The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: www.tandfonline.com/journals/utas20

Statistical Principles for Interpreting Clinical Reference Distributions

Berk A. Alpay & Michael M. Desai

To cite this article: Berk A. Alpay & Michael M. Desai (15 Jul 2026): Statistical Principles for Interpreting Clinical Reference Distributions, The American Statistician, DOI: 10.1080/00031305.2026.2691971

To link to this article: https://doi.org/10.1080/00031305.2026.2691971

THE AMERICAN STATISTICIAN
2026, VOL. 00, NO. 0, 1–11: Interdisciplinary
https://doi.org/10.1080/00031305.2026.2691971

Statistical Principles for Interpreting Clinical Reference Distributions

Berk A. Alpay[a] and Michael M. Desai[b,c]

[a]Department of Systems Biology, Harvard Medical School, Boston, MA; [b]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA; [c]Department of Physics, Harvard University, Cambridge, MA

ABSTRACT

Reference distributions quantify the extremeness of clinical test results, typically relative to those of a healthy population. Intervals of these distributions are used in medical decision-making, but while there is much guidance for constructing them, the statistics of interpreting them have been less explored. Here we work directly in terms of the reference distribution, defining it as the likelihood in a posterior calculation of the probability of disease. We thereby identify assumptions of conventional interpretations, criteria for combining tests, and considerations for personalization.

ARTICLE HISTORY
Received February 2025
Accepted June 2026

KEYWORDS
Bayesian regression; Clinical chemistry; Reference ranges

Disease can perturb the abundances of chemical substances, called analytes, measurable by clinical lab technology. However, even when an analyte is physiologically expected to be associated with a disease, it is not immediately clear how to translate lab measurements to clinical decisions: how should these numbers dictate action (McNeil, Keeler, and Adelstein 1975; Murphy and Abbey 1967)? Reference distributions, which represent the variation in test results among healthy people, provide one way to interpret clinical results on the assumption that it is noteworthy if a patient has an extreme enough value relative to this population (Clinical and Laboratory Standards Institute 2008). Test results are commonly displayed alongside reference intervals, thresholds of extremeness with respect to the reference distribution—a familiar sight in blood reports. One very large-scale study of more than half a million people, for example, inferred a reference interval for albumin-adjusted serum calcium of 2.19–2.56 mmol/L (Schini et al. 2021). If used clinically, test results outside this interval would be flagged. The authors speculated that adopting this interval over the previous consensus interval of 2.20–2.60 mmol/L (Berg and Lane 2011) would affect rates of calcium metabolism disorder diagnosis.

Studies of simulated and clinical data over several decades have informed guidelines (Solberg 1987; Clinical and Laboratory Standards Institute 2008) about how to set these reference intervals. Topics of statistical recommendations include inferring intervals from data (Reed, Henry, and Mason 1971; Lott et al. 1992; Horn and Pesce 2003), subgrouping reference data to create more specific intervals (Harris and Boyd 1990; Lahti et al. 2004), and transferring intervals between labs (Estey et al. 2013). With increasing data has come a growing concern to create new references and tune existing ones. Reference intervals are reconsidered (Surks, Goswami, and Daniels 2005; Ceriotti et al. 2008), new ones constructed (Malka et al. 2020), and sources of healthy variation accounted for (Miller et al. 2014; Ozarda 2016), with an enduring goal being references personalized to the patient (Cohen et al. 2021).

Although less abundant than studies on constructing intervals, important work has been done on how to interpret them for diagnosis (Murphy and Abbey 1967; Schoenberg 1970; Sunderman Jr. and Van Soestbergen 1971; Wilson 1973). Because, in clinical practice, results are compared to reference intervals, previous theory has framed questions of interpretation by viewing reference intervals as thresholds of diagnosis. Indeed, a reference distribution is generally seen as an intermediate object whose purpose is to yield a reference interval (Solberg 1987; Wright and Royston 1999; Horn and Pesce 2003; Ichihara, Boyd, and IFCC Committee on Reference Intervals and Decision Limits (C-RIDL) 2010). It has been shown, critically, that the diagnostic predictiveness of intervals depends on the prevalence of the disease (Sunderman Jr. and Van Soestbergen 1971) and the distributions of results among both the healthy and non-healthy population (Murphy and Abbey 1967; Schoenberg 1970; Wilson 1973).

While it is reasonable that thresholds be valued endpoints in clinical application since clinical diagnoses and actions are often binary choices, the underlying level of concern that results should elicit is continuous. The probability of disease is a mathematically more valuable quantity than a yes-or-no prediction. In focusing directly on diagnostic thresholds, current theory largely gives up this intermediate continuousness, which can make it difficult to reason precisely about downstream questions regarding clinical practice.

Here, instead, we formalize the connection between reference values and the probability of disease in terms of population-level distributions of test results. This approach allows us to expressively analyze aspects of reference distributions and their

CONTACT Berk A. Alpay *berk_alpay@g.harvard.edu* Department of Systems Biology, Harvard Medical School, 210 Longwood Avenue, Armenise Building Room 623, Boston, MA 02115.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/TAS*.

resultant intervals. We identify assumptions of current clinical interpretation and reason about how to interpret test results in a personalized manner and with respect to other analytes. Our analysis often complements and expands on prior work. We synthesize previous findings, recapitulate and counter various theoretical arguments, and highlight new considerations for constructing references. For the sake of generality, we make few assumptions and examine a wide range of examples.

1. Reference Distributions Express Likelihoods of Results

A patient's test reports a concentration y of an analyte. How should this result be interpreted?

The test might have been ordered to rule out or settle on some set of diagnoses, according to which additional tests may be performed or treatment adjusted. So, the test result y could be interpreted in terms of the probability of each possible diagnosis d given that result (Ledley and Lusted 1959). Applying Bayes' theorem shows that $p(d|y)$ depends on $p(y|d)p(d)$, a factor that would need to be computed for separate diagnoses. Thus, for each diagnosis considered, a prior probability, or "pre-test suspicion," (Boyd 2010) would be required (Sunderman Jr. 1975), as well as the likelihood of the result given that diagnosis, estimable by measuring the frequency of that result among a sample of patients with the diagnosis.

Grouping all diagnoses together as "non-healthy" and asking only whether the patient is "healthy" (H) or "non-healthy" (H') simplifies the matter. We need only compute

$$p(H|y) = \frac{p(y|H)p(H)}{p(y)}, \tag{1}$$

as illustrated in Figure 1. (An intermediate simplification is to ask whether the patient should be diagnosed with a broad class of disease informed by the test, an event D.) In Bayesian terms, $p(H|y)$ is the posterior, $p(y|H)$ is the likelihood, and $p(H)$ is the prior. Indeed, there is now only one prior to calculate: define criteria for who qualifies as healthy and ask what proportion of the population is healthy. Testing this healthy population will estimate $p(y|H)$. To sample

$$p(y) = p(y|H)p(H) + p(y|H')(1 - p(H)), \tag{2}$$

the overall distribution of results in the population, both healthy and non-healthy people should be tested. Both distributions are important for interpreting the result (Murphy and Abbey 1967; Schoenberg 1970; Wilson 1973).

In practice, statistical interpretation of quantitative lab results is often done with just $p(y|H)$, what is called a (health-associated) reference distribution, representing results among apparently healthy people (Gräsbeck 2004). Much has been written about how to empirically construct reference distributions. The standard procedure is that a sample of results from a healthy population be collected (Clinical and Laboratory Standards Institute 2008). Then different methods, accounting for outliers (Dixon 1953; Reed, Henry, and Mason 1971) and the shape of the distribution, may be used to estimate quantiles of $p(y|H)$ (Horn and Pesce 2003). Parametric methods assume a form of the distribution while nonparametric methods are interested directly in the sample quantiles. To calculate the commonly used 95% interval, sampling on the order of 200 reference values is recommended (Reed, Henry, and Mason 1971; Lott et al. 1992; Miller et al. 2016).

Studies can apply different criteria to decide who is healthy, the definition of which can be controversial (Callahan 1973). Some authors use a definition of health that seems aligned with computing $p(y|D)$ (Schneider 1960), for example, that "A person may be suitable as a healthy reference individual for one test …but unsuited for another" (Gräsbeck 2004), while others use health criteria that are uniform across tests (Horn and Pesce 2003), which aligns with computing $p(y|H)$. We will use the H notation for simplicity, but in practice the particular criteria should be specified (Clinical and Laboratory Standards Institute 2008).

