Title: Rethinking Clinical Trial Outcomes: Use of a Standardized Metric to Enhance Design and Interpretation

Authors: Alyssa Ylescupidez†, Henry T. Bahnson†, Colin O’Rourke†, Sandra Lord†, Cate Speake‡, Carla J. Greenbaum†*‡

Affiliations:

1Center for Interventional Immunology, Benaroya Research Institute at Virginia Mason; Seattle, WA, USA.

*Corresponding author: Carla J. Greenbaum. email: cjgreen@benaroyaresearch.org

†Authors contributed equally.

‡Authors contributed equally.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract: The use of a standardized outcome metric enhances clinical trial conduct, interpretation, and cross-trial comparison. If a disease course is predictable, comparing modeled predictions with outcome data affords the precision and confidence needed to accelerate precision medicine. We demonstrate the power of this approach in type 1 diabetes (T1D) trials aiming to preserve endogenous insulin secretion measured by C-peptide. C-peptide is a predictable outcome given an individual’s age and baseline value; quantitative response (QR) adjusts for these variables and represents the difference between the observed and predicted outcome. Validated across 13 trials, the QR metric reduces each trial’s variance and markedly increases statistical power. As smaller studies are especially subject to random sampling variability, using QR as the outcome metric introduces alternative interpretations of previous clinical trial results analyzed by traditional statistical methods. QR can provide model-based estimates that quantify whether individuals or groups did better or worse than expected. QR also provides a purer metric to associate with biomarker data, enabling improved mechanistic insights. Using data from more than 1,300 participants, we demonstrate the value of QR in advancement of disease-modifying therapy (DMT) in T1D. QR applies to any disease where outcome is predictable by baseline covariates, rendering it useful for defining responders to therapy, comparing the effectiveness of different therapies, and understanding causal pathways in disease.
Main Text:

Introduction

Ideally, a metric for studying disease should be clinically and scientifically meaningful, objective, predictable, and able to be standardized across individuals and cohorts. When applied in the context of clinical trials for any disease, such a standardized metric may enable acceleration of trials through increased statistical power and aid in interpretation of clinical trial data by regulators, clinicians, investigators, translational scientists, and study participants.

If a disease course or outcome is predictable using baseline factors, analysis should adjust for these factors. This approach is advantageous over traditional unadjusted analysis, which essentially compares an average treated individual to an average control individual. In unadjusted analysis, results may be more strongly impacted by chance covariate imbalances at baseline, especially when there is evidence that the covariate is associated with the outcome, obscuring the effect of treatment. Despite the known benefits of baseline-adjusted analyses, reviews have found that only 24 to 34% of trials use covariate adjustment for the primary analysis.

A standardized quantitative response (QR) metric that adjusts for baseline covariates can be developed for any reproducible outcome measure for which the natural history is known and predictable. This is the case for trials of disease-modifying therapy (DMT) in type 1 diabetes (T1D) aiming to preserve endogenous pancreatic beta cell function, as there is a wealth of natural history data on the loss of insulin secretion over time measured by the C-peptide response to a mixed meal tolerance test (MMTT). Moreover, though it is noteworthy that a therapy to delay onset of clinically apparent disease was recently approved for clinical use, there are still no DMTs that preserve endogenous insulin secretion in individuals recently diagnosed with T1D. To date, there have been a few dozen randomized controlled trials (RCT) of immune therapy in recently (<3 months) diagnosed T1D and almost all are phase 2 trials led by academic investigators. Further, the time to conduct such studies is long, given that trials are challenging to enroll and study endpoints are at least 1 year from randomization. While C-peptide response to a MMTT is accepted as the appropriate measure of endogenous insulin secretion, regulatory ambiguity for potential indications exists since there is no established clinical therapeutic threshold of C-peptide that definitively qualifies therapies or interventions as successes or failures. Additionally, by design, published trials of immune therapy express the average effect of therapy
on the randomized cohort and, though multiple definitions have been proposed, there is no accepted standardized criteria to define a responder to therapy. Together, these issues create limitations in understanding mechanisms of disease and response to therapy, hindering the ability to develop a precision medicine approach to DMT in T1D and other immune-associated diseases.

The QR, originally developed by Bundy and Krischer, leverages the well-known statistical property that model adjustment with prognostic baseline covariates increases precision and confidence by way of controlling for outcome heterogeneity. Bundy and Krischer used data from five trials in similar populations to develop an analysis of covariance (ANCOVA) model to predict the C-peptide area under the curve (AUC) mean value by adjusting for baseline C-peptide AUC mean and age. The resulting QR metric is a standardized measure of the difference between an individual’s observed and predicted C-peptide AUC mean one year after study entry. Values above zero indicate a better-than-expected outcome and values below zero indicate a worse-than-expected outcome.

Using data from 13 RCTs testing 14 different therapies aiming to preserve beta cell function in those with T1D, we demonstrate how the QR metric increases the precision and confidence of clinical trial results, thus enhancing interpretation of these studies while suggesting new concepts for future trial designs. We found that the QR metric reduced variance and standardized C-peptide outcomes across trials, leading to re-evaluation and new interpretations of both clinical and mechanistic results. In addition, we illustrate how the QR metric may be useful for design of future trials. Together, these findings represent a significant step towards precision medicine and serve as an example for the study of other diseases.

Results

The QR metric reduces variance and standardizes C-peptide outcomes across trials

We first validated the published QR metric using data from 13 studies: five TrialNet RCTs used in the development of QR (referred to as the development cohort), and eight additional RCTs (referred to as the validation cohort). Table 1 lists the key characteristics of each of the 13 trials, including number of subjects, median age, and baseline C-peptide AUC mean. We found that the mean QR matches closely between the development and validation cohorts, and centers around zero in both cohorts with similar distributions (Figure 1; p=0.44, two-sample t-test [t=-0.8, DF=1021.5]; p=0.62, Kolmogorov-Smirnov two-sample test).
We next updated the model to include all available data from all 13 studies. This revised ANCOVA model predicted 1-year C-peptide AUC mean values very similar to those predicted from the published QR model (Supplemental Figure 1A and B). Moreover, the predicted values between the original and revised models were strongly associated ($R^2 = 0.998$; Supplemental Figure 1C); thus, it was determined that the original QR model applied well to all 13 studies and all results hereafter use the original QR model. Additionally, the model is robust; predictions were counterintuitive for only eight of the 1306 individuals studied. For these eight older individuals with low C-peptide, the model predicted a very minor increase in C-peptide over time (Supplemental Figure 2).

