Estimating Predictive Probability of Success

SHAUN COMFORT

PREVIEW *This article illustrates how Kahneman-Tversky's (KT) original reference class forecasting (RCF) for calibrating subjective forecasts can be reformulated using the language of Bayesian inference. Shaun Comfort shows a simple implementation for estimating the probability of success for Bernoulli outcomes such as clinical trials, contract bids, and medical devices. The approach uses the Beta conjugate model. The reference class distribution becomes an informative "prior" and the team forecast is treated as new data. The predictive validity determines the effective sample size weight for the new data, in order to generate a posterior probability of success distribution. The resulting posterior mean is identical to the RCF corrective procedure point estimate. In addition, the Bayesian implementation provides the entire posterior probability distribution, from which useful statistics such as credible intervals can be calculated with ease. This approach can be a useful method for individuals external to development teams responsible for capturing, tracking, calibrating, and presenting forecasts for decision makers, such as portfolio leaders or decisional analysis professionals.*

KAHNEMAN-TVERSKY (KT) AND REFERENCE CLASS FORECASTING

In 1977, Daniel Kahneman and Amos Tversky published an influential paper describing the psychological issues leading to poor accuracy with expert intuitive forecasts, such as bankers assessing a new business or a counselor predicting the likely achievement of a student. This procedure has been subsequently discussed in *Thinking, Fast and Slow* (Kahneman, 2011) and *Noise* (Kahneman and colleagues, 2021). In the original paper they show that forecasters have access to two general types of information: singular information (also referred to as the "inside" view) about a project or situation, and distributional information (also referred to as the "outside" view) about the outcomes of similar cases. Unfortunately, individuals frequently rely on singular information such as the unique details of a business startup while ignoring relevant statistical information on the success of similar startups. Their general finding is that individuals predict by matching the prediction to the impression a case makes upon them. For example, the impression a job candidate makes in an interview (e.g., "She was smart, well-spoken, and definitely in the top 5% of interviews I've had!") becomes the corresponding prediction of success in the role.

Extreme forecasts should be adjusted by considering the predictability of the task based on the diagnostic value of the information used to produce the estimate. Specifically, predictions should only match impressions when a task is perfectly predictable based on the impression. In contrast, predictions in situations with little or no predictability (e.g., fair coin flips) should simply match the class average or base rate frequency of heads or tails. For situations in between these extremes, the prediction or forecast should be regressive and fall between the class average and the matching impression in direct proportion to the predictability, as measured by the correlation between predictions and outcomes.

Their paper outlined a formal corrective procedure, which became the basis for reference class forecasting as shown in equation 1. Their equation works equally well for binary probability forecasts such as probability of success, or continuous forecasts such as annual sales. In this paper I focus on binary forecasting and will refer to this equation as Kahneman-Tversky Calibration (KTC):

Key Points

- Kahneman and Tversky's original corrective procedure for intuitive subjective forecasts is the foundation of reference class forecasting which can be restated in Bayesian form.
- The resulting mathematics use the closed form Beta conjugate model, which is easily implemented on spreadsheets for business use.
- A key benefit of the Bayesian approach is that it produces the same point estimate results as Kahneman and Tversky's original procedure with the associated uncertainty estimates, to inform business decisions.

$$F_c=(F_u\times P_v)+(1-P_v)\times BR_{rc} \quad \textbf{Equation 1}$$

Where F_c = corrected forecast, F_u = uncorrected forecast, P_v = predictive validity (the historically observed correlation between uncorrected forecasts and outcomes), BR_{rc} = base rate for the relevant reference class. Note that F_c is a weighted average of the base rate and the original uncorrected forecast.

Kahneman and Tversky then outlined five steps for implementing the corrective procedure (i.e., recalibrating) for forecasts:

1. Select an appropriate reference class of similar cases or projects with binary outcomes (e.g., probability of success rate for new employees, probability of successful rocket launch, medical device approval, etc.).
2. Estimate distributional information for the reference class.
3. Obtain the intuitive prediction from the relevant expert.
4. Determine the predictive validity (e.g., correlation between predictions and outcomes).
5. Correct the initial forecast.