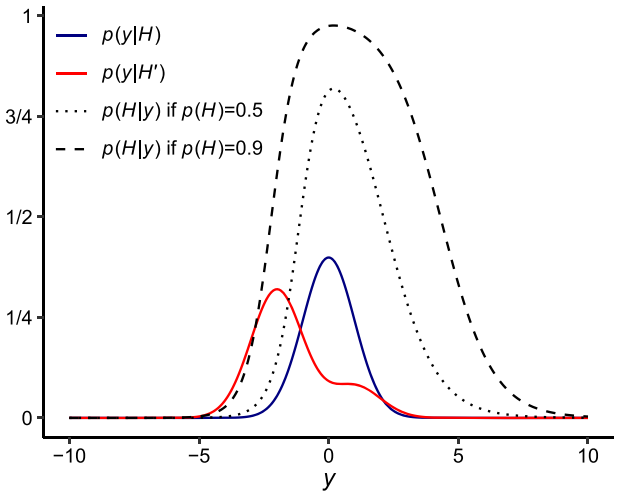

Figure 1. The posterior probability of health is determined by the prior probability and the distribution of results among the healthy and non-healthy. In this example, healthy results are normally distributed about zero, while non-healthy results are a mixture of two normal distributions on either side of the healthy mode. Due to the shape of the non-healthy distribution, results to the left of the healthy mode become concerning more quickly than to the right.

2. Using Reference Distributions Requires Assumptions on Non-healthy Results

The theoretical simplifications above result in an object, the reference distribution, that is relatively easy to estimate but does not yield all the probabilities of different diagnoses (Boyd 2010). An assumption that underpins current clinical use is that the extremeness of a result with respect to the reference distribution corresponds to how concerning it is. This assumption is true, for example, if $p(y|H')$ is more diffuse than $p(y|H)$, as in both limits of Figure 1. But the non-healthy distribution flexibly affects the implications of a patient's result. If, to take an exceptional example, non-healthy results were for some reason concentrated at a high-probability interval of $y|H$, then extreme results could unintuitively be reassuring (Fig. S1A). And using (1) and (2), we can also see that extremes of an analyte unassociated with health (i.e., when $y|H$ and $y|H'$ are identically distributed) would not be informative with respect to disease no matter how extreme the result, though we should not expect this to ever be exactly the case especially since analytes are often chosen based on physiological connection with disease.

It is also often implicitly assumed that extremes are only at the left and/or right limit of y. If true, the extremeness of the result can be quantified using the cumulative distribution function $F(y|H)$, with high $F(y|H)$, for example, corresponding to high extremes. Depending on the analyte, a low result, a high result, or both (as in Figure 1) could be concerning. If both are concerning, one might quantify the extremeness of the result as, say, $\min(F(y|H), 1 - F(y|H))$. Consider, however, a case in which there are two modes of healthy results, and that non-healthy results also exhibit these two modes but more diffusely. (Perhaps these modes are explained by certain covariates, the sort of situation we will analyze later.) Then concerning extremes are not exclusively at the left and right of the graph: they can also lie between the two modes (Figure S1B). Directly applying a nonparametric method, using the rank order of reference values rather than their parametric form, to construct a reference interval would in this case obscure such nuances. It is sometimes held that nonparametric methods make no "assumptions as to the specific form of the underlying distribution of the data" (Horn and Pesce 2003) and that "they can be applied to any set of data, regardless of how the parent population of values is distributed" (Schneider 1960). As these cases show, however, there are assumptions in considering only the tails of the reference distribution as regions of concerning extremes.

Even if it is known where the concerning extremes lie with respect to the reference distribution, the risk of disease can vary continuously with the extremeness of the test result; there are often no clean boundaries on the result that dictate the occurrence of disease. But clinical decisions are often discrete: an action is taken or not. And so thresholds of extremeness, based on the available information, are set between courses of action. A 95% interval is typically used to define thresholds for concern (Bold 1976; Ceriotti, Hinzmann, and Panteghini 2009; Boyd 2010), as it often is for other statistical analyses and illustrations, including here. It is a convenient but ultimately arbitrary choice (Bold 1976; Albert 1981; Vickers and Lilja 2009; Boyd 2010)—why not use, say, 89% as a default for statistical analyses (McElreath 2018, chap. 4)? Ideally, rather, the probability of health, combined with the potential costs and benefits of actions should dictate the optimal threshold (Murphy and Abbey 1967; Wilson 1973; Sunderman Jr. 1975; Pauker and Kassirer 1975; Vickers and Lilja 2009; Miller et al. 2016). As Figure 1 illustrates, the former depends on $p(y|H')$, as do the sensitivity and specificity of a threshold in detecting disease (Wilson 1973). And indeed, a threshold on $\min(F(y|H), 1 - F(y|H))$ may have different diagnostic ability according to the tail at which the extreme result lies.

Statistical decision theory (DeGroot 2005) can be used to understand how to act on a test result. Since reference distributions are typically used to interpret safe, noninvasive, and relatively inexpensive tests performed at the discretion of the physician, it is reasonable to exclude the cost of the test from the decision-making process, as opposed to, say, biopsies (Springfield and Rosenberg 1996; Burnside, Chhatwal, and Alagoz 2012), which might call for additional consideration of what it will reveal about the proper course of treatment in light of risk to the patient (Gelman et al. 2013, chap. 9). Let us focus instead on whether to flag the result. Given the expected costs of flagging it incorrectly, L_{H} for a false positive and $L_{\mathrm{H}'}$ for a false negative, the result should be flagged if $L_{\mathrm{H}'}p(H'|y) > L_{\mathrm{H}}p(H|y)$, or equivalently if $p(H|y)/p(H'|y) < L_{\mathrm{H}'}/L_{\mathrm{H}}$ (Parmigiani 2002, chap. 2), when the posterior odds of health fail to exceed the cost of a false negative relative to a false positive. The Neyman-Pearson Lemma, without setting explicit costs of flagging the result, similarly provides that the most powerful α-level test rejects H for H' if $p(y|H)/p(y|H') < c$ for a constant c. This again requires knowledge of the non-healthy distribution.

To decide on an action, such as to assign a medical treatment, the potential outcomes of each treatment should be taken into account. Suppose a utility function $U(z)$ assigns a numerical desirability to a given clinical outcome z. A treatment may have different outcomes; the (random) utility given a treatment t can be called $U(z)|t$. Presumably the outcome given the treatment also depends on the patient's state of health. This state is unknown, so to calculate the expected utility of a treatment given the test result, one can take y as an intermediary of the state of health, and compute the weighted sum of the utility across these states to yield the expected utility of the treatment:

$$\begin{aligned} E(U(z)|t,y) = {} & E(U(z)|t,H)p(H|y) \\ & + E(U(z)|t,H')(1 - p(H|y)). \end{aligned} \tag{3}$$

Note that again these calculations require the posterior probability of health, beyond simply the reference distribution. Also note that additional data x that may be useful in calculating the posterior probability of health can in principle be incorporated into these calculations to make decisions more precise, so that wherever one conditions on y, one can also condition on x. In general this is done via regression, which we will discuss later with respect to inferring better-personalized reference distributions.

3. Jointly Interpreting Results Depends on their Relationship with Each Other (and with Health)

Bayes' theorem also allows us to interpret the levels of several analytes, say y_1 and y_2, together with respect to health:

$$\begin{aligned} p(H|y_1,y_2) &= \frac{p(y_1,y_2|H)p(H)}{p(y_1,y_2)} \\ &= \frac{p(y_1|y_2,H)p(y_2|H)p(H)}{p(y_1|y_2)p(y_2)}. \end{aligned} \tag{4}$$

Conditioning on results of multiple tests in principle enables more precise estimation of the probability of health, but when using solely reference distributions we would again miss the non-healthy variation to exactly compute this probability. Thus, it is not always advantageous to use the joint reference distribution of y_1 and y_2. We illustrate a variety of circumstances in Figures 2 and S2.

