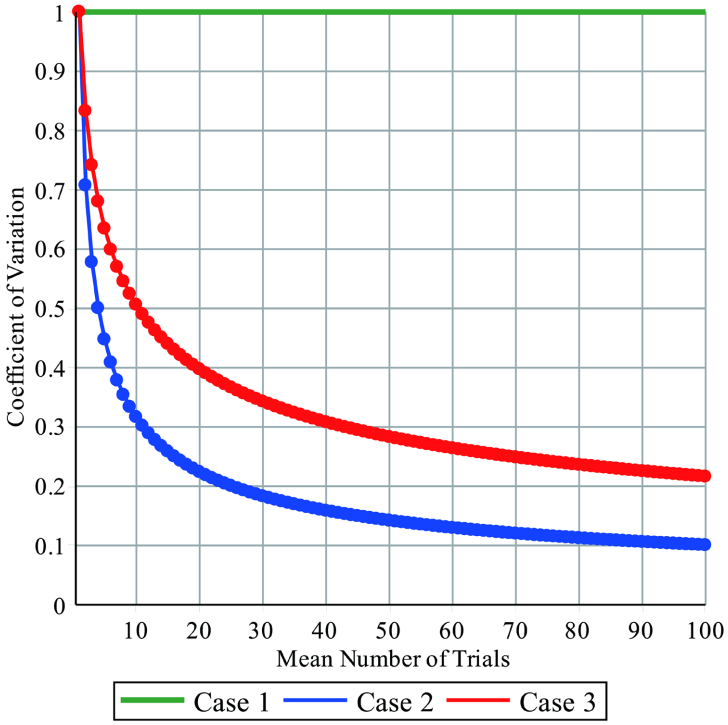

Figure 1. Coefficient of variation as a function of number of trials per sample for Case 1 (green), Case 2 (blue), and Case 3 (red). The cases are distinguished by (1) no number reported, (2) exact number reported, (3) mean number reported.

The blue trace in Figure 1 shows the variation of CV with N for Case 2. Theory predicts that about 10 trials suffice to reduce the CV from an initial value of 1 in Case 1 (a presumably single trial governed by an exponential distribution) to 0.3. However, to achieve a CV of 0.1 would require about 100 trials because of the slow asymptotic decrease of Equation (32). The CV for Case 1 is shown in the Figure as a horizontal solid green trace without discrete markers spanning all values of N since no prior information regarding the number of trials was provided. Thus, while it is natural to *presume* only 1 trial was made, the PME makes no use of speculative assumptions.

The blue trace in Figure 2 shows the variation of P_{in} with N, a perspective complementary to the probability values given in Table 1. From Figure 2, one infers that there is a 50% probability that 10 trials suffice to produce a mean outcome that falls within the reference limits. If only a single trial is actually (rather than presumably) performed, then this probability is about 15%, consistent with the complementary probability given in Table 1. However, to achieve a 90% probability that a set of measurements yield a mean within the reference limits would require about 100 trials. The horizontal solid green trace in the figure represents

the outcome of Case 1, as discussed in the preceding paragraph for Figure 1.

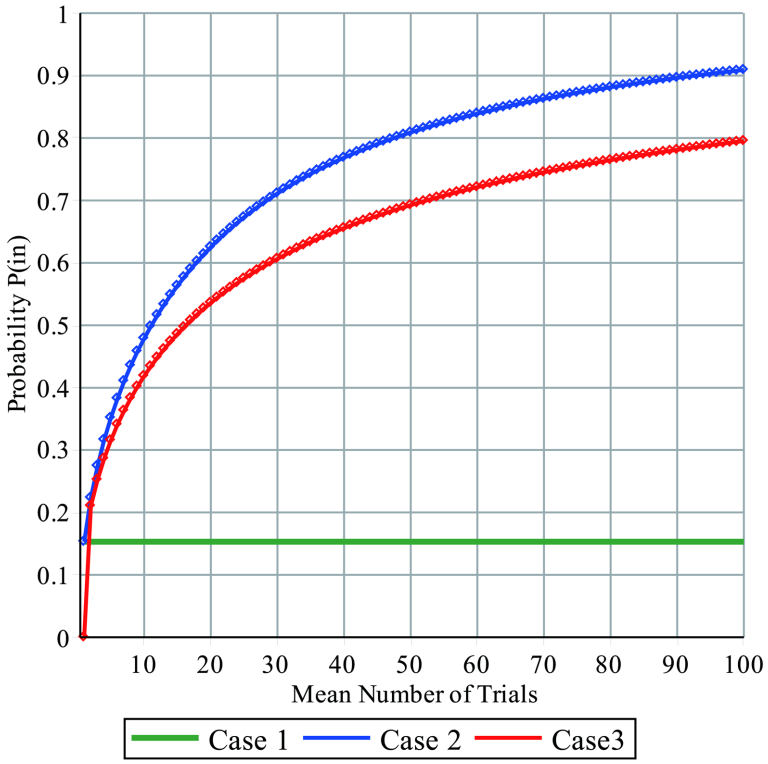

Figure 2. Probability, as a function of number of trials per sample, that the mean outcome of a repeated set of trial falls within the reference interval for Case 1 (green), Case 2 (blue), and Case 3 (red). Case description is the same as for Figure 1.

It is to be borne in mind that the point estimates of the risk factor in both Cases 1 and 2 are exactly the same number x_0. In Case 1, the physician was totally uninformed as to how the number was arrived at. In Case 2, the physician learned additionally that the number was the mean of N trials—hence the initial notation $\bar{x}$ for that number. As a consequence of that extra bit of information, the most probable prediction of risk dramatically changed from a simple exponential distribution to the more informative gamma distribution. This transition is again illustrative of the fact that probability is a mathematical quantity reflecting one's knowledge of a system and not intrinsic properties of the system itself.

5. Case 3: Point Estimate of Risk Factor and Mean Number of Trials

Consider the case of a non-negative, uncorrelated risk factor X for which the information from a clinical lab comprises a point estimate $\bar{x}$ of the mean of an *indeterminate* number of trials of estimated mean $\bar{n}$, rather than an *exact* number N. Such a situation could arise, for example, if tests of different analytes in a panel were performed with different numbers of trials, and the lab simply reported a mean number of trials for the panel. The exact trial number labeled N in Case 2 is now a random variable, and the probability density (26) must now take the form

$$p_{\bar{X}}\left(\bar{x}\,|\,\bar{n}\right)=\sum_{n=1}^{\infty}p_{\bar{X}}\left(\bar{x}\,|\,n\right)p_N\left(n\,|\,\bar{n}\right)=\sum_{n=1}^{\infty}\frac{n\lambda\left(n\lambda\bar{x}\right)^{n-1}\mathrm{e}^{-n\lambda\bar{x}}}{\Gamma\left(n\right)}p_N\left(n\,|\,\bar{n}\right) \quad (36)$$

in which the conditional probability $p_N\left(n\,|\,\bar{n}\right)$ of n trials, given the mean $\bar{n}$, must be determined.

