

RESEARCH ARTICLE

Novel application of a discrete time-to-event model for randomized oral immunotherapy clinical trials with repeat food challenges

Natasha Purington¹ | Sandra Andorf^{2,3,4}  | Bryan Bunning² | Sharon Chinthrajah^{2,5} | Kari Nadeau^{2,5} | Manisha Desai^{1,6} 

¹Quantitative Sciences Unit, Division of Biomedical Informatics Research, Department of Medicine, Stanford University, Palo Alto, California

²Department of Medicine, Sean N. Parker Center for Allergy and Asthma Research, Stanford, California

³Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio

⁴Divisions of Biomedical Informatics and Allergy & Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

⁵Department of Pediatrics, Stanford University, Stanford, California

⁶Department of Biomedical Data Science, Stanford University, Stanford, California

Correspondence

Manisha Desai, Department of Medicine, Quantitative Sciences Unit, Stanford University, 1701 Page Mill Road, Palo Alto, CA 94304, USA.

Email: manishad@stanford.edu

Funding information

Ping Li and Kim Li Endowment; Clinical and Translational Science Award

Abstract

The evaluation of double-blind, placebo-controlled food challenges (DBPCFC) generally focuses on a participant passing a challenge at a predetermined dose, and does not consider the dose of reaction for those who fail or are censored due to study discontinuation. Further, a number of food allergy trials have incorporated multiple DBPCFCs throughout the duration of the study in order to evaluate changes in reaction over time including sustained unresponsiveness from treatment. Outcomes arising from these trials are commonly modeled using Chi-squared or Fisher's exact tests at each time point. We propose applying time-to-event methodology to food allergy trials in order to exploit the inherent granularity of challenge outcomes that additionally accommodates repeated DBPCFCs. Specifically, we consider dose-to-failure for each study challenge and extend the cumulative tolerated dose across challenges to result in a dose-time axis. A discrete time-to-event framework is applied to the dose-time outcome to assess the efficacy of treatment across the entire study period. We illustrate ideas with data from the Peanut Oral Immunotherapy Study: Safety, Efficacy and Discovery (POISED) trial, conducted at Stanford University, which evaluated the efficacy of oral immunotherapy on desensitization and sustained unresponsiveness in peanut allergic children and adults. We demonstrate the advantages of time-to-event approaches for assessing the efficacy of treatment over time and incorporating information for those who failed or were lost to follow up. Further, we introduce a dose-time outcome that is interpretable to clinicians and allows for examination of such outcomes over time.

KEYWORDS

clinical trials, cumulative tolerated dose, food allergy, oral food challenge, oral immunotherapy, time-to-event analysis

1 | INTRODUCTION

The double-blind, placebo-controlled food challenge (DBPCFC) is the gold standard for diagnosis of food allergy.^{1,2} While the DBPCFC was introduced to accurately diagnose food allergies in individual patients, it has gained momentum as a standard protocol in randomized controlled trials (RCT) to assess the efficacy of oral immunotherapy (OIT) on patients with diagnosed food allergies.^{3,4} For example, in the Peanut Oral Immunotherapy Study: Safety, Efficacy and Discovery (POISED) trial, a three-arm randomized double blind, placebo-controlled clinical trial conducted at Stanford University, DBPCFC success was used as a primary outcome.

In recent RCTs of OIT, such as POISED, commonly reported study outcomes are (a) persistent clinical non-reactivity in food allergic patients following discontinuation of maintenance dosing after successful OIT, referred to as sustained unresponsiveness, and (b) temporary clinical non-reactivity based on continued dosing of OIT, referred to as desensitization.⁵ Both outcomes are defined by reaching and passing a pre-specified allergen dose during a challenge by a certain study week (coded as a binary “pass/fail”), but differ in terms of treatment administration. More generally, sustained unresponsiveness is assessed after a participant has avoided an allergen for a set amount of time, while desensitization is assessed while a participant consumes the allergen daily. As with other RCTs of OIT, the POISED trial characterized and compared rates of desensitization and sustained unresponsiveness across treatment arms using Chi-squared and Fisher’s exact tests along with presentation of 95% confidence intervals (CIs) for response, as appropriate for binary outcomes.⁶⁻⁹

Defining study outcomes in such a way provides a snapshot of success - defined dichotomously—at a specific study timepoint and ignores valuable information including the time of study failure, particularly if the failure itself was not observed during the challenge or study period (known as right-censoring).¹⁰ To illustrate, take two participants (A and B) who were successfully desensitized by the study week of interest, and participate in subsequent study DBPCFCs to assess changes in clinical outcomes. While the above framework would treat these two participants equivalently (eg, both are study successes at that time point), we may see that person A is able to sustain desensitization for an additional 8 weeks while person B may only remain desensitized for an additional 4 weeks. This additional information obtained on subsequent challenges allows us to assess changes in clinical outcomes over the entire study period. Alternatively, take two participants (C and D) who failed to reach desensitization by the study week of interest; in the framework described above, they also contribute identical information (ie, both are study failures). However, when tolerated dose at failure is incorporated, we may see that person C failed at the very first dose of the challenge, while person D failed at the last dose preceding success. These nuances are not unlike those encountered in RCTs evaluating mortality, which motivated introducing time-to-event outcomes in 1958.¹¹ Analytic tools including the Kaplan-Meier method, the Cox proportional hazards model, and parametric regression that models the hazard function were subsequently established for modeling time-to-event outcomes and therefore may be appropriate for this context.¹²

A natural proposal came from Chinchilli et al. to consider discrete-time survival analysis for the purpose of characterizing the comparative efficacy of an intervention on such events of interests, focusing on the change in maximum tolerated dose from baseline to a single end-of-study DBPCFC.¹³ They were the first to apply a time-to-event framework to DBPCFC-based RCTs and to highlight its advantages of handling right-censored observations including its ability to control for covariates of interest. This approach has improved upon the less informative binary outcome assigned to a given DBPCFC and the authors provide a detailed guide on implementing and interpreting results from this model, which have been incorporated in proceeding studies.^{14,15} While the authors suggest this method is adaptable to longitudinal DBPCFC data, neither they nor others expand on how to incorporate outcomes in the presence of multiple challenges over a study period. Gamage et al. assess a time-to-desensitization outcome in their elegant study of OIT,¹⁶ although the design and framework are unique and not applicable to the RCT with multiple DBPCFCs, which are increasingly prevalent in the OIT literature today.^{17,18} The extension of these discrete-time methods to RCTs with multiple DBPCFCs poses unique challenges to fit the common framework in such a way that is easily translatable for clinicians. Thus, there is a need for a more granular, intuitive definition of outcomes that incorporates longitudinal information and right-censoring to describe both time to failure and dose at which failure occurs.