Table 1. Simple Recalibration Result using KTC Method

Base Rate FOS	Team PPOS	Predictive Validity	Recalibrated PPOS
40%	70%	29%	48%

I illustrate this approach with a simple example. Assume that we are a consulting firm for a biotechnology company and are asked to evaluate a team's predicted probability of success (PPOS) for an important upcoming study of a new biologic treatment. We learn that the company keeps detailed records of all development team forecasts and final study outcomes that we will use as our reference class (step 1). Analysis of the historical data shows that over the last seven years the company conducted 81 proof-of-concept studies with 32 successes, for 32/81 ≈ 40% frequency of success (FOS) sample mean for our Bernoulli reference class distribution (step 2). Analysis of the corresponding team forecasts shows an average predicted success rate of 60%. The current team's PPOS for the planned study is 70% (step 3). Assuming that the distribution of prior PPOS predictions and outcomes represents the best available reference class for future studies, we have the relevant data for the steps 1-3, up to the point of estimating predictive validity. That is our next challenge.

Estimating the predictive validity (step 4) can be approached from several aspects, the simplest being adopting an average correlation observed in human predictive performance in the social sciences of $P_v \approx 0.28$ (Kahneman and colleagues, 2021). While this is easy, it is a crude approximation that I would recommend using only when no other data are available or alternatives are possible. An alternative approach is to estimate the correlation by using a percent concordance based on subject matter expert inputs as described by Kahneman and Tversky (1977) and Kahneman and colleagues (2021). Lastly, whenever possible, I recommend calculating the actual correlation from relevant forecasts and outcome data if they are available. For this example, I will use the hypothetical 81 records of predictions and outcomes to calculate the standard Pearson correlation between the PPOS forecasts and outcomes.

Based on the hypothetical data, the resulting correlation is 0.29, which is quite close to the average correlation estimate

of 0.28. We now combine this information with the KTC procedure to correct the team's prediction (step 5) with the results shown in **Table 1**.

Note that the effect of the calibration moves the team forecast PPOS (70%) towards the base success frequency in proportion to the predictive validity. This effect is often initially surprising for experts and teams producing forecasts. A common complaint is that the results are conservative and are unlikely to forecast success for rare events – but this should be expected because of the nature of rare events. Forecast methods that correctly predict rare outliers generally over-predict in the much larger majority of typical cases, producing poor overall accuracy.

It is important to be aware that the cost of errors can vary based on the task (e.g., predicting success of clinical trials) and type (e.g., false positives and false negatives). For instance, consistently inflated probability of success estimates can lead to costly "false positive" investments in late-stage studies and needless patient exposure to treatments that may not succeed. Conversely, consistent pessimistic forecasts can lead to "false negative" missed opportunities to execute trials leading to new patient treatments as well as unrealized financial returns for investors. In general, the goal of the forecaster is to produce unbiased, calibrated estimates for decision makers unless there is a clear reason to do otherwise. Finally, while the recalibrated results are not guaranteed to be accurate, they are by definition more regressive, produce probability estimates much closer to the historical frequency of success (i.e., calibrated "in the large" [Harrell, 2022]), and provide plausible probabilities for the large majority of predicted events.

RECASTING KTC CORRECTION IN BAYESIAN FORM

One shortcoming of the original KTC calibration method is that while it produces recalibrated point estimates, it does not explicitly account for uncertainty. Ideally, we want an approach that can provide both a point estimate and the associated uncertainty.