5.1. Distribution of the Number of Trials

Since N is an independent random variable, the solution to finding $p_N\left(n\,|\,\bar{n}\right)$ is to apply the PME again, given only the probability completeness relation

$$\sum_{n=1}^{\infty}p_N\left(n\,|\,\bar{n}\right)=1 \quad (37)$$

and the expectation of trial number

$$\sum_{n=1}^{\infty}n\,p_N\left(n\,|\,\bar{n}\right)=\bar{n}\,. \quad (38)$$

As in Case 1, the solution is an exponential function of the variable n and a Lagrange multiplier here designated α

$$p_N\left(n\,|\,\alpha\right)=\frac{\mathrm{e}^{-\alpha n}}{Z\left(\alpha\right)}. \quad (39)$$

Unlike Case 1, the variable n is discrete, and the partition function evaluates to

$$Z\left(\alpha\right)=\sum_{n=1}^{\infty}\mathrm{e}^{-\alpha n}=\left(\mathrm{e}^{\alpha}-1\right)^{-1}. \quad (40)$$

From Equation (11), which relates the Lagrange multiplier to the expectation of the variable, one finds

$$\mathrm{e}^{\alpha}=\left(\bar{n}-1\right)^{-1} \quad (41)$$

from which follows the sought-for conditional probability

$$p_N\left(n\,|\,\bar{n}\right)=\frac{\left(\bar{n}-1\right)^{n-1}}{\bar{n}^{n}}, \quad (42)$$

explicitly expressed in terms of the given information $\bar{n}$, instead of the Lagrange multiplier α. To a reader familiar with physics, Equation (42) is recognized as the distribution of photons in thermal equilibrium [26]. In the case of photons, however, n can take the lowest value 0, whereas in the case of testing a risk factor, the lowest meaningful value is $n=1$.

Upon substitution of Equation (42) into Equation (36), with replacement of the dimensioned coordinate $\bar{x}$ by a dimensionless coordinate u

$$u=\lambda\,\bar{x}=\bar{x}/x_0\,, \quad (43)$$

one can re-express the distribution $p_{\bar{X}}\left(\bar{x}\,|\,\bar{n}\right)$ more simply in the form

$$p_{\lambda\bar{X}}\left(u\,|\,\bar{n}\right)\mathrm{d}u=\sum_{n=1}^{\infty}\left(\frac{n}{\bar{n}}\right)^{n}\frac{\left(\bar{n}-1\right)^{n-1}}{\Gamma\left(n\right)}u^{n-1}\mathrm{e}^{-nu}\mathrm{d}u \quad (44)$$

where the differential $\mathrm{d}u=\lambda\mathrm{d}\bar{x}$ is explicitly included.

The above notational change facilitates the calculation of the lowest moments as dimensionless quantities

$$\langle u \rangle = \langle \lambda \bar{x} \rangle = \int_0^\infty u \, p_{\lambda \bar{X}} (u \mid \bar{n}) \, \mathrm{d}u = 1, \tag{45}$$

$$\langle u^2 \rangle = \langle (\lambda \bar{x})^2 \rangle = \int_0^\infty u^2 p_{\lambda \bar{X}} (u \mid \bar{n}) \, \mathrm{d}u = 1 + \frac{\ln(\bar{n})}{\bar{n} - 1}, \tag{46}$$

from which follow the predicted uncertainty relations

$$\Delta^2 \bar{u} = \sigma_U^2 = \langle \bar{u}^2 \rangle - \langle \bar{u} \rangle^2 = \frac{\ln(\bar{n})}{\bar{n} - 1} \tag{47}$$

and

$$\mathrm{CV}(\bar{n}) = \frac{\sigma_U}{\bar{u}} = \frac{\sigma_{\bar{X}}}{\bar{x}} = \sqrt{\frac{\ln(\bar{n})}{\bar{n} - 1}}. \tag{48}$$

The integrations over u in Equation (44) can be performed generally, leading to any desired k^{th} moment $(k = 0, 1, 2, \cdots)$

$$\langle u^k \rangle = \langle (\bar{x} / x_0)^k \rangle = \sum_{n=1}^\infty \left(\frac{n}{\bar{n}} \right)^n \frac{(\bar{n} - 1)^{n-1}}{n^{n+k}} \frac{\Gamma(n + k)}{\Gamma(n)} \tag{49}$$

where $k = 0$ yields the correct normalization constant 1.

5.2. Prediction Reliability

The conditional probability $P_{\mathrm{in}} (\bar{x}_0 \mid \bar{n})$ that a mean of $\bar{n}$ trials yields an outcome $\bar{x}_0$ that falls between the lower (s_L) and upper (s_U) boundaries of the reference range is given by

$$\begin{aligned} P_{\mathrm{in}} (\bar{x}_0 \mid \bar{n}) &\equiv P(s_U \geq \bar{x}_0 \geq s_L \mid \bar{n}) = \int_{s_L}^{s_U} p_{\bar{X}, \bar{N}} (\bar{x} \mid \bar{n}) \, \mathrm{d}\bar{x} \\ &= \sum_{n=1}^\infty \frac{(\bar{n} - 1)^{n-1}}{\bar{n}^n \, \Gamma(n)} \left[\Gamma\left(n, \frac{n s_L}{x_0} \right) - \Gamma\left(n, \frac{n s_U}{x_0} \right) \right]. \end{aligned} \tag{50}$$

Before considering an application to a specific example, it is necessary to clarify what may appear to be an ambiguity in Equation (50).

Being of greater generality than Equation (33), Equation (50) must therefore reduce correctly in the previous case where the number of trials is constant, including the value $\bar{n} = 1$. Substitution of $\bar{n} = 1$, however, leads to $P_{\mathrm{in}} = 0$, which is not correct. A mean value of 1 can only occur if there is only 1 trial. What is required is to evaluate the right side of Equation (50) in the limit $n \to 1$, which leads to

$$P_{\mathrm{in}} (x_0 \mid \bar{n} = 1) = \mathrm{e}^{-s_L / x_0} - \mathrm{e}^{-s_U / x_0} \tag{51}$$

which correctly agrees with Equation (33) for $N = 1$.

The red traces in Figure 1 and Figure 2 respectively show the variation with $\bar{n}$ of the functions CV and $P_{\mathrm{in}} (\bar{x}_0 \mid \bar{n})$. By comparison with the blue traces for Case 2, one sees immediately the effects of having less precise information regarding

the number of trials. For example:

- A mean of 10 trials leads to a CV of 0.5, compared with 0.3 for Case 2. And 100 trials reduce the CV to 0.2, rather than 0.1 for Case 2.
- Similarly, to achieve a 50% probability that the mean outcome of the trials falls within the reference range would require a mean of about 15 trials, rather than exactly 10 as for Case 2.
- A mean of 100 trials yields a probability P_{in} of 80%, in contrast to Case 2 for which 100 trials yields a probability P_{in} of 90%.
- And last, although beyond the range displayed in **Figure 2**, to achieve a probability P_{in} of 90% according to Equation (50) would require close to 300 trials.