This paper aims to provide methodology to expand the discrete-time model proposed by Chinchilli et al. into the longitudinal setting and to introduce a clinical outcome that allows for modeling and testing of the trajectory of study success in RCTs with multiple study DBPCFCs. The remainder of the paper is organized as follows: Section 2 provides an overview of the DBPCFC study design and commonly applied statistical methods to assess study outcome; Section 3 describes discrete time-to-event methods and their specific applications to food allergy RCTs; Section 4 introduces our

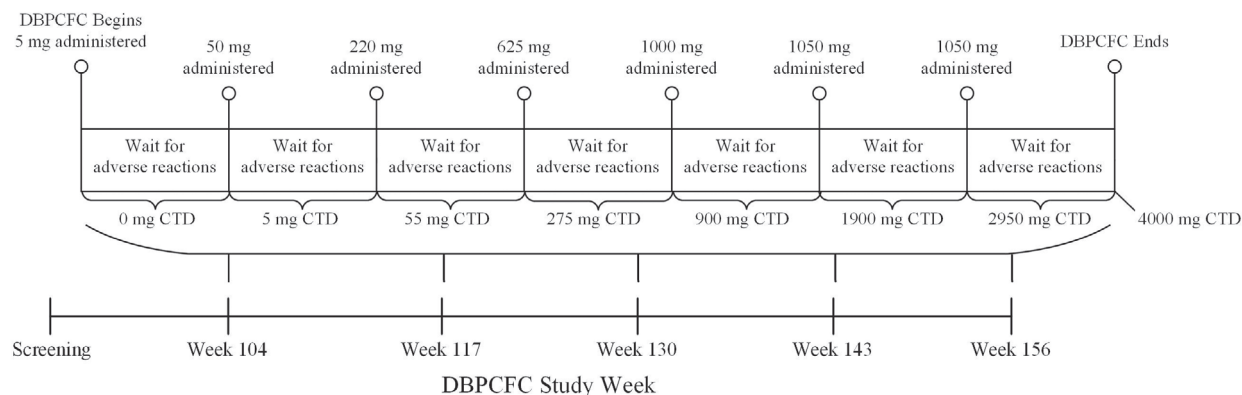


FIGURE 1 Example of a double-blind placebo-controlled food challenge (DBPCFC) dose escalation timeline that are repeated throughout the study. Doses during each DBPCFC are administered at equally spaced time intervals in order to assess the effect of ingested dose on occurrence of adverse events. CTD = cumulative tolerated dose

novel dose-time outcome; Section 5 illustrates the proposed methodology using the POISED study; and Section 6 provides critical discussion of the described methods.

2 | STUDY DESIGN

2.1 | POISED—A motivating example

The POISED trial was conducted at the Sean N. Parker Center for Allergy and Asthma Research at Stanford university with the primary goal of evaluating the superiority of OIT with short maintenance followed by avoidance vs placebo. More specifically, participants were randomized into one of three arms: (a) build up to and maintenance of 4000 mg peanut OIT through 104 weeks followed by avoidance (peanut-0), (b) build up to and maintenance of 4000 mg peanut OIT through 104 weeks followed by continued OIT dosing of 300 mg daily (peanut-300), or (c) placebo throughout. A screening DBPCFC was administered where participants consumed doses periodically up to 500 mg of cumulative peanut protein (steps: 5, 20, 50, 100, 100, 100, 125) to confirm peanut allergy and establish a baseline threshold. After 104 weeks of build up to and maintenance of 4000 mg OIT (either peanut or placebo), participants had their first post-screening DBPCFC of up to 4000 mg of cumulative peanut protein (steps: 5, 50, 220, 625, 1000, 1050, 1050) at week 104 (Figure 1). The week 104 challenge assessed desensitization, while the week 117 challenge assessed sustained unresponsiveness among those in the peanut-0 arm. Challenges were repeated every 3 months up through week 156 for participants who were able to tolerate a cumulative dose of 4000 mg at a challenge without experiencing dose-limiting adverse events. Participants who failed to reach a cumulative tolerated dose (CTD) of 4000 mg at a DBPCFC were considered failures and not re-challenged at later time points.

The primary endpoint of reaching and passing the DBPCFC to a CTD of 4000 mg at both 104 and 117 weeks was used to assess the efficacy of peanut OIT on sustained unresponsiveness. Demographic and clinical features such as peanut skin prick test, peanut-specific IgE and IgG4 were collected at baseline and at each study challenge. The study design and results have been previously published.¹⁹ Due to variations in the design of OIT RCTs, the following sections are written in the context of the POISED trial framework.

2.2 | Double-blind placebo-controlled food challenge

Confirmation of food allergy is most commonly conducted via skin prick tests (SPT), evaluation of food allergen-specific immunoglobulin E (IgE), and trial elimination diets³. However, these tests frequently produce false positive and negative results.² Therefore, the DBPCFC remains the gold standard for food allergy identification. The following outlines the design of DBPCFCs for the POISED trial. While study designs may vary from center to center, the fundamental principles described here can be generalized.

All participants with a suspected food allergy, defined as an allergen-specific IgE > 0.35 kU/L and/or an SPT > 3 mm above the negative control, undergo a qualifying screening DBPCFC to confirm allergy diagnosis and baseline reactivity.²⁰ A DBPCFC - standardized in methodology with a pre-established maximum cumulative food protein dose - involves sequentially escalating the dose of food protein ingested by the participant at equally spaced time intervals as tolerated (Figure 1). At each dose step, vital signs and any dose-related allergic adverse events are evaluated by a clinician. If a participant reaches the maximum cumulative dose without any dose-related adverse events, a diagnosis of food allergy is ruled out to that CTD and the participant becomes ineligible for trial inclusion. If a participant has a dose-related adverse event, the challenge is stopped and the maximum ingested dose, cumulative ingested dose, and CTD are recorded, along with adverse event information.

2.3 | Build up, maintenance, and study outcomes

Given a participant is eligible based on screening DBPCFCs and other eligibility criteria, each participant begins OIT and undergoes build up and maintenance to a pre-specified dose amount. Then, based on treatment assignment, participants will either discontinue OIT (to assess sustained unresponsiveness in the peanut-0 arm) or maintain OIT on a lower dose (to assess desensitization in the peanut-300 arm).

After a pre-determined amount of time, participants return to the clinic to undergo additional DBPCFCs in order to evaluate the efficacy of treatment over time. These challenges are designed similarly to that of the screening challenge, with the exception that participants who reach the maximum dose without dose-related adverse events are considered to successfully pass the challenge. Note that within the POISED trial, the maximum cumulative dose is higher for the post-screening DBPCFC.

2.4 | Traditional framework

2.4.1 | Outcomes

As mentioned above, the primary outcome is generally defined as binary - as an indicator for whether or not a participant is successful at passing a DBPCFC to a pre-determined CTD by a certain study week. The primary outcome is then compared across the study treatment arms using appropriate methods.

In the RCT framework, we compare study arms using intent-to-treat principles (ITT) - ie, in the so-called ITT population, defined as all participants who underwent randomization.²¹ To that end, any participant that drops out prior to being challenged may be given an outcome value reflecting failure in order to include them in the analysis and apply ITT principles. Common reasons for study drop out in RCTs of OIT include withdrawal of consent, non-compliance, and general loss to follow up.^{6,19} Reasons are not anticipated to be related to the outcome.

