Personalizing attribution of adverse events in Phase I oncology studies

Alexia Iasonos
John O'Quigley

Available at: https://works.bepress.com/alexia_iasonos/4/
Personalizing attribution of adverse events in Phase I oncology studies

Alexia Iasonos (1) and John O'Quigley (2)

(1) Department of Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA.

(2) Université Paris VI, 75005 Paris, France

**Corresponding Author:**
Alexia Iasonos, PhD
Associate Attending Biostatistician
Department of Epidemiology and Biostatistics,
Memorial Sloan Kettering Cancer Center.
307 E. 63rd Street, 3rd Floor.
New York, NY 10065, USA.
Tel 646 735 8115. E-mail:iasonosa@mskcc.org

Word count 2938; 2 tables; 4 figures; 16 pages

Role of Funding Source: This was work was partially supported by a grant from the National Institutes of Health, R01-CA142859, which funded Iasonos A and O'Quigley J.
ABSTRACT

Background: In Phase I trials, attribution of adverse events (AE) is a complicated process. The current endpoint of dose limiting toxicity (DLT) assumes attribution is known. Such a simplified endpoint ignores individual-specific or drug-specific information. We propose an approach that allows for personalized scores that account for the uncertainty in AE attribution.

Methods: The proposed design takes into account possible errors when attributing AE and allows investigators to explicitly express their degree of uncertainty in whether any AE is drug-related or disease-related. A score of 100% means that the AE is certainly drug-related, whereas a score of 0% means that the AE is not drug-related. We evaluated individual scores that range from 20-100% (mean=50% and 75%, variance=0-30%), and attribution error rates of 5%, 10% to 15%. The scores vary by patient, by dose, and by the trial’s accumulated safety history.

Results: When error rates are small, the method using scores can lead to improvements in the accuracy of the maximum tolerated dose (MTD) on the order of 25%-100% (e.g. 88% vs 11% of trials find the MTD in the presence of errors, error rate is 15%). Thus, using such refined personalized score compensates for potential errors in attribution and results in significant improvements in MTD accuracy.

Conclusion: Our proposed method represents a new paradigm for Phase I trials in which each AE may be weighted in accordance with the drug’s mechanism of action from pre-clinical studies, the individualized patient risk and other data, producing more accurate identification of the MTD.
Introduction

Phase I trials are often first in man (FIM) trials of an experimental agent or combination of regimens, and these trials aim to find the maximum tolerated dose (MTD) that is considered safe for further testing. The rate of dose limiting toxicity (DLT), defined as a drug related toxicity pre-specified according to the protocol is the Phase I endpoint. However, accurate attribution of toxicities to the experimental drug, which requires determining whether the drug or a patient’s disease is responsible for specific toxicities, is not always clear. AttrIBUTIONS can be inaccurate due to the potential presence of other competing factors, such as concomitant chemotherapy or medications, cumulative toxicities from previous treatment, disease progression, and/or various co-morbidities. The challenge is that Phase I trials typically test novel agents or combination regimens, and investigators must rely on the drug's proposed mechanism of action, the toxicities observed in animal studies, and temporal associations as the basis for evaluating toxicities. These sources of information might not strongly correlate to DLTs observed in humans.

It is known from subsequent phase III trials that there is considerable error in attribution.[1, 2] A retrospective review of attribution rates in randomized Phase III trials found that 50% of adverse events (AE) were reported as drug related on the placebo arm [2]. Among those AEs reported more than once for the same patient, 36% changed in attribution over time and certain toxicities were overestimated as attributable to treatment. While this review was in the Phase III randomized setting with low incidence of grade 3 to 5 AE's, investigators agree that attribution of certain toxicities can be subjective [3-5]. Other reports [6] suggest that there might have been over-reporting of serious adverse events (SAE), since until recently there were no specific guidelines for drug causality [7]. For this reason, the FDA has issued a new regulation that
clarifies the definitions of adverse events [8]. The FDA guidelines provide examples of the types of evidence that suggest causal relationship for investigators reporting a suspected adverse reaction.

