Where's the Evidence?

Veronica J. Vieland

Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital,
Columbus, Ohio, USA

Key Words

Calibration • Evidence • Likelihood ratios • p values •
Statistical inference • Thermodynamics • Thermometry

Abstract

Science is in large part the art of careful measurement, and a fixed measurement scale is the sine qua non of this art. It is obvious to us that measurement devices lacking fixed units and constancy of scale across applications are problematic, yet we seem oddly laissez faire in our approach to measurement of one critically important quantity: statistical evidence. Here I reconsider problems with reliance on p values or maximum LOD scores as measures of evidence, from a measurement-theoretic perspective. I argue that the lack of an absolute scale for evidence measurement is every bit as problematic for modern biological research as was lack of an absolute thermal scale in pre-thermodynamic physics. Indeed, the difficulty of establishing properly calibrated evidence measures is strikingly similar to the problem 19th century physicists faced in deriving an absolute scale for the measurement of temperature. I propose that the formal relationship between the two problems might enable us to apply the mathematical foundations of thermodynamics to establish an absolute scale for the measurement of evidence, in statistical applications and possibly other areas of mathematical modeling as well. Here I begin to sketch out what such an endeavor might look like.

Copyright © 2011 S. Karger AG, Basel

'Nearly all the grandest discoveries of science have been but the rewards of accurate measurement...'

Lord Kelvin, 1871

What is science if not evidence based? It goes without saying that assessment of evidence is a core component of all biomedical research. And in quantitative research, the mathematical measurement of evidence is, arguably, the principal desired outcome of most analyses. It is on the basis of the strength of evidence for or against hypotheses that we interpret our findings, adjust our understanding, and make decisions regarding future studies. Yet something is seriously amiss in the methods we use to measure evidence in human genetics, as I will argue below. To be evidence based without a proper measure of evidence is patently problematic. We need to fix this problem.

How We Measure Evidence

There are many areas of biological research in which evidence is measured completely informally. We look under the microscope and judge whether what we see conforms or not to our hypotheses regarding what should be happening, etc. But in human genetics and many other fields, the far more common procedure for measuring evidence is via statistical analysis. Even the simplest genetic principles, such as Mendel's 1st law, involve stochas-

tic, or probabilistic, variation, so that determining whether data conform to one genetic model or another requires statistical reasoning. Moreover, we are now capable of generating vast quantities of ‘-omics’ data (e.g., genomics or proteomics data) relatively quickly and ever less expensively. Both the quantity and the complexity of such data defy assessment of evidence based on simple inspection. Accordingly, our literature contains a multitude of statistical methods for analyzing such data, that is, for measuring the evidence that the data convey.

Take for instance the workhorse of human genetics, the simple genome scan. I will illustrate with genome-wide association studies (GWAS), but the same principles apply to linkage and many other types of analyses as well. We collect cases and controls, and genotype them at large numbers of SNPs spanning the genome. We then perform a statistical analysis at each SNP, looking for the SNPs with the strongest evidence of being associated with disease. Then to be sure that any findings to emerge are not false positives, we attempt to replicate the salient results in an independent sample.

The way we measure the evidence in GWAS is typically via the *p* value, based on whatever particular form of calculation is chosen, with smaller *p* values taken to indicate stronger evidence in favor of an association. So, what properties must the *p* value have for this whole procedure to be evidence based? Minimally, the *p* value must effectively rank-order the SNPs from strongest evidence for association to weakest; otherwise, the entire enterprise would not be evidence based at all. Moreover, we must have reasonable grounds for the expectation that if a finding is true positive, a second sample should also yield an appropriately small *p* value. Equivalently, we must have reasonable grounds to believe that if we were to increase the sample size, the *p* value at a truly associated SNP would continue to get smaller, that is, yield stronger evidence. If the second sample is equally or more likely not to replicate a true finding, then non-replication becomes moot. Similarly, if increasing the sample size might well make the *p* value larger at an associated SNP, indicating less evidence, then the genome scan design would be thwarted by our inability to clarify the evidence by collecting more data.

The Problem with the Way We Measure Evidence

The dirty little secret of statistics (to paraphrase Siegfried [1]) is that we know that these properties do not hold for the *p* value. The fact that *p* values do not rank-order

results by the evidence is widely accepted among practitioners of GWAS, given that the *p* value depends upon such things as allele frequencies. But there is a deeper issue too.