2.4.2 | Statistical methods applied in traditional framework

Fisher's exact or Chi-squared tests are commonly used to determine whether there are significant differences in the proportion of subjects with success between treatment arms at the challenge or timepoint of interest.⁶⁻⁹ Odds ratios and 95% CIs are generally reported.

While these approaches address the issue of failure or success at a certain dose threshold and pre-specified treatment duration, they do not incorporate the more granular dose information for participants who fail the challenge, distinguish between participants who drop out of the study vs those that fail, or provide information on overall study success. Applying time-to-event methodology to a combined binary pass/fail and dose outcome not only allows for further quantification of treatment sensitivity within those that fail, but addresses data truncation issues due to drop out or absence of event occurrence, both historically common events in RCTs of OIT.

Further, while the likelihood function for the logistic model approximates the partial likelihood of the Cox model when the event rate is rare and short follow-up time is observed, coefficients of the logistic model become less certain as follow-up time increases.^{22,23}

3 | TIME-TO-EVENT ANALYSIS

3.1 | Censoring

The key feature that distinguishes time-to-event analysis from other statistical methods is that time-to-event data commonly have censoring or partially observed data. In the context of POISED, either failure during a DBPCFC is observed to occur and the dose steps between which it occurred are measured, or failure did not occur during the study period (ie, participant dropped out between two DBPCFC time points) and it is only known that a participant was failure-free up to a certain time. In the former case, failure is considered interval-censored as we only know that failure occurred between two observed time points during the DBPCFC (ie, dose steps) and not the exact timing of failure.²⁴ In the latter case, the participant would be right-censored.¹⁰ While we are unable to determine whether a participant would have failed if they were followed longer, we are able to make use of the fact that they were failure-free up to a specific time.

Regardless of the type of censoring, examination of the censoring mechanism needs careful consideration. A common assumption made in analysis of time-to-event data is noninformative censoring, which occurs when the distribution of survival times provides no additional information about the distribution of censoring times, that is, subjects who drop out of the study do so for reasons unrelated to the study. More formally, we define the noninformative interval censoring assumption as follows:

$$P(T \leq t | L = l, R = r, L < T \leq R) = P(T \leq t | l < T \leq r)$$

Where the interval $(L, R]$ does not provide any additional information for survival time T , except that T lies somewhere between l and r (which are the realizations of L and R). Consider the following two scenarios in which informative censoring would be of concern: in a study comparing disease-free survival, a subject on placebo may be more likely to experience disease recurrence and be lost to follow up due to declining health. Alternatively, a subject could be cured of the disease due to the treatment drug and therefore no longer feel the need to continue in the trial. Disease-free survival estimates would be calculated on patients who continued to be followed-up in the study, which would be overestimated in the placebo arm and underestimated in the treatment arm. Specifically for interval-censored data, an examination scheme is considered as satisfying the noninformative censoring assumption if all subjects are examined at preassigned intervals, such as seen in the POISED trial.¹⁰

3.2 | Discrete time-to-event analysis

While the timing of an event is generally regarded on a continuum, there are situations in which measuring time (or any ordinal value) on a discrete scale is more appropriate. Two primary forms of discrete time-to-event data occur when (a) measurements represent natural numbers that form intrinsically discrete measurements, and (b) events are represented in underlying time intervals that form grouped or interval-censored data, as seen in the POISED trial. In the latter case, we only know that the true failure dose (“time”) is greater than the last dose at which no reaction was observed and less than or equal to the dose at which the reaction was first observed. In a grouped data framework, there are often observations that have the same survival times (ie, “ties”), which violate assumptions of some models used for continuous time. While the interpretation of the estimates are different, the modeling approaches used for both types of data are comparable, particularly when grouped data intervals are equal in length.²⁵

Even if data are truly discrete, a common approach is to assume the event occurred at an observed time and apply standard continuous time-to-event models. However, this approach has been shown to underestimate the standard errors of the estimated parameters of interest.²⁴ Further, statistical methods designed for truly discrete event times have the advantage of formulating hazards as conditional probabilities, avoiding issues arising from tied event times, and having the ability to be embedded into a generalized linear model framework.²⁵ This motivates application of such methods to the clinical trial setting described here.

More formally, let discrete time take values in $\{1, \dots, k\}$ where a participant has k underlying intervals $[a_0, a_1)$, $[a_1, a_2)$, $\dots$, $[a_{q-1}, a_q)$, $[a_q, \infty)$, with $q = k - 1$. The first observed time a_0 is generally assumed $= 0$, while a_q is represented by the last follow-up time. Let T denote the survival time of interest, where $T = t$ is observed if failure occurs within the interval $[a_{t-1}, a_t)$. The hazard function $h(t)$ in the discrete setting can be expressed as

$$h(t|\mathbf{x}) = P(T = t \mid T \geq t, \mathbf{x})$$

or the conditional probability for failure in the interval $[a_{t-1}, a_t]$ granted a participant reaches the interval, given covariates $\mathbf{x} = (x_1, \dots, x_p)$.²⁵ At a fixed t , the hazard function can be expressed as a binary response that differentiates whether or not the event takes place in category t , given $T \geq t$. In other words, we can examine the conditional probability of experiencing the event at time t given event-free status at time $t-1$. Using the complementary log-log (clog-log) link function, the discrete hazard model can be written as

$$h(t|\mathbf{x}) = 1 - e^{(-e^{(\gamma_{0t} + \beta_1 x_1 + \dots + \beta_p x_p)})}.$$

This model is referred to as the grouped proportional hazards model which is widely seen as the discrete version of the Cox proportional hazard model.²⁵ In contrast to the discrete logit link model that treats the levels of the outcome as truly discrete, the grouped proportion hazards model is appropriate when the outcome is a discretized version of an underlying continuous value defined by intervals.²⁶ The model can also be expressed as

$$\log(-\log(P(T > t \mid T \geq t, \mathbf{x}))) = \gamma_{0t} + \beta_1 x_1 + \dots + \beta_p x_p$$

with separate intercepts for each time interval (γ_{0t}), which can be interpreted as the baseline hazard. Note that while γ_{0t} depends on time, change in covariates does not. Here, the discrete hazard is the conditional probability that a participant will experience the event in category t given the participant was event-free in the previous category. Model coefficients are represented as log-hazard ratios, defined as the log-transformed ratio of the predicted hazard function under two levels of an explanatory variable. Exponentiating the coefficient allows interpretation of the estimate as a hazard ratio (HR). An HR greater than 1 indicates an increased hazard of experiencing the event in one study arm relative to another, whereas an HR less than 1 indicates a decreased hazard of the arm relative to the other. As with the Cox proportional hazards model, the discretized version formally defines the hazard ratio as

$$\frac{\gamma_{0t} \exp(\beta_{1a})}{\gamma_{0t} \exp(\beta_{1b})} = \exp(\beta_{1(a-b)}),$$

which is independent of time t , and thus assumes the hazard is proportional over time.