The degrees to which y_1 and y_2 are each associated with health influence whether to jointly consider them. For the sake of intuition, take the extreme case that the healthy and non-healthy distributions of y_2 are identical, that is, that y_2 is independent of health. This fact would be taken into account when computing the posterior, but not in the joint reference distribution $p(y_1,y_2|H)$. In Figure 2(A) and (C), for example, extreme y_2 is not in itself concerning, but the degree to which a particular y_1 is extreme can be inappropriately altered by the inclusion of y_2

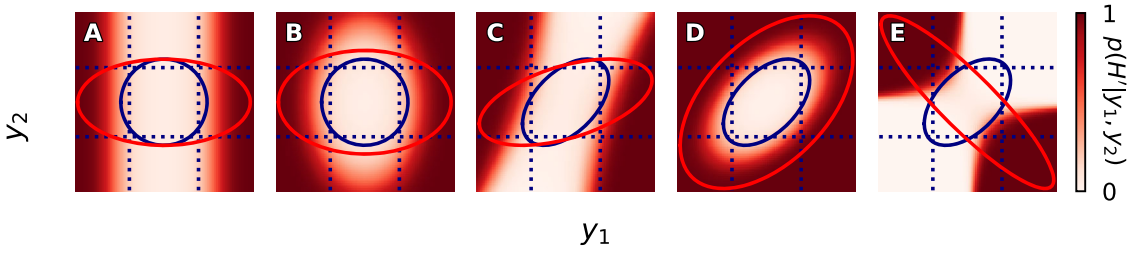

Figure 2. Examples of how the joint distributions of two analytes y_1 and y_2 are related to the posterior probability of health. Shown are 95% joint highest density regions of the healthy (blue ellipses) and non-healthy (red ellipses) distributions, as well as the 95% marginal highest density intervals of the healthy distributions (dotted lines). Blue ellipses are thus joint reference regions while dotted lines mark univariate reference ranges. The prior probability of health is $p(H) = 0.9$. In (A–B), y_1 and y_2 are independent, and while y_1 is here always associated with health, y_2 is unassociated with health in (A) and (C).

(Healy 1973). It can thus be misleading to combine the reference distribution of y_1 with y_2, an effect alleviated as the association of y_2 with health increases (Figure 2(B), (D)).

Whether it is advantageous to consider the joint distribution also depends on the relationship between y_1 and y_2. By the Neyman-Pearson lemma, one should reject the hypothesis of health according to the ratio

$$\frac{p(y_1, y_2|H)}{p(y_1, y_2|H')} = \frac{p(y_1|y_2, H)p(y_2|H)}{p(y_1|y_2, H')p(y_2|H')}. \quad (5)$$

If y_2 is independent of health, that is, $p(y_2|H) = p(y_2|H')$, the ratio still depends on y_2. It is when, also, $p(y_1|y_2, H) = p(y_1|H)$ and $p(y_1|y_2, H') = p(y_1|H')$, that is, when the two analytes are also independent of each other under both states, that the dependence on y_2 is lost and the ratio simplifies to $p(y_1|H)/p(y_1|H')$, the univariate ratio. Meanwhile, if y_1 and y_2 are independent given health (Figure 2(A)–(B)), the joint reference distribution simply separates into $p(y_1|H)p(y_2|H)$, the product of the individual reference densities.

The advantage of using joint reference distributions can be seen when y_1 and y_2 similarly covary among healthy and non-healthy people and are both associated with health; in Figure 2(D), the values of the posterior are diagonally symmetric and the reference distribution captures its shape. If only healthy or only non-healthy y_1 and y_2 covary, the posterior can follow a diagonal, sometimes behaving in ways unpredicted by the shape of the reference distribution (Figure S2). The general advice that tests that are combined be "systematically" (Grams, Johnson, and Benson 1972) or "physiologically" (Boyd and Lacher 1982) related is sound: it implies they covary. But these are not sufficient conditions because if they do not covary in a similar way between healthy and non-healthy populations, the reference distribution could be a poor proxy of the shape of the posterior distribution (Figure 2(E)).

The following three blood measurements constitute a relevant example. The product of mean corpuscular hemoglobin (MCH) and red blood cell (RBC) count determines the hemoglobin concentration (HGB), a common measure of anemia as it quantifies how much oxygen a volume of blood can deliver. Based on this logic, MCH and RBC would better diagnose anemia together than apart, but is it better to use the reference distribution of HGB (a function of the MCH and RBC) or the joint distribution of MCH and RBC? The latter was observed in one study to be better diagnostic of mortality risk (Malka et al. 2020), suggesting it can be important to know whether MCH and RBC are jointly extreme even if HGB is not.

As others have noted, jointly considering analytes affects how one interprets the extremeness of a set of results. For example, a set of results that have no extremes apart may be extreme together (Boyd and Lacher 1982), and results extreme apart may not be extreme together (Winkel and Lyngbye 1972). Proponents of joint reference regions have argued that results should be combined because they make for more specific interpretation (Grams, Johnson, and Benson 1972). Other proponents have argued that they can control for false positives that arise from multiple testing (Winkel and Lyngbye 1972; Harris et al. 1982; Boyd 2004), while it has also been regretted that joint reference regions can bury the extremeness of a single test (Boyd and Lacher 1982). But as we have shown, the diagnostic impact of jointly considering analytes depends on their relationship with each other and with health. Joint reference distributions can be useful in some cases and misleading in others.

4. Conditioning on Features Personalizes Reference Distributions

What if y depends on some random variable, a feature x unassociated with health, that is, for which $x|H$ and $x|H'$ are identically distributed? An implication of our analysis above is that jointly considering two covarying analytes where y is associated with health and x is not (as in Figure 2(C), taking y to be y_1 and x to be y_2), risks a misleading interpretation of the extremeness of x with respect to health. The joint reference distribution is $p(y, x|H) = p(y|x, H)p(x|H)$. But from (4) we see that $p(x|H) = p(x|H')$ implies

$$p(H|y, x) = \frac{p(y|x, H)p(x|H)p(H)}{p(y|x)p(x)} = \frac{p(y|x, H)p(H)}{p(y|x)}, \quad (6)$$

canceling $p(x|H)$ with $p(x)$ and making it irrelevant in the calculation of the posterior probability of health. The reference distribution that remains is $p(y|x, H)$, which we can call the reference distribution of y specified to x. This logic provides the theoretical justification for when the reference distribution of x can be disregarded. There may be other reasons to do so, including to avoid risks in interpreting joint distributions that we highlighted earlier, or simply to keep interpretation focused on y. It may be useful, for example, to measure the extremeness of a result adjusted to the patient's age (Oesterling et al. 1993; Nichols et al. 2007) while not compounding this quantity with age itself as an indicator of concern.

Even if the reference distribution of x can be disregarded, this x should not necessarily be disregarded with respect to

the reference distribution of y. Suppose such a feature x exists: a feature of patients or their test conditions (e.g., the testing instrument or the time of day (Fraser 2001)) not associated with health, or whose prevalence among healthy people one would like to disregard, but which is associated with the result y. Then while a random sample of the healthy population would exhibit the healthy variation in results in general (Schneider 1960), $p(y|H)$, it might not estimate well the healthy variation $p(y|x, H)$ one should expect given the feature (Boyd 2010). Suppose the patient is an athlete and exercises far more than most people sampled. The greater the effect of exercise on healthy results, the more likely is a false positive for the patient when using a reference interval derived from the general healthy population. Creatine kinase, for example, is elevated under both muscular dystrophy and following intense physical exercise (Baird et al. 2012).

Note that (6) shows that the interpretation of the reference distribution with respect to the probability of health may change depending on the features conditioned on. This would occur if, as features are conditioned on, the distribution of results among non-healthy people does not mirror changes in the reference distribution. Thus, two results of the same analyte, under different feature values, which each fall outside their specified 95% reference interval may not necessitate the same level of concern (Ceriotti and Henny 2008). (One consequence of this fact is that the argument that intra-person reference intervals, derived only from previous results from the patient, are preferable to inter-person ones (Harris 1974) may not be so straightforward: how should thresholds be calibrated for each patient?)

With this caveat in mind, to adjust for a dichotomy in test results, perhaps if certain statistical criteria (Harris and Boyd 1990; Lahti et al. 2004; Ichihara, Boyd, and IFCC Committee on Reference Intervals and Decision Limits (C-RIDL) 2010) indicate to do so, we might opt to consider only samples from people who share the differentiating feature, such as fellow athletes. This process is called subgrouping and it is the current practice for making reference intervals specific to certain features (Clinical and Laboratory Standards Institute 2008). In reality, however, there are many features about the patient and the conditions under which the test was performed (Schneider 1960; Fraser 2001). Progressively conditioning on each such x further limits the people who can be sampled and the test conditions under which they can be sampled.