In viewing the KT recalibration procedure, a number of aspects appear similar to, and could potentially be approached from, the perspective of Bayesian statistics. For example, a standard version of Bayes' rule illustrates the approach:

$$P(H \mid D) \propto (ie, is\ proportional\ to)\ P(H) \times P(D \mid H)$$

Equation 2

The separate parts of equation 2 have specific definitions:

a. P(H | D) = the probability that hypothesis "H" is true given the data "D" ≡ Posterior

b. P(H) = the probability that hypothesis "H" is true ≡ Prior knowledge (equivalent to base rate)

c. P(D | H) = probability that data "D" is observed if hypothesis "H" is true ≡ Likelihood (this is represented by new information from an intuitive forecast based on singular information)

The key point is that the posterior probability is a combination of the prior knowledge and some additional information supplied. While not immediately apparent in the mathematics, Bayes' rule can be viewed as a weighted average of the prior knowledge and new information (e.g., the team forecast), analogous to equation 1.

In their original paper, Kahneman and Tversky state that in their approach the concept of the distributional base rate data is not equivalent to the prior probability because it is defined by the nature of the data, and the Bayesian prior is defined in terms of the sequence of data acquisition.

However, by recasting the approach in a sequential manner, and assuming that the historic distributions of PPOS predictions and outcomes are a reasonable estimate of our prior state of knowledge, we can recast their approach using the language of Bayes' rule. So the base rate becomes our prior that we want to update given new information (e.g., the development team's forecast) for the planned study. This produces a final posterior distribution for the

PPOS that will be our best estimate of the probability distribution of success for the study.

In determining the weight that new information provides, Kahneman and Tversky's predictive validity can be used to generate estimates of the Bayesian effective sample sizes for the new information and resulting posterior. From the posterior distribution we can compute any desired point estimate, range, or percentile. **Table 2** illustrates the conceptual mapping between KT Recalibration and the Bayesian approach.

INTRODUCING THE BETA DISTRIBUTION

The Bayesian approach presented here uses the conjugate Beta distribution for binomial observations. This means that there is no need to perform complicated mathematics to find the posterior distribution. We simply use the observation (e.g., a forecast estimate of PPOS) to update our prior PPOS to find the final conjugate Beta posterior distribution (Bolstad, 2017).

Our Beta conjugate model is easily calculated using standard Excel or other spreadsheet functions, or even by hand. I will use this nomenclature to specify the prior and posterior functions:

a. Beta Distribution = $Beta(a,b)$, where a and b are hyperparameters representing the number of successes and failures observed, respectively

b. The Beta equivalent sample size (ESS) = $a + b$ **Equation 3**

c. The Beta expected value (mean) = $a/(a + b)$ **Equation 4**

d. The posterior distribution is simply the sum of the "new hyperparameters" (a') and (b') based on the new information and the prior hyperparameters

Posterior = $Beta(a+a',b+b')$ **Equation 5**

Note that the hyperparameters (a' and b') can be interpreted as the number of "pseudo successes and failures" implied by the Bayesian equivalent sample size for the development team forecast, based on the predictive validity. This can be a confusing concept when first encountered but should become clear as I work through the calculations discussed below.

To illustrate, let's start with the reference class information from the earlier example (49 successes and 32 failures) to construct an informed prior distribution for the probability of success. This represents our "prior" knowledge about what the probability of success might be and is termed an "informed" prior because it is based on relevant data. This is in contrast to what is sometimes termed an "uninformed" prior typically represented by a flat, uniform distribution representing our state of ignorance about the possible probability of success.

Table 2. Kahneman-Tversky Calibration and Bayes Rule Formulation

Number	Kahneman – Tversky Concept	Bayes Rule	Comment(s)
1	Reference Class (Base Rate) Distributional Data	Informed Prior Distribution for Updating	The prior is an "informed" distribution based on previous or relevant data in contrast to a flat or uniform prior
2	Subjective Forecast	New Information	The subjective forecast provides new information from the team that can be used to calculate a posterior distribution
3	Predictive Validity	Effective Sample Size	Predictive validity can be used to estimate the effective sample size (i.e., impact weight) for the new information (i.e., the uncorrected forecast)
4	Recalibrated Forecast	Posterior Distribution	The posterior distribution can be used to calculate point estimates, uncertainty ranges, etc.