The accumulation of evidence with increasing amounts of data follows certain patterns that we can all agree upon. For instance, suppose that we flip a coin to see whether it is ‘fair’ (lands heads 50% of the time) or biased, and we do so in two stages, with data set D_1 comprising n_1 tosses and having evidence E_1 , and D_2 comprising n_2 tosses with evidence E_2 . Let E be the total evidence, considering both D_1 and D_2 . Given what we mean by evidence, I believe that with a little thought everyone will agree that if D_1 and D_2 each individually support ‘bias’, then the total evidence E must be greater in favor of bias than the larger of E_1 and E_2 , each considered on its own. Similarly, if D_1 and D_2 each support ‘no bias’ then the total evidence E must be less (more strongly in support of ‘no bias’) than the smaller of E_1 and E_2 . That’s how evidence behaves. But it is easily shown that *p* values do not necessarily follow this pattern [2]. They will under many circumstances tend to behave more like averages of E_1 and E_2 . (The same is true of the maximum likelihood ratio (LR) or its logarithm, the maximum LOD score [3].) We can think of the overall evidence E as resulting from some particular concatenation operation performed on E_1 and E_2 . We may not have a precise understanding of what this concatenation operation is, but we can see that whatever it is, it is not the operation that occurs when we combine the data sets and recompute the *p* value. This means that the *p* value is in some very fundamental way simply not tracking with the evidence at all. The point is not that the *p* value never tracks with the evidence, but rather that there is no intrinsic connection between the *p* value and the evidence such that we can rely upon it to do so.

In fact, the widely accepted ‘winner’s curse’ – a form of regression to the mean – is simply a manifestation of this problem. It has been shown that it is reasonable to expect, when following up on the smallest *p* value under fairly ordinary circumstances and assuming a true association, that increasing the sample size will tend to lead to larger, not smaller, *p* values [2]. We can express this by saying that the *p* value can indicate less evidence for association even while the evidence itself is going up. Thus the smallest *p* values do not necessarily replicate, even if they represent true positives, even setting aside all other complications [4, 5]. This of course means that failure to replicate in no way establishes the falsehood of a prior finding, as is well known. What is less appreciated is that the winner’s curse is symptomatic of an underlying mea-

surement problem, viz., the failure of the *p* value to consistently track with the evidence as data accrue.

One important additional feature of our current practice is that we can only accumulate evidence (no matter how badly measured) in favor of genetic effects, but we have no mechanism for accumulating evidence against genetic effects. We can of course use failure to replicate for this purpose, but this begs the question of how we will know when a replication sample weakly supports an initial finding versus when it actually gives evidence against this initial finding, bringing us back to the problem of measuring evidence against a hypothesis. Once a result is reported in the literature, subsequent literature reviews will find the result without any systematic way of assessing whether the preponderance of evidence actually weighs against it. The consequence of this is that ‘findings’ are seldom removed from our (real or virtual) databases of genomic results, leading to a monotonically increasing proportion of the genome becoming implicated for any given disorder over time. In view of this, it should come as no surprise that in many cases the more literature we accumulate on a given topic, the murkier is the picture that emerges of the evidence in its entirety. If we do science properly, however, the more data we accumulate the clearer things should become. This is after all the purpose of collecting data.

In statistical circles, problems with using *p* values to measure evidence have been widely discussed (e.g. [6–9]) and the very practical pitfalls of relying on them in biomedical research are drawing increasing attention even in the popular press [10], thanks in part to recent work by Ioannidis [11]. Arguably, many, though not all, of the problems that this work is uncovering stem from the reliance on *p* values and significance testing in lieu of proper evidence measures. But in any case, it is hard to see the justification for continuing to use *p* values for a purpose for which they are clearly unsuited.

I have also thus far skipped over one standard feature of GWAS analysis that provides a singularly telling clue that what we have on our hands is an underlying measurement problem. Before one uses the *p* value as a measure of evidence, one ‘corrects’ it for multiple tests. This seems on the face of it innocuous. The *p* value is the probability of seeing a result as extreme as the observed result under the hypothesis of ‘no association’, but clearly seeing an event with probability of, say, 0.001 means something quite different in the context of a single SNP than in the context of evaluating a million SNPs. Thus we use that number – 0.001 – as a measure of evidence, but only after correcting it for the total number of tests.