6. Case 4: Correlated Risk Factors

As a final set of conditions, consider two continuous, non-negative risk factors, X and Y, that are correlated. A timely example, given that the leading cause of death in the US [27] and globally [28] is heart disease, is the pair of high-density lipoprotein cholesterol (HDL-C) and low-density lipoprotein cholesterol (LDL-C). In the author's experience, clinical laboratory reports have routinely included only the respective point estimates x_0 and y_0, together with associated reference ranges (x_1, x_2) and (y_1, y_2). Since HDL-C and LDL-C are part of the Lipid Panel, the reports would also include point estimates of total cholesterol and triglycerides along with their associated reference ranges. However, the reports generally do *not* include the number of trials, any measure of the variability (*i.e.* uncertainty) of the tests, or any measure of the correlation of the variables.

If no measure of correlation is reported, then, from the perspective of the Principle of Maximum Entropy (PME), HDL-C and LDL-C are independent random variables, such as treated in Case 1 with most probable risk prediction following an exponential distribution. The inclusion of a measure of correlation, such as covariance, markedly changes the distribution function, as will be shown shortly. Before deriving this presumably more informative distribution, it is worth examining *why* the variables $X \equiv \text{HDL-C}$ and $Y \equiv \text{LDL-C}$ can be correlated.

As a matter of terminology, a lipoprotein is a complex biochemical particle that transports lipids (*i.e.* molecules of fat) in the liquid component (plasma) of blood [29] [30]. The center of the particles consists of cholesterol and triglyceride molecules; the hydrophilic outer shell contains a special kind of protein (apolipoprotein) that stabilizes the assembly. Lipoprotein particles have been classified into five groups [29] (chylomicrons, very low density VLDL, low density LDL, intermediate density IDL, high density HDL) ranging from 1) large and low density to 2) small and high density. Low density lipoproteins (LDL) ferry cholesterol (LDL-C) and other molecules of fat around the body and are associated with atherosclerotic cardiovascular disease (ASCVD) by deposition of cholesterol on the inside walls of blood vessels. By contrast, high density lipoproteins (HDL) reduce the risk of heart disease by collecting cholesterol (HDL-C) and other molecules of fat

from cells and tissues for transport to the liver for eventual removal. The LDL-C and HDL-C therefore serve (although not perfectly) as markers of "bad" (LDL) and "good" (HDL) lipoprotein particles.

There are basically two general procedures for obtaining plasma concentrations of HDL-C and LDL-C. The first is to measure HDL-C and LDL-C directly by one or more of a variety of physical methods [31] that make use of processes and materials such as centrifugation, chemical precipitation, surfactants, ionic polymers, and other components that facilitate extraction of cholesterol selectively from specific classes of lipoproteins present in the blood. In such a direct approach for example, HDL and LDL are together separated from chylomicrons and VLDL, and concentrations of total cholesterol and HDL-C are then measured, whereupon the LDL-C concentration is obtained by subtracting the latter from the former. By virtue of such a sequential separation—whereby the LDL-C depends on the subtracted concentration of HDL-C, as well as the potential misidentification of some lipoproteins in samples from dyslipidemic patients—the measurements of LDL-C and HDL-C are correlated.

The second approach is to measure HDL-C directly and then estimate LDL-C by calculation from an empirical formula that contains HDL-C and other lipid-panel constituents [32] [33]. In this case, LDL-C is correlated with HDL-C through the mathematical equation connecting the two variables. The following subsections examine the predictive distributions arising respectively from the direct and indirect procedures.

6.1. Case 4A: Risk Factors Correlated by Measurement

In this section we examine the most probable distribution of two variables that are correlated with one another, but independent of any other variables. This is not exactly the system of HDL-C and LDL-C, which are part of a lipid panel comprising four constituents, but it provides a simple tractable system to address the seminal question of interest here: Does knowledge of a measure of correlation, in addition to just the point estimates of the two mean values, significantly reduce predictive uncertainty? A model more closely representative of the variables HDL-C and LDL-C will be taken up afterward.

Consider, therefore, two risk factors, X and Y, continuous over the non-negative real axis, for which there is a non-zero covariance

$$\mathrm{cov}_{X,Y} \equiv \left\langle \left(X - x_0\right)\left(Y - y_0\right)\right\rangle = \left\langle X\,Y\right\rangle - x_0\,y_0\,. \tag{52}$$

Expectations of products such as $X^a Y^b$ for real powers a, b are calculated with a bivariate PDF $p_{X,Y}\left(x,y\right)$ as follows

$$\left\langle X^a Y^b\right\rangle = \int_0^\infty\int_0^\infty x^a y^b\, p_{X,Y}\left(x,y\right)\mathrm{d}x\mathrm{d}y\,. \tag{53}$$

As in the previous cases, $p_{XY}\left(x,y\right)$ is the sought-for maximum entropy (MaxEnt) distribution obtained by solving a variational equation based on prior known information. In the present case we take this information to be the proba-

bility completeness relation

$$\int_0^\infty\int_0^\infty p_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y = 1, \tag{54}$$

the means of X and Y

$$\int_0^\infty\int_0^\infty x\, p_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y = x_0, \tag{55}$$

$$\int_0^\infty\int_0^\infty y\, p_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y = y_0, \tag{56}$$

and the covariance of X and Y, more conveniently reported as a dimensionless covariance coefficient

$$\kappa \equiv \frac{\mathrm{cov}_{X,Y}}{x_0\, y_0} \tag{57}$$

from which follows the mean of the product XY

$$\int_0^\infty\int_0^\infty xy\, p_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y = x_0 y_0 (1+\kappa). \tag{58}$$

If $\kappa = 0$, then the integrand in Equation (58) simply factors into a product of two independent univariate distributions, each being of exponential form as treated in Section 3 for Case 1. If $\kappa \neq 0$, then following the procedure of Section 2.1 leads to the MaxEnt solution

$$p_{X,Y}(x,y) = \frac{\exp(-\lambda_1 x - \lambda_2 y - \lambda_3 xy)}{Z(\boldsymbol{\lambda})} \tag{59}$$

with partition function

$$Z(\boldsymbol{\lambda}) = Z(\lambda_1,\lambda_2,\lambda_3) \equiv \int_0^\infty\int_0^\infty \mathrm{e}^{(-\lambda_1 x - \lambda_2 y - \lambda_3 xy)}\,\mathrm{d}x\mathrm{d}y = \lambda_3^{-1}\exp(\Lambda)\Gamma(0,\Lambda) \tag{60}$$

in which, for convenience, Λ has been defined as

$$\Lambda \equiv \lambda_1\lambda_2/\lambda_3\ . \tag{61}$$

The unknown Lagrange multipliers can be determined from Equation (11) as applied below,

$$\begin{aligned} -\partial\ln(Z)/\partial\lambda_1 &= x_0, \\ -\partial\ln(Z)/\partial\lambda_2 &= y_0, \\ -\partial\ln(Z)/\partial\lambda_3 &= x_0 y_0 (1+\kappa), \end{aligned} \tag{62}$$

provided the prior information is consistent.