The use of the QR metric both reduced the variance and standardized the mean of the C-peptide outcome within each trial. Among the placebo-treated individuals in the 13 trials, there were noteworthy variations with respect to both age and baseline C-peptide between studies (Figure 2A). The 1-year C-peptide AUC mean value varied (mean range 0.36 to 0.69 nmol/L) between trials, and within each trial demonstrated wide heterogeneity (Figure 2B). In contrast, by accounting for baseline C-peptide and age, the mean QR value of placebo-treated individuals centered around zero for each trial (mean range -0.07 to 0.08) (Figure 2B). For 12 of these 13 trials, the mean QR value was not statistically different from zero; the only exception being the TrialNet ATG/GCSF trial (mean of -0.072, $p=0.031$; 30 placebo participants). For each individual trial, the standard deviation of the QR was markedly lower than the standard deviation using the C-peptide AUC mean (Figure 2B, annotated in blue).

We next investigated whether the QR metric would increase statistical power since covariate adjustment in randomized trials leads to greater power and better control of type I and type II error. We chose to examine this in the Immune Tolerance Network (ITN) AbATE trial of teplizumab, which had a positive outcome in demonstrating efficacy of teplizumab in preserving beta cell function in recently diagnosed individuals relative to controls. In the original analysis, the primary outcome used the difference in 4-hour C-peptide AUC mean between baseline and 2 years, adjusted for baseline C-peptide, with a $p$-value of 0.002 for the difference between treatment groups. In our re-analysis, we used the 2-hour C-peptide AUC mean at 1 year and found that the difference in C-peptide AUC mean in control (0.364 nmol/L) compared to teplizumab-treated (0.647 nmol/L) at 1 year was statistically significant with a $p$-value of 0.009 ($t=2.7$, $DF=46.4$).
using a two-sample t-test assuming unequal group variances (Figure 3A). When using the change from baseline C-peptide AUC mean as in the published trial report, the effect size and precision increased with a mean change from baseline of -0.321 nmol/L in the control group and -0.086 nmol/L in the teplizumab-treated group (p=0.0002, t=4.0, DF=57.0) (Figure 3B). Controlling for both baseline C-peptide and age by using QR as the outcome further increased the statistical significance of the result (-0.015 nmol/L control vs 0.141 nmol/L teplizumab-treated, p<0.0001, t=4.3, DF=47.3) (Figure 3C).

We next assessed whether the QR approach could be utilized to predict trial outcomes beyond 1 year, using varying baseline reference points and outcome timepoints. The original QR model used baseline values from individuals within three months of diagnosis to predict the outcome at 12 months after treatment initiation and demonstrated an R² value of 53%. Predictions further in the future are more challenging, thus it is not surprising that the R² drops when this same baseline is used to predict a 24-month outcome (Figure 4). However, using 6-, 12-, or 18-month data as baseline C-peptide, the R² value at two years is high (74%, 85%, 87%, respectively), suggesting that QR can be used as an outcome measure for trials enrolling individuals further from diagnosis.

**Applying the QR metric to previous published clinical trials can change the interpretation of both clinical and mechanistic results**

Since the QR metric incorporates historical data from many placebo/control individuals, it minimizes the random sampling variability often present in individual studies with small sample sizes. We determined whether using the QR metric would alter the interpretation of clinical trial results, compared to the originally published reports. In Figure 5, we show the mean QR (± 95% confidence intervals) for the active treatments (Figure 5A) and placebo/control arms (Figure 5B), as well as the treatment effect expressed as the difference in QR between arms (Figure 5C) for 13 published trials.

Expressing the overall treatment effect and results of each arm using QR altered the interpretation of some of the published results. For example, the primary outcome of the alefacept trial was the 2-hour C-peptide AUC at 1 year, and the difference between treatment arms did not reach statistical significance in the original analysis. However, applying the QR metric to the alefacept trial dataset demonstrated a large effect in the active treatment group (Figure 5A), strongly suggesting that alefacept, or drugs working in the same pathway, are worth pursuing in future trials.
For the canakinumab trial, assessing the overall trial result by the difference in treatment arms using the QR metric finds no treatment effect (Figure 5C), consistent with the published outcome. Yet, the QR of the active arm of the canakinumab trial suggests a positive effect of this therapy on C-peptide (Figure 5A). Moreover, the QR of the placebo arm allowed us to further interpret this result, revealing that the lack of statistical significance between the groups may be driven by higher-than-expected C-peptide response in the 22 individuals in the placebo arm of the study (Figure 5B).

Lastly, applying the QR metric to the two studies testing anti-thymocyte globulin (ATG) also suggests a different interpretation than originally published. The TrialNet ATG/GCSF trial reported a positive outcome using lower-dose therapy than the ITN ATG study, which did not meet its primary outcome, leading to the interpretation that dose level was the key variable in the effectiveness of the drug. However, the mean QR in the treated participants was notably similar between studies: mean (95% CI) QR was 0.08 (0.02-0.13) in the higher-dose ITN ATG trial, 0.09 (0.03-0.15) in the lower-dose TrialNet ATG/GCSF trial. This suggests that the reported difference in treatment effect between the studies was driven by the placebo participants: those in the TrialNet ATG/GCSF study had a fairly low mean QR (-0.07) while those in the ITN ATG study had a higher mean QR (0.05). Of note, across the 13 studies, only the TrialNet ATG/GCSF control arm was statistically significantly different from zero. This further demonstrates how random sampling variability in smaller studies can complicate interpretation of RCT results.