From a purely measure-theoretic point of view, this is an extraordinarily peculiar procedure! The fact that we don’t balk at ‘correcting’ the *p* value for multiple tests prior to interpreting it is symptomatic of the fact that the *p* value is simply not a proper measure and should not be used as one. Imagine if carpenters had to divide the length of each board by the total number of boards going into a new house before placing an order at the lumberyard, or if a nurse had to divide your temperature by the total number of patients in your ward before assessing your fever. This would raise some serious questions regarding the nature of the measurement technique and the interpretability of measurement results.

A Precedent for Our Measurement Problem

Speaking of temperature, we have an excellent model for our predicament in the history of thermometry. Note that this section continues a line of reasoning first explored by Vieland [2], and there is some redundancy with the arguments given there. However, here I frame the central issues somewhat differently and follow this reasoning to a very different, and entirely new, conclusion.

Throughout the 18th and 19th centuries, the topic of temperature permeated a great deal of work in physics, ranging from the very practical problem of designing efficient steam engines to reconciliation of the various conceptions of caloric, a substance postulated as the medium of heat transfer that we now know to be nonexistent. But physicists of the day had no recourse to thermometers in their modern form to assist them in these pursuits. Instead, they relied on instruments called thermoscopes, which resembled ordinary thermometers but for one key feature: they were, in essence, uncalibrated.

For example, some thermoscopes contained alcohol, while others contained mercury. In each case, when heated, the liquid expanded and the amount of expansion could be used as an indication of increased temperature. But because different liquids expand at different rates when heated, the unit of measurement – the degree – did not mean the same thing across the different devices. Furthermore, within each thermoscope the rate of expansion was a function of the range in which the temperature was varied, among other factors, further complicating comparisons across devices and even across the scale for a single device.

This posed several problems. First, the measured temperature could be both 50 and 60° at the same time and place, if two different thermoscopes were used, or it

could be 50° at each of two places, while the actual temperatures were quite different. Second, it was unclear whether an increase in temperature of 1° at the low end of the scale meant the same thing as an increase of 1° at the high end, even for measurements made with a single thermoscope. And third, above all, physicists lacked an objective procedure for resolving these difficulties. This represented both a significant practical challenge and a deep conceptual obstacle: It is easy to verify that a particular thermoscope is accurately measuring the temperature if there exists an established thermometer for comparison, but how does one go about calibrating the first thermometer?

What makes the calibration of thermoscopes so difficult is that it poses a problem of what Chang [12] calls nomic measurement, or measurement of a quantity that is itself not directly observable, based on application of an underlying law. In this case, we want to measure a quantity X (temperature) on the basis of some observable phenomenon Y (expansion of a liquid when heated), through a law $f(Y) = X$, expressing X as a function of Y . But such a law cannot be empirically discovered or verified, because this would require knowing the true value of X , which is obtained only subsequent to application of the law. The hallmark of nomic measurement is thus its apparent circularity. This circularity led many early physicists to eschew talk of the temperature as something with unique physical existence, advocating instead a purely relative calibration in which one thermoscope was selected and others were simply calibrated against it. Of course, ‘the temperature’ presents no such dissonance to the modern ear. This represents a fundamental triumph of physics over metaphysics. And interestingly, it was in fact relative – rather than absolute – calibration that proved intractable, due to the difficulty of establishing the degree as a constant unit in the absence of a deeper theoretical understanding of the nature of temperature. As Chang [12] notes, the derivation of the absolute measurement scale gave rise to the rigorous thermodynamic definition of temperature, and not the other way around.

The history of thermodynamics provides a useful analogy in thinking about evidence and the nature of measurement. Moreover, I would like to propose that the mathematics of thermodynamics provides the foundations upon which a coherent theory of evidence measurement might actually be constructed. But before considering this proposal in more detail, we need to reframe our statistical problem as a measurement problem.

Measurement of Evidence as a Nomic Measurement Problem

If we view statistical analyses from a measure-theoretic perspective, the analogy with pre-thermodynamic physics is obvious. We have a number of statistical ‘devices’ for measuring evidence in biological settings. The most familiar outcome measures are the p value and its relatives (e.g., q values in sequence alignment). Other evidence measures include the LR, the LOD score ($\log_{10}$ LR) in statistical genetics, integrated LRs, posterior probabilities, and various distance and correlation measures. Additionally, multiple statistical models can be applied to any set of data, yielding one or more of these outcome measures.