The operations in Equation (62) lead to relations

$$x_0 = -\frac{1}{\lambda_1}\left(\Lambda - \frac{1}{\mathrm{e}^\Lambda\Gamma(0,\Lambda)}\right) \tag{63}$$

$$y_0 = -\frac{1}{\lambda_2}\left(\Lambda - \frac{1}{\mathrm{e}^\Lambda\Gamma(0,\Lambda)}\right) \tag{64}$$

$$x_0 y_0 (1+\kappa) = \frac{1}{\lambda_3}\left(1+\Lambda-\frac{1}{\mathrm{e}^{\Lambda}\Gamma(0,\Lambda)}\right) \tag{65}$$

Manipulation of Equations (63)-(65) results in relations

$$\lambda_1 x_0 = \lambda_2 y_0 = 1-\lambda_3 x_0 y_0 (1+\kappa) \tag{66}$$

which express λ_1 and λ_2 in terms of λ_3. Λ is then a function only of λ_3, which can be found by solving the equation

$$1+\Lambda(\lambda_3)-\frac{1}{\mathrm{e}^{\Lambda(\lambda_3)}\Gamma(0,\Lambda(\lambda_3))}-x_0 y_0 (1+\kappa)\lambda_3 = 0. \tag{67}$$

Equation (67) is highly nonlinear and generally requires numerical solution.

Upon solution of the three Lagrange multipliers, the joint PDF $p_{X,Y}(x,y)$ is uniquely determined. One can then calculate the marginal distributions of X and Y by integration

$$p_X(x) = \int_0^{\infty} p_{X,Y}(x,y)\mathrm{d}y = \frac{\lambda_3 \exp(-\lambda_1 x-\Lambda)}{(\lambda_3 x+\lambda_2)\Gamma(0,\Lambda)} \tag{68}$$

$$p_Y(y) = \int_0^{\infty} p_{X,Y}(x,y)\mathrm{d}x = \frac{\lambda_3 \exp(-\lambda_2 y-\Lambda)}{(\lambda_3 y+\lambda_1)\Gamma(0,\Lambda)}. \tag{69}$$

From Equations (59), (68), and (69) then follow the conditional distributions of interest that encode the predictive statistics of one variable in light of knowledge of a corresponding value of the other variable

$$p_X(x\,|\,y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = (\lambda_1+\lambda_3 y)\mathrm{e}^{-(\lambda_1+\lambda_3 y)x} \tag{70}$$

$$p_Y(y\,|\,x) = \frac{p_{X,Y}(x,y)}{p_X(x)} = (\lambda_2+\lambda_3 x)\mathrm{e}^{-(\lambda_2+\lambda_3 x)y}. \tag{71}$$

One sees from Equations (70) and (71) that the conditional distributions are of exponential form, just as in Case 1 for *un*correlated variables, except that now the conditional mean values

$$x_0(y) \equiv \int_0^{\infty} x\, p_X(x\,|\,y)\mathrm{d}x = (\lambda_1+\lambda_3 y)^{-1} \tag{72}$$

$$y_0(x) \equiv \int_0^{\infty} y\, p_Y(y\,|\,x)\mathrm{d}y = (\lambda_2+\lambda_3 x)^{-1}, \tag{73}$$

are correlated through the multiplier λ_3. Moreover, as shown in the analysis of Case 1, the standard deviation of an exponentially distributed variable is equal to the mean. It therefore follows that the coefficients of variation predicted by the conditional probability densities are

$$\mathrm{CV}_{X|Y}(x_0\,|\,y_0) = \frac{x_0(y_0)}{x_0} = x_0^{-1}(\lambda_1+\lambda_3 y_0)^{-1} \tag{74}$$

$$\mathrm{CV}_{Y|X}(y_0\,|\,x_0) = \frac{y_0(x_0)}{y_0} = y_0^{-1}(\lambda_2+\lambda_3 x_0)^{-1}. \tag{75}$$

The expressions in Equations (74) and (75) are identical by virtue of Equation (66).

Regarding the forms of $\mathrm{CV}_{X|Y}$ and $\mathrm{CV}_{Y|X}$, it is to be recalled that the information actually reported by diagnostic labs includes the point estimates (x_0, y_0), not the conditional means. If $\lambda_3 = 0$ in Equations (74) and (75), then λ_1 and λ_2 respectively equal the reciprocals of the reported point estimates, whereupon the coefficients of variation reduce to 1.

Knowing the conditional distributions of X and Y, one can then predict the probability that a repeat measurement of one of them, given knowledge of the other, would fall within the reference interval (x_1, x_2) or (y_1, y_2):

$$P_{\text{in}}(x_1, x_2 \mid y) = \int_{x_1}^{x_2} p_{X|Y}(x \mid y)\,\mathrm{d}x = \mathrm{e}^{-(\lambda_1+\lambda_3 y)x_1} - \mathrm{e}^{-(\lambda_1+\lambda_3 y)x_2} \tag{76}$$

$$P_{\text{in}}(y_1, y_2 \mid x) = \int_{y_1}^{y_2} p_{Y|X}(y \mid x)\,\mathrm{d}y = \mathrm{e}^{-(\lambda_2+\lambda_3 x)y_1} - \mathrm{e}^{-(\lambda_2+\lambda_3 x)y_2}. \tag{77}$$

The preceding relations are to be compared with the relations for uncorrelated exponential variables:

$$P_{\text{in}}(x_1, x_2) = x_0^{-1}\int_{x_1}^{x_2} \mathrm{e}^{-x/x_0}\,\mathrm{d}x = \mathrm{e}^{-x_1/x_0} - \mathrm{e}^{-x_2/x_0} \tag{78}$$

$$P_{\text{in}}(y_1, y_2) = y_0^{-1}\int_{y_1}^{y_2} \mathrm{e}^{-y/y_0}\,\mathrm{d}x = \mathrm{e}^{-y_1/y_0} - \mathrm{e}^{-y_2/y_0}. \tag{79}$$

6.2. Application to Measurements of Cholesterol

For the purpose of illustration, the system of HDL-C and LDL-C are here treated as an isolated pair (X, Y) of correlated variables. Since the given prior information took no account of other lipid panel variables, the PME placed no upper limits on the ranges of X and Y. Reproduced below for convenience are the pertinent patient lipid panel data from Table 1 expressed in units of mg/dL except for the dimensionless covariance coefficient κ:

Mean HDL-C	x_0	70	
Mean LDL-C	y_0	81	
Cov. Coeff.	κ	–0.33	(80)
HDL-C Ref. Limits	(x_1, x_2)	$(40, 90)$	
LDL-C Ref. Limits	(y_1, y_2)	$(40, 100)$	

Since no empirical value of the correlation of HDL-C and LDL-C is known to the author, the coefficient κ in Equation (80) was set to correspond to the empirical covariance coefficient of Case 4B, which is taken up in the following subsection.