4 | THE DOSE-TIME AXIS

In some RCTs of OIT, such as POISED, there are repeated DBPCFCs over time throughout the study period, and within each DBPCFC, food allergen doses are administered at equally spaced time intervals (Figure 1). Each repeat challenge has equivalent dose steps, which can be seen as a natural extension of time. Instead of considering a time-to-event outcome, we consider a discrete dose-to-event outcome. Specifically, we use the CTD - the last dose at which a participant tolerated the allergen without presence of an allergic reaction during a challenge - as a proxy for time. The event of interest is challenge failure, defined as not reaching a pre-specified CTD during a specific challenge. Thus, the outcome of interest is dose-to-failure.

Sustained unresponsiveness and desensitization are evaluated after a specific period of time by conducting repeat DBPCFCs. Instead of evaluating these outcomes at one study time point of interest, we exploit information on multiple challenge outcomes throughout the study. Thus, we are able to assess dose-to-loss of desensitization or sustained unresponsiveness, that is, failure, by considering data from all of a participant's challenges.

Each participant will have multiple CTDs corresponding to the number of challenges they have completed. We evaluate dose-to-failure at any challenge over the entire study period by extending the CTD at each challenge, resulting in a dose-time axis that covers the entire study period. The CTD is summed across all a participant's post-screening challenges (1 to n) through their last challenge:

$$CTD \text{ time} = \sum_{i=1}^n CTD_i$$

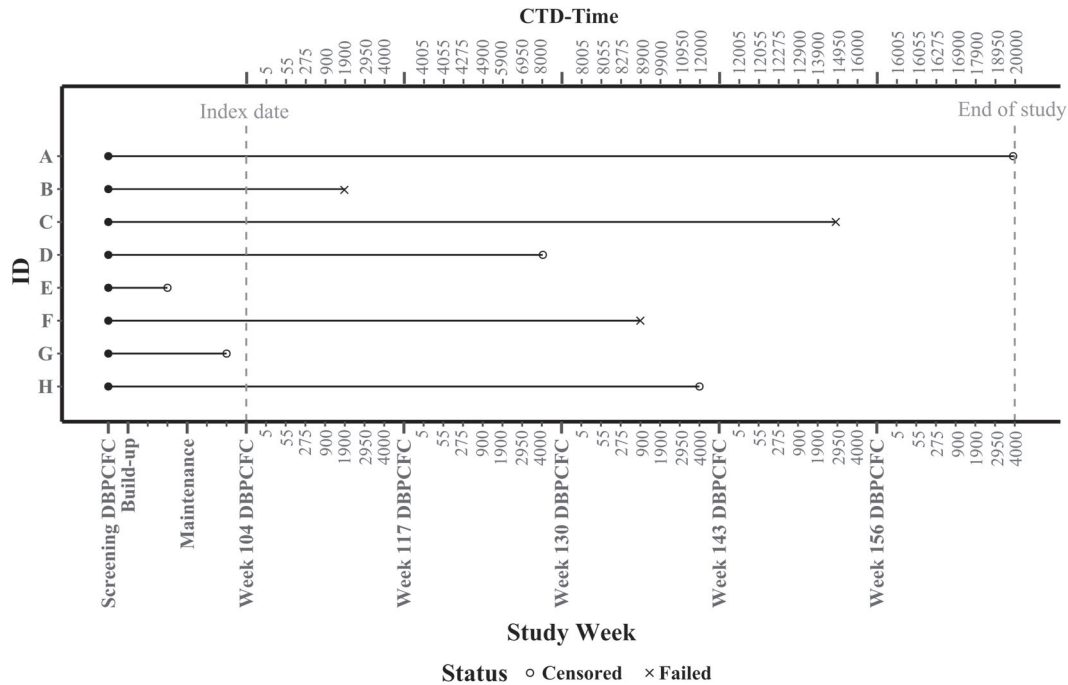


FIGURE 2 Illustration of various outcome scenarios based on the study design. Values between the doubleblind placebo-controlled food challenge (DBPCFC) study weeks represent the cumulative tolerate doses (CTD)(mg) administered within each challenge, with the corresponding CTD-time on the upper x-axis. Censoring occurs if a participant reaches the end of the study without experiencing failure, or does not return for subsequent challenges after completion of a challenge

Therefore, participants who progress through the study will have a higher CTD-time than those who fail or drop out earlier. The maximum CTD-time is the highest challenge CTD multiplied by the number of possible challenges. To illustrate, if the maximum CTD at a given challenge is 4000 mg and there are 5 post-screening challenges, then the maximum CTD-time a participant can have is $4000 \times 5 = 20\,000$. The discrete-time model for CTD-time is specified below, using the clog-log link:

$$\log(-\log(P(T > t \mid T \geq t, \mathbf{x}))) = \beta_1 CTDtime_1 + \dots + \beta_p CTDtime_p + \beta_{p+1} treatment,$$

which represents the conditional probability that a participant will experience challenge failure at CTD-time t given the participant did not fail at CTD-time $t-1$.

Failure is defined as not passing the pre-specified CTD threshold at any challenge. Participants are censored if they pass all challenges and never experience the event or if they pass a challenge and do not return for subsequent challenges. Figure 2 below illustrates possible scenarios. Participants E and G both drop out of the study prior to the index date (ie, first post-screening challenge) and are therefore excluded from the dose-to-failure analysis. Participants B, C, and F all experience failure quantified by a CTD-time of 1900 mg, 11 350 mg, 8900 mg, respectively. Participants D and H successfully complete their week 117 and 130 challenges, but do not return for subsequent challenges and are therefore censored at their CTD-time of 8000 mg and 12 000 mg, respectively. Finally, participant A successfully completes every study challenge and is censored at the end of the study with a CTD-time of 20 000 mg. Summing the CTD over all challenges provides an intuitive longitudinal outcome and allows for back-transformation to the dose-challenge scale.

5 | APPLICATION

To illustrate the application of the proposed discrete-time tools, the following analysis focuses on comparing the two active treatment arms (peanut-0 and peanut-300) from the POISED trial. Participants in the placebo arm were excluded as only one participant passed the primary endpoint (ie, the DBPCFC at weeks 104 and 117). An overview of the original

TABLE 1 Data set up for discrete-time time-to-event analysis

a. Person-period					
ID	Treatment	DBPCFC	CTD	CTD-time	Failure
A	Peanut-0	Screening	175	–	–
A	Peanut-0	Week 104	4000	4000	0
A	Peanut-0	Week 117	4000	8000	0
A	Peanut-0	Week 130	4000	12 000	0
A	Peanut-0	Week 143	4000	16 000	0
A	Peanut-0	Week 156	4000	20 000	0
B	Peanut-300	Screening	25	–	–
B	Peanut-300	Week 104	1900	1900	1
–	–	–	–	–	–
E	Peanut-0	Screening	275	–	–
F	Peanut-300	Screening	175	–	–
F	Peanut-300	Week 104	4000	4000	0
F	Peanut-300	Week 117	4000	8000	0
F	Peanut-300	Week 130	900	8900	1
b. Person-level					
ID	Treatment	CTD-time	Failure		
A	Peanut-0	20 000	0		
B	Peanut-300	1900	1		
–	–	–	–		
E	Peanut-0	–	–		
F	Peanut-300	89 000	1		