We also looked at the applicability of the QR model to timepoints prior to 1 year. We found that a mean QR value above zero at six months was indicative of the trial outcome: the mean 6-month QR value for those in the active treatment arm was above zero in all trials with a positive outcome by QR at 1 year (abatacept, low-dose ATG, rituximab, teplizumab, and alefacept) (Figure 6).

Next, we investigated whether using the QR metric would impact the results of immune marker studies aiming to explore mechanisms of response to therapy. Using C-peptide and mechanistic results obtained from the ITN AbATE trial of teplizumab, we confirmed previous reports of the positive association between the frequency of treatment-induced KLRG1+TIGIT+ CD8+ T cells, a known signature of T cell exhaustion, and C-peptide outcome (Figure 7A). However, adjusting for baseline C-peptide and age by using the QR metric revealed that the association between treatment-induced KLRG1+TIGIT+ CD8+ T cells and outcome may be weaker than originally
thought (Figure 7B). This observation is likely accounted for in part by a moderate association between treatment-induced exhausted T cells and age (Figure 7C). As previously noted, age is one of two key variables in the QR metric; age is also known to be important in defining setpoints and responsiveness for many immune cell populations (recently reviewed in 24-27). This novel analysis implies that therapy-induced exhaustion of T cells unveils mechanistic insights about age itself, which may or may not be causally related to a particular therapy, but is important to our understanding of the role age plays in disease progression and response to therapy. Supplemental Figure 3 graphically illustrates why QR is a more powerful metric for identification of a biomarker that is causally related to therapy.

**The QR metric better quantifies responders to therapy**

As in other diseases, not all individuals recently diagnosed with T1D will respond to a given therapy. Although continuous measurements are preferred to minimize loss of statistical power, historically, analyses of clinical trial results across many diseases frequently stratify treated individuals as responders and non-responders to therapy. In T1D trials, varying definitions of response to therapy using C-peptide have been used 11,28,29. Reasoning that previously published responder definitions may be associated with baseline variables, we investigated whether the standardized QR metric could better quantify responders to therapy.

We first explored the relationship between baseline C-peptide, age, and the previously published categories of a C-peptide responder/non-responder 11,28,29. As shown in Figure 8, among placebo/control participants, the probability of meeting each of the four responder definitions is strongly associated with age (Figure 8 panels A [p=0.0007], B [p=0.0008], C [p=0.0001], D [p=0.0007]); two of these definitions are also associated with baseline C-peptide (Figure 8A [p=0.02] and C [p=0.02]). In contrast, the probability of being a responder using the QR-based definition of above or below zero, is, as expected, not associated with either age (p=0.18) or baseline C-peptide (p=0.62) (Figure 8E).

To further illustrate the consequences of not accounting for baseline variables in classifying responders, we benchmarked the probability of being a responder for each definition using the data from an “average” treated individual. The average age (16.4 years) and average baseline C-peptide AUC mean for treated individuals across all 13 studies was determined, yielding a QR value of 0.039 for an average individual. However, the probability that this individual is defined as a
treatment responder varies widely using different responder definitions (Figure 8B and C). Most importantly, the probability of being a treatment responder is strongly associated with age, baseline C-peptide, or both of these metrics for all non-QR definitions. Since the QR metric adjusts for age and baseline C-peptide, the probability of being a responder is not conditional on these factors, as can also be seen from the annotated p-values (Figure 8E).

Given that the probability of being a treatment responder is not conditional on age and baseline C-peptide, we asked how the QR metric could be used to select a threshold for classifying responders and non-responders to therapy. In selecting a threshold for a continuous measure such as QR, it is useful to understand the variability or confidence intervals around a given QR value, reflecting the probability that a given QR value represents a true treatment responder. Here, we observed that the distribution of the QR scores of all placebo/control individuals is symmetrical (Figure 9A), leading to quantile statistics whereby the QR value can be assigned to a percentile (e.g., a QR of 0.10 corresponds to the 75th percentile of the distribution). Figure 9A also illustrates that while there is a symmetrical distribution around the mean, heterogeneity is also apparent; a placebo-treated individual may have a QR value ranging from -0.58 to 0.45. Similarly, though the mean QR value among all individuals in the treatment arms of the positive studies is above zero, there is also a wide range of values in each treated group (Figure 9B), many of which overlap with the distribution of placebo-treated individuals.

The overlapping QR values between treatment and placebo groups demonstrate that using a particular QR cutoff will not necessarily distinguish individuals who received an efficacious therapy from placebo individuals. We reasoned that these distributions can be used to understand the probability that a specific QR value is associated with a successful treatment (Figure 9C). As shown in Figure 9C, the probability of identifying a treatment responder or non-responder increases at the extremes of the distribution. For example, the probability that an individual with QR=0.4 received an effective therapy exceeds 80% (Figure 9A). Conversely, the probability that an individual with QR= -0.4 received an effective therapy is only 15%: at this threshold, individuals are more likely to be placebo participants, and thus we can infer that a treated individual with this QR value was likely a non-responder to therapy. Choosing less extreme cut points introduces greater uncertainty. For example, selecting a threshold of treatment response of QR=0.2 would yield only a 65% probability that this individual received an effective therapy.
Despite decades of clinical trials of DMT in individuals recently diagnosed with T1D, there are no drugs currently in clinical practice. Here, we have demonstrated that the QR metric may address many challenges to the field, facilitating the identification of potentially effective therapies. Importantly, standardization of outcomes enables a uniform method of analysis across trials, and thus a manner for comparing therapies and identifying responders to therapies through a consistent responder definition.