This leads to fundamental difficulties. For example, one may have p values of both, say, 0.01 and 0.06 for the same data, if one uses two different methods (e.g., variance components vs. regression) for calculating the p value. Or one may obtain 0.01 and 0.06 for two different sets of data, based on two types of analysis, with no way to say which represents the stronger evidence. Additionally, different evidence measures (e.g., LRs and p values) are on incomparable scales. Moreover, it is unclear whether a change in LR from 1 to 2 as new data are added means the same thing as a change from 10 to 11, or perhaps 10 to 20 (similarly for p values). Even assuming that one experiment could succeed in rank-ordering its own results with respect to the strength of the evidence, across experiments, results would remain incomparable and the units in which evidence is measured obscure.

This ought to bother us deeply. Think again about thermoscopes. Would we be satisfied if the measured temperature at a particular time and place was both 50 and 60°? Or if a measure of 50° at a second location gave us no information as to which was the warmer locale? Science is in large part the art of careful measurement, and a fixed measurement scale is the sine qua non of this art. It is obvious to us that thermal measurement devices lacking fixed units and constancy of scale across applications are problematic, yet we seem oddly *laissez faire* in our approach to evidence.

Of course, if we are interested merely in establishing the warmest room in a building, we can carry a single thermoscope from room to room, waiting in each location for it to reach equilibrium with its environment and recording the results. Imagine, however, trying to integrate large numbers of temperature readings into a unified model for purposes of weather prediction, with those readings coming from reporting stations around

the globe and based on a multitude of different thermoscopes. Such data would be useless. Just so, in simple biological settings, any measure that can reasonably rank-order data with respect to relative evidence may be sufficient (though how we would verify that the measure was correctly rank-ordering by the evidence is itself a conundrum), but any attempt to simultaneously interpret measures derived from many distinct approaches, or to layer multiple levels of analysis into unified experimental outcomes, requires a system for calibrating the various and sundry evidence measures against one another on an absolute basis, i.e., independent of context.

But the critical point here is to recognize that evidence, like temperature, represents a nomic measurement problem. We wish to measure the strength of the evidence based on some observable, or computable, quantity, for instance, the p value or LR (or perhaps a non-stochastic measure). Here the LR, say, is analogous to the volume of liquid within a thermometer, in the sense that, at least in principle, as the underlying evidence goes up or down, the LR should decrease or increase accordingly. We seek a definition of the degree of evidence that has the same meaning across measurement ‘devices’ (alternative statistical models), across different experimental domains (multiple biological levels of analysis), and across the range of the evidential scale. But how will we know whether we have correctly defined the law mapping the evidence onto these computable quantities if we can only infer what is the true evidence from the application of that same law? We are confronted with the same circularity that stymied thermometry for so long.

Statistical analyses seem to us so intrinsically method and context dependent that we may at first balk at the notion that evidence and temperature could possibly share a metaphysical basis. Surely evidence is always relative to a particular model or form of analysis, is it not? How then is it possible to disentangle evidence from the form of the equation used to measure it, or to talk about *the* evidence as a thing apart from the method one uses to measure it? But neither did it seem possible for 250 years of serious work in thermometry to craft a coherent concept of the temperature, one that disentangled the temperature from the thermometer used to measure it. Of course we now know that physicists eventually solved their nomic measurement problem, and I believe we can solve ours. In fact, I believe we can do so by adopting their methods.

Towards the Measurement of Evidence on the Absolute Scale

While the analogy with thermometry is instructive, I would propose that there is an even deeper connection between the physicists’ underlying measurement problem and ours, and I believe we can adapt the established foundations of thermodynamics to accelerate the development of an absolute evidence scale. Others have already made various relevant connections: see for instance work by Cox [13], Jaynes [14], and Shannon [15], all of whom borrow methods (e.g., differential equations) and individual concepts (e.g., entropy) from physics in grappling with the foundations of inferential and informational systems. But I think we can take an even more radical approach, and harness the actual foundations of thermodynamics to derive an evidential analogue of the true thermometer. Of course I am not claiming that evidence and temperature literally share a physical basis, but neither am I invoking thermodynamics as mere metaphor. Rather, I am proposing that the two are related by virtue of our ability to represent them within the same underlying mathematical framework. Indeed, work within physics itself has already shown the generality of the framework (see Callen’s derivation of thermodynamics [16] from simple symmetry relationships, applicable to all types of systems in macroscopic aggregation, and Caratheodory’s axiomatic development [17]).

