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

Authors: Alyssa Ylescupidez[1†], Henry T. Bahnson[1†], Colin O'Rourke[1], Sandra Lord[1], Cate Speake[1‡], Carla J. Greenbaum[1*‡]

Affiliations:

[1]Center 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.

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 [1].

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) [2-16]. Moreover, though it is noteworthy that a therapy to delay onset of clinically apparent disease was recently approved for clinical use [17], 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 [18,19], 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 [20]. 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 [21], leverages the well-known statistical property that model adjustment with prognostic baseline covariates increases precision and confidence by way of controlling for outcome heterogeneity [1]. Bundy and Krischer used data from five trials [4-6,9,10] 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 [21].

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) [4-6,9,10], and eight additional RCTs (referred to as the validation cohort) [7,8,11-16]. 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 [1]. 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 [11]. 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)

121 using a two-sample t-test assuming unequal group variances (Figure 3A). When using the change
122 from baseline C-peptide AUC mean as in the published trial report, the effect size and precision
123 increased with a mean change from baseline of -0.321 nmol/L in the control group and -0.086
124 nmol/L in the teplizumab-treated group (p=0.0002, t=4.0, DF=57.0) (Figure 3B). Controlling for
125 both baseline C-peptide and age by using QR as the outcome further increased the statistical
126 significance of the result (-0.015 nmol/L control vs 0.141 nmol/L teplizumab-treated, p<0.0001,
127 t=4.3, DF=47.3) (Figure 3C).

128 We next assessed whether the QR approach could be utilized to predict trial outcomes beyond 1
129 year, using varying baseline reference points and outcome timepoints. The original QR model
130 used baseline values from individuals within three months of diagnosis to predict the outcome at
131 12 months after treatment initiation and demonstrated an R^2 value of 53%. Predictions further in
132 the future are more challenging, thus it is not surprising that the R^2 drops when this same baseline
133 is used to predict a 24-month outcome (Figure 4). However, using 6-, 12-, or 18-month data as
134 baseline C-peptide, the R^2 value at two years is high (74%, 85%, 87%, respectively), suggesting
135 that QR can be used as an outcome measure for trials enrolling individuals further from diagnosis.

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

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

145 Expressing the overall treatment effect and results of each arm using QR altered the interpretation
146 of some of the published results. For example, the primary outcome of the alefacept trial was the
147 2-hour C-peptide AUC at 1 year, and the difference between treatment arms did not reach statistical
148 significance in the original analysis [12]. However, applying the QR metric to the alefacept trial
149 dataset demonstrated a large effect in the active treatment group (Figure 5A), strongly suggesting
150 that alefacept, or drugs working in the same pathway, are worth pursuing in future trials.

151 For the canakinumab trial, assessing the overall trial result by the difference in treatment arms
152 using the QR metric finds no treatment effect (Figure 5C), consistent with the published outcome
153 [6]. Yet, the QR of the active arm of the canakinumab trial suggests a positive effect of this therapy
154 on C-peptide (Figure 5A). Moreover, the QR of the placebo arm allowed us to further interpret
155 this result, revealing that the lack of statistical significance between the groups may be driven by
156 higher-than-expected C-peptide response in the 22 individuals in the placebo arm of the study
157 (Figure 5B).

158 Lastly, applying the QR metric to the two studies testing anti-thymocyte globulin (ATG) also
159 suggests a different interpretation than originally published. The TrialNet ATG/GCSF trial
160 reported a positive outcome [7] using lower-dose therapy than the ITN ATG study, which did not
161 meet its primary outcome [22], leading to the interpretation that dose level was the key variable in
162 the effectiveness of the drug. However, the mean QR in the treated participants was notably similar
163 between studies: mean (95% CI) QR was 0.08 (0.02-0.13) in the higher-dose ITN ATG trial, 0.09
164 (0.03-0.15) in the lower-dose TrialNet ATG/GCSF trial. This suggests that the reported difference
165 in treatment effect between the studies was driven by the placebo participants: those in the TrialNet
166 ATG/GCSF study had a fairly low mean QR (-0.07) while those in the ITN ATG study had a
167 higher mean QR (0.05). Of note, across the 13 studies, only the TrialNet ATG/GCSF control arm
168 was statistically significantly different from zero. This further demonstrates how random sampling
169 variability in smaller studies can complicate interpretation of RCT results.

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

174 Next, we investigated whether using the QR metric would impact the results of immune marker
175 studies aiming to explore mechanisms of response to therapy. Using C-peptide and mechanistic
176 results obtained from the ITN AbATE trial of teplizumab, we confirmed previous reports of the
177 positive association between the frequency of treatment-induced KLRG1+TIGIT+ CD8+ T cells,
178 a known signature of T cell exhaustion, and C-peptide outcome (Figure 7A) [23]. However, adjusting
179 for baseline C-peptide and age by using the QR metric revealed that the association between
180 treatment-induced KLRG1+TIGIT+ CD8+ T cells and outcome may be weaker than originally

181 thought (Figure 7B). This observation is likely accounted for in part by a moderate association
182 between treatment-induced exhausted T cells and age (Figure 7C). As previously noted, age is one
183 of two key variables in the QR metric; age is also known to be important in defining setpoints and
184 responsiveness for many immune cell populations (recently reviewed in [24-27]). This novel analysis
185 implies that therapy-induced exhaustion of T cells unveils mechanistic insights about *age* itself,
186 which may or may not be causally related to a particular therapy, but is important to our
187 understanding of the role age plays in disease progression and response to therapy. Supplemental
188 Figure 3 graphically illustrates why QR is a more powerful metric for identification of a biomarker
189 that is causally related to therapy.

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

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

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

206 To further illustrate the consequences of not accounting for baseline variables in classifying
207 responders, we benchmarked the probability of being a responder for each definition using the data
208 from an "average" treated individual. The average age (16.4 years) and average baseline C-peptide
209 AUC mean for treated individuals across all 13 studies was determined, yielding a QR value of
210 0.039 for an average individual. However, the probability that this individual is defined as a

211 treatment responder varies widely using different responder definitions (Figure 8B and C). Most
212 importantly, the probability of being a treatment responder is strongly associated with age, baseline
213 C-peptide, or both of these metrics for all non-QR definitions. Since the QR metric adjusts for age
214 and baseline C-peptide, the probability of being a responder is not conditional on these factors, as
215 can also be seen from the annotated p-values (Figure 8E).

216 Given that the probability of being a treatment responder is not conditional on age and baseline C-
217 peptide, we asked how the QR metric could be used to select a threshold for classifying responders
218 and non-responders to therapy. In selecting a threshold for a continuous measure such as QR, it is
219 useful to understand the variability or confidence intervals around a given QR value, reflecting the
220 probability that a given QR value represents a true treatment responder. Here, we observed that
221 the distribution of the QR scores of all placebo/control individuals is symmetrical (Figure 9A),
222 leading to quantile statistics whereby the QR value can be assigned to a percentile (e.g., a QR of
223 0.10 corresponds to the 75th percentile of the distribution). Figure 9A also illustrates that while
224 there is a symmetrical distribution around the mean, heterogeneity is also apparent; a placebo-
225 treated individual may have a QR value ranging from -0.58 to 0.45. Similarly, though the mean
226 QR value among all individuals in the treatment arms of the positive studies is above zero, there
227 is also a wide range of values in each treated group (Figure 9B), many of which overlap with the
228 distribution of placebo-treated individuals.