We applied the QR metric, which adjusts for baseline age and C-peptide AUC mean, to data from 13 clinical trials of DMT in individuals with recently diagnosed T1D. Since these 13 trials occurred over a 10-year period, included individuals from 3 to 46 years of age, were conducted at multiple locations, and included data from both academic trials and a phase 3 industry-sponsored study, the strength of the model is sufficiently robust to be considered for regulatory purposes.

Whereas traditional unadjusted analysis may be impacted by chance imbalances in covariates at baseline (especially those known to be associated with outcome), baseline-adjusted analysis can lead to individual-specific (conditional) estimates which conceptually match individuals in the intervention group and control group who are similar with respect to the adjusted variables. Baseline-adjusted analysis increases statistical power, allowing for robust comparisons between studies.

In T1D, more than half of the heterogeneity in the natural history of disease can be explained by age and baseline C-peptide. While several T1D trials used ANCOVA models adjusted for baseline metrics, this was inconsistent between studies. Computing a QR further utilizes those ANCOVA predictions to determine a standardized score, which enables cross-trial analysis. Analyzing treatment effects in terms of QR also allows for evaluation of treatment groups in a standardized manner, with comparisons to a large number of controls. In any setting (e.g., cancer, other autoimmune diseases) in which the natural history of the outcome is known and predictable by baseline covariates, application of a QR metric would be expected to provide similar benefits to understanding trial results, mechanistic studies, and responses to therapy.

Using data from the ITN trial of teplizumab, we demonstrate that using the QR metric reduces the variance of the outcome, resulting in increased power and the potential for reducing sample size.
However, it is not clear that reducing the sample size for a phase 2 RCT is the optimal approach to select promising therapies, or to identify responders for T1D or other diseases. Placebo-controlled randomized trials have clear advantages, as randomization can account for potential differences in variables that are known to impact outcome. However, when small sample sizes are used, random sampling variation can significantly impact inferences about trial outcomes. In the case of trials of DMT aiming to preserve C-peptide, the known factors are baseline C-peptide and age; QR adjusts for these factors.

Using data from almost 500 control/placebo individuals in the 13 trials studied, we show that untreated individuals’ outcomes are reliably predicted. Utilizing QR as an outcome measure implies comparison of a treatment arm to this large number of historical controls, and may obviate the need for a contemporaneous control group. Single- or multiple-active arm trials are likely to conserve resources by eliminating or minimizing placebo participants while accelerating recruitment (as more participants agree to trials without a placebo arm). This paradigm, common in clinical trial settings including cancer research, increases the likelihood that participants will benefit from prospective therapeutic interventions. The QR framework achieves these advantages by leveraging large amounts of historical data to create “synthetic controls” to test promising interventions in phase 2 trials, which could then be reproduced in larger phase 3 placebo-controlled pivotal randomized trials.

Furthermore, using a QR outcome allows for adaptive study designs of multiple active agents, as we found that a mean QR value above zero in the treatment arm at six months after randomization completely predicted the success of all the tested RCTs with a positive outcome at 1 year. A trial with multiple active agents could drop ineffective therapeutic arms at six months. Using QR could enable shorter clinical trials, which would reduce burden on study staff and participants, reduce cost, and reduce the time that participants in the active arm are exposed to ineffective therapies. By pre-specifying a QR threshold of interest at an early timepoint, adaptive re-randomization designs, such as the sequential parallel comparison design (SPCD; Supplemental Figure 4), are also feasible. This design identifies placebo participants with a QR value below zero early in a trial, and re-randomizes those individuals to treatment or placebo, allowing a larger number of participants to potentially benefit from therapy. QR can also enable enrollment of individuals...
outside the traditionally used period of 3 months post-diagnosis in new-onset T1D trials, as it can reliably predict a 2-year outcome when a baseline timepoint is > 6 months from diagnosis.

Assessing treatment effect as the difference in QR between treatment and control arms may alter interpretation of prior trial results. Additionally, it can aid in prioritizing therapies for further study. Since there were similarities between the 13 trials with respect to baseline C-peptide and age, and many of the trials used baseline-adjusted ANCOVA models, it is not surprising that using QR to express the trial result is similar to that seen in published reports; that is, the teplizumab, abatacept, rituximab, and low dose ATG trials all demonstrate that the QR of the actively treated group is higher than that of the control group. However, while the published primary outcome of the ITN alefacept trial was negative, when considering the outcomes of the active treatment arms for each trial, teplizumab and alefacept both stand out as therapies with the greatest QR values, which suggests both therapies (or similar drugs) are worth pursuing. Furthermore, while the published results of two trials using ATG differ (ITN “higher” dose trial being negative and TrialNet “lower” dose trial being positive), the QR point estimates of the active arms in each of these trials are similar, indicating that the differences in clinical trial outcomes reported were perhaps impacted by differences in the placebo arms rather than differences in efficacy between the two doses.

The canakinumab trial exemplifies the risks of comparison to a small control cohort. The originally reported canakinumab negative trial result had a detrimental impact on future studies; despite preclinical and mechanistic data suggesting a role of IL-1 in T1D, there has been reluctance to test this type of therapy further. However, we show here that the QR of individuals in the canakinumab-treated arm was positive, suggesting therapeutic effect. Our data indicates that the negative result in the originally published trial was due to the small number of placebo participants who performed much better than expected. Of note, in a retrospective analysis of the original study, Bundy et al. addressed this issue by comparing canakinumab-treated individuals to a larger placebo group, and also concluded that canakinumab may be effective. The integration of large amounts of historical data here provides added context to robustly interpret studies.

Perhaps the most powerful use of QR is its ability to determine the extent to which an individual responded to therapy. Participants are typically informed of clinical trial results with information about their own insulin secretion and the mean values for treatment and control groups. QR allows for standardized, subject-specific estimates to be provided to each participant; study staff can
describe the probability that the participant did better or worse than expected while on treatment (i.e., a responder or non-responder).

