Recommendations for the Definition of Clinical Responder in Insulin Preservation Studies

Craig A. Beam¹, Stephen E. Gitelman², Jerry Palmer³

¹Division of Epidemiology and Biostatistics, Department of Pediatrics, University of South Florida, Tampa FL, ²Department of Pediatrics, University of California San Francisco, CA, and Department of medicine, VA Puget Sound Health Care System, ³University of Washington, Seattle, WA

Running title: Clinical Responder Studies in T1D
Keywords: Type 1 Diabetes, Insulin Preservation, Clinical Responder, FDA, EMA

Word Count: 3,463; Tables: 3; Figures: 2

Corresponding Author:

Craig A. Beam, PhD
Department of Pediatrics
University of South Florida
3650 Spectrum Blvd.
Tampa FL, 33612
Tel/FAX: 813-396-2682/9601
beamc@epi.usf.edu
ABSTRACT

Clinical responder studies should contribute to the translation of effective treatments and interventions to the clinic. Since ultimately this translation will involve regulatory approval, we recommend that clinical trials pre-specify a responder definition that can be assessed against the requirements and suggestions of regulatory agencies. In this paper we propose a clinical responder definition to specifically assist researchers and regulatory agencies interpret the clinical importance of statistically significant findings for studies of interventions intended to preserve β-cell function in newly diagnosed Type 1 diabetes. We focus on studies of six-month β-cell preservation in type I diabetes as measured by 2-hour stimulated C-peptide. We introduce criteria (bias, reliability and external validity) for the assessment of responder definitions to ensure they meet FDA and European Medicines Agency guidelines. Using data from several published TrialNet studies, we evaluate our definition (no decrease in C-peptide) against published alternatives and determine that our definition has minimum bias with external validity. We observe that reliability could be improved by using changes in C-peptide later than six months beyond baseline. In sum, to support efficacy claims of β-cell preservation therapies in Type I diabetes submitted to US and European regulatory agencies, we recommend use of our definition.
BACKGROUND

The European Medicines Agency (EMA) recently recommended that “responder analyses” be included as either primary or secondary analyses in applications submitted for the approval of medicinal products for diabetes treatment and/or prevention.¹ For example, the EMA guidelines require that studies of glucose reduction in type 2 diabetes “…should justify the clinical relevance of the observed effect by presenting responder analyses comparing the proportion of patients who reached an absolute (HbA1c) value of ≤7 and/or 6.5% (≤53 and/or 48 mmol/mol) across the different treatment groups.” For treatment studies based on β-cell preservation in newly diagnosed Type 1 diabetics, the EMA recommends using several co-primary endpoints one of which could be, if appropriately justified, “…the percentage of patients with (stimulated) C-peptide increases above a clinically meaningful threshold…”

Although, as of the time of the writing of this paper, the FDA has not required responder analyses in their draft “Guidance Document for Industry related to Diabetes”², the agency did endorse the practice as part of the International Conference on Harmonization guidelines for statistical principles for clinical trials³. Since the agency does advise sponsors of therapeutic interventions for preservation of endogenous β-cell function in patients with newly diagnosed Type 1 diabetes to demonstrate the maintenance of C-peptide from baseline, it is reasonable to expect that the agency would be interested in responder analyses defining a clinical responder as someone who maintains C-peptide.

The term “responder analysis” refers to the dichotomization of a continuous primary efficacy measure into "responder" and "non-responder" categories⁴. Such responder classifications help us interpret data clinically and speak directly to the question of fundamental interest in clinical
science and practice: “Is this therapy benefitting the patient?” Clinical responder classifications have been used as endpoints in clinical trials to compare treatments and to define subgroups which are then compared for differences in immunologic, metabolic or mechanistic characteristics. Examples from the diabetes literature are numerous: Chiasson\textsuperscript{5} compared the efficacy of miglitol and metformin on glycemic control in T2D and defined a “clinically significant response” to be either a 15% or greater reduction from baseline in HbA1c or an HbA1c level less than 7.0%; Luque\textsuperscript{6} compared hypertension control in T2D patients given either manidipine or enalapril monotherapy, defining a responder to be a subject having at the end of the study either a diastolic blood pressure (DBP) less than 90 mmHg or a reduction in DBP of at least 10 mmHg; and Shaibani\textsuperscript{7} defined a responder to be a subject achieving a 30% reduction in mean daily pain score in patients with painful diabetic neuropathy.

A recent position paper\textsuperscript{8} by the Pharmaceutical Research and Manufacturers of America (PhRMA) organization points out that the role of responder analysis in regulatory guidance documents is to help the regulatory agency interpret the clinical relevance of statistically significant results. and put forth the following requirements for responder analyses: “…(1) the criteria for ‘responder’ should be generally accepted, fully validated; (2) the responder analysis should be reliable, robust, sensitive to change of disease and also be able to discriminate an experimental treatment compared to a control in a clinical trial; (3) the cut-point for dichotomization should be pre-specified in the study protocol and agreed up front with regulatory agencies.”