Structure of person-period and person-level challenge data using scenarios illustrated in Figure 2. DBPCFC: double-blind, placebo-controlled food challenge; CTD: cumulative tolerated dose (mg). Participants are censored (0) if they never return after a successful challenge (defined as reaching 4000 mg) or reaching study end without failing a challenge. Once a challenge is failed, a participant cannot return for subsequent challenges.

data structure is displayed in Table 1. As the maximum CTD at each post-screening challenge was 4000 mg, participants who passed all 5 DBPCFCs had a maximum CTD-time of 20 000 mg. Structure of person-period and person-level challenge data using scenarios illustrated in Figure 2. DBPCFC: double-blind, placebo-controlled food challenge; CTD: cumulative tolerated dose (mg). Participants are censored (0) if they never return after a successful challenge (defined as reaching 4000 mg) or reaching study end without failing a challenge. Once a challenge is failed, a participant cannot return for subsequent challenges.

5.1 | Data format

Cumulative tolerated dose-time (CTD-time) and the censoring indicator (Failure) were expanded to form a sequence of binary responses, where Failure indicated whether or not a reaction was elicited in the dose-time interval. This structure is similar to the format for the Cox model, but needs to be manually reformatted for the grouped time setting. Therefore, the person-level data structure outlined in Table 1b was expanded to the person-period dataset shown in Table 2. Because each participant now has multiple records, the original n records (equal to the sample size) becomes $N = (a_1 + a_2 + \dots + a_n)$ records in the person-period data set. To illustrate, Participant A passed every challenge in the study, yielding a row for every distinct CTD-time and no indicator for failure. Participant B failed the study at a CTD-time of 1900 mg and therefore

TABLE 2 Analytic data structure for discrete-time analysis

ID	Treatment	CTD-time	Failure
A	Peanut-0	0	0
A	Peanut-0	5	0
A	Peanut-0	55	0
A	Peanut-0	220	0
A	Peanut-0	275	0
A	Peanut-0	–	0
A	Peanut-0	8000	0
A	Peanut-0	–	0
A	Peanut-0	20 000	0
B	Peanut-300	0	0
B	Peanut-300	5	0
B	Peanut-300	55	0
B	Peanut-300	275	0
B	Peanut-300	900	0
B	Peanut-300	1900	1
–	–	–	–
E	Peanut-0	–	–
F	Peanut-300	0	0
F	Peanut-300	5	0
F	Peanut-300	55	0
F	Peanut-300	220	0
F	Peanut-300	275	0
F	Peanut-300	–	0
F	Peanut-300	8000	0
F	Peanut-300	8005	0
F	Peanut-300	8055	0
F	Peanut-300	8275	0
F	Peanut-300	8900	1

Participants are censored (failure = 0) if they never return after a successful challenge (defined as reaching 4000 mg) or reach the study end without failing a challenge. Once a challenge is failed, a participant cannot return for subsequent challenges.

has 6 rows corresponding to the distinct CTD-time values up through 1900 mg, with the failure indicator only present in the last row. Participants are censored (failure = 0) if they never return after a successful challenge (defined as reaching 4000 mg) or reach the study end without failing a challenge. Once a challenge is failed, a participant cannot return for subsequent challenges.

5.2 | Characterizing time-to-study failure

Figure 3 shows the Kaplan-Meier plot for dose-to-failure by the two active treatment arms. Fourteen participants dropped out of the study prior to the first post-screening challenge and are therefore censored at the “Pre-104” time point. Only one participant (in peanut-300) failed the study during the week 104 challenge, while 21 participants (8 in peanut-0 and

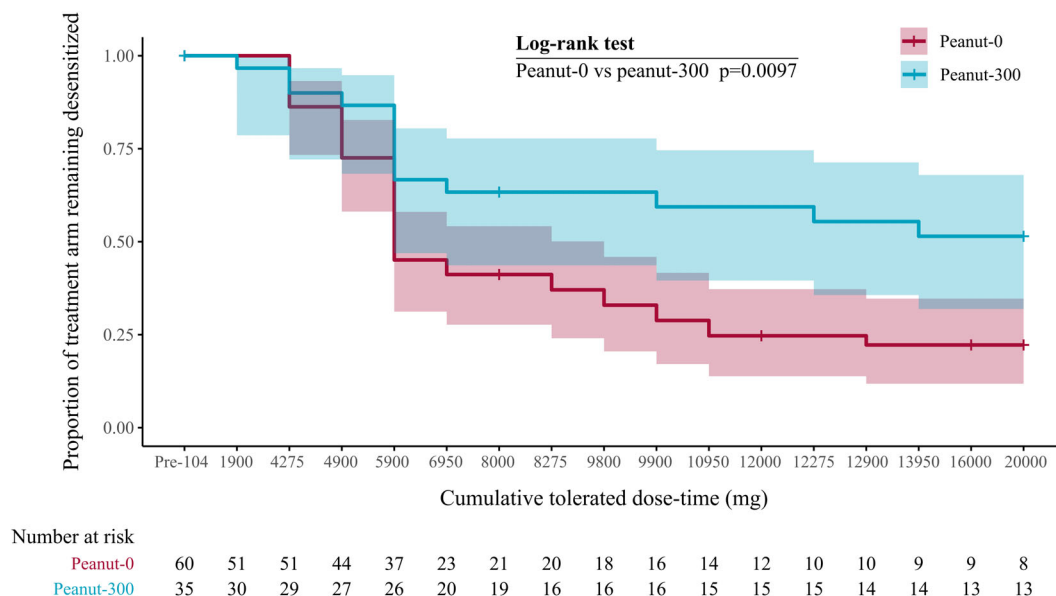


FIGURE 3 Time until challenge failure. Drop in risk at the pre-104 time point refers to participants who did not reach the week 104 challenge (ie, dropped out). To condense the figure, some dose steps are not shown as no censoring or failures occurred in either treatment arm. Participants who passed a challenge but did not return for a subsequent challenge or dropped out before the week 104 challenge were censored at their last cumulative tolerated dose (denoted by the vertical tick in the Kaplan-Meier curve). Colored bands represent 95% CIs [Colour figure can be viewed at wileyonlinelibrary.com]