Solution of Equation (67) with subsequent use of Equation (66) leads to the following Lagrange multipliers

$$\begin{aligned}\lambda_1 &= 9.1519\times10^{-3}\\ \lambda_2 &= 7.9090\times10^{-3}\\ \lambda_3 &= 9.4598\times10^{-5}\end{aligned} \tag{81}$$

truncated to 4 significant figures. (Calculations were made with the *Maple* Computer Algebra System set to 20-digit precision.)

Although there are numerous statistics that can be predicted by means of the joint, marginal, and conditional PDFs, the focus in this section (and in the following section) is on the predictive statistics of LDL-C. The reason for this emphasis is that numerous studies have concluded that lowering a hypercholesterolemic patient's serum LDL-C should be a primary objective of preventative and therapeutic measures to lower the risk of ASCVD and other cardiovascular disease [34]-[36].

As examined in the previous cases, the coefficient of variation (CV) and probability P_{in} of a point estimate falling within the reference range are two statistics indicative of the predictive uncertainty (and therefore reliability) of the numbers furnished by clinical labs. The conditional distributions based on the data in relation (80) and the solutions (81) lead to

$$\begin{cases}\mathrm{CV}_{Y|X}\left(81\,|\,70\right)=0.850\\ P_{Y|X}^{(\text{in})}\left(40,100\,|\,70\right)=32.5\%\end{cases} \tag{82}$$

from Equations (75) and (77) respectively, instead of $\mathrm{CV}_Y = 1.000$ and $P_Y^{(\text{in})}\left(40,100\right)=31.9\%$ for uncorrelated LDL-C. The differences are small, but do indicate a lower uncertainty when information regarding correlation is provided. The more information that goes into the MaxEnt analysis, the greater is the expected reliability of a diagnostic inference.

6.3. Case 4B: Risk Factors Correlated by Empirical Formula

Methods to measure LDL-C directly are in general complex, time-consuming, and expensive. Consequently, clinical labs in the U.S. and elsewhere routinely prefer to estimate LDL-C concentrations by means of an empirical formula such as the Friedewald equation [37], Martin equation [38], Sampson equation [39], and various others [32].

The Friedewald equation, which is linear in the variables, is the simplest of the empirical equations and has been in use the longest. It was proposed in 1972 as a means of estimating LDL-C without the need for ultracentrifugation, and has provided satisfactory results for patients with plasma lipid levels that are not too high or low. Subsequent elaborations to extend the formula to include patients with extreme lipid levels entail inclusion of numerous empirical parameters and/or powers of variables greater than 1. For the analysis in this section with lipid data within reference levels, the use of the Friedewald equation will suffice.

The Friedewald equation takes the simple form

$$Y = T - \frac{1}{5}G - X = X_m - X \tag{83}$$

in which

$$\begin{aligned} Y &= \text{LDL-C} & y_0 &= 83 \\ X &= \text{HDL-C} & x_0 &= 70 \\ T &= \text{Total-C} & t_0 &= 165 \\ G &= \text{Triglycerides} & g_0 &= 60 \end{aligned} \tag{84}$$

and

$$X_m \equiv T - \frac{1}{5}G \qquad x_m = 153\,. \tag{85}$$

X_m is the maximum value that X can take, since concentrations of the substances in Equation (83) must be non-negative quantities. The term involving triglycerides is a proxy for VLDL-C, *i.e.* the cholesterol carried by very low density lipoproteins. Thus T is representative of the total cholesterol concentration. It is apparent from the second equality in Equation (83), that LDL-C anti-correlates with HDL-C in a given sample.

The symbols and numerical values (in units of mg/dL) to the right of the variables in Equations (84) and (85) are the reported point estimates of a patient's Lipid Panel taken from Table 1 with one minor exception. A reader may notice that the point estimate y_0 is here taken to be 83, rather than 81 as in Table 1 and as used in Equation (80). The reason for the small change is that the numbers in Equation (84) satisfy the Friedewald equation, whereas the actual LDL-C point estimate reported by the clinical lab was obtained by the Martin equation. Nevertheless, the closeness of the two values is indicative of how well the Friedewald equation works for a non-hyperlipidemic patient.

Although a clinical lab is not likely to provide a correlation coefficient, knowledge of the point estimates of each of the four constituents in the Lipid Panel makes it possible, by means of the PME, to derive the most probable distribution function of the variables Y and X. Rigorous application of the PME in the present context would lead to a trivariant MaxEnt probability distribution for X, Y, G with constraint on their sum T. One could then integrate over unwanted variables to obtain the marginal distributions of X and Y. However, a much simpler and more transparent approximate approach is to take X_m constant for the given sample and derive univariate probability densities for X and Y directly.

In what follows, Y is treated as a function of the independent random variable X. Figure 3, based on the Friedewald Equation (83) and data in Equation (84), shows the linear variation of Y over the physically allowed range $[0, x_m]$. From Equation (83) and the basic principles of calculus, it follows that the probability densities of X and Y are related by the following transformation

$$p_Y(y) = \frac{p_X(x(y))}{|\mathrm{d}y/\mathrm{d}x|} = p_X(x_m - x)\,. \tag{86}$$

Thus, one need only find $p_X(x)$ to obtain $p_Y(y)$.

Implementation of the MaxEnt procedure, making use only of the available

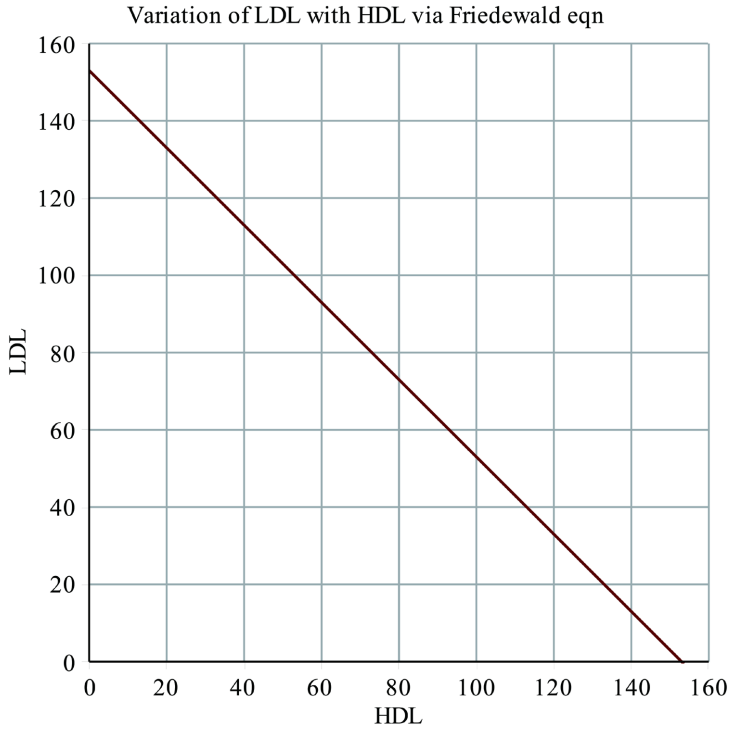

Figure 3. Variation of a patient's low-density lipoprotein cholesterol as a function of the concentration of high-density lipoprotein cholesterol as calculated from the Friedewald equation. Concentrations are in units of mg/dL.

information.