Figure 1 shows our resulting prior as a Beta (32 successes, 49 failures) probability distribution function, (PDF) where the mean equals the mean point estimate (shown as a vertical dashed line near the peak of the curve) of the reference class, and uncertainty is illustrated by the spread of the probability distribution function.

PREDICTIVE VALIDITY AND EFFECTIVE SAMPLE SIZE

As stated earlier, we can use predictive validity to "weigh" the new information provided by the team forecast. Specifically, we want to estimate a sample size that we can separate into a number of "pseudo" success (*a'*) and failure (*b'*) hyperparameters, in order to update our prior. Like the KTC approach, our Bayesian inference will be a weighted average.

Using the predictive validity of 0.29 and equation 1, we know that our prior distribution must have a weight of (1 – 29%) or 71%. Since we know the prior total sample size, we can use this information to determine the sample size of our posterior. Then the difference between the prior and posterior will be the equivalent sample size of our new information (i.e., from the team's forecast). I illustrate this mathematically in the equations below:

$$n_{post}=n_{prior}/((1-P_v)) \quad \textbf{Equation 6}$$

$$n_{new}=n_{post}-n_{prior} \quad \textbf{Equation 7}$$

For our clinical trial example, we can easily solve equation 6 and round the result to obtain an equivalent sample size for the posterior of $n_{post}=81/.71\approx114$. (While rounding results is not required, I am doing this to obtain integer values consistent with the concept of an integer sample size of discrete successes and failures.) Using this value in equation 7 produces a final effective sample size of $n_{new}\approx33$. This is the "weight" of the development team's forecast for probability of success.

Lastly, using the original team PPOS estimate of 70%, we can restate this as a sample estimate and use equations 3 and 4 to solve for the corresponding "pseudo successes" *a'* and "pseudo failures" *b'* Beta hyperparameters:

Figure 1. Example Beta Prior PDF

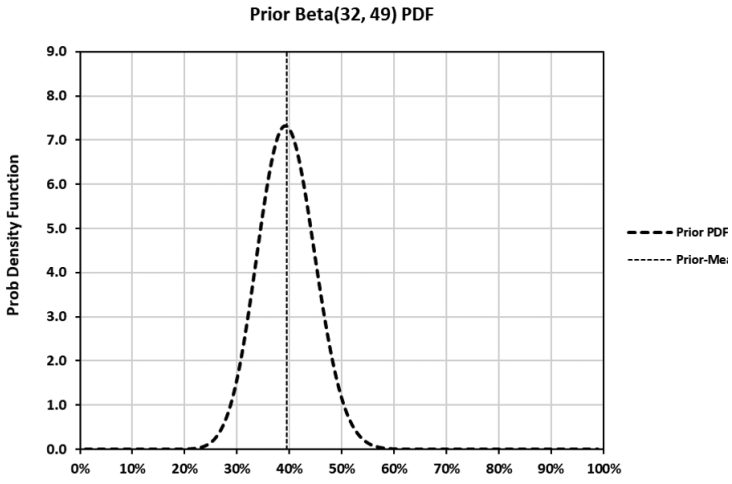
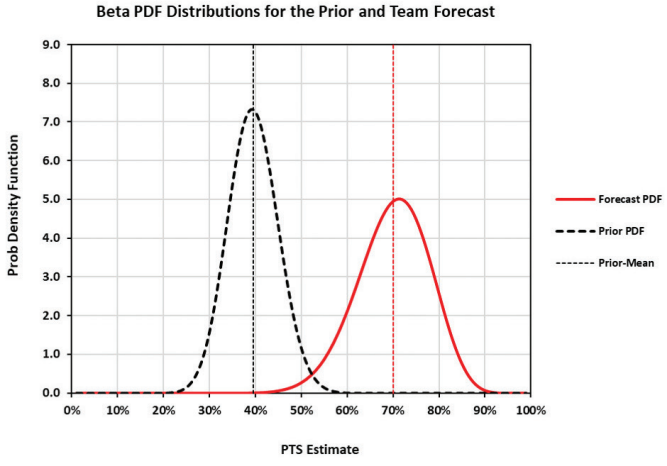