Rather, a single sample of a general population is simpler to collect and more widely applicable. In such a sample, features will only sometimes match the patient's; to subset the data by even two levels of specificity—say, to athletes over 25 years old—greatly reduces the number of samples with which to construct a more specific reference distribution. Framed in this way, it seems we must be satisfied with reference distributions in which we are confident, being derived from a large reference sample, but which are perhaps not as specific as we would like.

5. Reference Distributions Can be Seen as Posterior Predictives of Regression

Let us consider the problem of specifying reference distributions more formally. We have thus far examined how to interpret reference distributions when they are exactly known. We will adjust our notation to speak of inferring them from data. Let X be a matrix, the ith row of which comprises the ith reference individual's values of the k features $x_1, \ldots, x_k$, and let y be the vector of reference values, the ith element of which is the test result of the ith reference individual. We can call X and y together the reference data. Now let $\tilde{x}$ represent the features of the patient's test (which was not part of the reference sample). An $\tilde{x}$ of the patient consisting of $k = 3$ features could be $(1, 1, 0)$, maybe meaning: over 25 years old, athlete, not ambidextrous.

The reference distribution specified to the patient's test is then the distribution of the result conditional on its features and the reference data: in probabilistic terms, $p(\tilde{y}|\tilde{x}, X, y, H)$. To infer a random variable's relationship with features is to perform regression, and the distribution of new data based on observed data is called the posterior predictive. Thus, a reference distribution can be interpreted as the posterior predictive distribution of regression.

To compute this distribution, it is necessary to specify a model that sets the form of the relationship between X_i and y_i. A possible regression model is one that disregards the features and explains the result as simply a mean value β_0 plus normal random error ϵ:

$$y_i|H = \beta_0 + \epsilon, \tag{7}$$

where $\epsilon \sim \text{Normal}(0, \sigma)$. Reference distributions are usually constructed assuming this model, computing only $p(\tilde{y}|y, H)$. Uncertainty in the parameters is ignored if $p(\tilde{y}|\beta_0 = \hat{\beta}_0, \sigma = \hat{\sigma}, H)$ is computed, where $\hat{\beta}_0$ and $\hat{\sigma}$ represent estimates of the parameters, for example, when the reference distribution is assumed to be normal around the sample mean with variance equal to the sample variance (Horn and Pesce 2003).

Subgrouping to make reference distributions more specific would seem again to necessitate too much data to be feasible beyond a few levels of conditioning. To calculate separate reference distributions for all combinations of k binary subgroups, it seems 2^k of these models must be created, β_0 and σ inferred separately for each possible X_i. One would need to separately sample test results arising from 2^k combinations of features.

Using regression to construct reference distributions (Virtanen et al. 1998; Suominen et al. 2001; Holmes et al. 2021), however, can be considerably more efficient than subgrouping. With subgrouping, the statistical effect of a feature is inferred anew for each combination of the remaining features; the effect of being an athlete would be inferred independently for younger and older people. (We use "effects" to refer to regression coefficients; features need not be causal.) But instead of modeling subgroups separately, one can treat them as features which contribute, independently at least to some degree, to the result (Ichihara, Boyd, and IFCC Committee on Reference Intervals and Decision Limits (C-RIDL) 2010). For example, in the form of a linear model in which

$$y_i|H = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon. \tag{8}$$

This kind of regression infers the effect of being an athlete from both younger and older reference individuals, potentially extracting more value from the data (Holmes et al. 2021).

There are risks in adding features to a model. First, reference distributions can always be made more specific, and at some level

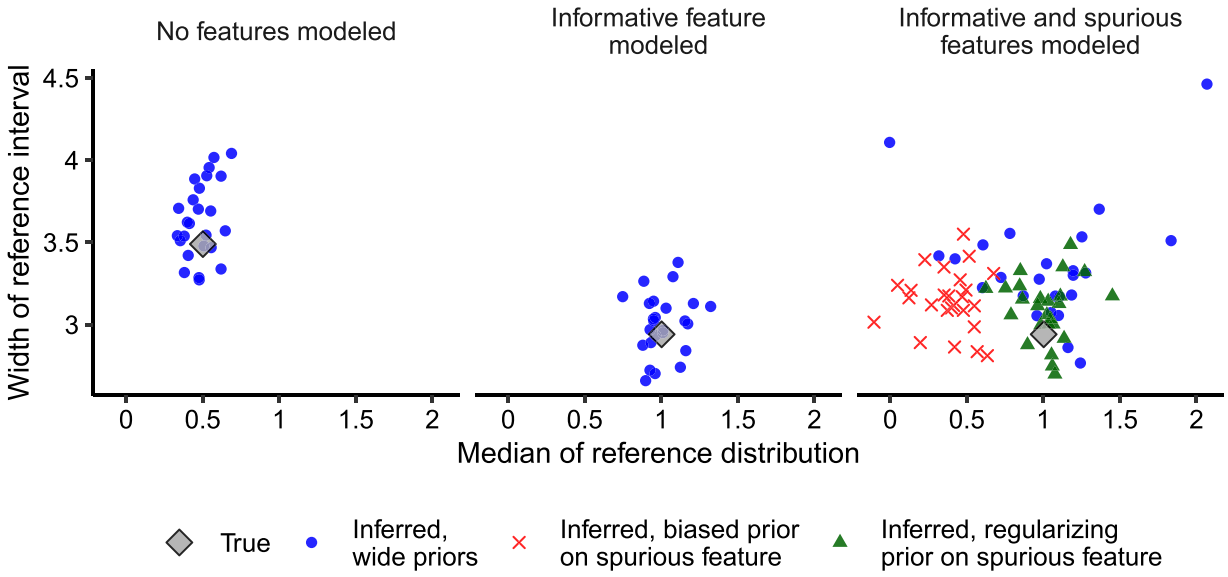

Figure 3. Inferring the 95% reference interval as the presence of two binary features are progressively accounted for. Each point represents one reference interval, inferred from a sample of 100 simulated observations. The informative feature x_1 occurs among 50% of people and its presence is associated with a unit increase in the result. The other, spurious feature x_2 occurs among 5% of people and has no association with the result. The model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ and the true coefficients are $\beta_0 = \beta_2 = 0$ and $\beta_1 = 1$ with $\epsilon \sim \text{Normal}(0, \frac{3}{4})$. Different priors over β_2 are considered: Normal(0, 10) (wide prior), Normal$(-1, \frac{1}{4})$ (biased prior), and Normal$(0, \frac{1}{4})$ (regularizing prior).

of specificity the data will be too limited to precisely infer effects. Uncertainty in the model parameters θ increases the variance of the posterior predictive distribution, which integrates over uncertainty in the parameters:

$$p(\tilde{y}|\tilde{x}, X, y, H) = \int p(\tilde{y}|\tilde{x}, \theta, H) p(\theta|X, y, H) d\theta. \tag{9}$$

This integral essentially computes an aggregate reference distribution that combines the possible densities of the reference distribution under different parameter settings, weighted by the posterior probabilities of those settings. The equation reflects, for example, that if we are uncertain that the effect of a feature is as negative as it seems to be, then we should not be as surprised to see a large result than if we were certain in the effect. Well-considered, informative priors (Van Dongen 2006) allow features to be incorporated into the model without unnecessarily increasing uncertainty in the reference distribution or overfitting the reference data, leading to more accurate reference intervals (Figure 3). Setting priors such that in total they agree with the expected prior predictive variance (Gelman, Hill, and Vehtari 2020, chap. 12), for example, could be useful; we should not expect variance in results to be arbitrarily large.

A second risk in incorporating additional features is that models are always wrong to some degree, passing inaccuracies downstream to the posterior predictive distribution. For example, we thus far assumed homoscedasticity, that is, that the variance of the error is constant across values of the feature. Suppose instead that error varies with some binary feature x_j, lower when it is present and higher when absent. Fitting a homoscedastic model to this data, we may find that the error when $x_j = 1$ is overestimated and that it is underestimated when $x_j = 0$, the posterior predictives being too wide and too narrow, respectively. The functional form of the feature effects may also be incorrect. This can occur, for example, when effects of two binary features are modeled additively while in fact their combined effect is distinct. This misspecification skews coefficient estimates of the features that are accounted for (Figure S3).