QR is also an improvement over previously used responder/non-responder definitions. Incorporating historical data via QR provides a greater level of certainty when identifying treatment responders. Standardized estimates based on historical placebo data can be used to understand the probability of observing a specific QR value in the absence of a treatment effect. Higher QR values are associated with increased confidence that an individual’s response is related to treatment. In the absence of a QR framework, we would be less certain about these predictions at both the individual and group level. Because QR has been defined for hundreds of untreated individuals, this increased confidence in defining treatment response can also enable design of clinical trials that use only active treatments. Under this scenario, assuming two active treatments with differing mechanisms of action, the QR can be used as a standardized instrument to discriminate biomarkers hypothesized to be causally related to treatment with the objective of personalizing immune therapies to specific endotypes.

QR is particularly useful for a more principled analysis of mechanistic data seeking to explain whether a mechanistic marker lies in the hypothesized causal pathway for the therapy. This concept is exemplified by analysis of exhausted T cells in individuals treated with teplizumab. Although the increase in these cells post-therapy is associated with C-peptide, data suggest that there is also a relationship between T cell exhaustion and age. Therefore, when evaluated by QR, which adjusts for age, the relationship between exhausted cells and response to therapy is less apparent. This provides an important insight, suggesting that understanding the causal pathway between teplizumab therapy and the induction of exhausted T cells must consider age as a factor, while also helping the field to consider the general phenomenon of why children may be more likely to respond to therapy.

Using data from 13 clinical trials and more than 1,300 participants, we demonstrate the significant value of using QR to advance the field of DMT in T1D. Our study serves as an example for applying the QR approach in other diseases that lack clarity in defining responders to therapy, for comparing the effectiveness of different therapies, and for understanding causal pathways in disease. Our analysis shows that the QR metric of insulin secretion measured by C-peptide is clinically and scientifically meaningful, objective, predictable, and standardized across individuals.
and cohorts, thus accelerating and aiding in interpretation of trials and providing a framework for precision medicine in T1D and other diseases.

**Methods**

**Datasets**

De-identified data were obtained from 13 clinical trials of DMT in individuals with recently diagnosed T1D (Table 1). These include six studies conducted by Diabetes TrialNet (TrialNet.org), an NIH-sponsored clinical trial network: MMF/DZB (TN02 NCT00100178 10), Rituximab (TN05 NCT00279305 9), GAD-Alum (TN08 NCT00529399 5), Abatacept (TN09 NCT00505375 4), Canakinumab (TN14 NCT009474276), Low Dose ATG/GCSF (TN19 NCT02215200 7); four studies conducted by the NIH funded Immune Tolerance Network (ITN; Immunetolerance.org): Teplizumab (AbATE NCT02067923 11), Alefacept (T1DAL NCT00965458 12), ATG (START NCT00515099 16), Tocilizumab (EXTEND NCT02293837 13); one investigator-initiated study sponsored by JDRF, Imatinib/Gleevec (NCT01781975 8); and two industry led studies, Diamyd Therapeutics AB: Phase 2 (NCT00435981 15) and Phase 3 (NCT00723411 14) GAD-Alum.

**Statistical Methods**

The ANCOVA model developed by Bundy and Krischer 21, using data from recently-diagnosed T1D individuals, $QR_i = \ln(Cp_{\text{1year},i} + 1) - 0.812 \cdot \ln(Cp_{0,i} + 1) - 0.00638 \cdot \text{Age}_i + 0.191$, was used to compute the individual’s QR, where $Cp_{0,i}$ and $Cp_{\text{1year},i}$ represent 2-hour C-peptide AUC mean (AUC divided by 120 minutes, in nanomoles per liter) at baseline and one year post treatment, respectively; $\text{Age}_i$ is the age at study entry, in years. Since the model assumes a linear relationship between baseline and 1 year C-peptide, we additionally computed QR at 3, 6, and 9 months post-randomization by deriving the expected C-peptide values at these timepoints from the original QR equation, and determining the difference between the expected and observed values at each timepoint.

To validate the QR method, we tested the model performance by applying the published ANCOVA model to data from 8 new studies not used for the development of QR (Figure 1). The Kolmogorov-Smirnov two-sample test was used to compare QR distributions between the development (n=5 studies) and validation (n=8 studies) cohorts. Additionally, an ANCOVA model was developed using all control participants to assess the association between actual 1 year C-peptide AUC mean...
values and predicted values from both the formula derived from the ANCOVA developed from our dataset and the published QR formula.

Participants were classified as either active treatment or placebo/control. Two-sample, two-sided t-tests were used for comparison of means between groups. For responder analyses, participants were dichotomized based on historical thresholds from the literature used to define responders and non-responders, and using a QR responder definition, where responders are individuals with positive QR and non-responders are individuals with negative QR. Generalized linear models with a binomial distribution and a logit link were fit among placebo/control participants, to each responder definition with adjustments for baseline age and baseline C-peptide. For biomarker analyses, Pearson correlations were computed to examine the association of KLRG1+TIGIT+ CD8+ T cells with baseline metrics (age and C-peptide), and with outcome metrics (QR and C-peptide).

Using control group data, additional ANCOVA models were developed to expand the utility of the QR method to different time intervals. Specifically, post-baseline predictions ranging from 3 months to 2 years were created using different baseline reference points and prediction horizons. Analysis was performed using SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA) and JMP Pro 16 (SAS Institute Inc., Cary, NC, USA).

**Data availability**

Data and code for analyses are available upon reasonable request from the authors.

**Supplementary materials**

Supplemental Figure 1-4.


All rights reserved. No reuse allowed without permission.


**Acknowledgments:** The authors gratefully acknowledge the access to data from Diamyd trials of GAD-alum from Diamyd Medical and for the Imatinib trial from Principal Investigator Stephen Gitelman, UCSF. Publicly available data was obtained from trials conducted by the Immune Tolerance Network (an international clinical research consortium headquartered at the Benaroya Research Institute and supported by the National Institute of Allergy and Infectious Diseases and JDRF) and from The Type 1 Diabetes TrialNet Study Group. We also thank Anne Hocking and Taylor Lawson (BRI) for thoughtful comments on this manuscript.