A series of new onset type 1 diabetes clinical trials have been conducted in the past decade to identify safe and effective means to preserve β-cell function. Interest stems from prior
observations in the Diabetes Control and Complications Trial, showing that even modest
preservation of endogenous insulin secretion confers clinical benefit, with better overall
glycemic control and lower risk for long term complications. Earlier studies usually reported
aggregate analyses and usually did not identify “responder” sub-groups with more (or less)
marked benefit from the therapy being evaluated. However, several recently published new
onset T1D clinical trials have presented responder analyses, but apparently without consideration
of regulatory guidelines or the recommendations discussed above, and these have not been pre-
specified but rather post-hoc analyses. As the ultimate goal in diabetes research is to bring
effective treatments and interventions to the clinic, and as this goal inevitably will require
regulatory approval, clinical researchers in diabetes should consider responder analyses as part of
their pre-specified clinical research strategies and must ensure as well that regulatory
requirements or recommendations are met by adhering to the standards stated by the PhRMA. In
this paper we propose a clinical responder definition that would assist researchers and regulatory
agencies interpret the clinical importance of statistically significant findings for studies of
interventions to preserve β-cell function in patients with newly diagnosed Type 1 diabetes. We
then compare this definition with two previously published alternatives. In addition, we
evaluate our definition against the PhRMA requirements stated above.

METHODS

Concepts and Definitions

Table 1 summarizes the key concepts used in this paper. We describe below our implementation
of these concepts.

Resonder Criterion
A primary clinical objective in the treatment of Type 1 diabetes (T1D) is the preservation of β-cell function with resultant preservation of endogenous insulin production. An objectively definable criterion of clinical response is, therefore, “preservation or increase from baseline in insulin secretion” Mixed meal tolerance testing is often used to assess insulin secretion by measurement of the endogenous insulin metabolite C-peptide over a 2 to 4-hour period. Responder definitions are then typically based on changes in C-peptide exceeding some threshold (referred in this paper as the “responder criterion”).

Since C-peptide preservation or improvement is the goal, a logical criterion to define a responder would be a post-baseline value that is 100% or more of the baseline value of C-peptide. However, two modifications of this responder criterion have been offered, and used, in the diabetes literature to address the problems that arise from intra-subject variability (test-retest variability). Each definition effectively lowers the 100% threshold in an attempt to counteract the fact that a true responder might occasionally have an observed change slightly less than 100% due to intra-subject variation. One definition lowers the threshold to 92.5% an amount equal to one-half the inter-assay variation of the C-peptide assay. The other definition essentially lowers the threshold to 87.2% which is an amount equal to the median subject-level coefficient of variation found in a test-retest study of C-peptide. In this paper we compare all three criteria, namely; 100%, 92.5% and 87.2%.

**Bias and Reliability**

However, it stands to reason that lowering the threshold from 100% will lead to over-estimates of the percentage of subjects maintaining or improving C-peptide. In addition, the presence of
intra-subject variability can also result in the over-, or even under-, estimation of responder percent. The net amount of this over- or under- estimation, referred to in this paper as “bias”, is evaluated for each of the responder definitions using the statistical method described below classifying individual subjects and including intra-subject variability of the C-peptide assay can also lead to the misclassification of true responders as non-responders and vice-versa. The rates at which these misclassifications occur are important characteristics of a responder definition. We refer to the percent of responders who are wrongly classified as non-responders as the “False Non-responder Rate”. Similarly, we refer to the percent of non-responders wrongly classified as responders as the “False Responder Rate”. We refer to these two rates collectively as the “reliability” of the responder classification.

**Evaluation Populations**

We used two evaluation populations, one to assess bias and reliability and the other to determine external validity. For the former, we used parameters estimated from Greenbaum, who combined data from three TrialNet onset treatment studies to represent the decline in C-peptide to be expected in newly diagnosed untreated subjects (Table 2). This evaluation population is largely white (93%), and male (61%), with mean (SD) age of 18.1 y (8.8 y). The mean (SD) C-peptide AUC at baseline was 0.49 pmol/ml (0.50 pmol/ml). As can be seen from this table, subsequent AUC values decline across time. In addition, a TrialNet “test-retest” study provided an estimate of intra-subject variability in C-peptide AUC of 0.4167, which is the variance in C-peptide AUC measured on the same individual on two occasions within 10 days of each other. The estimate was derived from a Variance Components Analysis in which there were random
between and within subject effects. Intra-subject variability was then taken to be the within subject variance. This estimate was required for the simulations described below.