Figure 2. Example Beta Prior and Forecast PDFs



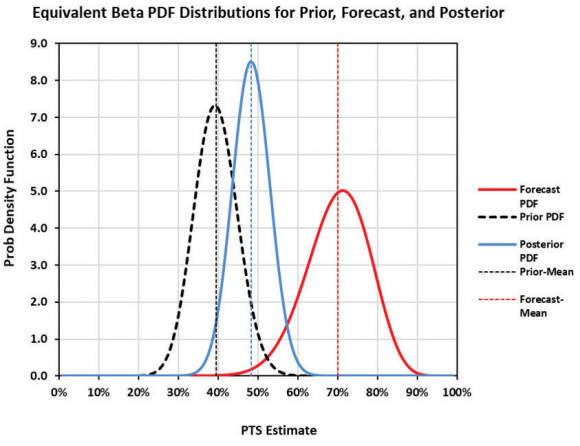
$$Beta\ Mean = a'/(a'+b') \equiv a'/ESS$$

where ESS is the effective sample size estimate for the team POS forecast = 33. By equating the mean with the subjective forecast of 70%, we can solve for the a' hyperparameter as the product of the sample size and the forecast estimate, which after rounding becomes a'≈23. The corresponding b' hyperparameter is simply the difference between the effective sample size and the a' estimate, which becomes 10.

At this point, we can graph the beta function representations with the respective means (shown as vertical dashed lines) of the team forecast information with the prior, as shown in **Figure 2**.

Note that the wide range of the team forecast PDF (from approximately 50% to 90%) reflects the imprecision of the

Figure 3. Example Beta Prior, Forecast, and Posterior PPOS Distribution



team's forecast based on the smaller effective sample size than for the prior distribution. In addition, the graph clearly shows that the team forecast distribution has little overlap with the prior. This likely represents optimism bias, which is commonly observed with "inside" view forecasts. This visualization of the discrepancy between the reference class distribution of frequency of success and the team forecast can be an excellent entry point for a discussion with leadership regarding why the team believes the proposed new study has a much higher probability of success than the reference class. Perhaps the team has a strong rationale based on additional information that is convincing – or after further thought and discussion, they may reconsider their forecast.

RECALIBRATED STUDY PPOS FORECAST USING BAYESIAN INFERENCE

We now have a beta model for our prior knowledge of clinical trials success and the team forecast for likely success. We are in position to use the conjugate beta binomial model to compute our posterior prediction of the probability of success (POS), using equation 5:

$$\text{Posterior POS} = Beta(32+23, 49+10) = Beta(55, 59)$$

Our final PPOS distribution is shown in **Figure 3** below, along with the prior and team forecast, in what is sometimes referred to as a "Bayesian Triplot" graph. Note that the posterior PDF is close to the prior, with a mean POS estimate of 48% (or 55/[55 + 59]), which equals the KTC PPOS estimate shown in Table 2. This is due to the prior sample size of 81 being more than twice the sample size of 33 for the team's forecast, based on the historic predictive validity. In so many words, the team historical forecasts have low predictive validity, and consequently the recalibrated forecast is anchored more on the prior reference class distribution.

Using our posterior beta distribution, we can easily calculate point estimates for the mean, 10th percentile, and 90th percentile (i.e., 80% Credible Interval) for direct comparison with the original approach by Kahneman and Tversky. The results are shown in **Table 3** where the means are identical. In addition, I've also provided the relevant credible intervals. While an uncertainty range is not generally requested by decision makers, it is an important element in forecasting that illustrates how likely or unlikely a particular point estimate may be. That said, showing the uncertainty visually as in Figure 3 is likely to be much more impactful to decision makers than listing tabular data.