**Funding:** TrialNet is a clinical trials network funded by the National Institutes of Health (NIH) through the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Allergy and Infectious Diseases, and The Eunice Kennedy Shriver National Institute of Child Health and Human Development, through the cooperative agreements U01 DK061010,
U01 DK061034, U01 DK061042, U01 DK061058, U01 DK085461, U01 DK085465, U01 DK085466, U01 DK085476, U01 DK085499, U01 DK103180, U01 DK103153, U01 DK103266, U01 DK103282, U01 DK106984, U01 DK106994, U01 DK107013, U01 DK107014, UC4 DK106993, UC4DK117009.

JDRF grant 3-SRA-2019-791-S-B (to CS).

**Author contributions:**

Conceptualization: HTB, CJG

Data curation: AY, HTB

Formal analysis: AY, HTB, CO

Visualization: AY, HTB

Funding acquisition: CJG, CS

Writing – original draft: CJG, AY, HTB, CS

Writing – review & editing: CJG, AY, HTB, CO, SL, CS

**Competing interests:** The authors declare no conflicts for the submitted work. Outside of the submitted work CJG has served on Advisory Board for Merck, Viela Bio, Enthera Pharmaceuticals, GentiBio, Altheia Science and has research support from Bristol-Myers-Squibb, Janssen, and gave an educational presentation to AstraZeneca. CS serves on an Advisory Board for Vertex Pharmaceuticals. HTB has received contract research funding from DBV Technologies and MyOR, and is now an employee of DBV Technologies.
FIGURE LEGENDS

Figure 1. Validation of QR model: comparison of development and validation cohorts.
Placebo/control participants used for QR development (n=5 studies) and validation (n=8 studies).
The published QR model (QR = ln (Cp1year,i +1)−0.812⋅ln (Cp0,i +1)−0.00638⋅Agei +0.191) from the development studies (n=5) was applied to the validation studies (n=8). No significant difference was observed between the development and validation cohorts when comparing group means using a two-tailed t-test (p=0.44); both cohorts have similar distributions centered on zero as evaluated by a Kolmogorov-Smirnov test. Mean and SD are indicated by black lines and dark shaded region.

Figure 2. Use of model adjusting for baseline C-peptide and age reduces variance in outcome measure among placebo/control individuals.
Distribution across trials of baseline variables (A) of age and C-peptide AUC mean, and outcome measures (B) of 1 year C-peptide AUC mean and 1-year QR. Use of QR instead of 1-year C-peptide AUC mean reduces standard deviation (SD; blue) for each trial. Mean and SD are indicated by black lines and dark shaded region. Reference lines in panel A indicate average value in combined placebo participants: 16.4 years and baseline C-peptide of 0.73 nmol/L.

Figure 3. More statistical precision with use of QR in teplizumab trial.
Outcome data from teplizumab-treated vs control arm using outcome defined as (A) one-year C-peptide AUC mean (nmol/L) difference, p=0.009; (B) change from baseline C-peptide AUC mean (nmol/L) difference, p=0.0002; or (C) one-year QR difference, p<0.0001. Mean and SD are indicated by black lines and dark shaded region. P-values determined using two-tailed t-test.

Figure 4. Relationship between observed and expected outcome according to time of baseline and time of outcome allows for different trial durations.
R² using baseline C-peptide (within 3 months of diagnosis) and month 12 outcome from original QR model is 53%. Additional ANCOVA models were developed, using different baseline reference points and prediction horizons ranging from 3 months to 2 years. Using C-peptide and age at different time points from diagnosis can also predict outcome; the R² decreases the longer
from the initial measurement. Using predictors at later time points has strong association with outcomes illustrated up to 24 months.

**Figure 5. Use of QR to interpret across clinical trials.**

Mean QR values and 95% confidence interval (CI) for (A) active treatment, (B) control/placebo group, and (C) the difference between treatment arms for each trial. N for each treatment group is reported in Table 1.

**Figure 6. Effect of therapy determined at 6 months after randomization.**

Mean QR (± 95% confidence interval (CI)) at interval time points. At 6 months, all positive trials at 1 year had QR greater than 0 and negative trials below zero. Since the QR metric is based on a linear model, where baseline is predictive of 1 year C-peptide, QR at interim timepoints was computed by deriving the expected C-peptide values at specified timepoints from the original QR equation, and determining the difference between the expected and observed values.

**Figure 7. %KLRG1+TIGIT+ of CD8 T cells associated with age, not outcome of teplizumab therapy.**

C-peptide and %KLRG1+TIGIT+ of CD8 T cells (exhausted T cells) in actively-treated individuals (n=31) from ITN trial of teplizumab. Dashed line and shaded region indicate linear regression line and 95% confidence intervals. (A) There is a weak, but positive relationship between increase of exhausted T cells and one-year C-peptide; Pearson correlation © of 0.38, p=0.06. (B) No significant relationship between exhausted T cells and QR; r=0.27, p=0.20. (C) Correlation between exhausted T cells and age; r=0.47, p=0.007.

**Figure 8. Proportion of placebo/control individuals meeting responder definitions across trials.**

Generalized linear models fit in placebo/control participants, where baseline age (years) and baseline C-peptide (ln(x+1), nmol/L) are used to predict responder outcomes for QR-based responder definition and responder definitions reported in the literature. Mutually adjusted estimates for each predictor on the x-axis correspond to model predictions and 95% confidence intervals on the y-axis. The probability of a responder is benchmarked (indicated by vertical dashed red lines and annotated in red on x-axis) to an individual of average age (16.4 years) and baseline...
C-peptide (ln(Cp0+1) of 0.53 or 0.70 nmol/L); QR of 0.039. The probability (horizontal dashed red line) and 95% confidence interval of a placebo participant with these baseline characteristics being identified as a responder are annotated where dashed red lines intersect. (A) Inter-test SD definition: responders identified as those whose C-peptide change from baseline is nonnegative or negative but no more than 1 inter-test SD of 0.087 nmol/L below baseline (from 18,29. (B) 7.5% definition: responders defined as those with C-peptide decline of no more than 7.5% below baseline 28; (C) 40% definition: responders are those with <40% loss of baseline C-peptide 11; (D) CV definition: responders are those with nonnegative change from baseline or negative but coefficient of variation (CV) < 0.097 (median CV from 18,29. (E) QR >0 definition: responder is defined as positive QR.