We used data from the Rituximab study to evaluate external validity. This study, in fact, was included with the evaluation population collated by Greenbaum. Yet, since the evaluation of bias and reliability within the placebo treated population do not affect, nor are affected by, the evaluation of treatment effect (external validity), there are no issues of possible confounding results between the two analyses.

**External Validity**

As stated previously, one of the requirements suggested by the PhRMA is that the responder definition should be “reliable,…, and also to be able to discriminate an experimental treatment compared to a control in a clinical trial.” Our evaluation of reliability has been described above. We will refer to the ability to discriminate treatment from control as **External Validity** of the responder definition which will need to be evidenced by application in a real world setting. As described above, in our evaluation, we used data from the TrialNet Rituximab study, to test the external validity of the responder definition by comparing the difference in percent responders between the two treatment groups. An externally valid responder definition will also detect a difference between treatment groups.

**Statistical Methods**

In this section we provide a brief overview of our statistical methods. The online appendix provides greater detail.
We used Monte Carlo simulation to generate samples from the Greenbaum evaluation population. We first developed a multivariate normal measurement error model with parameters from the Greenbaum\textsuperscript{18} data and the intra-subject variability in C-peptide that was estimated in the test-retest study\textsuperscript{19}. The multivariate vector consisted of four measurements of C-peptide: two at baseline and two at follow-up. At each of these time-points, there are two measurements: one the actual C-peptide value and the other is the apparent value, which is the actual value plus measurement error. Inspection of the raw data led us to use the log transformed C-peptide data as this gave approximate normality in the univariate distributions. We then used Monte Carlo simulation to generate samples from the joint density function in order to estimate responder proportions, bias and reliability.

We ran five separate Monte Carlo simulations representing different levels of preservation. Each simulation used a different assumed percent responder (i.e., the percent with preserved C-peptide) in the population and generated 100,000 observations. One simulation assumed a 20% responder percent which is representative of the rate (21.2%) seen in the untreated population of Greenbaum\textsuperscript{18}. The other simulations used increasing rates of responders, namely 30%, 40%, 50% and 60%, and are therefore reflective of increasing levels of preservation.

For each sample of 100,000 we computed responder percentages for each criterion and false positive and false negative percentages using simple proportions. This number of simulations enabled us to estimate each of these quantities within, at most, 0.3%, using a 95% confidence interval.
Bias was computed as the difference between the apparent responder percentage and that assumed in the simulation. Reliability was measured by computing the false responder and false non-responder proportions using the method described in the appendix. External validity was evaluated with the Rituximab study using conventional methods for estimating and comparing (chi-square test) two independent proportions of responders between the treatment groups.

RESULTS

Bias: Over- or Under- estimating C-peptide preservation

Table 3 compares bias among the three criteria at each of three times of follow-up after baseline. In every case, the 100% criterion is nearly unbiased, with over-estimation of the assumed responder percent by no more than 0.4%. On the other hand, the other two criteria have greater bias ranging from 1.8% to 9.3. In addition, comparison of bias across time suggests that less bias occurs in later time points.

Based on the comparison of bias, we conclude that lowering the criterion from 100% leads to substantial overestimation of percent responders in the population and that the two alternative criteria have unacceptably large bias. We therefore focus the remainder of our evaluation on the 100% criterion for C-peptide preservation.

Reliability: Incorrect classification of the individual subject

Figure 1 summarizes the rates of the incorrect classification of individual subjects for the responder criterion of 100% with different assumed values of percent responders in the population and at different lengths of follow-up. False responder rates ranged from 6% to 36%. False non-responder rates ranged from 3% to 22%. These two rates change in opposite directions
as the assumed percent responders increases from 20% to 60%: Rates of false responders decrease and rates of false non-responders increase as the percent of true responders in the population increases. While the greatest rates of incorrect classification occur at six-months of follow-up, the rates at one- and two-year follow-up were nearly identical.

**External Validity: Detection of a treatment difference when one exists**

The Rituximab study\(^\text{15}\) found statistically significant higher mean C-peptide AUC in the treated subjects at six-months, one-year and two years of follow-up. Figure 2 presents responder percentages using the 100% responder criterion in the Rituximab evaluation population. This responder definition also determined there to be a benefit from rituximab in that the preservation of C-peptide was greater in the treated than the untreated subjects at each follow-up period. Although the percentage responders were statistically different only at six month follow-up, we note that the differences were smaller at later points (though still substantive) and lack of statistical significance might be simply due to limited sample size. Thus, we conclude that the proposed responder criterion of 100% was able to detect a previously-determined efficacious treatment and so is externally valid.