1) probability completeness,

2) point estimates x_0 and y_0 of the variables,

3) Friedewald value of cholesterol maximum x_m

Leads to the exponential PDFs

$$\begin{aligned} p_X(x) &= \exp(-\lambda x)/Z(\lambda) \\ p_Y(y) &= \exp(-\lambda(x_m - y))/Z(\lambda) \end{aligned} \tag{87}$$

in which the partition function

$$Z(\lambda) = \int_0^{x_m} \exp(-\lambda x)\mathrm{d}x = \lambda^{-1}\left(1 - \mathrm{e}^{-\lambda x_m}\right) \tag{88}$$

differs from that of Case 1 because the upper limit of the integral is no longer infinity, but constrained to x_m.

The Lagrange multiplier λ is obtained by applying Equation (11) in the form

$$-\partial \ln(Z(\lambda))/\partial\lambda = x_0, \tag{89}$$

which leads to the nonlinear equation

$$u\left(\left(\mathrm{e}^u - 1\right)^{-1} + \frac{x_0}{x_m}\right) = 1 \tag{90}$$

in the dimensionless variable

$$u \equiv \lambda x_m . \tag{91}$$

From Equations (87) and (88) can be derived the pertinent statistics of interest, including the population means $\overline{x}$, $\overline{y}$

$$\begin{cases} \overline{x}/x_m = u^{-1} - \left(\mathrm{e}^u - 1\right)^{-1} \\ \overline{y}/x_m = -u^{-1} + u\,\mathrm{e}^u \left(\mathrm{e}^u - 1\right)^{-1} \end{cases}, \tag{92}$$

variances

$$\frac{\mathrm{var}_X}{x_m^2} = \frac{\mathrm{var}_Y}{x_m^2} = u^{-2} - \mathrm{e}^u \left(\mathrm{e}^u - 1\right)^{-2}, \tag{93}$$

covariance

$$\frac{\mathrm{cov}_{X,Y}}{x_m^2} = -u^{-2} + \mathrm{e}^u \left(\mathrm{e}^u - 1\right)^{-2}, \tag{94}$$

and the integrated probabilities

$$\begin{cases} P_X^{(\mathrm{in})}\left(x_1, x_2\right) = \dfrac{\mathrm{e}^{-\lambda x_1} - \mathrm{e}^{-\lambda x_2}}{1 - \mathrm{e}^{-\lambda x_m}} \\ P_Y^{(\mathrm{in})}\left(y_1, y_2\right) = \dfrac{\mathrm{e}^{\lambda y_2} - \mathrm{e}^{\lambda y_1}}{\mathrm{e}^{\lambda x_m} - 1} \end{cases} \tag{95}$$

corresponding to Equations (76) and (77), that a repeat trial of each variable will fall within its respective reference interval.

It is to be noted from Equations (93) and (94) that the variance is the same for variables X and Y, and that the covariance is the negative of this variance. Thus, the Pearson correlation coefficient [40]

$$r_{X,Y} \equiv \frac{\mathrm{cov}_{X,Y}}{\sqrt{\left(\mathrm{var}_X\right)\left(\mathrm{var}_Y\right)}} = -1, \tag{96}$$

which is a widely used measure of linear correlation of two variables, takes the maximum negative value of the range $\left(1 \geq r_{X,Y} \geq -1\right)$.

6.4. Application to the Lipid Panel

Solution of Equation (90) for the variable u with $x_m = 153$ yields the Lagrange multiplier

$$\lambda = 3.3466 \times 10^{-3} . \tag{97}$$

Substitution of λ into relations (92)-(95) provides the statistics for judging the reliability of reported lipid panel results employing the Friedewald equation.

First, the results verify that the population means $\left(\overline{x}, \overline{y}\right)$ in Equation (92) do indeed reproduce the point estimates $\left(x_0, y_0\right)$ in the panel (84). Second, from the means (92) and variances (93) follow the coefficients of variation

$$\begin{cases} \mathrm{CV}_X = 0.627 \\ \mathrm{CV}_Y = 0.529 \end{cases} \tag{98}$$

which are significantly lower than the value 1 for uncorrelated exponential varia-

bles. Third, evaluation of the integrated probabilities (95) yield

$$\begin{cases} P_X^{(\text{in})}(40,90) = 33.6\% \\ P_Y^{(\text{in})}(40,100) = 38.0\% \end{cases}. \tag{99}$$

Although the value for LDL-C in Equation (99) exceeds that of Equation (82) for directly measured LDL-C, it still represents a large uncertainty.

Finally, for purposes of comparison with κ in relation (80), one can calculate the covariance coefficient

$$\kappa_0 \equiv \frac{\text{cov}_{X,Y}}{x_0\, y_0} = -0.331. \tag{100}$$

Note that the value of κ_0 in relation (100) is an actual outcome of the statistical analysis of Case 4B, whereas the value κ assigned as prior knowledge in Case 4A, relation (80), was intentionally chosen to be close to κ_0.

The reason for use of a "covariance coefficient" in Case 4A, rather than the standard Pearson correlation coefficient, is that the latter contains the variances (or standard deviations) of the two variables whose values are ordinarily *not* provided by the clinical laboratories. As pointed out previously, the purpose of examining Case 4A was to see what improvement in predictability would ensue from inclusion of the most minimal correlation information. Had that minimal information comprised the Pearson coefficient without also including the variances, there would have been insufficient information to derive a maximum entropy distribution. In such a case, one would have to resort to a specific model, which, in essence, is precisely what the Friedewald equation provides. In this case, minimal prior knowledge required to carry out a maximum entropy analysis did not have to include variances and/or covariances, since this information is retrievable from the empirical equation and full Lipid Panel.

7. Summary, Questions, and Conclusions

On the basis of the outcomes of many billions of tests of medical risk factors requisitioned from clinical labs each year, physicians in the U.S. and elsewhere make decisions on whether and how to provide therapeutic and preventative care to patients. This paper addressed quantitatively—for the first time, to the author's knowledge—the critical question of the predictive uncertainty of test results comprising just the point-estimates of analyte concentrations with *no* accompanying information as to the number of test trials, the intrinsic variability of the tests, or correlations with other analytes.

The analyses in this paper employed the Principle of Maximum Entropy (PME) to determine the least biased, most probable statistical distribution—referred to as the MaxEnt distribution—of test results in 4 fundamental cases involving incomplete information.

Case 1 examined the prevailing situation whereby no information beyond point-estimates of analyte concentrations are provided by the clinical labs. The resulting

MaxEnt distribution is of exponential form, leading to very high uncertainty (standard deviation equals the mean) and integrated probabilities well below 50% that a repeat measurement of the same sample would fall within the reference interval nominally regarded as "normal" or "low risk".

Cases 2 and 3 examined situations in which, respectively, the *exact* number or the *mean* number of test trials supplemented the reported point estimates. This additional information sharpened the uncertainty about the point estimate and increased the integrated probability over the reference range. Nevertheless, to achieve a satisfactorily high degree of predictive reliability would require a large number of trials, which perhaps is a feasible improvement for tests that can be automated.