Since a reference distribution is effectively a posterior predictive distribution, the regression model used to infer it can most directly be assessed via predictive checks (Box 1980; Gelman et al. 2013; Gelman, Hill, and Vehtari 2020), which evaluate the predicted distribution of new data $\tilde{y}$ under the model. For one, prior predictive checks can be used to evaluate whether the model is a priori reasonable. The prior predictive is the reference distribution before any data is observed, computed as in (9) but without conditioning on reference data, that is, $p(\tilde{y}|\tilde{x}, X, H) = \int p(\tilde{y}|\tilde{x}, \theta, H) p(\theta|H) d\theta$. A simple prior predictive check is to visualize prior reference distributions under different settings of the covariates and to assess whether they are reasonably consistent with one's knowledge of the analyte. Such checks are especially important when working with more complicated models or when data is limited and the posterior is liable to be dominated by the prior (Gelman et al. 2013, chap. 2). Following inference of parameters from actual data, posterior predictive checks can be used to assess the model by comparing statistics of the inferred reference distributions to distributions of empirical test results; greater deviation between these two distributions indicates greater misspecification of the model (Blei 2014). Models at risk of overfitting would benefit from cross-validation, that is, iteratively holding out portions of the data, inferring the reference distribution using the remainder, and testing how well the inferred reference distribution matches the holdout data. Ideally, a model should be iteratively adjusted and criticized until it passes such checks (Blei 2014).

6. An Example: Serum Albumin

We will use measurements of serum albumin from the National Health and Nutrition Examination Survey (NHANES) of 2017

to March 2020 (Akinbami et al. 2022) to illustrate principles we have discussed, about the difference between binary and probabilistic interpretations of results, assumptions on the form of the reference data, the behavior of the posterior probability of health, incorporating additional variables (namely body mass index), and evaluating predictiveness.

We restricted the original dataset of about 15,000 participants to those with complete measurements of serum albumin and body mass index, and questionnaire responses we use to determine health status, leaving 7757 participants. For the purposes of our analysis, we define a healthy participant as one who reported no history of chronic illness, specifically arthritis, heart disease, thyroid problems, lung disease, liver conditions, and cancer or malignancy as inquired by the questionnaire. Under these criteria, 53% of the participants were healthy.

Serum albumin, measured in NHANES at a resolution of tenths of a gram per deciliter, is a protein that plays a role in a variety of physiological processes including molecular transport and regulation of blood pressure (Belinskaia, Voronina, and Goncharov 2021). Low levels of serum albumin are associated with various diseases (Phillips, Shaper, and Whincup 1989), including of the liver (Spinella, Sawhney, and Jalan 2016), kidney (Haller 2006), and heart (Arques 2018). It is often used in intensive care as an indicator of the severity of illness (Goldwasser and Feldman 1997; Erstad 2021).

Although serum albumin is strongly associated with mortality in intensive care (Kendall, Abreu, and Cheng 2019; Cao et al. 2023), in this survey data there is considerable overlap in the healthy and non-healthy distributions (Figure 4(A)). The empirical 95% reference range is 3.5–4.7 g/dL; for comparison, a common reference book states a range of 3.5–5 g/dL (Pagana, Pagana, and Pagana 2015). As the upper range of results does not appear to be concerning, consider using the lower bound of this range, 3.5 g/dL, to predict the participant is non-healthy with a result less than this value and healthy otherwise. The non-healthy data allows us to calculate such statistics of classification as the false positive rate, the fraction of true negatives predicted wrongly as positive. On the whole NHANES data, the false positive rate using this threshold is just 3% but the true positive rate, that is, the fraction of true positives predicted as such, is only 5%. One can adjust the threshold to yield different rates, but due to the limited separation of the healthy and non-healthy distributions of serum albumin, one faces a steep tradeoff; to reach a true positive rate of 55% would yield a false positive rate of 43% (Figure S4A). In terms of classical hypothesis testing (Lehmann and Romano 2005), the threshold essentially determines the decision rule used to reject the null hypothesis of health. The choice of threshold can therefore be seen as a tradeoff between statistical power to detect disease and Type I error of incorrectly rejecting the healthy hypothesis.

Though it is difficult to predict whether a participant is healthy, the probability of health can be calculated fairly precisely using a simple model. Let us first assume both the healthy and non-healthy values of serum albumin are normally distributed. The sample standard deviation of the non-healthy distribution is in this case slightly less than that of the healthy distribution, and so under this assumption an extreme enough serum albumin result will not be of concern, as in Figure S1A. There is in principle, however, always a high enough concentration of an analyte that would be cause for concern, nor is severely low albumin auspicious in light of its various physiological roles; there may, in reality, and in contradiction to the normality assumption, be those with rare diseases whose results would populate other non-healthy modes and thereby greatly decrease the posterior probability of health at the tails. A simple remediation of the normality assumption is to restrict the range of analysis to the range of observed test results, as we do here, and acknowledge ignorance outside it.

Another consideration in calculating the probability of health is whether to integrate over uncertainty in the mean and

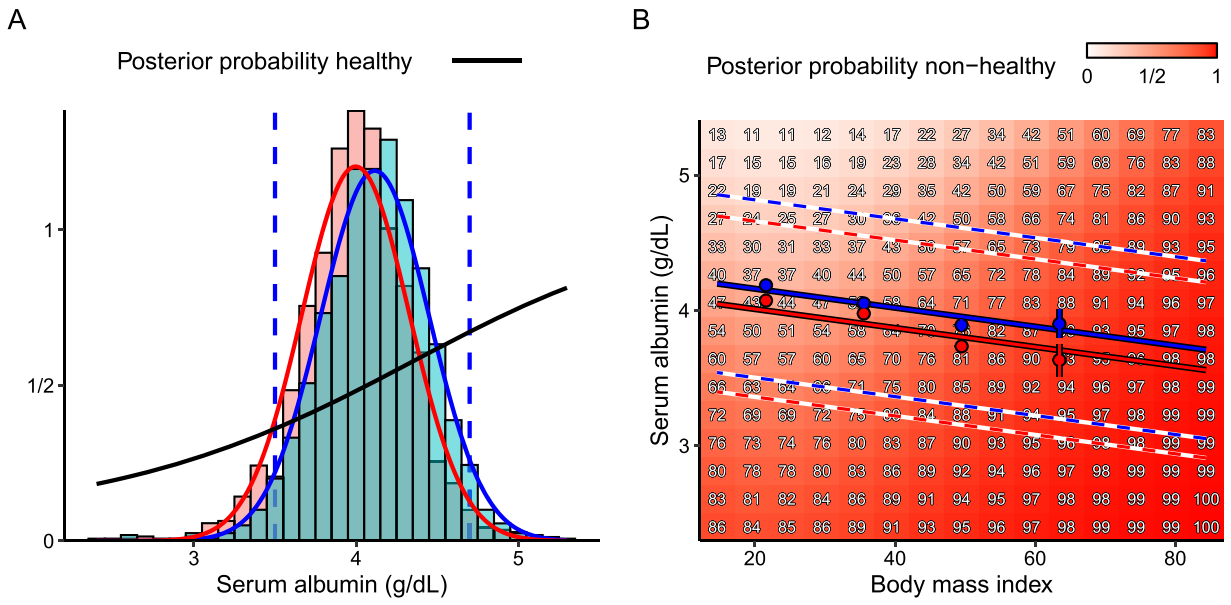

Figure 4. (A) The posterior probability of health with respect to serum albumin among a sample of three thousand NHANES 2017–March 2020 participants labeled using questionnaire data as healthy, blue, or unhealthy, red, with respect to chronic illness. Included are density histograms and normal approximations of serum albumin levels. The empirical 95% reference range is 3.5 to 4.7 g/dL as shown. The prior probability of health is set to 0.5. (B) The posterior probability of health with respect also to body mass index. Overlaid are the mean and standard error of serum albumin in bins of body mass index, and the modeled central 95% intervals of serum albumin results.

variance of each distribution, for example assuming $y|H \sim \text{Normal}(\mu, \sigma^2)$ with a noninformative $p(\mu, \sigma^2) \propto \sigma^{-2}$ prior. The posterior predictive distribution $p(\tilde{y}|y, H)$ would then be t-distributed with $n-1$ degrees of freedom, location equal to the sample mean, and scale equal to the sample standard deviation times $\sqrt{1+1/n}$ (Gelman et al. 2013, chap. 3). As n increases, the scale approaches the standard deviation and the distribution approaches normality. Since there are on the order of thousands of participants, let us simply assume that the posterior predictive distribution is well-approximated as normal with mean equal to the sample mean and variance equal to the sample variation. The corresponding posterior probability of health, calculated using (1), is shown in Figure 4(A).