**DISCUSSION**

Responder studies should contribute to the translation of effective treatments and interventions to the clinic. Since ultimately this translation will involve regulatory approval, we recommend that every clinical trial pre-specify a responder analysis as part of the statistical analysis plan, and that the responder definition reflect the requirements and suggestions of regulatory agencies. Consequently, we suggest that responder definitions be viewed as being akin to any assay or test
and be required, therefore, to be evaluated for bias, reliability and external validity prior to its use in clinical research. Further, in any published study reporting the results of the analysis of responders we recommend that such evaluations need to either be incorporated with the introduction of a new responder definition or, otherwise, a citation to the source of such evaluation.

In this paper we have introduced a clinical responder definition specifically targeted at studies of the six-month preservation of β-cell mass and/or function in Type I Diabetes as measured by 2-hour stimulated C-peptide AUC determined from mixed meal tolerance testing. Our focus on six-month change is relevant to studies needing an early endpoint to show efficacy. We show that this definition meets criteria recently recommended by the Pharmaceutical Research and Manufacturers of America (PhRMA) and is less biased than published alternatives. In fact, it can be shown mathematically that this definition has the least bias amongst alternative responder definitions that are based on changing threshold from 100% of baseline out to 6 months. We therefore recommend its use in studies submitted for regulatory approval in both the US and Europe.

However, a fundamental question faces every study aiming to preserve β-cell mass/function: whether or not it is realistic to expect no decline after diagnosis, particularly as one pushes the primary endpoint out further from diagnosis, often to 12 or 24 months. It might be more realistic to consider “preservation” to be the slowing of decline. Responder definitions based on this definition of preservation would then require specification of a level of decline that would be still be considered a positive response to treatment. Without clinical studies tying c-peptide to
definitive clinical endpoint, however, the selection will be arbitrary. This challenge will be faced if we consider other possible endpoints as well. For example, one might consider using the time to peak C-peptide falling below 0.2 nmol as an endpoint and then define a responder as someone whose time exceeds some threshold. Here, again, selection of the threshold is arbitrary until the extension of time can be directly linked to clinical outcome. Similar challenges face the use of composite endpoints, e.g. using both HbA1c and insulin use, to define the clinical responder. Yet, the use of arbitrary thresholds to define a responder is found in other medical settings. Responder criterion selection is often arbitrary because typically there exists no objectively definable single point in the measurement of therapeutic response which unequivocally demarcates favorable response from unfavorable response. For example, the RECIST threshold of partial response (30% reduction in tumor volume) was established by expert consensus and, according to Therasse is an “arbitrary convention”. We recommend the field of Type I Diabetes similarly establish, by consensus, a definition of responder related to the slowing of β-cell decline. (In addition, a definition of “deleterious effect” of treatment should be established as well). Of course, such a definition will need to also be assessed to ensure that it meets the Pharmaceutical Research and Manufacturers of America recommendations if it is to be used to support efficacy claims in insulin preservation studies submitted to US and European regulatory agencies. We recommend the methodology used in this paper be considered for that purpose.

The “placebo responder” presents a dilemma when interpreting responder analyses. In our analysis of the Rituximab data, we observed a 32% responder rate in the subjects treated with placebo. But, how can a placebo-treated subject be declared a “responder”-what are they
responding to? We recommend that a complete responder analysis include a clinical interpretation of the placebo responder (e.g. the responses reflect the effectiveness of insulin therapy) and assess as well the clinical significance of the increase in percent responders above the placebo responder rate that is due to treatment. For example, recalling our validation data, was the 10% increase in the percentage of subjects with C-peptide preservation at six-months by Rituximab above the percentage expected to preserve without treatment (as seen in the placebo group) large enough to clinically justify the adoption of Rituximab as first-line therapy?

We have shown that the bias of the 100% C-peptide preservation criterion is, for practical purposes, negligible (0.2%) and that it is sensitive enough to detect an efficacious treatment. Therefore, to support efficacy claims of β-cell preservation therapies in Type I diabetes submitted to US and European regulatory agencies, we recommend use of the 100% C-peptide preservation criterion. Individual classification is, however, unreliable and, thus, we agree with the PhRMA position paper which also advises that responder analysis should be done only after statistical significance of the original continuous variable (in our case, C-peptide change from baseline) has first been established and then the responder analysis should be considered as supportive analysis.  

Caution must still be exercised even when a responder analysis is relegated to “supportive” or “secondary” status. It is well known in the statistical literature that the dichotomization of a continuous measurement reduces statistical power. Moreover, in the presence of “measurement error”, as is the case with the C-peptide assay, the estimated treatment difference between responder percentages will be underestimated. Further issues arise from assay
variation by the misclassification of true responders as non-responders and true non-responders as responders. Such misclassification leads to estimation bias and lowers the power of tests of association between categorical variables, such as the presence of a gene allele or gender of the subject, with responder status.\textsuperscript{24,25}

One limitation of our study is that we have not considered the use of responder analyses which attempt to identify factors, such as genotype or phenotype, that determine response to therapy. That is, responder studies restricted to only treated subjects. However, such studies are typically exploratory and not intended to support efficacy claims. Although regulatory considerations are not directly relevant, we do recommend that the responder definition used in such studies be at least assessed for bias and reliability so that a realistic interpretation of findings can be made. Another limitation is that our evaluations required computer simulations with assumed values. In the case of the estimate of intra-subject variability, we note that the parent study collected C-peptide measurements from subjects having a broad range of diabetes duration at baseline. It is conceivable that data from a study with more homogeneous diabetes experience might have led to different estimates of within-subject variability and altered some of our findings.