229 The overlapping QR values between treatment and placebo groups demonstrate that using a
230 particular QR cutoff will not necessarily distinguish individuals who received an efficacious
231 therapy from placebo individuals. We reasoned that these distributions can be used to understand
232 the probability that a specific QR value is associated with a successful treatment (Figure 9C). As
233 shown in Figure 9C, the probability of identifying a treatment responder or non-responder
234 increases at the extremes of the distribution. For example, the probability that an individual with
235 QR=0.4 received an effective therapy exceeds 80% (Figure 9A). Conversely, the probability that
236 an individual with QR= -0.4 received an effective therapy is only 15%: at this threshold,
237 individuals are more likely to be placebo participants, and thus we can infer that a treated
238 individual with this QR value was likely a non-responder to therapy. Choosing less extreme cut
239 points introduces greater uncertainty. For example, selecting a threshold of treatment response of
240 QR=0.2 would yield only a 65% probability that this individual received an effective therapy.

Discussion

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 [21,30] 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.

270 However, it is not clear that reducing the sample size for a phase 2 RCT is the optimal approach
271 to select promising therapies, or to identify responders for T1D or other diseases. Placebo-
272 controlled randomized trials have clear advantages, as randomization can account for potential
273 differences in variables that are known to impact outcome. However, when small sample sizes are
274 used, random sampling variation can significantly impact inferences about trial outcomes. In the
275 case of trials of DMT aiming to preserve C-peptide, the known factors are baseline C-peptide and
276 age; QR adjusts for these factors.

277 Using data from almost 500 control/placebo individuals in the 13 trials studied, we show that
278 untreated individuals' outcomes are reliably predicted. Utilizing QR as an outcome measure
279 implies comparison of a treatment arm to this large number of historical controls, and may obviate
280 the need for a contemporaneous control group. Single- or multiple-active arm trials are likely to
281 conserve resources by eliminating or minimizing placebo participants while accelerating
282 recruitment (as more participants agree to trials without a placebo arm). This paradigm, common
283 in clinical trial settings including cancer research, increases the likelihood that participants will
284 benefit from prospective therapeutic interventions. The QR framework achieves these advantages
285 by leveraging large amounts of historical data to create "synthetic controls" to test promising
286 interventions in phase 2 trials, which could then be reproduced in larger phase 3 placebo-controlled
287 pivotal randomized trials [31,32].

288 Furthermore, using a QR outcome allows for adaptive study designs of multiple active agents, as
289 we found that a mean QR value above zero in the treatment arm at six months after randomization
290 completely predicted the success of all the tested RCTs with a positive outcome at 1 year. A trial
291 with multiple active agents could drop ineffective therapeutic arms at six months. Using QR could
292 enable shorter clinical trials, which would reduce burden on study staff and participants, reduce
293 cost, and reduce the time that participants in the active arm are exposed to ineffective therapies.
294 By pre-specifying a QR threshold of interest at an early timepoint, adaptive re-randomization
295 designs, such as the sequential parallel comparison design (SPCD; Supplemental Figure 4) [33], are
296 also feasible. This design identifies placebo participants with a QR value below zero early in a
297 trial, and re-randomizes those individuals to treatment or placebo, allowing a larger number of
298 participants to potentially benefit from therapy. QR can also enable enrollment of individuals

299 outside the traditionally used period of 3 months post-diagnosis in new-onset T1D trials, as it can
300 reliably predict a 2-year outcome when a baseline timepoint is >6 months from diagnosis.

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

315 The canakinumab trial exemplifies the risks of comparison to a small control cohort. The originally
316 reported canakinumab negative trial result had a detrimental impact on future studies; despite pre-
317 clinical and mechanistic data suggesting a role of IL-1 in T1D[34-37], there has been reluctance to
318 test this type of therapy further. However, we show here that the QR of individuals in the
319 canakinumab-treated arm was positive, suggesting therapeutic effect. Our data indicates that the
320 negative result in the originally published trial was due to the small number of placebo participants
321 who performed much better than expected. Of note, in a retrospective analysis of the original study,
322 Bundy et al. addressed this issue by comparing canakinumab-treated individuals to a larger placebo
323 group, and also concluded that canakinumab may be effective [30]. The integration of large amounts
324 of historical data here provides added context to robustly interpret studies.

325 Perhaps the most powerful use of QR is its ability to determine the extent to which an individual
326 responded to therapy. Participants are typically informed of clinical trial results with information
327 about their own insulin secretion and the mean values for treatment and control groups. QR allows
328 for standardized, subject-specific estimates to be provided to each participant; study staff can

329 describe the probability that the participant did better or worse than expected while on treatment
330 (i.e., a responder or non-responder).

331 QR is also an improvement over previously used responder/non-responder definitions.
332 Incorporating historical data via QR provides a greater level of certainty when identifying
333 treatment responders. Standardized estimates based on historical placebo data can be used to
334 understand the probability of observing a specific QR value in the absence of a treatment effect.
335 Higher QR values are associated with increased confidence that an individual's response is related
336 to treatment. In the absence of a QR framework, we would be less certain about these predictions
337 at both the individual and group level. Because QR has been defined for hundreds of untreated
338 individuals, this increased confidence in defining treatment response can also enable design of
339 clinical trials that use only active treatments. Under this scenario, assuming two active treatments
340 with differing mechanisms of action, the QR can be used as a standardized instrument to
341 discriminate biomarkers hypothesized to be causally related to treatment with the objective of
342 personalizing immune therapies to specific endotypes [38].

343 QR is particularly useful for a more principled analysis of mechanistic data seeking to explain
344 whether a mechanistic marker lies in the hypothesized causal pathway for the therapy. This
345 concept is exemplified by analysis of exhausted T cells in individuals treated with teplizumab.
346 Although the increase in these cells post-therapy is associated with C-peptide, data suggest that
347 there is also a relationship between T cell exhaustion and age. Therefore, when evaluated by QR,
348 which adjusts for age, the relationship between exhausted cells and response to therapy is less
349 apparent. This provides an important insight, suggesting that understanding the causal pathway
350 between teplizumab therapy and the induction of exhausted T cells must consider age as a factor
351 [39], while also helping the field to consider the general phenomenon of why children may be more
352 likely to respond to therapy.

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

359 and cohorts, thus accelerating and aiding in interpretation of trials and providing a framework for
360 precision medicine in T1D and other diseases.

361 **Methods**

362 *Datasets*

363 De-identified data were obtained from 13 clinical trials of DMT in individuals with recently
364 diagnosed T1D (Table 1). These include six studies conducted by Diabetes TrialNet (TrialNet.org),
365 an NIH-sponsored clinical trial network: MMF/DZB (TN02 NCT00100178 [10]), Rituximab (TN05
366 NCT00279305 [9]), GAD-Alum (TN08 NCT00529399 [5]), Abatacept (TN09 NCT00505375 [4]),
367 Canakinumab (TN14 NCT00947427[6]), Low Dose ATG/GCSF (TN19 NCT02215200 [7]); four
368 studies conducted by the NIH funded Immune Tolerance Network (ITN; Immunetolerance.org):
369 Teplizumab (AbATE NCT02067923 [11]), Alefacept (T1DAL NCT00965458 [12]), ATG (START
370 NCT00515099 [16]), Tocilizumab (EXTEND NCT02293837 [13]); one investigator-initiated study
371 sponsored by JDRF, Imatinib/Gleevec (NCT01781975 [8]); and two industry led studies, Diamyd
372 Therapeutics AB: Phase 2 (NCT00435981 [15]) and Phase 3 (NCT00723411 [14]) GAD-Alum.