Case 4 examined two situations of *correlated* risk factors: Case 4A in which the additional information was a point estimate of the covariance of two risk factors; Case 4B included point estimates of all constituents of a selected panel, together with an empirical formula connecting them. The example chosen for analysis was the pair of high (HDL-C) and low (LDL-C) density lipoprotein cholesterol without (Case 4A) or with (Case 4B) inclusion of other risk factors in the Lipid Panel. In both cases this minimal extra information sharpened the uncertainties of the point estimates of HDL-C and LDL-C by a small to modest amount, but the probability of a repeat test result falling within the reference interval was still disconcertingly low.

The preceding outcomes evoke several questions. Foremost is this: How *believable* are the MaxEnt projections in the chosen examples? For example: Is it *really* true that a repeat test of a risk factor, which is first measured to be within normal range, is more likely than not to fall *outside* the reference interval? A strictly accurate reply is to say "That's unknowable", since the clinical lab never provided results of concurrent retrials, and most likely none was made. A more expansive reply might include a suggestion to check the patient's records for results of past tests in order to look for any trends. However, such tests were made during annual exams and do not constitute concurrent retrials. Thus, the patient's data cannot be used to test the MaxEnt predictions.

Suppose, however, that concurrent retrials of a test *were* made and all results were again within reference levels. Would this indicate that the PME procedure could not be trusted? The answer is "No". MaxEnt distributions are consistent with the (presumed valid) prior information imported into the analysis. If there is a discrepancy between the statistics predicted by MaxEnt and the empirically acquired statistics, then, assuming the observations are accurate, *the empirical results must be subject to constraints not yet taken account of in the theoretical analysis*. This is not a flaw in the MaxEnt procedure, but an indication to the analyst to find out what unaccountable constraint has affected the data. In the context of the above supposition, a likely source would be something (chemical? physical? biological?) constraining the variability of the test. The solution, therefore, would be to determine the intrinsic variance of the test and include that among

the known prior constraints in the MaxEnt analysis. As pointed out at the beginning of this article, clinical labs *already have* that information, but just neglect to share it with physicians and patients.

Of the cases examined in this paper *none* included the variances or standard deviations of the mean estimates as part of the prior information. The reason for this omission is that the author has already analyzed those cases in great detail elsewhere [9] [10] for both independent and correlated variables.

The outcomes of a MaxEnt analysis with inclusion of test means, variances, covariances, and recognition of probability completeness are distribution functions of Gaussian or Lognormal form [18] (depending on how the prior information is presented). The predictive reliability of these distributions can be very high if the intrinsic uncertainties of the variables are sufficiently low. Conceivably, the tests performed by clinical labs do have satisfactorily low intrinsic uncertainties. However, if that information is not reported, then for physicians to assume it is true is to risk seriously misinterpreting the reported results to the detriment of the patient. This is especially concerning for point estimates of risk factors that skirt the upper or lower border of the reference range.

Finally, let it be supposed that at some future time clinical labs reform their practices and provide test results that include numbers of trials and test uncertainties. What reply can then be made to the director of a clinical lab who, in the Introduction to this article, told the author that "Neither physicians nor patients would know what to do with it."? Here is such a reply:

The author does *not* expect physicians (and certainly not patients) to be able to carry out analyses and numerical computations like those in this article. Nevertheless, it is not uncommon for professional health and medical organizations to create online calculators for short- and long-term risk prediction of specific conditions, and numerous such calculators already exist [41]. Clinical labs could likewise create such calculators for the panels of medical risk factors that they test. Then physicians, supplied with adequately informative test results, can enter this information by computer or other digital device to learn immediately and with scientific rigor, rather than by ill-informed guesswork, the potential risks their patients face.

Acknowledgements

The author would like to thank Dr. Brad Sherburne for his helpful correspondence and an enjoyable, informative discussion at the time he was medical director of the Hartford Hospital Laboratory. He would also like to thank the anonymous reviewer for his thoughtful suggestions.

Conflicts of Interest

This article was conceived and written by the author alone with no reliance on artificial intelligence. The author has received no compensation for the composition of this article and declares there are no conflicts of interest to disclose.

References

[1] Centers for Disease Control and Prevention (CDC) (2024) Clinical Standardization Programs. https://www.cdc.gov/clinical-standardization-programs/about/index.html

[2] Silverman, M.P. (2026) Communicating Medical Numbers. *JAMA*, **335**, 721-722. https://doi.org/10.1001/jama.2025.23446

[3] International Organization for Standardization (ISO/TAG 4/WG 3) (2023) Joint Committee for Guides in Metrology (JCGM), The Guide to the Expression of Uncertainty in Measurement. https://physics.nist.gov/cuu/Uncertainty/international2.html

[4] National Accreditation Center (2024) Measurement Uncertainty in Laboratories: The Hidden Aspect of Reliable Results. https://nac-us.org/faqs

[5] White, D.G. (2022) Hitchhikers Guide to Measurement Uncertainty in Medical Laboratories. *Asia-Pacific Federation for Clinical Biochemistry and Laboratory Medicine*, **1**, 68-75. https://doi.org/10.62772/apfcb-news.2022.2.1

[6] Milinković, N., Ignjatović, S., Šumarac, Z. and Majkić-Singh, N. (2018) Uncertainty of Measurement in Laboratory Medicine. *Journal of Medical Biochemistry*, **37**, 279-288. https://doi.org/10.2478/jomb-2018-0002

[7] Jaynes, E.T. (1957) Information Theory and Statistical Mechanics. *Physical Review*, **106**, 620-630. https://doi.org/10.1103/physrev.106.620

[8] Silverman, M.P. (2023) Probability Distribution of SARS-CoV-2 (COVID) Infectivity Following Onset of Symptoms: Analysis from First Principles. *Open Journal of Statistics*, **13**, 233-263. https://doi.org/10.4236/ojs.2023.132013

[9] Silverman, M.P. (2025) Maximum Entropy Distribution of Correlated Variables: Application to Human Height and Weight. *Open Journal of Statistics*, **15**, 371-389. https://doi.org/10.4236/ojs.2025.154020

[10] Silverman, M.P. and Lipscombe, T.C. (2022) Exact Statistical Distribution of the Body Mass Index (BMI): Analysis and Experimental Confirmation. *Open Journal of Statistics*, **12**, 324-356. https://doi.org/10.4236/ojs.2022.123022

[11] Silverman, M.P. (2025) Perspective on the Body Mass Index (BMI) and Variability of Human Weight and Height. *Journal of Biosciences and Medicines*, **13**, 309-320. https://doi.org/10.4236/jbm.2025.136026

[12] Silverman, M.P. (2015) Cheating or Coincidence? Statistical Method Employing the Principle of Maximum Entropy for Judging Whether a Student Has Committed Plagiarism. *Open Journal of Statistics*, **5**, 143-157. https://doi.org/10.4236/ojs.2015.52018

[13] Jaynes, E.T. (1963) Brandeis Lectures. In: Rosenkrantz, R.D. and Jaynes, E.T., *Papers on Probability, Statistics, and Statistical Physics*, Kluwer, 39-76.