In summary, we recommend that the use of responder analysis in insulin preservation studies be considered circumspectly as a means to help interpret the clinical relevance of statistically significant results based on C-peptide measurement and not as a primary analytic endpoint. Every responder definition should be pre-specified and evaluated for bias, reliability and external validity prior to its use in clinical research. When insulin preservation is the focus, the logical threshold for defining a C-peptide responder is” no change or increase from baseline”. There are
no advantages to, or need for, alterations to this threshold to accommodate intra-subject variation. In addition, this” logical threshold” meets the recommendations of the PhRMA, FDA and EMA for responder analysis in beta cell preservation studies. This logical threshold provides nearly unbiased interpretations which were shown to be sensitive to the presence of a treatment effect in the case of the TrialNet Rituximab study. Finally, we recommend that a responder definition based on the notion of an “acceptable” reduction of β-cell decline be established via consensus by the Type I Diabetes community.
AUTHOR CONTRIBUTIONS:

All authors at the time of writing were members of the Type 1 Diabetes TrialNet study group and as such contributed to the data used in this paper. CB designed and conducted the statistical analyses. CB, SG and JP wrote manuscript. CB is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

ACKNOWLEDGMENTS— the data used in this paper was provided by the Type 1 Diabetes TrialNet Study Group. Type 1 Diabetes TrialNet Study Group 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 DK061016, U01 DK061034, U01 DK061036, U01 DK061040, U01 DK061041, U01 DK061042, U01 DK061055, U01 DK061058, U01 DK084565, U01 DK085453, U01 DK085461, U01 DK085463, U01 DK085466, U01 DK085499, U01 DK085505, U01 DK085509, and a contract HHSN267200800019C; the National Center for Research Resources, through Clinical Translational Science Awards UL1 RR024139, UL1 RR024139, UL1 RR024153, UL1 RR024975, UL1 RR024982, UL1 RR025744, UL1 RR025761, UL1 RR025780, UL1 RR029890, UL1 RR031986, and General Clinical Research Center Award M01 RR00400; the Juvenile Diabetes Research Foundation International (JDRF); and the American Diabetes Association (ADA). The contents of this Article are solely the responsibility of the authors and
do not necessarily represent the official views of the NIH, JDRF, or ADA. The authors have no conflicts of interests with publication of this manuscript.

**REFERENCES**


### TABLES

#### Table 1 Concepts and Assessment Criteria for Responder Definitions

<table>
<thead>
<tr>
<th>Concept</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Responder Criterion</strong></td>
<td>For a continuous response variable (e.g. C-peptide), the threshold demarcating clinically favorable values (e.g. preservation)</td>
</tr>
<tr>
<td><strong>Intra-subject Variability</strong></td>
<td>Sometimes referred to as “biologic variability”, this term refers to the phenomenon that repeating an assay on material taken from the same subject but at different times can lead to different assay values.</td>
</tr>
<tr>
<td><strong>Bias</strong></td>
<td>The amount by which a responder definition will, on average, over- or underestimate the responder percentage in a patient population.</td>
</tr>
<tr>
<td><strong>False Responder Rate</strong></td>
<td>The percentage of subjects that are classified as responders by the responder definition but who actually are non-responders.</td>
</tr>
<tr>
<td><strong>False Non-responder Rate</strong></td>
<td>The percentage of subjects that are classified as non-responders by the responder definition but who actually are responders.</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>Refers to the false responder and false non-responder rates collectively.</td>
</tr>
<tr>
<td><strong>External Validity</strong></td>
<td>In general, concordance with the findings of an external study that found statistically significant differences between groups (typically defined by therapy vs. control or placebo) in the underlying continuous variable.</td>
</tr>
</tbody>
</table>
### Table 2 Evaluation Populations