373 *Statistical Methods*

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

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

388 values and predicted values from both the formula derived from the ANCOVA developed from
389 our dataset and the published QR formula.

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

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

405 **Data availability**

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

407 **Supplementary materials**

408 Supplemental Figure 1-4.

409

410

411 **References**

412

413 1 Kahan, B. C., Jairath, V., Dore, C. J. & Morris, T. P. The risks and rewards of covariate
414 adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials* **15**,
415 139 (2014). https://doi.org:10.1186/1745-6215-15-139
416 2 von Herrath, M. *et al.* Anti-interleukin-21 antibody and liraglutide for the preservation of
417 beta-cell function in adults with recent-onset type 1 diabetes: a randomised, double-blind,
418 placebo-controlled, phase 2 trial. *Lancet Diabetes Endocrinol* **9**, 212-224 (2021).
419 https://doi.org:10.1016/S2213-8587(21)00019-X
420 3 Quattrin, T. H., M.J., Steck, A.K., Felner, E.I., Li, Y., Xia, Y., Leu, ,J.H., Zoka, R., Hedrick,
421 J.A., Rigby, M.R., Vercruysse, F.; T1GER Study Investigators. Golimumab and beta-cell
422 function in youth with new-onset type 1 diabetes. *NEJM* **383**, 2007-2017 (2020).
423 https://doi.org:10.1056/NEJMoa2006136.
424 4 Orban, T. *et al.* Co-stimulation modulation with abatacept in patients with recent-onset
425 type 1 diabetes: a randomised, double-blind, placebo-controlled trial. *Lancet* **378**, 412-419
426 (2011). https://doi.org:10.1016/S0140-6736(11)60886-6
427 5 Wherrett, D. K. *et al.* Antigen-based therapy with glutamic acid decarboxylase (GAD)
428 vaccine in patients with recent-onset type 1 diabetes: a randomised double-blind trial.
429 *Lancet* **378**, 319-327 (2011). https://doi.org:10.1016/S0140-6736(11)60895-7
430 6 Moran, A. *et al.* Interleukin-1 antagonism in type 1 diabetes of recent onset: two
431 multicentre, randomised, double-blind, placebo-controlled trials. *Lancet* **381**, 1905-1915
432 (2013). https://doi.org:10.1016/S0140-6736(13)60023-9
433 7 Haller, M. J. *et al.* Low-Dose Anti-Thymocyte Globulin (ATG) Preserves beta-Cell
434 Function and Improves HbA1c in New-Onset Type 1 Diabetes. *Diabetes Care* **41**, 1917-
435 1925 (2018). https://doi.org:10.2337/dc18-0494
436 8 Gitelman, S. E. *et al.* Imatinib therapy for patients with recent-onset type 1 diabetes: a
437 multicentre, randomised, double-blind, placebo-controlled, phase 2 trial. *Lancet Diabetes*
438 *Endocrinol* **9**, 502-514 (2021). https://doi.org:10.1016/S2213-8587(21)00139-X
439 9 Pescovitz, M. D. *et al.* Rituximab, B-Lymphocyte Depletion, and Preservation of Beta-Cell
440 Function. *N. Engl. J. Med* **361**, 2143-2152 (2009).
441 10 Gottlieb, P. A. *et al.* Failure to preserve beta-cell function with mycophenolate mofetil and
442 daclizumab combined therapy in patients with new- onset type 1 diabetes. *Diabetes Care*
443 **33**, 826-832 (2010). https://doi.org:10.2337/dc09-1349
444 11 Herold, K. C. *et al.* Teplizumab (anti-CD3 mAb) treatment preserves C-peptide responses
445 in patients with new-onset type 1 diabetes in a randomized controlled trial: metabolic and
446 immunologic features at baseline identify a subgroup of responders. *Diabetes* **62**, 3766-
447 3774 (2013). https://doi.org:10.2337/db13-0345
448 12 Rigby, M. R. *et al.* Targeting of memory T cells with alefacept in new-onset type 1 diabetes
449 (T1DAL study): 12 month results of a randomised, double-blind, placebo-controlled phase
450 2 trial. *Lancet Diabetes Endocrinol* **1**, 284-294 (2013). https://doi.org:10.1016/S2213-
451 8587(13)70111-6
452 13 Greenbaum, C. J. *et al.* IL-6 receptor blockade does not slow beta cell loss in new-onset
453 type 1 diabetes. *JCI Insight* **6** (2021). https://doi.org:10.1172/jci.insight.150074
454 14 Ludvigsson, J. *et al.* GAD65 antigen therapy in recently diagnosed type 1 diabetes mellitus.
455 *N Engl J Med* **366**, 433-442 (2012). https://doi.org:10.1056/NEJMoa1107096

15 Ludvigsson, J. *et al.* GAD treatment and insulin secretion in recent-onset type 1 diabetes. *N. Engl. J. Med* **359**, 1909-1920 (2008).
16 Gitelman, S. G., PA; Rigby, MR et al. Antithymocyte globulin treatment for patients with recent-onset type 1 diabetes: 12-month results of a randomised, placebo-controlled, phase 2 trial. *Lancet Diabetes and Endocrinology* **1**, 306-316 (2013).
17 Herold, K. C. *et al.* An Anti-CD3 Antibody, Teplizumab, in Relatives at Risk for Type 1 Diabetes. *N Engl J Med* **381**, 603-613 (2019). https://doi.org:10.1056/NEJMoa1902226
18 Greenbaum, C. J. *et al.* Mixed-meal tolerance test versus glucagon stimulation test for the assessment of beta-cell function in therapeutic trials in type 1 diabetes. *Diabetes Care* **31**, 1966-1971 (2008).
19 Palmer, J. P. *et al.* C-peptide is the appropriate outcome measure for type 1 diabetes clinical trials to preserve beta-cell function: report of an ADA workshop, 21-22 October 2001. *Diabetes* **53**, 250-264 (2004).
20 https://c-path.org/design-of-clinical-trials-in-new-onset-type-1-diabetes-regulatory-considerations-for-drug-development/. *Design of Clinical Trials in New-Onset Type 1 Diabetes: Regulatory Considerations for Drug Development*, 2021).
21 Bundy, B. N., Krischer, J. P. & Type 1 Diabetes TrialNet Study Group. A quantitative measure of treatment response in recent-onset type 1 diabetes. *Endocrinol Diabetes Metab* **3**, e00143 (2020). https://doi.org:10.1002/edm2.143
22 Gitelman, S. E. *et al.* Antithymocyte globulin therapy for patients with recent-onset type 1 diabetes: 2 year results of a randomised trial. *Diabetologia* **59**, 1153-1161 (2016). https://doi.org:10.1007/s00125-016-3917-4
23 Long, S. A. *et al.* Partial exhaustion of CD8 T cells and clinical response to teplizumab in new-onset type 1 diabetes. *Sci Immunol* **1** (2016). https://doi.org:10.1126/sciimmunol.aai7793
24 Mogilenko, D. A., Shchukina, I. & Artyomov, M. N. Immune ageing at single-cell resolution. *Nat Rev Immunol* **22**, 484-498 (2022). https://doi.org:10.1038/s41577-021-00646-4
25 Davenport, M. P., Smith, N. L. & Rudd, B. D. Building a T cell compartment: how immune cell development shapes function. *Nat Rev Immunol* **20**, 499-506 (2020). https://doi.org:10.1038/s41577-020-0332-3
26 Shaw, A. C., Goldstein, D. R. & Montgomery, R. R. Age-dependent dysregulation of innate immunity. *Nat Rev Immunol* **13**, 875-887 (2013). https://doi.org:10.1038/nri3547
27 Roe, K. NK-cell exhaustion, B-cell exhaustion and T-cell exhaustion-the differences and similarities. *Immunology* **166**, 155-168 (2022). https://doi.org:10.1111/imm.13464
28 Herold, K. C. *et al.* A single course of anti-CD3 monoclonal antibody hOKT3gamma1(Ala-Ala) results in improvement in C-peptide responses and clinical parameters for at least 2 years after onset of type 1 diabetes. *Diabetes* **54**, 1763-1769 (2005).
29 Hao, W. *et al.* Fall in C-Peptide During First 4 Years From Diagnosis of Type 1 Diabetes: Variable Relation to Age, HbA1c, and Insulin Dose. *Diabetes Care* **39**, 1664-1670 (2016). https://doi.org:10.2337/dc16-0360
30 Bundy, B. N., Krischer, J. P. & Type 1 Diabetes TrialNet Study Group. A model-based approach to sample size estimation in recent onset type 1 diabetes. *Diabetes Metab Res Rev* **32**, 827-834 (2016). https://doi.org:10.1002/dmrr.2800