Figure 9. QR values for control/placebo and actively-treated individuals in trials reported with positive outcomes inform ability to distinguish actively-treated from placebo.

(A) QR values for control/placebo (n=448) participants with corresponding percentiles. Blue line and shaded region indicate mean and standard deviation (SD). (B) QR values for active treatment individuals in trials reported with positive outcome (n=259). Blue line and shaded region indicate mean and SD. (C) Probability of identifying actively treated individuals with 95% confidence intervals. The distribution shown among participants treated with drugs from positive trials (B) or placebos (A) was used in a logistic regression model with active treatment or placebo as the outcome. The dataset was weighted to produce equal representation of active vs placebo participants to produce the probability curve shown on the right.
Figure 1. Validation of QR model: comparison of development and validation cohorts.

Placebo/control participants used for QR development (n=5 studies) and validation (n=8 studies).

The published QR model \( QR_i = \ln (C_{p1year,i} + 1) - 0.812 \cdot \ln (C_{p0,i} + 1) - 0.00638 \cdot \text{Age}_i + 0.191 \) from the development studies (n=5) was applied to the validation studies (n=8). No significant difference was observed between the development and validation cohorts when comparing group means using a two-tailed t-test (p=0.44); both cohorts have similar distributions centered on zero.
as evaluated by a Kolmogorov-Smirnov test. Mean and SD are indicated by black lines and dark shaded region.
Figure 2. Use of model adjusting for baseline C-peptide and age reduces variance in outcome measure among placebo/control individuals.

Distribution across trials of baseline variables (A) of age and C-peptide AUC mean, and outcome measures (B) of 1 year C-peptide AUC mean and 1-year QR. Use of QR instead of 1-year C-peptide AUC mean reduces standard deviation (SD; blue) for each trial. Mean and SD are indicated by black lines and dark shaded region. Reference lines in panel A indicate average value in combined placebo participants: 16.4 years and baseline C-peptide of 0.73 nmol/L.
Figure 3. More statistical precision with use of QR in teplizumab trial.

Outcome data from teplizumab-treated vs control arm using outcome defined as (A) one-year C-peptide AUC mean (nmol/L) difference, p=0.009; (B) change from baseline C-peptide AUC mean (nmol/L) difference, p=0.0002; or (C) one-year QR difference, p<0.0001. Mean and SD are indicated by black lines and dark shaded region. P-values determined using two-tailed t-test.
Figure 4. Relationship between observed and expected outcome according to time of baseline and time of outcome allows for different trial durations.

R² using baseline C-peptide (within 3 months of diagnosis) and month 12 outcome from original QR model is 53%. Additional ANCOVA models were developed, using different baseline reference points and prediction horizons ranging from 3 months to 2 years. Using C-peptide and age at different time points from diagnosis can also predict outcome; the R² decreases the longer from the initial measurement. Using predictors at later time points has strong association with outcomes illustrated up to 24 months.
**Figure 5. Use of QR to interpret across clinical trials.**

Mean QR values and 95% confidence interval (CI) for (A) active treatment, (B) control/placebo group, and (C) the difference between treatment arms for each trial. N for each treatment group is reported in Table 1.
Figure 6. Effect of therapy determined at 6 months after randomization.

Mean QR (± 95% confidence interval (CI)) at interval time points. At 6 months, all positive trials at 1 year had QR greater than 0 and negative trials below zero. Since the QR metric is based on a linear model, where baseline is predictive of 1 year C-peptide, QR at interim timepoints was computed by deriving the expected C-peptide values at specified timepoints from the original QR equation, and determining the difference between the expected and observed values.
Figure 7. %KLRG1+TIGIT+ of CD8 T cells associated with age, not outcome of teplizumab therapy.

C-peptide and %KLRG1+TIGIT+ of CD8 T cells (exhausted T cells) in actively-treated individuals (n=31) from ITN trial of teplizumab. Dashed line and shaded region indicate linear regression line and 95% confidence intervals. (A) There is a weak, but positive relationship between increase of exhausted T cells and one-year C-peptide; Pearson correlation (r) of 0.38, p=0.06. (B) No significant relationship between exhausted T cells and QR; r=0.27, p=0.20. (C) Correlation between exhausted T cells and age; r=0.47, p=0.007.
Figure 8. Proportion of placebo/control individuals meeting responder definitions across trials.

Generalized linear models fit in placebo/control participants, where baseline age (years) and baseline C-peptide ($\ln(x+1)$, nmol/L) are used to predict responder outcomes for QR-based responder definition and responder definitions reported in the literature. Mutually adjusted estimates for each predictor on the x-axis correspond to model predictions and 95% confidence intervals on the y-axis. The probability of a responder is benchmarked (indicated by vertical dashed red lines and annotated in red on x-axis) to an individual of average age (16.4 years) and baseline C-peptide ($\ln(Cp_0+1)$ of 0.53 or 0.70 nmol/L); QR of 0.039. The probability (horizontal dashed red line) and 95% confidence interval of a placebo participant with these baseline characteristics...
being identified as a responder are annotated where dashed red lines intersect. (A) Inter-test SD definition: responders identified as those whose C-peptide change from baseline is nonnegative or negative but no more than 1 inter-test SD of 0.087 nmol/L below baseline (from 18,29. (B) 7.5% definition: responders defined as those with C-peptide decline of no more than 7.5% below baseline 28; (C) 40% definition: responders are those with <40% loss of baseline C-peptide 11; (D) CV definition: responders are those with nonnegative change from baseline or negative but coefficient of variation (CV) < 0.097 (median CV from 18,29. (E) QR >0 definition: responder is defined as positive QR.
Figure 9. QR values for control/placebo and actively-treated individuals in trials reported with positive outcomes inform ability to distinguish actively-treated from placebo.