<table>
<thead>
<tr>
<th>PURPOSE:</th>
<th>Bias &amp; Reliability</th>
<th>External Validity*</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOURCES:</td>
<td>Test-Retest Study</td>
<td>Greenbaum Study</td>
</tr>
<tr>
<td></td>
<td>(n=148)</td>
<td>(n=191)</td>
</tr>
<tr>
<td>Age-yr</td>
<td>16.2±6*</td>
<td>18.1±8.8</td>
</tr>
<tr>
<td>Gender -%Female</td>
<td>39%</td>
<td>38%</td>
</tr>
<tr>
<td>Race -%White</td>
<td>86%</td>
<td>93%</td>
</tr>
<tr>
<td>Ethnicity-%Hispanic</td>
<td>7%</td>
<td>7%</td>
</tr>
<tr>
<td>C-Peptide 2 hr AUC (pmol/ml)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.49±0.50</td>
<td>0.71±0.34</td>
</tr>
<tr>
<td>Follow-up</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10 days</td>
<td>0.46±0.44</td>
<td></td>
</tr>
<tr>
<td>6 months</td>
<td>0.56±0.39</td>
<td>0.79±0.57</td>
</tr>
<tr>
<td>12 months</td>
<td>0.43±0.34</td>
<td>0.62±0.42</td>
</tr>
<tr>
<td>24 months</td>
<td>0.36±0.37</td>
<td>0.47±0.46</td>
</tr>
</tbody>
</table>

*mean±SD.

*Both datasets are from the Rituximab paper. The C-peptide data reported here are only from subjects used in the ITT (Intention to Treat) analysis. For baseline, mean and SD do not match the values reported in Table 1 of the Rituximab paper which included all subjects.
Table 3 Bias† in Proposed vs. Published Responder Criteria

<table>
<thead>
<tr>
<th>Assumed Percent Maintaining C-peptide at 6 Months</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>+0.2</td>
<td>+0.</td>
<td>-0.4</td>
<td>+0.0</td>
<td>-0.3</td>
</tr>
<tr>
<td>92.5%</td>
<td>+4.0</td>
<td>+4.9</td>
<td>+4.9</td>
<td>+5.3</td>
<td>0+4.9</td>
</tr>
<tr>
<td>87.2%</td>
<td>+7.4</td>
<td>+8.7</td>
<td>+9.1</td>
<td>+9.3</td>
<td>+8.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assumed Percent Maintaining C-peptide at 1 Year</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>+0.0</td>
<td>-0.3</td>
<td>+0.1</td>
<td>-0.2</td>
<td>-0.1</td>
</tr>
<tr>
<td>92.5%</td>
<td>+2.0</td>
<td>+2.0</td>
<td>+2.7</td>
<td>+2.5</td>
<td>+2.6</td>
</tr>
<tr>
<td>87.2%</td>
<td>+3.5</td>
<td>+3.9</td>
<td>+4.8</td>
<td>+4.5</td>
<td>+4.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assumed Percent Maintaining C-peptide at 2 Years</th>
<th>20%</th>
<th>20%</th>
<th>20%</th>
<th>20%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>-0.1</td>
<td>-0.3</td>
<td>-0.1</td>
<td>+0.0</td>
<td>-0.0</td>
</tr>
<tr>
<td>92.5%</td>
<td>+1.8</td>
<td>+2.0</td>
<td>+2.5</td>
<td>+2.8</td>
<td>+2.6</td>
</tr>
<tr>
<td>87.2%</td>
<td>+3.4</td>
<td>+4.0</td>
<td>+4.5</td>
<td>+4.9</td>
<td>+4.5</td>
</tr>
</tbody>
</table>

†Bias is defined to be proportion in excess (+) or in deficit (-) of the assumed percent retaining C-peptide in the population.

Values denoted by + (or –) “0.0” were such that they were slightly above (below) zero and rounded to zero.
FIGURE LEGENDS

Figure 1
False non-responder (open circles) and false responder (closed circles) percentages for the 100% criterion as a function of the percent responders assumed for the population and the time of the C-peptide endpoint (6 months, 1 and 2 years).

Figure 2
Responder percentages based on the 100% criterion in the rituximab study.
The image shows a bar chart illustrating the percentage of responders over different follow-up periods for two treatments: placebo and rituximab. The follow-up periods are six months, one year, and two years.

- **Six months**: The diagram indicates that the percentage of responders is higher for rituximab compared to placebo. The p-values for the comparison are p = 0.304 for placebo and p = 0.5190 for rituximab.
- **One year**: The percentage of responders is also higher for rituximab. The p-value for the comparison is p = 0.5190.
- **Two years**: The percentage of responders remains higher for rituximab, with a p-value of p = 0.2457.

The chart includes sample sizes for each group: n = 28 for placebo and rituximab at six months, n = 51 for placebo, n = 28 for rituximab, n = 49 for placebo at one year, n = 28 for rituximab, n = 49 for placebo at two years, and n = 51 for placebo and rituximab.
In this supplemental material, we introduce the statistical model underpinning our approach, detail how characteristics of a responder definition are thus defined, and describe how we implemented the model in practice.