31 Viele, K. *et al.* Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat* **13**, 41-54 (2014). https://doi.org:10.1002/pst.1589

32 Freidlin, B. & Korn, E. L. Augmenting Randomized Clinical Trial Data with Historical Control Data: Precision Medicine Applications. *J Natl Cancer Inst* (2022). https://doi.org:10.1093/jnci/djac185

33 Fava, M., Evins, A. E., Dorer, D. J. & Schoenfeld, D. A. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychother Psychosom* **72**, 115-127 (2003). https://doi.org:10.1159/000069738

34 Wang, X. *et al.* Identification of a molecular signature in human type 1 diabetes mellitus using serum and functional genomics. *J Immunol* **180**, 1929-1937 (2008). https://doi.org:10.4049/jimmunol.180.3.1929

35 Cabrera, S. M. *et al.* Interleukin-1 antagonism moderates the inflammatory state associated with Type 1 diabetes during clinical trials conducted at disease onset. *Eur J Immunol* **46**, 1030-1046 (2016). https://doi.org:10.1002/eji.201546005

36 Padgett LE, B. K., Hansen PA, Corbett JA, Tse HM. The role of reactive oxygen species and proinflammatory cytokines in type 1 diabetes pathogenesis. *Ann N Y Acad Sci* **1281(1)**, 16-35 (2013).

37 Mandrup-Poulsen, T., Pickersgill, L. & Donath, M. Y. Blockade of interleukin 1 in type 1 diabetes mellitus. *Nat Rev Endocrinol* **6**, 158-166 (2010). https://doi.org:10.1038/nrendo.2009.271

38 Battaglia, M. *et al.* Introducing the Endotype Concept to Address the Challenge of Disease Heterogeneity in Type 1 Diabetes. *Diabetes Care* **43**, 5-12 (2020). https://doi.org:10.2337/dc19-0880

39 Pearl, J. & Mackenzie, D. *The book of why : the new science of cause and effect*. First edition. Basic Books, 2018.

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 DK085509, 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.

562 **FIGURE LEGENDS**

563 **Figure 1. Validation of QR model: comparison of development and validation cohorts.**

564 Placebo/control participants used for QR development (n=5 studies) and validation (n=8 studies).
565 The published QR model ($QR_i = \ln(Cp_{1year,i} + 1) - 0.812 \cdot \ln(Cp_{0,i} + 1) - 0.00638 \cdot Age_i + 0.191$) from
566 the development studies (n=5) was applied to the validation studies (n=8). No significant
567 difference was observed between the development and validation cohorts when comparing group
568 means using a two-tailed t-test (p=0.44); both cohorts have similar distributions centered on zero
569 as evaluated by a Kolmogorov-Smirnov test. Mean and SD are indicated by black lines and dark
570 shaded region.

571 **Figure 2. Use of model adjusting for baseline C-peptide and age reduces variance in outcome**
572 **measure among placebo/control individuals.**

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

578 **Figure 3. More statistical precision with use of QR in teplizumab trial.**

579 Outcome data from teplizumab-treated vs control arm using outcome defined as (**A**) one-year C-
580 peptide AUC mean (nmol/L) difference, p=0.009; (**B**) change from baseline C-peptide AUC mean
581 (nmol/L) difference, p=0.0002; or (**C**) one-year QR difference, p<0.0001. Mean and SD are
582 indicated by black lines and dark shaded region. P-values determined using two-tailed t-test.

583 **Figure 4. Relationship between observed and expected outcome according to time of baseline**
584 **and time of outcome allows for different trial durations.**

585 R^2 using baseline C-peptide (within 3 months of diagnosis) and month 12 outcome from original
586 QR model is 53%. Additional ANCOVA models were developed, using different baseline
587 reference points and prediction horizons ranging from 3 months to 2 years. Using C-peptide and
588 age at different time points from diagnosis can also predict outcome; the R^2 decreases the longer

589 from the initial measurement. Using predictors at later time points has strong association with
590 outcomes illustrated up to 24 months.

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

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

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

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

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

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

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

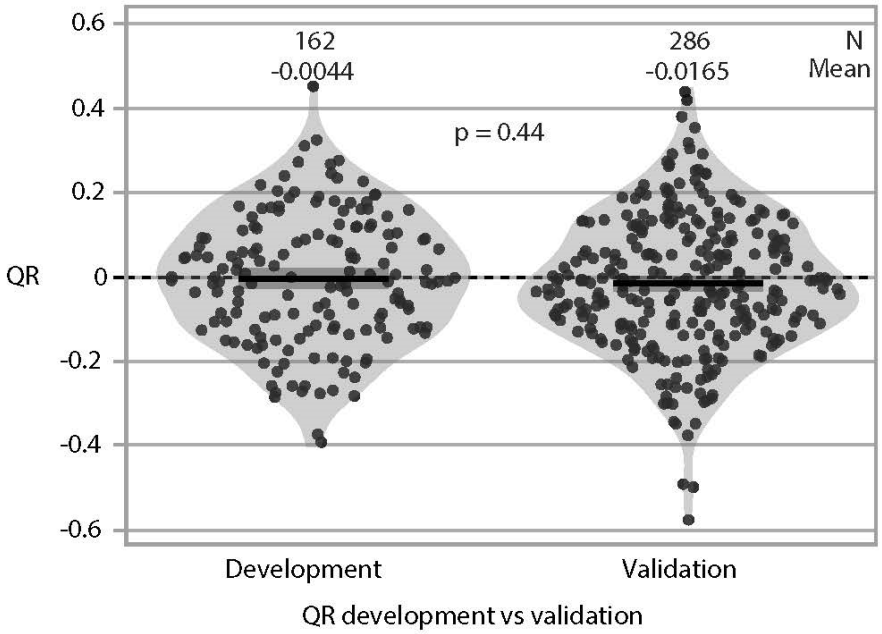
611 Generalized linear models fit in placebo/control participants, where baseline age (years) and
612 baseline C-peptide (ln(x+1), nmol/L) are used to predict responder outcomes for QR-based
613 responder definition and responder definitions reported in the literature. Mutually adjusted
614 estimates for each predictor on the x-axis correspond to model predictions and 95% confidence
615 intervals on the y-axis. The probability of a responder is benchmarked (indicated by vertical dashed
616 red lines and annotated in red on x-axis) to an individual of average age (16.4 years) and baseline

617 C-peptide ($\ln(Cp_0+1)$ of 0.53 or 0.70 nmol/L); QR of 0.039. The probability (horizontal dashed
618 red line) and 95% confidence interval of a placebo participant with these baseline characteristics
619 being identified as a responder are annotated where dashed red lines intersect. (**A**) Inter-test SD
620 definition: responders identified as those whose C-peptide change from baseline is nonnegative or
621 negative but no more than 1 inter-test SD of 0.087 nmol/L below baseline (from [18,29]. (**B**) 7.5%
622 definition: responders defined as those with C-peptide decline of no more than 7.5% below
623 baseline [28]; (**C**) 40% definition: responders are those with <40% loss of baseline C-peptide [11]; (**D**)
624 CV definition: responders are those with nonnegative change from baseline or negative but
625 coefficient of variation (CV) < 0.097 (median CV from [18,29]. (**E**) QR >0 definition: responder is
626 defined as positive QR.