Let us now consider an additional variable of interest. Mean serum albumin decreases fairly linearly with respect to body mass index (Figure 4(B)). Assuming for simplicity that the rate of decrease is identical between both groups and that variance in serum albumin is constant across body mass index within each group, we can say, roughly, that $y|H \sim \text{Normal}(4.3-0.007x, s_H^2)$ and $y|H' \sim \text{Normal}(4.15-0.007x, s_{H'}^2)$ where x is the body mass index of the participant and s_H^2 and $s_{H'}^2$ are the sample healthy and non-healthy variations in serum albumin. Assuming body mass index is gamma-distributed (Figure S4B), we can then use (4) to compute the posterior probability of health given both serum albumin and body mass index (Figure 4(B)).

The predictions of the two models are similar; the rank correlation of their predictions on holdout data is 0.92. The greatest change in posterior probability of health moving from the albumin-only model to the combined model is for those with relatively low and high body mass index (Figure S4C). While it is still inherently difficult to confidently predict whether a participant is healthy using these variables, the model is reasonably well-calibrated in that a fairly precise calculation of the probability of disease can be made (Figure S4D). A more detailed model might handle the tails of the distributions more precisely to improve the probability calibration, and more variables could be incorporated to understand their implications with respect to health.

7. Discussion

How then should a patient's test result be interpreted? Ideally, based on the results and prevalence of people with and without the diagnoses of interest, one could weigh the probability of the diagnoses. It is less demanding, in terms of data collection and modeling, to just consider the likelihood that the result arose under healthy conditions, assuming the more extreme the result, the more concerning it is. A reference distribution, estimated from reference data of a sample of healthy people, provides these likelihoods. If other health-associated analytes covary with the one measured, it may, depending on the nature of the covariance, be useful to consider their joint reference distribution. Finally, it is possible that the effects of features recorded in the reference data could be inferred using a carefully specified regression model and be used to generate a reference distribution more specific to the patient.

In contrast to reference intervals, decision limits use clinical outcomes to define thresholds, thus, taking both the healthy and non-healthy into account. Decision limits are now used to interpret some tests (Prati et al. 2002; Ozarda et al. 2018), but data to construct them is limited (Miller et al. 2016) and not many of them have been set (Ceriotti and Henny 2008). It is generally agreed that they are preferable to their reference interval counterparts (Ceriotti and Henny 2008; Clinical and Laboratory Standards Institute 2008; Sikaris 2012; Miller et al. 2016). Our theoretical reasoning supports this conclusion; when the underlying healthy and non-healthy distributions and prior are exactly known, computation of the posterior probability of health (or a particular diagnosis) as in (1) is more precise and avoids the interpretive risks of relying solely on the reference distribution. Likelihood ratios (Van der Helm and Hische 1979), for example, $p(y|d)/p(y|d')$, capture one component of (1) and are useful as a test via the Neyman-Pearson lemma (Neyman and Pearson 1933) to fix the probability of a false positive to a desired level while maximizing the power to detect true positives. However, likelihoods should be adjusted by the prior, for example, $p(d)$ (Boyd 2010), when computing the posterior probability of disease, and as we have shown, it is useful to continue with full Bayesian calculation as far as possible before a decision threshold must be made.

We simulated univariate regression with respect to binary features to illustrate general considerations of using regression to specify reference distributions. With additional model assumptions, joint reference distributions could be derived by multivariate regression (Gelman et al. 2013, chap. 16). There are opportunities and risks in incorporating continuous features (Royston 1991), which we did in our analysis of serum albumin. Assuming a functional form of the effect of a feature risks increasing model misspecification. But if the function is correct, this approach has advantages over the traditional practice (Nichols et al. 2007) of binning continuous features and modeling results independently within each bin. For one, intervals could be specified to the exact feature value, for example, age rather than age bin. There can be additional complications in regression to the ones we showed, such as covariance between modeled features, which increases uncertainty about their effects (Gelman, Hill, and Vehtari 2020, chap. 10) and thereby widens posterior predictives. We repeatedly assumed results are normally distributed for sake of demonstration, but this should not hastily be assumed in practice (Elveback, Guillier, and Keating 1970).

The Bayesian formulation of reference distributions could be applied to additional problems in constructing them. One such problem is reconciling differences between different laboratories (Panteghini 2012; Higgins, Nieuwesteeg, and Adeli 2020). For example, multi-level modeling—which splits samples into groups, between which parameter estimates are partially pooled (Gelman and Hill 2006)—could be used to capture how effects vary across labs. Current practice in transference of reference intervals between laboratories, important for reducing costs, is to draw a small sample to validate that an existing interval applies to the new lab (Clinical and Laboratory Standards Institute 2008). Although intervals can align well with one another (Estey et al. 2013), there is a dead end in the procedure if they do not. Multi-level modeling would allow data from other laboratories to be used to reduce the sample burden even if intervals are not strictly consistent.

A theme of our work is the importance of the marginal likelihood, $p(y)$, or in the case of regression, $p(y|x)$, in interpreting reference distributions. If it is not available, reference distributions can be used provided assumptions on the missing factors, though this limits the statistical rigor with which decision thresholds can be set. Regression can be applied to take more and more features into account, gradually personalizing interpretation of results from limited data. Tracking technical and physiological details of tests (Sunderman Jr. 1975), pooling data (Manrai, Patel, and Ioannidis 2018), and iterating through careful modeling and criticism (Blei 2014) will be useful to realize clinical lab diagnostics of growing precision.

8. Methods

Linear regression was performed using rstanarm (Goodrich et al. 2020), built on the Stan programming language (Carpenter et al. 2017). We set a Normal(0, 10) prior over coefficients unless otherwise stated, and an Exponential(1) prior over the standard deviation of the error. Code can be accessed at *https://github.com/berkalpay/reference-distributions*.

Supplementary Materials

The supplemental material illustrates additional examples of the posterior probability of health, inferences of a linear model that neglects an interaction term, and further analysis of serum albumin.

Acknowledgments

We thank John Higgins, Rachel Petherbridge, Thomas Dupic, Wendy Valencia-Montoya, and Derek Aguiar for useful discussions and comments.

Disclosure Statement

The authors declare no conflicts of interest.

Funding

B.A.A. was supported by an NSF Graduate Research Fellowship under Grant No. DGE-2140743.

ORCID

Berk A. Alpay https://orcid.org/0000-0002-1022-2163
Michael M. Desai https://orcid.org/0000-0002-9581-1150

References

Akinbami, L. J., Chen, T.-C., Davy, O., Ogden, C. L., Fink, S., Clark, J., Riddles, M. K., and Mohadjer, L. K. (2022), "National Health and Nutrition Examination Survey, 2017–March 2020 Prepandemic File: Sample Design, Estimation, and Analytic Guidelines," Technical Report, National Center for Health Statistics. [7]

Albert, A. (1981), "Atypicality Indices as Reference Values for Laboratory," *American Journal of Clinical Pathology*, 76, 421–425. [3]

Arques, S. (2018), 'Human Serum Albumin in Cardiovascular Diseases," *European Journal of Internal Medicine*, 52, 8–12. [7]

Baird, M. F., Graham, S. M., Baker, J. S., and Bickerstaff, G. F. (2012), "Creatine-Kinase-and Exercise-Related Muscle Damage Implications for Muscle Performance and Recovery," *Journal of Nutrition and Metabolism*, 2012, 960363. [5]

Belinskaia, D., Voronina, P., and Goncharov, N. (2021), "Integrative Role of Albumin: Evolutionary, Biochemical and Pathophysiological Aspects," *Journal of Evolutionary Biochemistry and Physiology*, 57, 1419–1448. [7]