To give the flavor of how this might go, consider first the fundamental physical construct known as the equation of state. An equation of state gives a complete description of a system in terms of a (generally small) number of interrelated parameters describing the system at a macroscopic level. For instance, the behavior of m moles of an ideal gas can be described in terms of the absolute temperature T (that is, temperature measured on the absolute scale), volume V , and pressure P through the equation of state

$$PV = mRT, \quad (1)$$

(with R = gas constant). This equation represents a complete description of the system in the sense that knowing the value of any two of the three parameters (T , V , P) determines the value of the third, while any additional macroscopic properties of the system (e.g., entropy) can be derived once this equation is known. Equation 1 is therefore a particular instance of an equation of state, which can be represented in the more general form

$$f_{IG}(T, V, P) = 0, \quad (2)$$

where the subscript *IG* (for ideal gas) indicates that the function *f* will take on a particular form specific to ideal gases.

A great deal of fundamental theory can be worked out without an explicit definition of *T*, based only on the assumption that an equation of state in the form of equation 2 exists. This theory can then be related to the behavior of actual gases provided only that we have access to a device for measuring empirical temperature *t* (that is, temperature as it would be measured by a thermoscope), such that *T* equals some function *g* of *t*. Here the specific law *g(t) = T* need not be known, but there is a presumption that a device for measuring *t* exists such that this law could be established in principle. In practice, this simply means that the empirical temperature *t* must ‘track’ with the absolute temperature, going up when *T* goes up and down when *T* goes down.

Statistical models can also be viewed as equations of state, similarly framed in terms of an empirical (‘thermoscopic’) measure of evidence *e*, even before we have a definition of what would be the absolute evidence *E*. All we need is an empirical measure that behaves like a thermoscope, that is, one that goes up or down as the evidence goes up or down, at least in simple settings and under normal circumstances (see also [2] for further discussion of how we know when an evidence measure is behaving like a thermoscope). For instance, think of binomial data (*N* coin tosses of which *X* land heads) and the simple hypotheses $P[\text{heads}] = q = 0.05$ versus $q = 0.5$. Let us assume at least for the moment that the LR itself behaves like a thermoscope for this simple system, that is, that it correctly tracks with the evidence, or in other words, that $e = \text{LR}$. (This is probably a safe assumption for comparisons between simple hypotheses [3, 7], but does not necessarily assist us in measuring evidence for compound hypotheses, where unknown parameters can take multiple values.) In this case we have

$$e = 2^N (0.05)^X (1 - 0.05)^{N-X} \quad (3)$$

Now if we hold *e* constant and increase *N*, *X* will have to change by a compensatory amount; similarly for other relationships among *e*, *N*, and *X*. Indeed, if we fix any two of the three variables in this equation, the value of the third is known and simple to calculate. Thus equation 3 is an equation of state in *e*, *N*, and *X*.

But what can we do with this equation of state? In physics, the equation of state allows construction of a Carnot engine, a device for quantifying the transformation of heat into work in the absence of any prior operational definition of heat [18]. Note that the Carnot engine

is itself a purely mathematical device, based on certain assumptions – e.g., perfect reversibility – which cannot be realized by physical systems. Kelvin and Joule defined the absolute temperature scale by deriving it from mathematical features of the Carnot engine, rather than defining it *ab initio* [18]. It is hard to overstate the ingenuity of this maneuver. Without knowing what heat was (this was still the era of caloric, after all) or what precisely was meant by *T*, mathematical insights based on the Carnot engine yielded both a definition of *T* and an absolute scale for its measurement. The details are beyond the scope of this essay, but the interested reader is referred to Chang [12] for a lucid presentation that is quite accessible to non-physicists.