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

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

637 **Figures**

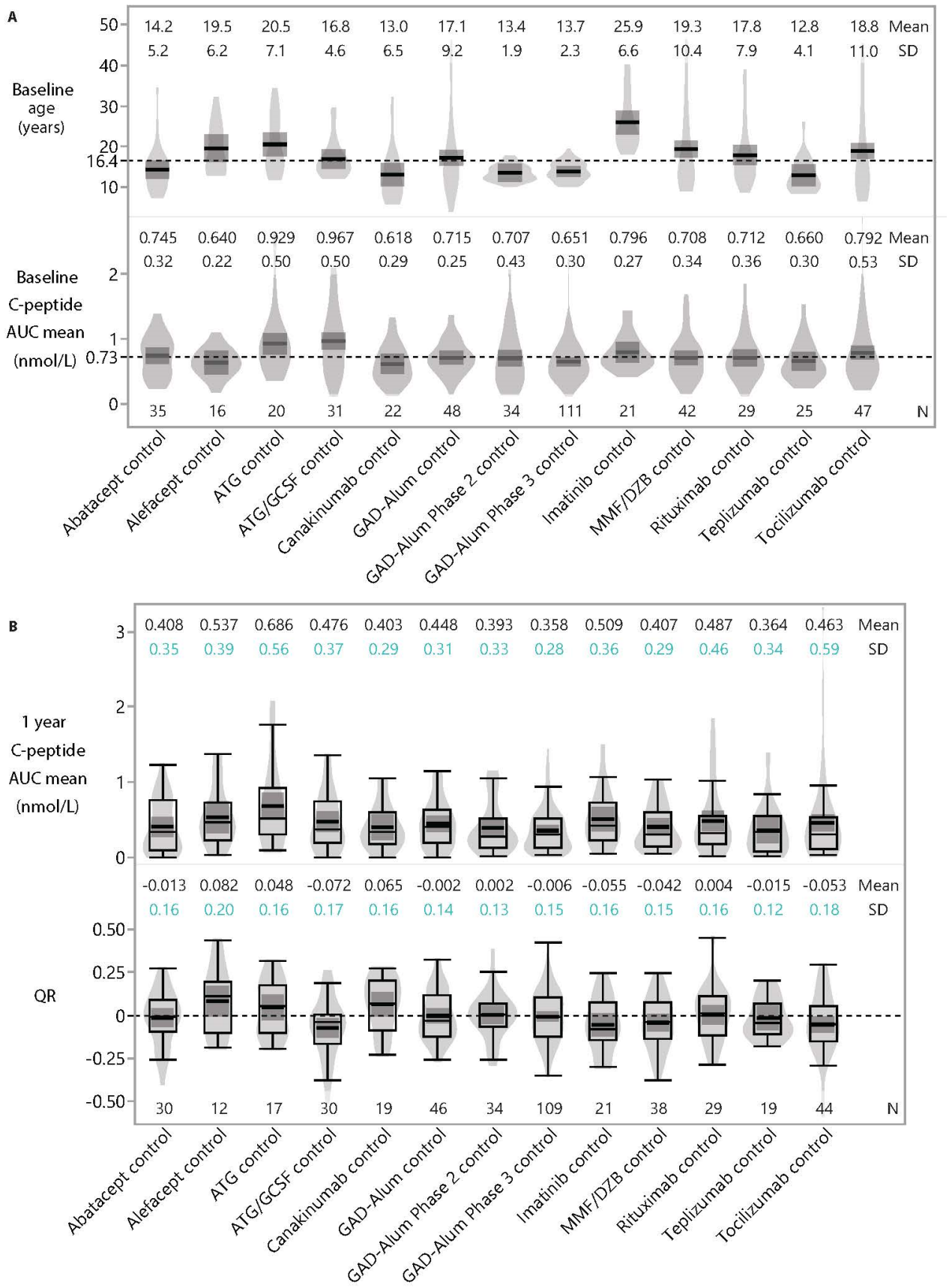


QR development vs validation	Trial	N
Development	TrialNet: Abatacept	30
Development	TrialNet: Canakinumab	19
Development	TrialNet: GAD-Alum	46
Development	TrialNet: MMF/DZB	38
Development	TrialNet: Rituximab	29
Validation	Diamyd Therapeutics AB: GAD-Alum Phase 2	34
Validation	Diamyd Therapeutics AB: GAD-Alum Phase 3	109
Validation	ITN: Alefacept	12
Validation	ITN: ATG	17
Validation	ITN: Teplizumab	19
Validation	ITN: Tocilizumab	44
Validation	JDRF: Imatinib	21
Validation	TrialNet: Low Dose ATG/GCSF	30

638

639 **Figure 1. Validation of QR model: comparison of development and validation cohorts.**
640 Placebo/control participants used for QR development (n=5 studies) and validation (n=8 studies).
641 The published QR model ($QR_i = \ln(Cp_{1year,i} + 1) - 0.812 \cdot \ln(Cp_{0,i} + 1) - 0.00638 \cdot Age_i + 0.191$) from
642 the development studies (n=5) was applied to the validation studies (n=8). No significant
643 difference was observed between the development and validation cohorts when comparing group
644 means using a two-tailed t-test (p=0.44); both cohorts have similar distributions centered on zero

645 as evaluated by a Kolmogorov-Smirnov test. Mean and SD are indicated by black lines and dark
646 shaded region.


A
Baseline age (years)
14.2 19.5 20.5 16.8 13.0 17.1 13.4 13.7 25.9 19.3 17.8 12.8 18.8 Mean
5.2 6.2 7.1 4.6 6.5 9.2 1.9 2.3 6.6 10.4 7.9 4.1 11.0 SD
50
40
30
20
16.4
10
Baseline C-peptide AUC mean (nmol/L)
0.745 0.640 0.929 0.967 0.618 0.715 0.707 0.651 0.796 0.708 0.712 0.660 0.792 Mean
0.32 0.22 0.50 0.50 0.29 0.25 0.43 0.30 0.27 0.34 0.36 0.30 0.53 SD
2
1
0.73
0
35 16 20 31 22 48 34 111 21 42 29 25 47 N
Abatacept control
Alefacept control
ATG control
ATG/GCSF control
Canakinumab control
GAD-Alum control
GAD-Alum Phase 2 control
GAD-Alum Phase 3 control
Imatinib control
MMF/DZB control
Rituximab control
Teplizumab control
Tocilizumab control
B
1 year C-peptide AUC mean (nmol/L)
0.408 0.537 0.686 0.476 0.403 0.448 0.393 0.358 0.509 0.407 0.487 0.364 0.463 Mean
0.35 0.39 0.56 0.37 0.29 0.31 0.33 0.28 0.36 0.29 0.46 0.34 0.59 SD
3
2
1
0
QR
-0.013 0.082 0.048 -0.072 0.065 -0.002 0.002 -0.006 -0.055 -0.042 0.004 -0.015 -0.053 Mean
0.16 0.20 0.16 0.17 0.16 0.14 0.13 0.15 0.16 0.15 0.16 0.12 0.18 SD
0.50
0.25
0
-0.25
-0.50
30 12 17 30 19 46 34 109 21 38 29 19 44 N
Abatacept control
Alefacept control
ATG control
ATG/GCSF control
Canakinumab control
GAD-Alum control
GAD-Alum Phase 2 control
GAD-Alum Phase 3 control
Imatinib control
MMF/DZB control
Rituximab control
Teplizumab control
Tocilizumab control