[14] Wikipedia (2025) Boltzmann's Entropy Formula. https://en.wikipedia.org/wiki/Boltzmann's_entropy_formula

[15] Jeffreys, H. (1961) Theory of Probability. Oxford University Press, 33-34.

[16] Gordon, C.C., *et al.* (2012) Technical Report Natick/TR-15/007, Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics. U.S. Army Natick Soldier Research, Development and Engineering Center. https://apps.dtic.mil/sti/citations/ADA611869

[17] Jaynes, E. (1968) Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics*, **4**, 227-241. https://doi.org/10.1109/tssc.1968.300117

[18] Forbes, C., Evans, M., Hastings, N. and Peacock, B. (2011) Statistical Distributions. 4th Edition, John Wiley & Sons, 88-92, 109-113, 131-134, 143-148.

[19] Watson, S. (2024) What Is Good Cholesterol?

https://webmd.com/cholesterol-management/good-cholesterol-too-high

[20] Hussain, S.M., Robb, C., Tonkin, A.M., *et al.* (2024) Association of Plasma High-Density Lipoprotein Cholesterol Level with Risk of Incident Dementia: A Cohort Study of Healthy Older Adults. *The Lancet Regional Health*, **43**, Article 100963. https://doi.org/10.1016/j.lanwpc.2023.100963

[21] Ferguson, E.L., Zimmerman, S.C., Jiang, C., Choi, M., Swinnerton, K., Choudhary, V., *et al.* (2023) Low- and High-Density Lipoprotein Cholesterol and Dementia Risk over 17 Years of Follow-Up among Members of a Large Health Care Plan. *Neurology*, **101**, e2172-e2184. https://doi.org/10.1212/wnl.0000000000207876

[22] Barter, P. and Genest, J. (2019) HDL Cholesterol and ASCVD Risk Stratification: A Debate. *Atherosclerosis*, **283**, 7-12. https://doi.org/10.1016/j.atherosclerosis.2019.01.001

[23] Cleveland Clinic (2026) Hypocholesterolemia. https://my.clevelandclinic.org/health/diseases/hypocholesterolemia-low-cholesterol

[24] Mayo Clinic (2022) Cholesterol Level: Can It Be Too Low? https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/expert-answers/cholesterol-level/faq-20057952

[25] Silverman, M.P. (2014) A Certain Uncertainty: Nature's Random Ways. Cambridge University Press. https://doi.org/10.1017/cbo9781139507370

[26] Loudon, R. (1973) The Quantum Theory of Light. Oxford University Press, 9-12.

[27] CDC National Center for Health Statistics (2026) Leading Causes of Death. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

[28] World Health Organization (2024) The Top 10 Causes of Death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[29] Wikipedia (2026) Lipoprotein. https://en.wikipedia.org/wiki/Lipoprotein

[30] Cleveland Clinic (2022) Lipoproteins. https://my.clevelandclinic.org/health/articles/23229-lipoprotein

[31] Miller, W.G., Myers, G.L., Sakurabayashi, I., Bachmann, L.M., Caudill, S.P., Dziekonski, A., *et al.* (2010) Seven Direct Methods for Measuring HDL and LDL Cholesterol Compared with Ultracentrifugation Reference Measurement Procedures. *Clinical Chemistry*, **56**, 977-986. https://doi.org/10.1373/clinchem.2009.142810

[32] Brown, W.V. (2020) Methods of Calculating Low-Density Lipoprotein Cholesterol Level. *JAMA Cardiology*, **5**, 502-503. https://doi.org/10.1001/jamacardio.2020.0042

[33] Wolska, A. and Remaley, A.T. (2020) Measuring LDL-Cholesterol: What Is the Best Way to Do It? *Current Opinion in Cardiology*, **35**, 405-411. https://doi.org/10.1097/hco.0000000000000740

[34] Agrawal, N. (2026) Confused About the New Cholesterol Guidelines? Here's What to Know. The New York Times. https://www.nytimes.com/2026/04/09/well/cholesterol-guidelines-heart-disease.html

[35] Abbasi, J. (2026) What to Know about the New Lipid Guidelines. *JAMA*, **335**, 1285-1287. https://doi.org/10.1001/jama.2026.3968

[36] Blumenthal, R.S., *et al.* (2026) 2026 ACC/AHA/AACVPR/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Dyslipidemia: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*, **153**, e1-e123. https://www.ahajournals.org/doi/10.1161/CIR.0000000000001423

[37] Friedewald, W.T., Levy, R.I. and Fredrickson, D.S. (1972) Estimation of the Concentration of Low-Density Lipoprotein Cholesterol in Plasma, without Use of the Preparative

Ultracentrifuge. *Clinical Chemistry*, **18**, 499-502. https://doi.org/10.1093/clinchem/18.6.499

[38] Martin, S.S., Blaha, M.J., Elshazly, M.B., Toth, P.P., Kwiterovich, P.O., Blumenthal, R.S., *et al.* (2013) Comparison of a Novel Method vs the Friedewald Equation for Estimating Low-Density Lipoprotein Cholesterol Levels from the Standard Lipid Profile. *JAMA*, **310**, 2061-2068. https://doi.org/10.1001/jama.2013.280532

[39] Sampson, M., Ling, C., Sun, Q., Harb, R., Ashmaig, M., Warnick, R., *et al.* (2020) A New Equation for Calculation of Low-Density Lipoprotein Cholesterol in Patients with Normolipidemia and/or Hypertriglyceridemia. *JAMA Cardiology*, **5**, 540-548. https://doi.org/10.1001/jamacardio.2020.0013

[40] Altman, D.G. (1991) Practical Statistics for Medical Research. Chapman & Hall/CRC, 277-288.

[41] Rossello, J. (2025) Best Interactive Risk Calculators and Health Assessment Tools: A Comprehensive Analysis. Preventive Medicine Daily. https://www.preventivemedicinedaily.com/risk-assessment/best-interactive-risk-calculators-and-health-assessment-tools-a-comprehensive-analysis/

Summary of Abbreviations

ASCVD	Atherosclerotic cardiovascular disease
BIPM	International Bureau of Weights and Measures
BUN	Blood, Urea, Nitrogen
CDC	U.S. Centers for Disease Control and Prevention
CF	Characteristic function
CI	Confidence interval
CIPM	Comité international des Poids et Mesures
CV	Coefficient of variation
ESM	Equilibrium statistical mechanics
HDL	High density lipoprotein
HDL-C	High density lipoprotein cholesterol
IDL	Intermediate density lipoprotein
LDL	Low density lipoprotein
LDL-C	Low density lipoprotein cholesterol
MaxEnt	Maximum Entropy Probability Distribution
mg/dL	milligrams per deciliter
PDF	Probability density function
PME	Principle of Maximum Entropy
Total-C	Total cholesterol
VLDL	Very low density lipoprotein
VLDL-C	Very low density lipoprotein cholesterol
WHO	World Health Organization