Berg, J., and Lane, V. (2011), "Pathology Harmony; A Pragmatic and Scientific Approach to Unfounded Variation in the Clinical Laboratory," *Annals of Clinical Biochemistry*, 48, 195–197. [1]

Blei, D. M. (2014), "Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models," *Annual Review of Statistics and Its Application*, 1, 203–232. [6,9]

Bold, A. M. (1976), "Clinical Chemistry Reporting: Problems and Proposals," *The Lancet*, 307, 951–955. [3]

Box, G. E. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society*, , Series A, 143, 383–404. [6]

Boyd, J. C. (2004), "Reference Regions of Two or More Dimensions," *Clinical Chemistry and Laboratory Medicine*, 42, 739–746. [4]

——— (2010), "Defining Laboratory Reference Values and Decision Limits: Populations, Intervals, and Interpretations," *Asian Journal of Andrology*, 12, 83–90. [2,3,5,8]

Boyd, J. C., and Lacher, D. A. (1982), "The Multivariate Reference Range: An Alternative Interpretation of Multi-Test Profiles," *Clinical Chemistry*, 28, 259–265. [4]

Burnside, E. S., Chhatwal, J., and Alagoz, O. (2012), "What is the Optimal Threshold at Which to Recommend Breast Biopsy?" *PloS One*, 7, e48820. [3]

Callahan, D. (1973), "The WHO Definition of 'Health'," *Hastings Center Studies*, 1, 77–87. [2]

Cao, Y., Su, Y., Guo, C., He, L., and Ding, N. (2023), "Albumin Level is Associated with Short-Term and Long-Term Outcomes in Sepsis Patients Admitted in the ICU: A Large Public Database Retrospective Research," *Clinical Epidemiology*, 15, 263–273. [7]

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76, 1–32. [9]

Ceriotti, F., Boyd, J. C., Klein, G., Henny, J., Queralto, J., Kairisto, V., Panteghini, M., and IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). (2008), "Reference Intervals for Serum Creatinine Concentrations: Assessment of Available Data for Global Application," *Clinical Chemistry*, 54, 559–566. [1]

Ceriotti, F., and Henny, J. (2008), "'Are My Laboratory Results Normal?' Considerations to be Made Concerning Reference Intervals and Decision Limits," *eJIFCC*, 19, 106–114. [5,8]

Ceriotti, F., Hinzmann, R., and Panteghini, M. (2009), "Reference Intervals: The Way Forward," *Annals of Clinical Biochemistry*, 46, 8–17. [3]

Clinical and Laboratory Standards Institute. (2008), "EP28-A3c: Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline–Third Edition." [1,2,5,8]

Cohen, N. M., Schwartzman, O., Jaschek, R., Lifshitz, A., Hoichman, M., Balicer, R., Shlush, L. I., Barbash, G., and Tanay, A. (2021), "Personalized Lab Test Models to Quantify Disease Potentials in Healthy Individuals," *Nature Medicine*, 27, 1582–1591. [1]

DeGroot, M. H. (2005), *Optimal Statistical Decisions*, Hoboken, NJ: Wiley. [3]

Dixon, W. (1953), "Processing Data for Outliers," *Biometrics*, 9, 74–89. [2]

Elveback, L. R., Guillier, C. L., and Keating, F. R. (1970), "Health, Normality, and the Ghost of Gauss," *Journal of the American Medical Association*, 211, 69–75. [8]

Erstad, B. L. (2021), "Serum Albumin Levels: Who Needs Them?" *Annals of Pharmacotherapy*, 55, 798–804. [7]

Estey, M. P., Cohen, A. H., Colantonio, D. A., Chan, M. K., Marvasti, T. B., Randell, E., Delvin, E., Cousineau, J., Grey, V., Greenway, D., Meng, Q. H., Jung, B., Bhuiyan, J., Seccombe, D., and Adeli, K. (2013), "CLSI-based Transference of the CALIPER Database of Pediatric Reference Intervals from Abbott to Beckman, Ortho, Roche and Siemens Clinical Chemistry Assays: Direct Validation Using Reference Samples from the CALIPER Cohort," *Clinical Biochemistry*, 46, 1197–1219. [1,8]

Fraser, C. G. (2001), *Biological Variation: From Principles to Practice*, Washington, DC: American Association for Clinical Chemistry. [5]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, Boca Raton, FL: CRC Press. [3,6,8]

Gelman, A., and Hill, J. (2006), *Data Analysis using Regression and Multilevel/Hierarchical Models*, Cambridge: Cambridge University Press. [8]

Gelman, A., Hill, J., and Vehtari, A. (2020), *Regression and Other Stories*, Cambridge: Cambridge University Press. [6,8]

Goldwasser, P., and Feldman, J. (1997), "Association of Serum Albumin and Mortality Risk," *Journal of Clinical Epidemiology*, 50, 693–703. [7]

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020), "rstanarm: Bayesian Applied Regression Modeling via Stan." [9]

Grams, R. R., Johnson, E. A., and Benson, E. S. (1972), "Laboratory Data Analysis System: Section III — Multivariate Normality," *American Journal of Clinical Pathology*, 58, 188–200. [4]

Gräsbeck, R. (2004), "The Evolution of the Reference Value Concept," *Clinical Chemistry and Laboratory Medicine*, 42, 692–697. [2]

Haller, C. (2006), "Hypoalbuminemia in Renal Failure: Pathogenesis and Therapeutic Considerations," *Kidney and Blood Pressure Research*, 28, 307–310. [7]

Harris, E. K. (1974), "Effects of Intra-and Interindividual Variation on the Appropriate Use of Normal Ranges," *Clinical Chemistry*, 20, 1535–1542. [5]

Harris, E. K., and Boyd, J. C. (1990), "On Dividing Reference Data into Subgroups to Produce Separate Reference Ranges," *Clinical Chemistry*, 36, 265–270. [1,5]

Harris, E. K., Yasaka, T., Horton, M. R., and Shakarji, G. (1982), "Comparing Multivariate and Univariate Subject-Specific Reference Regions for Blood Constituents in Healthy Persons," *Clinical Chemistry*, 28, 422–426. [4]

Healy, M. (1973), "Multivariate Analysis in Medicine and Biology," in *Perspectives in Biomedical Engineering: Proceedings of a Symposium Organised in Association with the Biological Engineering Society and Held in the University of Strathclyde, Glasgow, June 1972*, pp. 261–265, Springer. [4]

Higgins, V., Nieuwesteeg, M., and Adeli, K. (2020), "Reference Intervals: Theory and Practice," in *Contemporary Practice in Clinical Chemistry* (4th ed.), eds. W. Clarke and M. A. Marzinke, pp. 37–56, New York: Academic Press. [8]

Holmes, D. T., van der Gugten, J. G., Jung, B., and McCudden, C. R. (2021), "Continuous Reference Intervals for Pediatric Testosterone, Sex Hormone Binding Globulin and Free Testosterone Using Quantile Regression," *Journal of Mass Spectrometry and Advances in the Clinical Lab*, 22, 64–70. [5]

Horn, P. S., and Pesce, A. J. (2003), "Reference Intervals: An Update," *Clinica Chimica Acta*, 334, 5–23. [1,2,3,5]

Ichihara, K., Boyd, J. C., and IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). (2010), "An Appraisal of Statistical Procedures Used in Derivation of Reference Intervals," *Clinical Chemistry and Laboratory Medicine*, 48, 1537–1551. [1,5]

Kendall, H., Abreu, E., and Cheng, A.-L. (2019), "Serum Albumin Trend is a Predictor of Mortality in ICU Patients with Sepsis," *Biological Research for Nursing*, 21, 237–244. [7]

Lahti, A., Hyltoft Petersen, P., Boyd, J. C., Rustad, P., Laake, P., and Solberg, H. E. (2004), "Partitioning of NonGaussian-Distributed Biochemical Reference Data into Subgroups," *Clinical Chemistry*, 50, 891–900. [1,5]

Ledley, R. S., and Lusted, L. B. (1959), "Reasoning Foundations of Medical Diagnosis: Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason," *Science*, 130, 9–21. [2]

Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistical Hypotheses*, Cham: Springer. [7]

Lott, J., Mitchell, L., Moeschberger, M. L., and Sutherland, D. E. (1992), "Estimation of Reference Ranges: How Many Subjects are Needed?" *Clinical Chemistry*, 38, 648–650. [1,2]