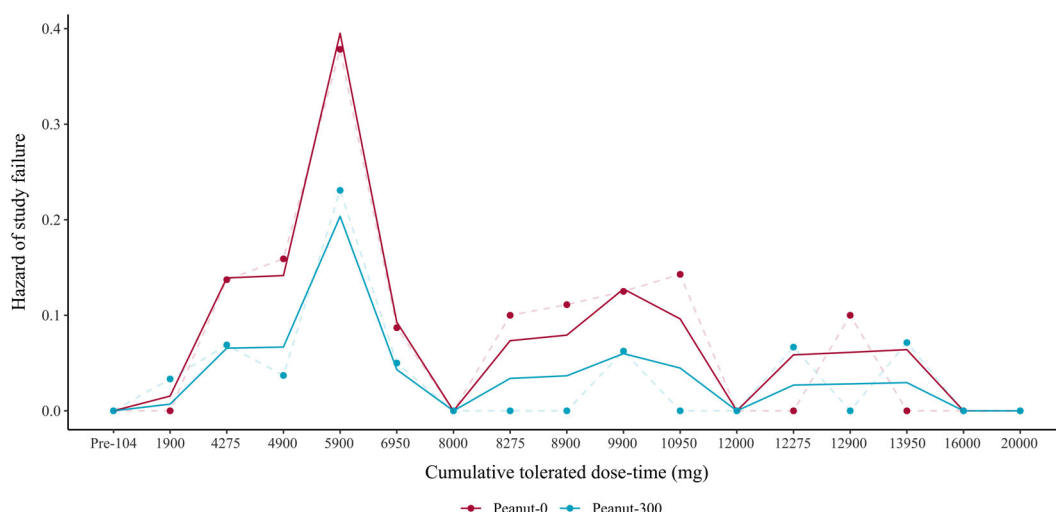


FIGURE 4 Hazard of study failure over dose-time by treatment arm. Hazards were extracted from the grouped time model (clog-log link). Dashed lines and dots represent the empirical hazard estimate by treatment arm at each dose-time. To condense the figure, some dose steps are not shown as no censoring or failures occurred in either treatment arm [Colour figure can be viewed at wileyonlinelibrary.com]

13 in peanut-300) succeeded in passing all subsequent DBPCFCs. Significant differences in time-to-study failure were found between the two treatment arms (log-rank $P = 0.0097$).

The highest risk of study failure occurs at a CTD-time of 5900 mg in both treatment arms, equivalent to failing the 1900 mg dose in the week 117 DBPCFC, with participants in the peanut-0 arm having a higher overall risk of study failure compared to those in the peanut-300 arm (Figure 4). A similar, although not as dramatic, rise in hazard of failure also occurs at the 1900 mg dose, but in the following challenge week (DBCPFC week 130).

While there is a second peak of risk of failure after week 117, risk generally decreases as participants progress through the study. Hazards that drop to zero at CTD-time of 8000, 12 000, and 16 000 mg correspond to participants who reach and

TABLE 3 Comparisons of time-to-event model estimates

	Cox proportional hazards	Grouped time (clog-log link)	Discrete time (logit link)
Treatment	Hazard ratio (95% confidence interval)		
Peanut-300 (vs Peanut-0)	0.46 (0.25, 0.85)	0.45 (0.24, 0.82)	0.49 (0.28, 0.86)

tolerate the maximum challenge CTD (4000 mg) in weeks 117, 130, and 143, respectively, and therefore cannot experience challenge failure. All participants who reached the week 156 challenge were able to reach the 4000 mg CTD without symptoms, thus resulting in a hazard of zero for week 156. Compared to participants in the peanut-0 arm, participants in the peanut-300 arm had approximately half the risk of failing the study at any point (HR: 0.45, 95% CI: 0.24, 0.82) (Table 3). Code for fitting these models can be found in the Appendix.

5.3 | Binary vs time-to-event outcome

The original clinical trial results reported no significant differences in odds of reaching 4000 mg between the peanut-0 and peanut-300 arms at weeks 104 (odds ratio [OR]: 1.2, 95% CI: 0.3-4.1) and 117 (0.5, 0.2-1.2) DBPCFCs.¹⁹ However, differences were found in subsequent weeks, with participants in the peanut-300 arm maintaining a higher proportion of success compared to those in the peanut-0 arm. While the originally published results provide an adequate snapshot of success at each challenge time point and suggest a decline in success over time, we are able to formally evaluate dose-to-failure across all time points through the use of the discrete-time modeling.

5.4 | Power estimation

Let us refer back to the primary outcome from the POISED trial, defined as the proportion of participants who reached and passed the DBPCFC to a CTD of 4000 mg at weeks 104 and 117 (ie, success). The study was designed to detect a difference in success rates of 35% in the peanut-0 arm vs a 5% success rate in the placebo arm. Reconstructed as failure rates, the assumed rate of failure for the peanut-0 and placebo arms would be 65% and 95%, respectively. Using a two-sided Fisher's Exact test, with alpha level of 0.05 and assuming 60 and 25 participants in the peanut-0 and placebo arms, respectively, the study would achieve 81% power to detect these differences in failure rates, corresponding to an OR of 0.098.²⁷ This calculation assumes no study drop out. Translating to a time-to-event framework, let us assume the same alpha and sample size per arm, with failure rates of 65% and 95% at the 4000 mg dose in week 117 (ie, dose-time period 14 out of 35), with a 10% drop out rate. Instead of comparing the proportion of failures in each arm at a single time point, we can use the log-rank test to compare survival across the entire study period.²⁸ Using the log-rank test, 90% power would be achieved to detect a hazard ratio for overall study failure of 0.35. If no drop out is assumed, similarly to what was done for the binary outcome, power increases to 99%. If we were to compare power and required sample size needed between these two approaches by varying censoring assumptions (assuming a sample size of 100 in each arm and the same OR and HR as described above), we see that while power slightly decreases as the proportion censored increases in the time-to-event outcome, total sample size needed exponentially increases in order to detect the same effect size (Table 4). Censoring for the binary outcome is represented by the dropout rate, assuming a complete cases analysis, where the adjusted sample size is calculated by dividing the total sample size by 1 - the dropout rate. This simple example shows the increase in power that time-to-event outcomes provide over simple analyses of proportions.

6 | DISCUSSION

OIT clinical trials today are becoming more complex as they increasingly involve food challenges to test response to therapy (either desensitization or sustained unresponsiveness) at multiple time points throughout the study duration. Current practice ignores the granularity of the data at a given timepoint where a simple pass/fail for a DBPCFC is typically defined as the primary outcome even though dose is recorded. Further the trajectory of sustained responsiveness is highly relevant to assessing the efficacy of therapy over time but needs a formal approach for analysis. Use of simple analytic tools

Censoring rate	Log-rank		Fisher's exact	
	Total N	Power	Total N	Power
0	200	1.0000	200	0.9997
0.1	200	0.9996	222	0.9999
0.2	200	0.9942	250	1.0000
0.3	200	0.9711	286	1.0000
0.4	200	0.9201	334	1.0000
0.5	200	0.8412	400	1.0000

TABLE 4 Comparison of sample size and power of time-to-event versus binary outcomes

to describe efficacy are accessible and easy to interpret, however these tools do not have the capacity to provide a formal framework for assessing longitudinal efficacy of treatment on allergic response, an important property of OIT that needs evaluation. Chinchilli and others have provided a means to incorporate the right-censored nature of DBPCFC data from baseline to a single end-of-study challenge. We extend these principles to the longitudinal setting by summing the CTD across all completed DBPCFC, demonstrating the ease of interpretation through graphical means of displaying results, and demonstrating advantages over the traditional approaches appropriate for binary data. We additionally provide example code for implementing the proposed models for ease of adopting the tools.