(A) QR values for control/placebo (n=448) participants with corresponding percentiles. Blue line and shaded region indicate mean and standard deviation (SD). (B) QR values for active treatment individuals in trials reported with positive outcome (n=259). Blue line and shaded region indicate mean and SD. (C) Probability of identifying actively treated individuals with 95% confidence intervals. The distribution shown among participants treated with drugs from positive trials (B) or placebos (A) was used in a logistic regression model with active treatment or placebo as the outcome. The dataset was weighted to produce equal representation of active vs placebo participants to produce the probability curve shown on the right.
<table>
<thead>
<tr>
<th>Trial</th>
<th>Start Year</th>
<th>N Treatment Arm(s)</th>
<th>N Control Arm</th>
<th>Age, median (range)</th>
<th>Baseline C-peptide AUC Mean (nmol/L), mean (SD)</th>
<th>Primary Outcome/Time Point</th>
<th>Primary Study Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diamyd Therapeutics AB: GAD-Alum Phase 2</td>
<td>2005</td>
<td>35</td>
<td>35</td>
<td>13.9 (10.1, 18.4)</td>
<td>0.66 (0.37)</td>
<td>Fasting C-peptide at 15 months</td>
<td>Negative</td>
</tr>
<tr>
<td>Diamyd Therapeutics AB: GAD-Alum Phase 3</td>
<td>2008</td>
<td>107 (2 doses), 109 (4 doses)</td>
<td>111</td>
<td>13.2 (10, 19.3)</td>
<td>0.66 (0.32)</td>
<td>2-hr MMTT AUC C-peptide at 15 months</td>
<td>Negative</td>
</tr>
<tr>
<td>Immune Tolerance Network (ITN): Anti-thymocyte globulin (ATG)</td>
<td>2007</td>
<td>38</td>
<td>20</td>
<td>17.5 (12, 35)</td>
<td>0.88 (0.42)</td>
<td>2-hr MMTT AUC C-peptide at 1 year</td>
<td>Negative</td>
</tr>
<tr>
<td>Immune Tolerance Network (ITN): Alefacept</td>
<td>2011</td>
<td>33</td>
<td>16</td>
<td>18 (12, 34)</td>
<td>0.78 (0.38)</td>
<td>2-hr MMTT AUC C-peptide at 1 year (secondary 4-hr MMTT AUC C-peptide at 1 year)</td>
<td>Negative primary; Positive secondary</td>
</tr>
<tr>
<td>Immune Tolerance Network (ITN): Teplizumab</td>
<td>2005</td>
<td>52</td>
<td>25</td>
<td>12 (8.3, 29.6)</td>
<td>0.7 (0.3)</td>
<td>4-hr MMTT AUC C-peptide at 2 years</td>
<td>Positive</td>
</tr>
<tr>
<td>Immune Tolerance Network (ITN): Tocilizumab</td>
<td>2015</td>
<td>88</td>
<td>47</td>
<td>14 (6, 45)</td>
<td>0.76 (0.43)</td>
<td>2-hr MMTT AUC C-peptide at 1 year</td>
<td>Negative</td>
</tr>
<tr>
<td>JDRF: Imatinib</td>
<td>2014</td>
<td>43</td>
<td>21</td>
<td>24.5 (18.3, 45)</td>
<td>0.84 (0.4)</td>
<td>2-hr MMTT AUC C-peptide at 1 year</td>
<td>Positive</td>
</tr>
<tr>
<td>TrialNet: Abatacept</td>
<td>2008</td>
<td>76</td>
<td>35</td>
<td>12.9 (6.4, 36.8)</td>
<td>0.75 (0.38)</td>
<td>2-hr MMTT AUC C-peptide at 2 years</td>
<td>Positive</td>
</tr>
<tr>
<td>TrialNet: Canakinumab</td>
<td>2010</td>
<td>48</td>
<td>22</td>
<td>11 (6, 31.9)</td>
<td>0.64 (0.33)</td>
<td>2-hr MMTT AUC C-peptide at 1 year</td>
<td>Negative</td>
</tr>
<tr>
<td>TrialNet: GAD-Alum</td>
<td>2009</td>
<td>49 (2 doses), 48 (3 doses)</td>
<td>48</td>
<td>15.1 (3.5, 45.7)</td>
<td>0.73 (0.32)</td>
<td>2-hr MMTT AUC C-peptide at 1 year</td>
<td>Negative</td>
</tr>
<tr>
<td>TrialNet: Low Dose ATG/GCSF</td>
<td>2014</td>
<td>29 (ATG only), 28 (ATG/GCSF)</td>
<td>31</td>
<td>16 (12, 42.5)</td>
<td>0.89 (0.44)</td>
<td>2-hr MMTT AUC C-peptide at 1 year</td>
<td>Positive</td>
</tr>
<tr>
<td>TrialNet: MMF/DZB</td>
<td>2004</td>
<td>31 (MMF only), 41 (MMF/DZB)</td>
<td>42</td>
<td>14.9 (8.7, 46.1)</td>
<td>0.69 (0.32)</td>
<td>2-hr MMTT AUC C-peptide at 1 year</td>
<td>Negative</td>
</tr>
<tr>
<td>TrialNet: Rituximab</td>
<td>2005</td>
<td>55</td>
<td>30</td>
<td>16 (8.3, 40.4)</td>
<td>0.75 (0.39)</td>
<td>2-hr MMTT AUC C-peptide at 1 year</td>
<td>Positive</td>
</tr>
</tbody>
</table>