647

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

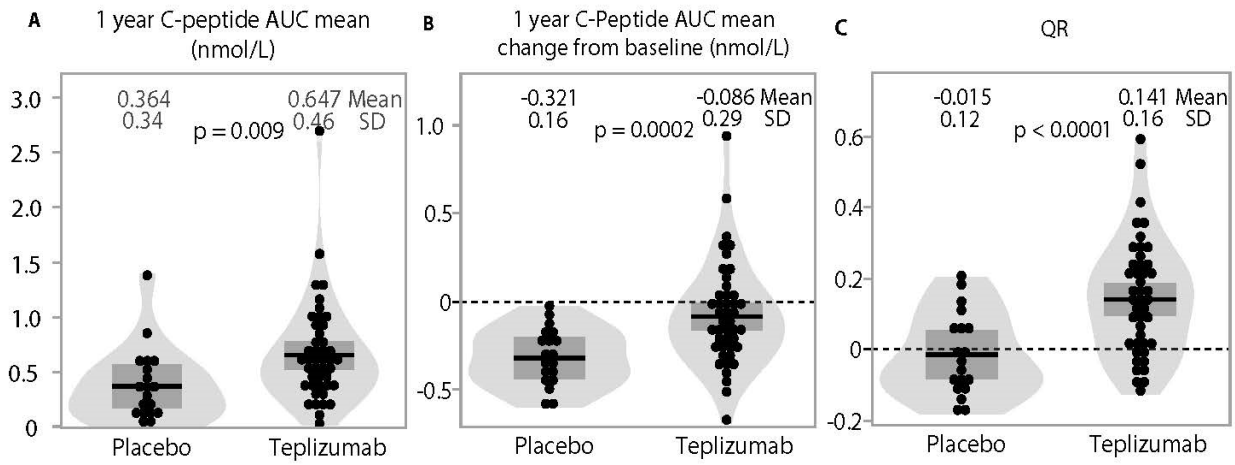


Figure 3. More statistical precision with use of QR in teplizumab trial.

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.

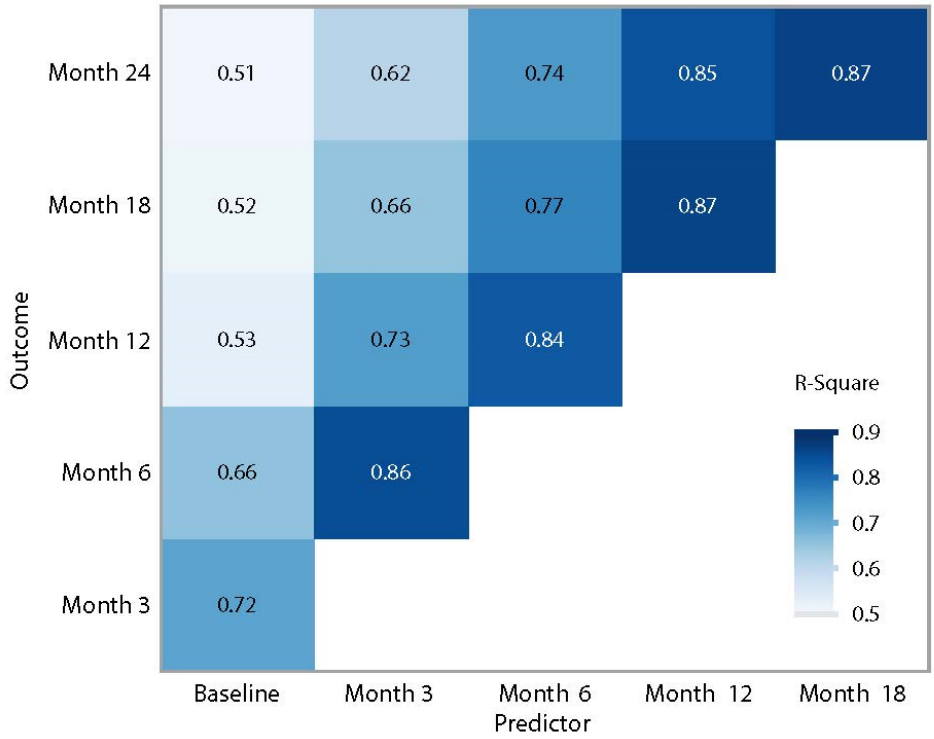


Figure 4. Relationship between observed and expected outcome according to time of baseline and time of outcome allows for different trial durations.

R^2 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 R^2 decreases the longer from the initial measurement. Using predictors at later time points has strong association with outcomes illustrated up to 24 months.

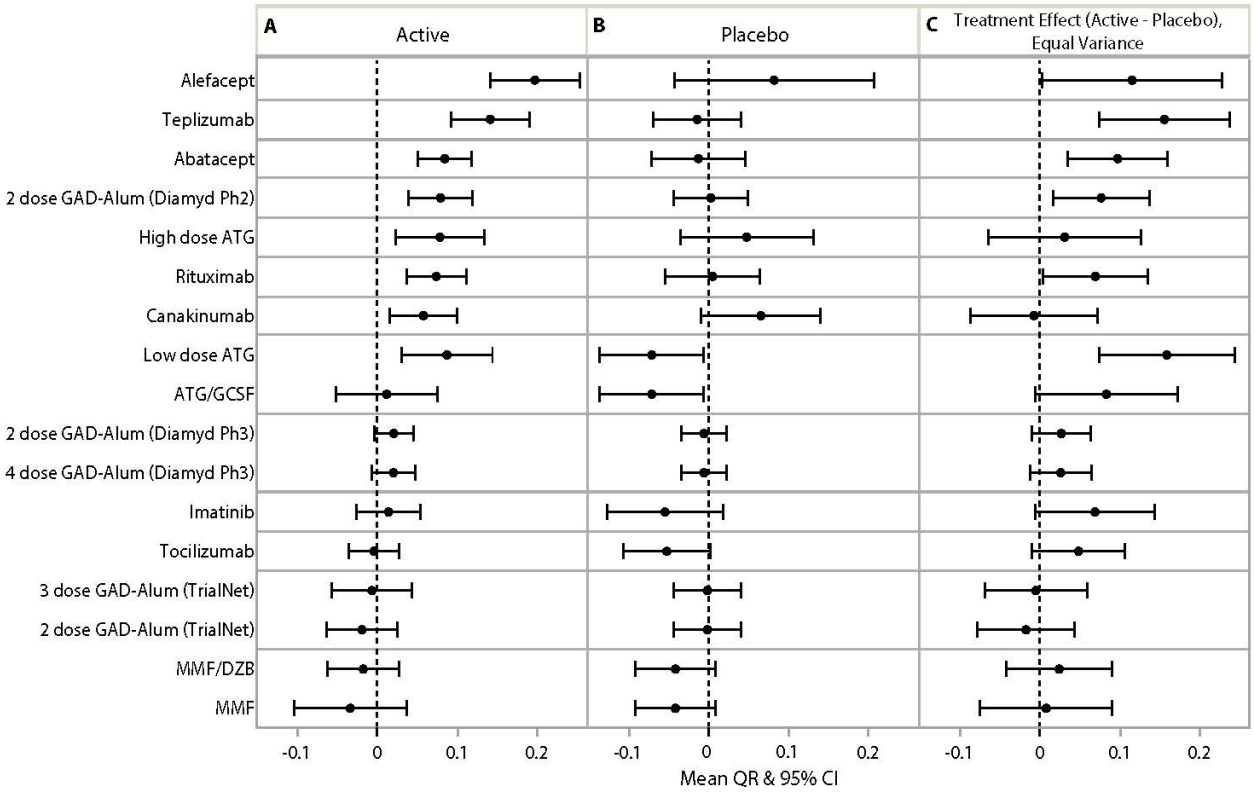


Figure 5. Use of QR to interpret across clinical trials.