1. Statistical Model

Broadly speaking, the model is a multivariate normal “errors in variables” model in which some of the variables are unobservable and the measurement error variance is assumed known. The errors in variables model has been proposed for responder analysis in the past, but only from a univariate or bivariate standpoint. We extend these past approaches via a multivariate model having unobservable variables. The assumption of known error variance has been utilized in the literature concerning misclassification arising from measurement error in which an independent study is conducted to estimate this variance (Byonaccorsi, JASA 1990).

For subject $i$ measured at time $t=1,2$ we observe the apparent value $y$ which arises from measurement error of the actual value $x$:

$$y_{ti} = x_{ti} + \epsilon_{ti}$$

We assume that the vector $X_i \equiv [x_{1i}, x_{2i}]'$ is multivariate normal having parameters

Equation 1

$$E(X_i) = [\mu_1, \mu_2]'$$

$$V(X_i) \equiv \Sigma_{xx} = \begin{bmatrix} \sigma_x^2 & \sigma_{x_1x_2} \\ \sigma_{x_1x_2} & \sigma_x^2 \end{bmatrix}$$
Result 1

Let $Y_i \equiv [y_{1i}, y_{2i}]'$. Assuming that $\epsilon_{ti} \sim iid N(0, \sigma^2_\epsilon)$ are independent of $x_{ti}$, the vector $[X_i, Y_i]$ is multivariate normal having parameters

$$E([X_i, Y_i]) = [\mu_1, \mu_2, \mu_1, \mu_2]'$$
$$V([X_i, Y_i]) = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} + \sigma^2_\epsilon \end{bmatrix}$$

Proof: This result follows immediately from the assumptions:

$$E(y_t) = E(x_t + \epsilon_t) = \mu_t$$
$$V(y_t) = V(x_t + \epsilon_t) = \sigma^2_x + \sigma^2_\epsilon$$
$$\text{Cov}(y_1, y_2) = E((y_1 - \mu_1)(y_2 - \mu_2))$$
$$= E(((x_1 + \epsilon_1) - \mu_1)((x_2 + \epsilon_2) - \mu_2))$$
$$= E\left(((x_1 - \mu_1) + \epsilon_1)((x_2 - \mu_2) + \epsilon_2)\right)$$
$$= E((x_1 - \mu_1)(x_2 - \mu_2) + (x_1 - \mu_1)\epsilon_2 + (x_2 - \mu_2)\epsilon_1 + \epsilon_1\epsilon_2 = \sigma_{x_1x_2}$$
$$\text{Cov}(x_1, y_1) = E((x_1 - \mu_1)(x_1 + \epsilon_1 - \mu_1))$$
$$= E((x_1 - \mu_1)^2 + E(x_1 - \mu_1)E(\epsilon_1) = \sigma^2_x$$
$$\Rightarrow \text{Cov}(x_2, y_2) = \sigma^2_x$$
$$\text{Cov}(x_1, y_2) = E((x_1 - \mu_1)(x_2 + \epsilon_2 - \mu_2))$$
$$= E(x_1 - \mu_1)(x_2 - \mu_2) + E(x_1 - \mu_1)E(\epsilon_2) = \sigma_{x_1x_2}$$
$$\Rightarrow \text{Cov}(x_2, y_1) = \sigma_{x_1x_2}$$

Result 2

By the preceding, if $|\Sigma_{xx} - \sigma^2_\epsilon I| > 0$ then the conditional distribution of the unobservable $X_i$ given the observable $Y_i$ is multivariate normal having parameters

$$E(X_i|Y_i) = [\mu_1, \mu_2]' + \Sigma_{xx}[\Sigma_{xx} + \sigma^2_\epsilon I]^{-1}(Y_i - [\mu_1, \mu_2]')$$
$$V(X_i|Y_i) = \Sigma_{xx} - \Sigma_{xx}[\Sigma_{xx} + \sigma^2_\epsilon I]^{-1}\Sigma_{xx}$$
This results follows from Result 1 and multivariate normal theory (see Result 4.6 on page 135 of Johnson⁴).

However, in practical applications, since x is unobservable, parameters of the joint distribution or conditional distribution cannot be directly estimated. Nonetheless, we can re-express the previous results in terms of the variance-covariance matrix of the observable y vector:

**Equation 2**

\[
V([X_i, Y_i]) = \begin{bmatrix}
\Sigma_{yy} - \sigma^2_{\epsilon} & \Sigma_{yy} - \sigma^2_{\epsilon} \\
\Sigma_{yy} - \sigma^2_{\epsilon} & \Sigma_{yy}
\end{bmatrix}
\]

Where,

**Equation 3**

\[
\Sigma_{yy} = \begin{bmatrix}
\sigma^2_x + \sigma^2_{\epsilon} & \sigma_{x_i x_2} \\
\sigma_{x_i x_2} & \sigma^2_x + \sigma^2_{\epsilon}
\end{bmatrix}
\]