Table 3. KT and Bayesian Data Comparison

KTC Recalibration		Bayesian Recalibration				
Variable	**Mean**	**Variable**	**Mean**	**P10**	**P50**	**P90**
Team PPOS	70%	**Team PPOS**	70%	60%	70%	80%
Base Rate	40%	**Prior PPOS**	40%	33%	39%	47%
Recall PPOS	48%	**Post PPOS**	48%	42%	48%	54%

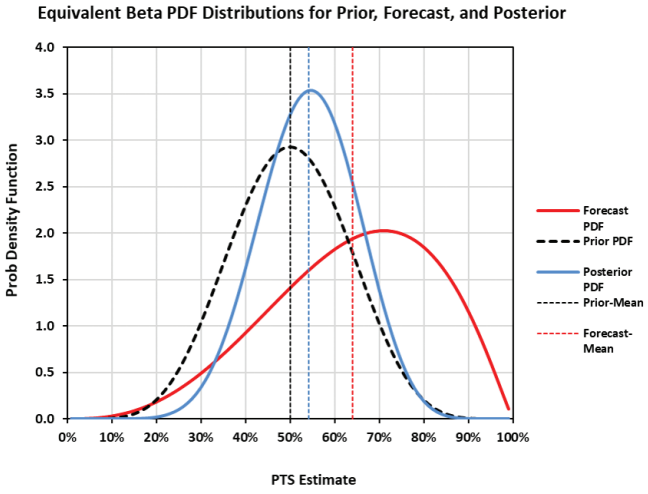
DISCUSSION

The example I have shown here is optimistic in that the hypothetical organization kept specific records of forecast PPOS and the corresponding outcomes (i.e., success or failure). Unfortunately, this is not a common occurrence, and some of you may be thinking: How do you calculate the relevant beta distributions with less data? For example, one potential scenario is where a company or organization does not keep detailed records, but has experts who can estimate the inputs for a prior distribution.

For example, let us imagine a biotech company with a new proposed Phase 2 study where the development team predicts a 64% probability of success. Interviews with the relevant experts conclude that the company has run at least 14 relatively similar Phase 2 studies over the previous eight years with about half being successful. Unfortunately, the actual forecast probabilities of success were not captured for later comparison with outcomes. How might one proceed? While there are no perfect rules, here I would estimate the prior distribution information at face value, with a total sample size of 14 with half of the studies being successful and half failing. That would imply a prior Beta (7, 7) probability density function.

To estimate predictive validity I would use the percent concordance task outlined in Kahneman and colleagues (2021), where relevant experts are asked to consider two studies at random. What proportion of the time could they correctly identify the study most likely to be successful? Using Kahneman's concordance chart, we can convert this to an equivalent correlation coefficient (i.e., predictive validity P_v). For example, imagine that an interviewed expert gives a common percent concordance estimate of 60%, which corresponds to a Pv ≈ 0.30 (Kahneman and colleagues, 2021, p. 104). Using this information, we can then estimate the posterior sample size using equation 6 as 20, and the team's forecast sample size then becomes 20 – 14 = 6. Together with the team forecast of 64%, this implies a corresponding Beta function for the team forecast ≈ Beta (4, 2), because 64% of 6 is about 4. The posterior distribution (using equation 5) then becomes Beta (11, 9) with a mean ≈ 55%.