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.

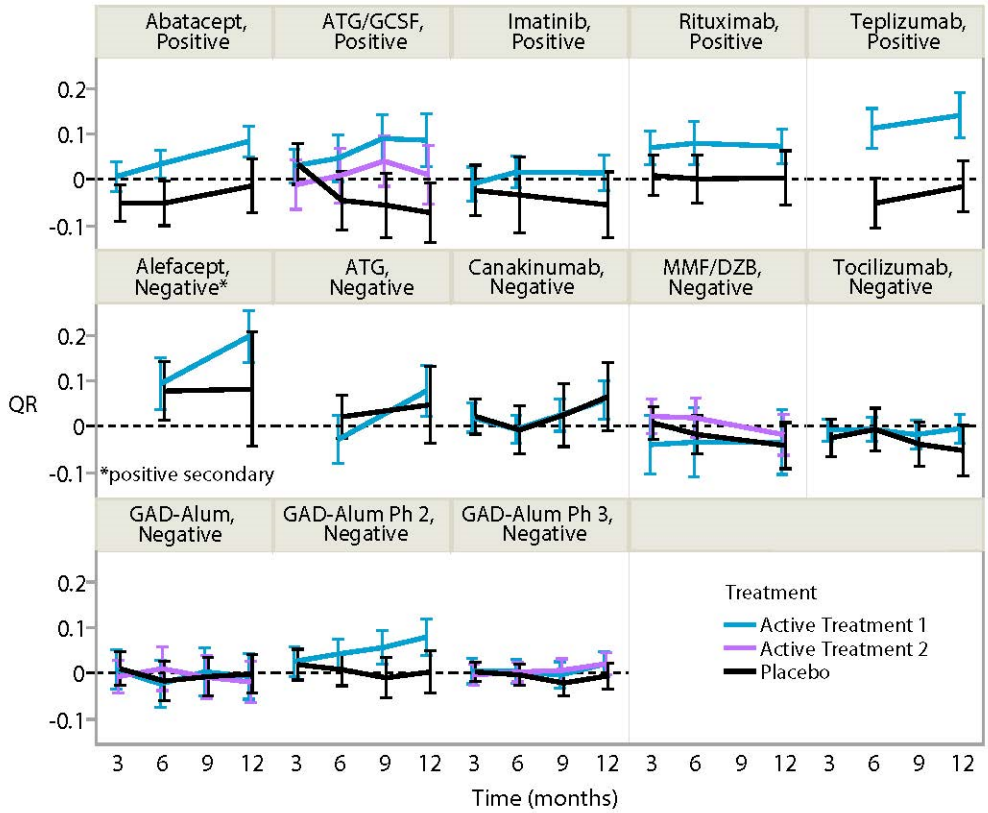


Figure 6. Effect of therapy determined at 6 months after randomization.

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.

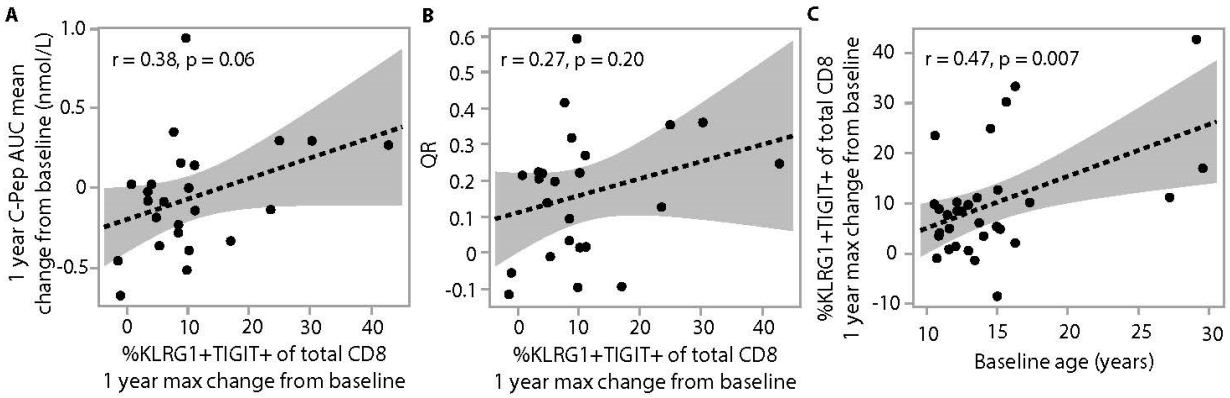


Figure 7. %KLRG1+TIGIT+ of CD8 T cells associated with age, not outcome of teplizumab therapy.

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.