So that,

**Equation 4**

\[
E(X_i|Y_i) = [\mu_1, \mu_2]' + (\Sigma_{yy} - \sigma^2_{\epsilon})\Sigma_{yy}^{-1}(Y_i - [\mu_1, \mu_2]')
\]

\[
V(X_i|Y_i) = (\Sigma_{yy} - \sigma^2_{\epsilon}) - (\Sigma_{yy} - \sigma^2_{\epsilon})\Sigma_{yy}^{-1}(\Sigma_{yy} - \sigma^2_{\epsilon})
\]

2. **Evaluating Characteristics of Responder Definitions**

Most published responder definitions represent conditions placed on the difference or ratio of two measurements separated in time. Such conditions can be expressed generally as a linear inequality in hyperspace, e.g. \{ (x₁, x₂, y₁, y₂): x₂ ≥ a + bx₁ \}. A responder definition based on the difference of x₁, x₂ would set b=1 and a definition based on their ratio would set a=0.

Therefore, actual and apparent responder proportions can be evaluated by integrating the multivariate density function over the appropriate region in \( R^4 \). Let \( \phi \) denote the multivariate density function having parameters given by equations Equation 1 and Equation 2, then the characteristics of interest are defined by the following integrals:
Actual responder proportion=

**Equation 5**

\[
\int_{\{(x_1, x_2, y_1, y_2): x_2 \geq a + bx_1\}} \phi d(x_1, x_2, y_1, y_2)
\]

Apparent responder proportion=

**Equation 6**

\[
\int_{\{(x_1, x_2, y_1, y_2): y_2 \geq a + by_1\}} \phi d(x_1, x_2, y_1, y_2)
\]

Misclassification error rates can be evaluated by integrating the conditional distribution defined in Equation 4: Let \( \phi_{x|y} \) denote the density function of \( X \) given \( Y \) and let \( \phi_y \) denote the marginal density of \( Y \). \( \phi_{x|y} \) is bivariate normal with parameters given by Equation 4. \( \phi_y \) is also bivariate normal with parameters \( E(Y) = E(X) \) and \( V(Y) = \Sigma_{yy} \) as defined in Equation 3. Recalling that \( \phi_{x|y} \) is a function of \( y \), we can then define the following quantities;

False responder proportion=

**Equation 7**

\[
\int_{\{(y_1, y_2): y_2 \geq a + by_1\}} \left\{ \int_{\{(x_1, x_2): x_2 \leq a + bx_1\}} \phi_{(x|y)} d(x_1, x_2) \right\} \phi_y d(y_1, y_2)
\]

False non-responder proportion=
Equation 8

\[
\int_{\{(y_1,y_2): y_2 < a + b y_1\}} \left\{ \int_{\{(x_1,x_2): x_2 \geq a + b x_1\}} \phi(x|y) \, d(x_1,x_2) \right\} \phi_y \, d(y_1,y_2)
\]

3. Practical Implementation

Although the computation of multivariate normal integrals can be done with functions such as “sadmvn” in the R package “mnormt”, these functions require limits of integration that can be expressed as fixed intervals. In our situation, in which the lower limit of integration is a function of one of the integrated variables (e.g. \(x_2 \geq a + b x_1\) when computing the actual responder proportion), the lower limit of integration is constantly changing. To address this problem, we randomly sampled 10,000 multivariate observations from the appropriate distribution and used the proportion falling into the region of interest as the value of the responder proportion. Random sampling of multivariate observations was accomplished with the R function “mvnorm”. We chose 10,000 observations in order to give adequate precision in estimation (standard error \(\leq 0.005\)) while reducing computational time to an acceptable level.

Computation of the conditional probabilities defining misclassification is made additionally difficult by the need to conduct nested integration with varying limits of integration. We used the previous method of estimating proportions via sampling from multivariate normal populations to achieve a solution to this problem as well.

Define the following unconditional probabilities:

False Positive Proportion \(=\)

\[
\int_{\{(x_1,x_2,y_1,y_2): x_2 < a + b x_1 \cap y_2 \geq a + b y_1\}} \phi \, d(x_1,x_2,y_1,y_2)
\]

and, False Negative Proportion \(=\)

\[
\int_{\{(x_1,x_2,y_1,y_2): x_2 \geq a + b x_1 \cap y_2 < a + b y_1\}} \phi \, d(x_1,x_2,y_1,y_2)
\]
Then, by the definition of conditional probabilities, it follows that the probabilities of interest can be computed by the following ratios:

False Responder Proportion = False Positive Proportion / Apparent Responder Proportion,

and,

False Non-responder Proportion = False Negative Proportion / (1 - Apparent Responder Proportion).
References