As previously described, discrete time-to-event models may be fit with various link functions. While both the clog-log and logit links produce similar estimates in practice when the number of discrete time periods is large, the estimates yield different interpretations. More specifically, implementation of the logit link results in the exponentiated estimates interpreted as odds ratios and assumes proportional odds between arms, whereas application of the clog-log link results in exponentiated estimates as hazard ratios with the assumption of proportional hazards between arms. While the logit link is preferable when modeling truly discrete measurements, such as number of doses, the clog-link should be used when modeling measurements that have an underlying continuous distribution but are measured or observed in discrete groups, such as dose amount. In situations where it is unclear which link function best describes the data generated, one may consider fitting both models and comparing fit statistics to determine the most appropriate choice of link.²⁹

This study is not without limitations. The present derivation of the newly proposed outcome does not account for the differential length of time between challenges in the study design, as “time” here is really “dose-time” which only progresses during a challenge. Thus, the current framework may overestimate the risk of failure when assessing the hazard at each study “time”. Additionally, while the timing of post-screening DBPCFCs were equally spaced in the POISED trial, future studies may consider methods to account for time between challenges or weighting challenge times differently. We recognize this extension as important future work. Further, we acknowledge that alternative modeling frameworks can also be implemented in this application. For example, while discrete-time frailty models and mixed effects linear regression models were considered, they were not pursued.³⁰ While both approaches model the heterogeneity across patients while estimating the risk of failure, our proposed approach using CTD-time can be back-transformed to the dose scale by week of challenge, thus resulting in ease of interpretation. The ability of our proposed approach to back-transform to an interpretable scale enables a full understanding of risk of failure at the specific time during the study, allowing for providers and patients to have a better perception of timing of risk. We recognize that the principles laid out in this paper can extend to alternative models in other studies. While our study does not provide power calculation tools specifically for designing studies that utilize this outcome and analytic model, tools for calculating power and sample size for discrete-time models are readily available.^{31–33}

We have provided an example using the log-rank test that shows an increase in power over Fisher's Exact test with similar assumptions. While increased power is an added benefit over approaches using binary outcomes, a stronger advantage of time-to-event approaches lies in its more flexible missing data assumptions. In conducting a complete case analysis (where challenges not observed are excluded from the analysis), validity of findings relies on an assumption that missing values are missing completely at random whereas imputation of missing values that assumes a worst-case scenario (ie, assuming failure if a participant does not return for a subsequent challenge) assumes data are missing not at random. Both scenarios are likely to introduce bias and yield misleading results, and are the most commonly applied assumptions by clinical investigators, as they are the easiest to apply. Even in scenarios where power in the binary setting is not greatly impacted, time-to-event approaches are superior because they rely on a much more flexible, and realistic, missing data assumption of the data being missing at random.

In order to obtain unbiased estimates from the proposed discrete time-to-failure models, the assumption of non-informative censoring needs to be fulfilled. In the DBPCFC trial setting, this assumption may be violated, for example, if a patient is symptomatic at home and therefore does not come into the clinic to be challenged. Alternatively, if a patient is asymptomatic and does not feel the need to continue in the study, the loss to follow up would be informative and thus threaten model validity. For the POISED trial, we believe this underlying assumption to hold.

We propose a novel application of time-to-event models to the OIT clinical trial setting. While the proposed time-to-event framework itself is not new, our approach provides the first derivation of a DBPCFC study outcome that incorporates multiple longitudinal challenges in an easily digestible way. We do not propose elimination of the traditional dichotomous framework as it lends itself to straightforward calculations of sample size and power particularly for those studies designed with one post-screening challenge; however, we advocate for the discrete-time methodology for OIT trials with multiple longitudinal challenges that we have shown can leverage a more granular examination of the efficacy of treatment on overall study performance by exploiting the presence of repeat challenge data.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Sandra Andorf  <https://orcid.org/0000-0002-3093-2568>

Manisha Desai  <https://orcid.org/0000-0002-6949-2651>

REFERENCES

1. Sampson HA, Aceves S, Bock SA, et al. Food allergy: a practice parameter update—2014. *J Allergy Clin Immunol*. 2014;134:1016-1025.e43.
2. Chapman JA, Bernstein L, Lee R, et al. Food allergy: a practice parameter. *Ann Allergy, Asthma Immunol*. 2006;96:S1-S68.
3. Allan Bock S, Sampson HA, Atkins FM, et al. Double-blind, placebo-controlled food challenge (DBPCFC) as an office procedure: a manual. *J Allergy Clin Immunol*. 1988;82:986-997.
4. Sampson HA, van Wijk RG, Bindslev-Jensen C, et al. Standardizing double-blind, placebo-controlled oral food challenges: American Academy of allergy, asthma & immunology–European academy of allergy and clinical immunology PRACTALL consensus report. *J Allergy Clin Immunol*. 2012;130:1260-1274.
5. Moran TP, Burks AW. Is clinical tolerance possible after allergen immunotherapy? *Curr Allergy Asthma Rep*. 2015;15:23.
6. du Toit G, Roberts G, Sayre PH, et al. Randomized trial of Peanut consumption in infants at risk for Peanut allergy. *N Engl J Med*. 2015;372:803-813.
7. Skripak JM, Nash SD, Rowley H, et al. A randomized, double-blind, placebo-controlled study of milk oral immunotherapy for cow's milk allergy. *J Allergy Clin Immunol*. 2008;122:1154-1160.
8. Andorf S, Purington N, Kumar D, et al. A phase 2 randomized controlled multisite study using Omalizumab-facilitated rapid desensitization to test continued vs discontinued dosing in Multifood allergic individuals. *EClinicalMedicine*. 2019;7:27-38.
9. Bird JA, Spergel JM, Jones SM, et al. Efficacy and safety of AR101 in Oral immunotherapy for Peanut allergy: results of ARC001, a randomized, double-blind, placebo-controlled phase 2 clinical trial. *J Allergy Clin Immunol Pract*. 2018;6:476-485.e3.
10. Leung KM, Elashoff RM, Afifi AA. Censoring issues in survival analysis. *Annu Rev Public Health*. 1997;18:83-104.
11. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457-481.
12. George B, Seals S, Aban I. Survival analysis and regression models. *J Nucl Cardiol*. 2014;21:686-694.
13. Chinchilli VM, Fisher L, Craig TJ. Statistical issues in clinical trials that involve the double-blind, placebo-controlled food challenge. *J Allergy Clin Immunol*. 2005;115:592-597.
14. Fernández-Rivas M, Garrido Fernández S, Nadal JA, et al. Randomized double-blind, placebo-controlled trial of sublingual immunotherapy with a Pru p 3 quantified peach extract. *Allergy*. 2009;64:876-883.
15. Garrido-Fernández S, García BE, Sanz ML, Echechipía S, Lizaso MT, Tabar AI. Are basophil activation and sulphidoleukotriene determination useful tests for monitoring patients with peach allergy receiving sublingual immunotherapy with a Pru p 3-enriched peach extract? *J Investig Allergol Clin Immunol*. 2014;24:106-113.
16. Withana Gamage PW, Chaudari M, McMahan CS, Kim EH, Kosorok MR. An extended proportional hazards model for interval-censored data subject to instantaneous failures. *Lifetime Data Anal*. 2020;26:158-182. <https://doi.org/10.1007/s10985-019-09467-z>.
17. Wright BL, Kulis M, Orgel KA, et al. Component-resolved analysis of IgA, IgE, and IgG4 during egg OIT identifies markers associated with sustained unresponsiveness. *Allergy*. 2016;71:1552-1560.
18. Caubet JC, Lin J, Ahrens B, et al. Natural tolerance development in cow's milk allergic children: IgE and IgG4 epitope binding. *Allergy*. 2017;72:1677-1685.
19. Chinthrajah RS, Purington N, Andorf S, et al. Sustained outcomes in oral immunotherapy for peanut allergy (POISED study): a large, randomised, double-blind, placebo-controlled, phase 2 study. *Lancet*. 2019;394:1437-1449.
20. Sindher S, Long AJ, Purington N, et al. Analysis of a large standardized food challenge data set to determine predictors of positive outcome across multiple allergens. *Front Immunol*. 2018;9:2689.