Figure 4. Example Beta Prior, Forecast, and Posterior PPOS Distribution



The point of this discussion is to illustrate the flexibility of this approach and the need for some level of judgment to obtain estimates that allow for useful computation. In situations such as my example here, the results are rough approximations as illustrated by the large spread in the beta distributions of the prior, forecast, and posterior shown in Figure 4. Lastly, using the Bayesian approach for this example illustrates the potential value of showing the underlying estimate uncertainty to decision makers, which is not apparent when using simple point estimates. In fact, simply showing estimates visually (as in **Figure 4**) to elicit further team and decision-maker discussions on risk and uncertainty may be an important benefit of using this approach. For example, while this example's mean posterior 55% point estimate of potential success for a trial (or project, etc.) may exceed some investment thresholds, an approximate 50% implied uncertainty range would likely foster requests for additional information and potential risk mitigations before endorsing investment.

One critique of this approach is that teams could counteract recalibration by

greatly inflating their PPOS estimates (e.g., to 100%) in order to keep the "recalibrated" estimates as high as possible. It is true that a motivated team could "game" this approach to give their programs the most optimistic assessment possible; the mathematics alone cannot easily prevent this. However, this is where good forecasting processes can help minimize gaming and support production of "decision grade" forecasts useful for leadership. For example, several potential processes could include

a. Tracking and recording team and recalibrated forecasts, as well as outcomes over time. This data can be used to determine base rates and show any large changes in the system (e.g., predictive validity or frequency of success) that could be due to gaming. Finally, capturing the forecasts and outcomes allows organizations to determine whether team forecasts, recalibrated Bayesian forecasts, or simple naïve base-rate forecasts are the best approach using proper scoring rules.

b. Recalibration should be performed by an independent group to avoid assessments by individuals with a vested interest in the outcome. In pharma and energy companies, this type of activity would typically be the responsibility of Portfolio and Decision Analysis functions outside development teams, but could also be done by others.

c. Decision makers should be shown the full Bayesian "triplot" graphs and not just single-point estimates. This allows leadership to see the original and recalibrated estimates along with the associated uncertainty. Grossly optimistic estimates will stand out and can be a catalyst for discussion between teams and the decision makers (e.g., "Why are you saying project X has a POS of 75% when our frequency of success has only been 40%? What is your rationale?").

In conclusion, I find this simple Bayesian approach quite useful. However, it is neither a perfect approach nor an automatic guarantee of accurate probability of success forecasts. Garbage in still equals garbage out, and without good forecasting processes this approach, as well as most others, can potentially be gamed or rendered ineffective. As Kahneman and Tversky stated in their original paper, the results are likely to be more realistic than forecasts produced with sole focus on the internal "inside" details of a project. There is nothing magic here, just a simple, rational procedure for producing useful forecasts for decision making. I hope you also find this to be the case.

Shaun Comfort holds an MD from the University of New Mexico and is a Principal Scientific Enablement Director for Roche-Genentech, leading scientific innovation work supporting pharmacovigilance. He is a neurologist with 20-plus years combined biopharma industry/regulatory experience including roles as former Medical Reviewer at the U.S. FDA and Clinical Research and Safety at J&J, Anesiva, and Genentech. Shaun's current work focuses on machine learning models and decision analysis techniques to evaluate drug-adverse event causality and predict project/phase probability of success. Recently he published his first book focused on using mathematical estimation in healthcare: *How Much Is that Cure in the Window? Simple Math Solutions for Complicated Problems in Biology, Medicine, and Healthcare*.

comfort.shaun@gene.com

REFERENCES

Bolstad, W. & Curran J. (2017). *Introduction to Bayesian Statistics* (3rd Ed.), John Wiley and Sons.

Harrell, F.E. Regression Modeling Strategies. ***hbiostat.org/doc/rms/book/***

Kahneman, D. & Tversky, A. (1977). *Intuitive Prediction: Biases and Corrective Procedures*, Decision Research, Perceptronics. ***apps.dtic.mil/sti/pdfs/ADA047747.pdf***

Kahneman, D., Sibony, O. & Sunstein, C.R. (2021). *Noise: A Flaw in Human Judgment*, Little, Brown Spark.

Kahneman D. (2011). *Thinking, Fast and Slow*, Farrar, Straus and Giroux.