Returning to the evidential problem, for a given set of data, the simple equation of state given above (equation 3) for the binomial case represents a ‘system’, which can be plotted as a curve with *q* on the x-axis and $e = \text{LR}$ on the y-axis. Insofar as equation 3 is an equation of state, this plot conveys all of the information in the data regarding the strength of the evidence for or against $q = 0.05$ for given data. The plot itself has some very real, physical properties, including for instance the area *A* under the curve and the maximizing value *Q* of *q*. We can, therefore, reformulate equation 3, which is written in terms of the data (*N*, *X*) to being about properties of the graph itself, such as *A* and *Q*. This yields an equation of state in the form

$$f_{\text{BIN}}(e, A, Q) = 0. \quad (4)$$

[Here the subscript *BIN* (for binomial) is a reminder that the particular form of *f* will be dictated by the behavior of binomial systems.] Equation 4 is a simple reparameterization of equation 3, which also encapsulates the property that if, say, we hold the evidence *e* constant and increase *A* by a specific amount, *Q* will have to shift by a compensatory amount (and similarly for other relationships among *e*, *A*, and *Q*).

In the physical system, *T*, *V*, and *P* constitute macroscopic properties of the system that are affected by changes in energy, in particular by the influx or outflux of heat. We can similarly think of the influx of new information – like the influx of heat – as performing ‘work’ on the graph associated with equation 4, that is, changing its physical properties *A* and *Q*. Viewing the statistical system in this way suggests that we should be able to run ‘evidential’ Carnot cycles and to study their properties. This opens the door to derivation of an absolute scale for the measurement of evidence, following the Kelvin-Joule template. There is an enormous amount of mathematical

detail remaining to be worked out, and, of course, the devil may well reside in the details. However, if this basic framework is even approximately correct, this means that evidential equations of state can be derived prior to defining what precisely is meant by the evidence, and deployed in 'evidentialism' just as they are in thermodynamics, to provide the basis for (and definition of) absolute measurement.

This requires of course that we accept evidential analogues of the laws of thermodynamics, but this is not so farfetched as it might at first appear. While there may not be a comparable physical basis for an evidential version of the 1st law (which stipulates the conservation of energy), certain elementary principles come to mind that could stand in nicely: for instance, the law of total probability, or perhaps, a comparable law constraining total evidential information. The analogue of the 2nd law (which, in one form or another, describes the tendency of systems to equilibrate in their maximum entropy configurations) is perhaps more obscure. However, given the close connection between thermodynamics, statistical mechanics, and the entropy-based information theoretic framework of Shannon [15], it seems reasonable to expect that an entropy-based formulation will also be forthcoming in the evidential case, and this would in turn allow concise formulation of an evidential 2nd law.

Not only would all of this give us a framework for solving our nomic measurement problem, but application of the theory to specific subject-matter domains could produce additional results paralleling those of physics. For example, the amount of heat required to raise the temperature of a liquid by a given amount depends on the liquid, or in other words, different liquids have different specific heats. This means that different liquids have different equations of state, so that the thermal meaning of a given change in volume within a thermoscope is not fixed. Just so, the quantity of data required to change the evidence by a fixed amount will depend upon the hypotheses of interest, among other things. We can think of this as a matter of different 'specific heats' for different statistical applications. We will therefore need to derive empirical adjustments to equations of state for particular applications, to ensure that a change of 1° in our measure of the evidence always means the same thing. The Carnot engine also yielded a theoretical upper bound on the efficiency with which heat could be converted to work, generating new metrics for investigating the relative efficiencies of various real engines. Just so, our evidential analogue of the Carnot engine could yield a theoretical upper bound on the efficiency with which data, or the informa-

tion conveyed by the data, can be converted to evidence, in turn leading to new ways to evaluate mathematical modeling methods by comparison with this upper bound.

It is also important to note that definition of the absolute (Kelvin) scale for temperature did not in any way necessitate replacement of thermoscopes. It simply provided the basis for absolute calibration of existing thermometric devices. Just so, this line of inquiry would *not* replace other research on statistical methods in biology, although it might call for some adjustments to ensure that all of our empirical outcome measures behave like thermoscopes. The object would be simply to harmonize the evidence scale on which results of a multitude of statistical approaches can be represented, in the process formulating a general theoretical framework and the many ancillary benefits that can come from having such a framework in place. Indeed, there is no reason to think that the benefits would be restricted to statistical modeling; non-stochastic methods, such as control theory, might also incorporate well into this framework.

But Why Bother?