Malka, R., Brugnara, C., Cialic, R., and Higgins, J. M. (2020), "Non-Parametric Combined Reference Regions and Prediction of Clinical Risk," *Clinical Chemistry*, 66, 363–372. [1,4]

Manrai, A. K., Patel, C. J., and Ioannidis, J. P. (2018), "In the Era of Precision Medicine and Big Data, Who is Normal?" *Journal of the American Medical Association*, 319, 1981–1982. [9]

McElreath, R. (2018), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Boca Raton, FL: Chapman and Hall/CRC. [3]

McNeil, B. J., Keeler, E., and Adelstein, S. J. (1975), "Primer on Certain Elements of Medical Decision Making," *New England Journal of Medicine*, 293, 211–215. [1]

Miller, W. G., Horowitz, G. L., Ceriotti, F., Fleming, J. K., Greenberg, N., Katayev, A., Jones, G. R. D., Rosner, W., and Young, I. S. (2016), "Reference Intervals: Strengths, Weaknesses, and Challenges," *Clinical Chemistry*, 62, 916–923. [2,3,8]

Miller, W. G., Tate, J. R., Barth, J. H., and Jones, G. R. (2014), "Harmonization: The Sample, the Measurement, and the Report," *Annals of Laboratory Medicine*, 34, 187–197. [1]

Murphy, E. A., and Abbey, H. (1967), "The Normal Range—A Common Misuse," *Journal of Chronic Diseases*, 20, 79–88. [1,2,3]

Neyman, J., and Pearson, E. S. (1933), "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London*, Series A, 231, 289–337. [8]

Nichols, B., Hentz, K., Aylward, L., Hays, S., and Lamb, J. (2007), "Age-Specific Reference Ranges for Polychlorinated Biphenyls (PCB) based on the NHANES 2001–2002 Survey," *Journal of Toxicology and Environmental Health, Part A*, 70, 1873–1877. [4,8]

Oesterling, J. E., Jacobsen, S. J., Chute, C. G., Guess, H. A., Girman, C. J., Panser, L. A., and Lieber, M. M. (1993), "Serum Prostate-Specific Antigen in a Community-based Population of Healthy Men: Establishment of Age-Specific Reference Ranges," *Journal of the American Medical Association*, 270, 860–864. [4]

Ozarda, Y. (2016), "Reference Intervals: Current Status, Recent Developments and Future Considerations," *Biochemia Medica*, 26, 5–16. [1]

Ozarda, Y., Sikaris, K., Streichert, T., Macri, J., and IFCC Committee on Reference Intervals and Decision Limits (C-RIDL) (2018), "Distinguishing Reference Intervals and Clinical Decision Limits–A Review by the IFCC Committee on Reference Intervals and Decision Limits," *Critical Reviews in Clinical Laboratory Sciences*, 55, 420–431. [8]

Pagana, K. D., Pagana, T. J., and Pagana, T. N. (2015), *Mosby's Diagnostic and Laboratory Test Reference*, St. Louis: Elsevier. [7]

Panteghini, M. (2012), "Implementation of Standardization in Clinical Practice: Not Always an Easy Task," *Clinical Chemistry and Laboratory Medicine*, 50, 1237–1241. [8]

Parmigiani, G. (2002), *Modeling in Medical Decision Making: A Bayesian Approach*, Chichester: Wiley. [3]

Pauker, S. G., and Kassirer, J. P. (1975), "Therapeutic Decision Making: A Cost-Benefit Analysis," *New England Journal of Medicine*, 293, 229–234. [3]

Phillips, A., Shaper, A. G., and Whincup, P. (1989), "Association between Serum Albumin and Mortality from Cardiovascular Disease, Cancer, and other Causes," *The Lancet*, 334, 1434–1436. [7]

Prati, D., Taioli, E., Zanella, A., Torre, E. D., Butelli, S., Del Vecchio, E., Vianello, L., Zanuso, F., Mozzi, F., Milani, S., Conte, D., Colombo, M., and Sirchia, G. (2002), "Updated Definitions of Healthy Ranges for Serum Alanine Aminotransferase Levels," *Annals of Internal Medicine*, 137, 1–10. [8]

Reed, A. H., Henry, R. J., and Mason, W. B. (1971), "Influence of Statistical Method Used on the Resulting Estimate of Normal Range," *Clinical Chemistry*, 17, 275–284. [1,2]

Royston, P. (1991), "Constructing Time-Specific Reference Ranges," *Statistics in Medicine*, 10, 675–690. [8]

Schini, M., Hannan, F. M., Walsh, J. S., and Eastell, R. (2021), "Reference Interval for Albumin-Adjusted Calcium based on a Large UK Population," *Clinical Endocrinology*, 94, 34–39. [1]

Schneider, A. J. (1960), "Some Thoughts on Normal, or Standard, Values in Clinical Medicine," *Pediatrics*, 26, 973–984. [2,3,5]

Schoenberg, B. S. (1970), "The 'abnormal" Laboratory Result: Problems in Interpreting Laboratory Data," *Postgraduate Medicine*, 47, 151–155. [1,2]

Sikaris, K. (2012), "Application of the Stockholm Hierarchy to Defining the Quality of Reference Intervals and Clinical Decision Limits," *The Clinical Biochemist Reviews*, 33, 141–148. [8]

Solberg, H. (1987), "Approved Recommendation (1987) on the Theory of Reference Values. Part 5. Statistical Treatment of Collected Reference

Values. Determination of Reference Limits," *Clinica Chimica Acta*, 170, S13–S32. [1]

Spinella, R., Sawhney, R., and Jalan, R. (2016), "Albumin in Chronic Liver Disease: Structure, Functions and Therapeutic Implications," *Hepatology International*, 10, 124–132. [7]

Springfield, D. S., and Rosenberg, A. (1996), "Biopsy: Complicated and Risky," *The Journal of Bone and Joint Surgery*, 78, 639–43. [3]

Sunderman Jr., F. W. (1975), "Current Concepts of 'Normal Values,' 'reference values,' and 'discrimination values' in Clinical Chemistry," *Clinical Chemistry*, 21, 1873–1877. [2,3,9]

Sunderman Jr., W. F., and Van Soestbergen, A. A. (1971), "Laboratory Suggestions: Probability Computations for Clinical Interpretations of Screening Tests," *American Journal of Clinical Pathology*, 55, 105–111. [1]

Suominen, P., Virtanen, A., Lehtonen-Veromaa, M., Heinonen, O. J., Salmi, T. T., Alanen, M., Mottonen, T., Rajamäki, A., and Irjala, K. (2001), "Regression-based Reference Limits for Serum Transferrin Receptor in Children 6 Months to 16 Years of Age," *Clinical Chemistry*, 47, 935–937. [5]

Surks, M. I., Goswami, G., and Daniels, G. H. (2005), "The Thyrotropin Reference Range Should Remain Unchanged," *The Journal of Clinical Endocrinology & Metabolism*, 90, 5489–5496. [1]

Van der Helm, H., and Hische, E. (1979), "Application of Bayes's Theorem to Results of Quantitative Clinical Chemical Determinations," *Clinical Chemistry*, 25, 985–988. [8]

Van Dongen, S. (2006), "Prior Specification in Bayesian Statistics: Three Cautionary Tales," *Journal of Theoretical Biology*, 242, 90–100. [6]

Vickers, A. J., and Lilja, H. (2009), "Cutpoints in Clinical Chemistry: Time for Fundamental Reassessment," *Clinical Chemistry*, 55, 15–17. [3]

Virtanen, A., Kairisto, V., Irjala, K., Rajamäki, A., and Uusipaikka, E. (1998), "Regression-based Reference Limits and their Reliability: Example on Hemoglobin During the First Year of Life," *Clinical Chemistry*, 44, 327–335. [5]

Wilson, J. (1973), "Current Trends and Problems in Health Screening," *Journal of Clinical Pathology*, 26, 555–563. [1,2,3]

Winkel, P., and Lyngbye, J. (1972), "The Normal Region—A Multivariate Problem," *Scandinavian Journal of Clinical and Laboratory Investigation*, 30, 339–344. [4]

Wright, E. M., and Royston, P. (1999), "Calculating Reference Intervals for Laboratory Measurements," *Statistical Methods in Medical Research*, 8, 93–112. [1]