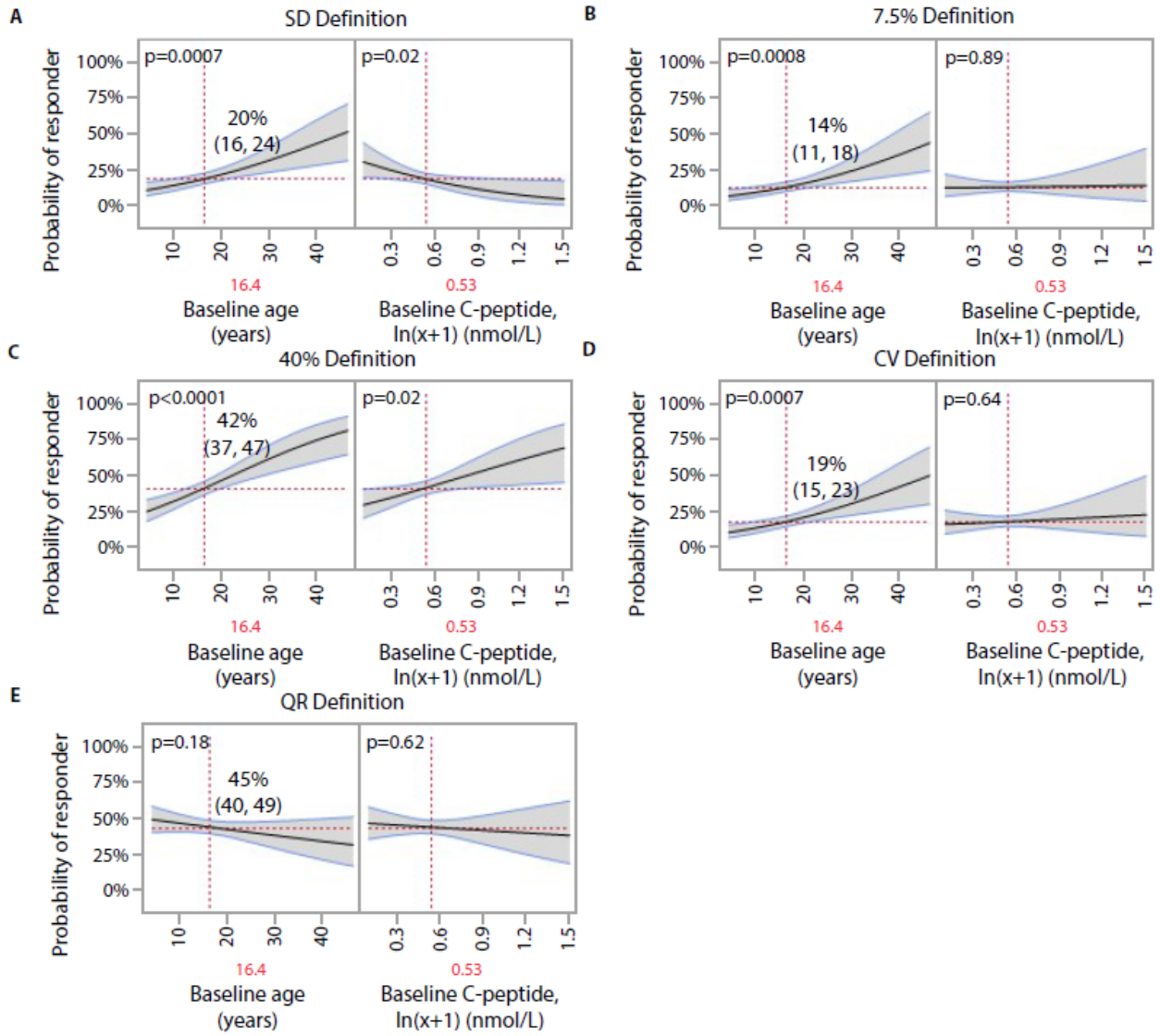


Figure 8. Proportion of placebo/control individuals meeting responder definitions across trials.

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.

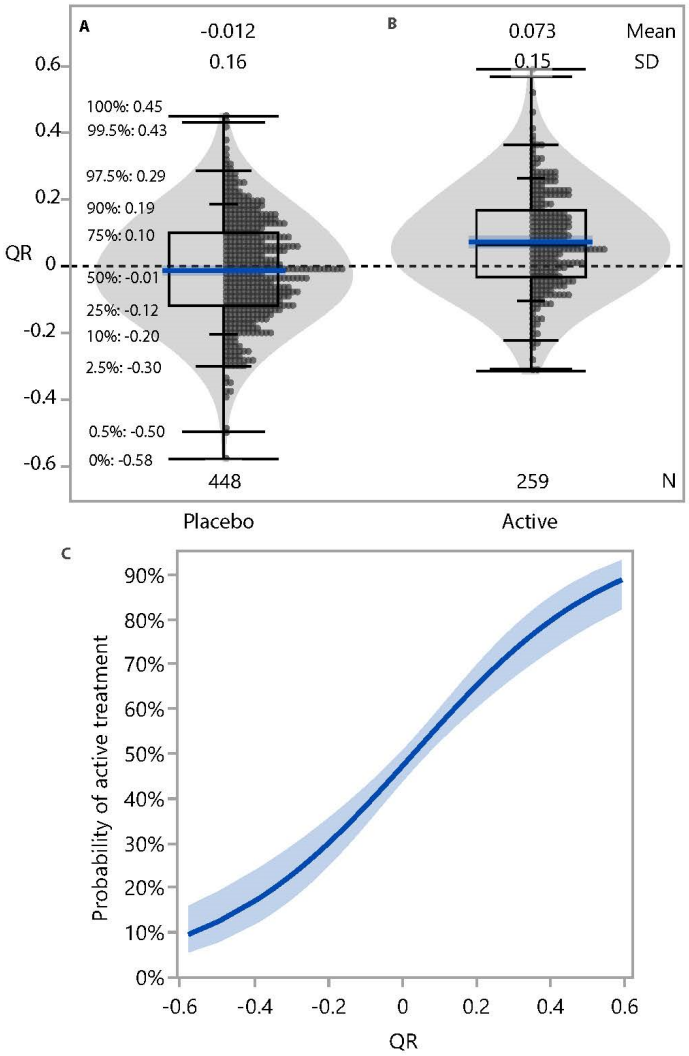


Figure 9. QR values for control/placebo and actively-treated individuals in trials reported with positive outcomes inform ability to distinguish actively-treated from placebo.

(**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.

731 **Table 1. Characteristics of trials used in analysis.**

Trial	Start Year	N Treatment Arm(s)	N Control Arm	Age, median (range)	Baseline C-peptide AUC Mean (nmol/L), mean (SD)	Primary Outcome/ Time Point	Primary Study Result
Diamyd Therapeutics AB: GAD-Alum Phase 2	2005	35	35	13.9 (10.1, 18.4)	0.66 (0.37)	Fasting C-peptide at 15 months	Negative
Diamyd Therapeutics AB: GAD-Alum Phase 3	2008	107 (2 doses), 109 (4 doses)	111	13.2 (10, 19.3)	0.66 (0.32)	2-hr MMTT AUC C-peptide at 15 months	Negative
Immune Tolerance Network (ITN): Anti-thymocyte globulin (ATG)	2007	38	20	17.5 (12, 35)	0.88 (0.42)	2-hr MMTT AUC C-peptide at 1 year	Negative
Immune Tolerance Network (ITN): Alefacept	2011	33	16	18 (12, 34)	0.78 (0.38)	2-hr MMTT AUC C-peptide at 1 year (secondary 4-hr MMTT AUC C-peptide at 1 year)	Negative primary; Positive secondary
Immune Tolerance Network (ITN): Teplizumab	2005	52	25	12 (8.3, 29.6)	0.7 (0.3)	4-hr MMTT AUC C-peptide at 2 years	Positive
Immune Tolerance Network (ITN): Tocilizumab	2015	88	47	14 (6, 45)	0.76 (0.43)	2-hr MMTT AUC C-peptide at 1 year	Negative
JDRF: Imatinib	2014	43	21	24.5 (18.3, 45)	0.84 (0.4)	2-hr MMTT AUC C-peptide at 1 year	Positive
TrialNet: Abatacept	2008	76	35	12.9 (6.4, 36.8)	0.75 (0.38)	2-hr MMTT AUC C-peptide at 2 years	Positive
TrialNet: Canakinumab	2010	48	22	11 (6, 31.9)	0.64 (0.33)	2-hr MMTT AUC C-peptide at 1 year	Negative
TrialNet: GAD-Alum	2009	49 (2 doses), 48 (3 doses)	48	15.1 (3.5, 45.7)	0.73 (0.32)	2-hr MMTT AUC C-peptide at 1 year	Negative
TrialNet: Low Dose ATG/GCSF	2014	29 (ATG only), 28 (ATG/GCSF)	31	16 (12, 42.5)	0.89 (0.44)	2-hr MMTT AUC C-peptide at 1 year	Positive
TrialNet: MMF/DZB	2004	31 (MMF only), 41 (MMF/DZB)	42	14.9 (8.7, 46.1)	0.69 (0.32)	2-hr MMTT AUC C-peptide at 1 year	Negative
TrialNet: Rituximab	2005	55	30	16 (8.3, 40.4)	0.75 (0.39)	2-hr MMTT AUC C-peptide at 1 year	Positive

732