21. Gupta SK. Intention-to-treat concept: a review. *Perspect Clin Res*. 2011;2:109-112.
22. Green MS, Symons MJ. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiological studies. *J Chron Dis*. 1983;36:715-723.
23. Annesi I, Moreau T, Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Stat Med*. 1989;8:1515-1521.
24. Lindsey JC, Ryan LM. Methods for interval-censored data. *Stat Med*. 1998;17:219-238.
25. Tutz G. Continuous versus discrete survival. *Modeling Discrete Time-to-Event Data*. Switzerland: Springer International Publishing; 2016:5-6.
26. Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. *Int J Epidemiol*. 1997;26:1323-1333.
27. D'agostino RB, Chase W, Belanger A. The appropriateness of some common procedures for testing the equality of two independent binomial populations. *Am Stat*. 1988;42:198-202.
28. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*. 1988;44:229-241.
29. Tutz G, Schmid M. Evaluation and model performance. *Modeling Discrete Time-to-Event Data*. Switzerland: Springer International Publishing; 2016:73-102.
30. Qiu F, Stein CM, Elston RC, Tuberculosis Research Unit (TBRU). Joint modeling of longitudinal data and discrete-time survival outcome. *Stat Methods Med Res*. 2016;25:1512-1526.
31. Jóźwiak K, Moerbeek M. Power analysis for trials with discrete-time survival endpoints. *J Educ Behav Stat*. 2012;37:630-654.
32. Li Z, Wang X, Wu Y, Owzar K. Sample size calculation for studies with grouped survival data. *Stat Med*. 2018;37:3904-3917.
33. Lawrence, J., Xu, J., Wang, S. J. & Hung, J.. SurvDisc: discrete time survival and longitudinal data analysis. 2018. <https://cran.r-project.org/package=SurvDisc>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Purington N, Andorf S, Bunning B, Chinthrajah S, Nadeau K, Desai M. Novel application of a discrete time-to-event model for randomized oral immunotherapy clinical trials with repeat food challenges. *Statistics in Medicine*. 2021;1–14. <https://doi.org/10.1002/sim.9019>

1 **SUPPLEMENTAL MATERIAL**

2 **R code implementation**

3 Assuming a person-level dataset where each challenge CTD has its own column, we derive
4 CTD-time by first summing the CTD across challenge variables (challenge1:challengeN) for
5 each participant. If a challenge was failed, no subsequent challenges were performed and NA
6 was recorded as the CTD for all future time points. We next assign all participants a value of 0 or
7 1 based on whether or not they reached the end of the study without failing (0=censored,
8 1=failed), i.e. whether or not a participant had the maximum CTD-time. We further censor
9 participants who passed a challenge but did not return for subsequent challenges
10 (participant_noreturn_list) or those who dropped out of the study prior to the first post-
11 screening DBPCFC.

```
12 KM$CTDtime <- rowSums(KM[c(challenge1:challengeN)], na.rm = T, dims=1)
13 KM$failure <- as.numeric(ifelse(KM$CTDtime==20000,0,1))
14 ## censor participants who pass and do not return
15 participant_noreturn_list = KM$participant_id[KM$CTDtime %in% c(4000, 8000, 12000, 16000)]
16 KM$failure[KM$participant_id %in% participant_noreturn_list] = 0
17 ## exclude participants who drop out pre-104
18 KM$failure[is.na(KM$CTD_week104)] = NA
19
```

20 This brings us to the person-period data structure described in Table 1b. A Kaplan-Meier curve is
21 fit using the npsurv function in the rms package, and visualized using the ggsurvplot function in
22 the survminer package:

```
23 surv1 <- npsurv(formula=Surv(CTDtime, failure)~treatment, data = KM, conf.type = "log-log")
24 ggsurvplot(surv1, size = 1, conf.int = TRUE, pval=T, risk.table = T, xlab = 'Cumulative
25 tolerated dose-time (mg)', ylab= 'Proportion of treatment arm remaining desensitized')
26
```

27 Next, the survsplit function in the survival package is used to create the person-period data set
28 presented in Table 3.

```
29 pp <- survsplit(KM, cut=c(unique(KM$CTDtime)), zero=-1, end="CTDtime", event="failure", episode="time")
30 pp <- pp[order(pp$participant_id, pp$time),]
31
```

32 Using the standard glm function, the grouped proportional hazards model (i.e. discretized Cox
33 PH model) is fit. Model summary, likelihood ratio test, and concordance are examined.

34

```
35 grp.m <- glm(failure ~ as.factor(CTDtime)+treatment-1, data=pp, family="binomial"(link="cloglog"))
36 summary(grp.m) # pay attention to the treatment estimate at the bottom
```

```
1 cbind(exp(coef(grp.m)), exp(confint(grp.m)))
2 anova(grp.m, test="LRT") # overall test of treatment
3 concordance(grp.m)
4 logLik(grp.m)
```

5

6 To plot the predicted hazards, a dataset is created in order to store the hazards based on CTD-

7 time and treatment arm. The hazards are then plotted over time by arm using ggplot.

```
8
9 dat<-expand.grid(CTDtime=unique(pp$CTDtime), treatment=c('A','B'))
10 dat$fitted <- predict(grp.m, type = "response", newdata=dat)
11 #plot the hazard
12 ggplot(data=dat, aes(x=CTDtime, y=fitted, color=treatment)) +
13   geom_line() +
14   labs(x="Cumulative Tolerated Dose Sum (mg)", y='Hazard of Study Failure', color='Treatment') +
15   scale_color_manual(values=c("#A80D37", "#00A1C0")) +
16   theme_minimal() +
17   theme(legend.position = "bottom")
```

18

19 More detailed R code, including a dummy dataset, is provided as additional material.

20