There are practical implications of relying on faulty evidence measures. In the first place, doing so is like relying on a piece of laboratory equipment known to be unreliable and uncalibrated across laboratories. Good science simply cannot be based on badly behaved measures. Secondly, as anyone who has ever attempted to summarize genetic or genomic findings based on the literature on a particular complex disease can attest, gauging the actual preponderance of the evidence across multiple published studies remains virtually impossible. This is a direct result of the fuzziness implicit in aggregating evidence across studies, when different investigators have used evidence measures that are on different scales. This is probably more obvious for linkage analyses, where the incommensurate scales of the various and sundry available statistics are widely recognized. The situation is no better for GWAS, although nearly universal reliance on *p* values obscures the fact that these relate differently to the underlying evidence under different circumstances and different modes of calculation. But as difficult as the situation is within either of these domains, we now have the opportunity to engage in the even more interesting systems biology task of consolidating evidence across multiple disorders and multiple experimental modalities. Without common and well-calibrated measures of evidence, this will be a largely futile endeavor.

Physics faced a formally very similar measurement problem, which presented equally daunting practical and conceptual challenges to scientific progress for a very long period of time. But the development of the absolute temperature measurement scale marked a turning point. This is in part because the significance of absolute temperature lay not merely in its application, but also in its essence. Its derivation required articulation of a deep and utterly new understanding of the very nature of temperature. Without this, physicists had faced fundamental hurdles in understanding the most basic physical phenomena; with it, physics itself took on a whole new cast. This is why the Kelvin scale is inexorably associated with the development of theoretical thermodynamics and eventually statistical and quantum mechanics. Just so, derivation of the absolute evidence scale should not merely provide a calibration technique, important as that will be in

practical applications, but also force us to develop a deep and general theoretical understanding of what evidence *is*. Given how essential the measurement of evidence is to research in human genetics and other branches of biology, isn't it time we too articulated – precisely and with rigor – just what it is we are measuring?

Acknowledgments

This work was supported in part by NIH grant MH086117. The section on evidential Carnot engines represents preliminary work from an ongoing collaboration involving Sang-Cheol Seok, Jayajit Das, and Susan E. Hodge, all of whom have influenced my thinking on this topic. Sue Hodge and Alberto Segre also provided valuable commentaries on earlier drafts of this essay, as did M. Anne Spence, who has been a tireless advocate for this work and whose support continues to propel the project forward.

References

- 1 Siegfried T: Odds are, it's wrong. *Sci News* 2010;177:26.
- 2 Vieland VJ: Thermometers: something for statistical geneticists to think about. *Hum Hered* 2006;61:144–156.
- 3 Hodge SE, Vieland VJ: Expected monotonicity – a desirable property for evidence measures? *Hum Hered* 2010;70:151–166.
- 4 Vieland VJ: The replication requirement. *Nat Genet* 2001;29:244–245.
- 5 Gorroochurn P, Hodge SE, Heiman GA, Durner M, Greenberg DA: Non-replication of association studies: 'pseudo-failures' to replicate? *Genet Med* 2007;9:325–331.
- 6 Edwards A: *Likelihood*. Baltimore, Johns Hopkins University Press, 1992.
- 7 Royall R: *Statistical Evidence: A Likelihood Paradigm*. London, Chapman & Hall, 1997.
- 8 Goodman SN: Toward evidence-based medical statistics. 1. The p value fallacy. *Ann Intern Med* 1999;130:995–1004.
- 9 Goodman SN: Toward evidence-based medical statistics. 2. The Bayes factor. *Ann Intern Med* 1999;130:1005–1013.
- 10 Lehrer J: The truth wears off. *The New Yorker* 2010;13:52.
- 11 Ioannidis JPA: Why most published research findings are false. *PLoS Med* 2005;2:e124.
- 12 Chang H: *Inventing Temperature: Measurement and Scientific Progress*. New York, Oxford University Press, 2004.
- 13 Cox RT: *The Algebra of Probable Inference*. Baltimore, Johns Hopkins University Press, 1961.
- 14 Jaynes ET: *Probability Theory: The Logic of Science*. New York, Cambridge University Press, 2003.
- 15 Shannon CE: A Mathematical theory of communication. *Bell System Tech J* 1948;27:379–423, 623–656.
- 16 Callen HB: *Thermodynamics and an Introduction to Thermostatistics*, ed 2. New York, Wiley, 1985.
- 17 Pippard AB: *The Elements of Classical Thermodynamics*. London, Cambridge University Press, 1960.
- 18 Fermi E: *Thermodynamics*. New York, Dover Publications, 1956.