Determining causality of AE remains a persistent problem in Phase I trials, particularly when investigators need to judge on the basis of an isolated incident observed in a FIM study at the setting of advanced disease. There are clinical scenarios where it is beyond the clinician's ability to determine for certain if a given toxicity is attributable to the study drug, especially early on in a trial. Despite the current practice where clinicians use five levels of attribution ranging from not related to definitely related, investigators need to dichotomize the five levels to drug related or not when determining DLTs. In this paper we introduce the framework of a patient-specific score that captures the variability inherent in toxicity attribution. This framework of personalized scores could replace the current one wherein investigators use a dichotomous outcome (the presence or absence of DLT) which is assumed to be known without error. Dose escalation designs can take into account the subjectivity in toxicity attribution by assigning a personalized score representing how likely it is that each toxicity is drug related, expanding the DLT outcome. This personalized score will quantify the chance that a particular toxicity is related to the experimental drug versus other causes as judged by the investigators, the disease management team, or the Data Safety Monitoring Board. For example, early on in a trial, when limited information is known about the experimental agent, the probability of drug related toxicity might be high but clinicians can also give a range of scores to reflect their uncertainty. Later on towards the end of the trial the probability could be smaller, given the accumulated knowledge during the trial. The scores we propose can be dynamic, potentially changing both during the trial and from
patient to patient, thus representing a patient specific score for drug related toxicity. The proposed design has the potential to dovetail with recent work that provides an individualized risk of DLT before a patient is being treated at baseline. [9] Our thesis is that this patient specific risk should be continuously evaluated at the time of enrollment, during the trial, and at the end of the trial, since this is a crucial aspect of the safety evaluation of phase I patients.

**Methods**

**Outcome Definition**

Instead of misclassifying patients’ AE when the true attribution is uncertain, clinicians can assign a score representing the probability of drug related toxicity in a given case. Table 1 shows an example of hypothetical scores and how they could correspond to the current attribution levels. Instead of grouping uncertain categories into a binary response, which is the current practice, the proposed design allows for a range of scores that captures the uncertainty in assessment of attribution. A score in the scale from 0 to 100 reflects the uncertainty in attribution. For example, if we know that neutropenia is specific to a drug, then the score of serious neutropenia should be close to 100, counting as DLT. By contrast, if liver failure is suspected but not known for certain to be associated with a drug, an instance of liver failure might receive a score of 60%. Patients who experience dermatologic toxicities could get a range of scores. For example, low scores (<30%) would correspond to an AE being “unlikely”, mid-range scores (30%<probability<60%) to “possible”, and high scores (probability>65%) to “probable” (Table 1). Assuming no attribution errors occur, all clinician-identified DLTs should be drug related SAEs that meet the definition of DLT, despite the uncertainty introduced by giving scores instead of dichotomous attribution.
**Trial Design and statistical model**

The foundation of model based, adaptive designs [10-12] provides us the methodological background to incorporate these personalized scores. A statistical model can determine the dose to be assigned to each patient and the model can be updated sequentially, based on the most current knowledge of toxicity. The flexibility provided by sequential updating enables us to reassess scores, which will reflect the likelihood of a toxicity being drug related or not, and the reassessment can occur in real time as the trial is ongoing. If attribution is known for any one subject, this can also be incorporated in the proposed framework by assigning a score of 0 or 100%. Figure 1 shows the dose finding algorithm. Details of the statistical model are provided in Appendix.

**Statistical Analysis**

There are two types of errors that might occur in toxicity attribution: 1) when an investigator incorrectly attributes a SAE as non drug related, when in fact it is related to the experimental drug (Type A error) and 2) when an investigator attributes toxicity to an experimental drug when in fact it is due to other causes (Type B error). Type A errors can result in additional patients being enrolled to a trial at higher, potentially toxic dose levels, thus adversely affecting patient morbidity and mortality; they can also lead to inappropriate dose expansion resulting in both the accrual of additional patients and the presence of heightened risks to those who participate. On the other hand, Type B errors can result in early termination of accrual, declaring all higher dosages as unsafe, and consequently recommending a sub-therapeutic dose for further studies.
For the purposes of illustration let us assume that in a particular trial toxicities classified by investigators as non-DLTs are truly non-drug related toxicities (i.e. Type A error=0). However, when investigators need to determine whether a SAE is a drug related DLT, some of these AE are erroneously counted as DLTs when in fact they are not drug-related. We can summarize the possible toxicity outcomes of such a scenario with probabilistic errors as shown in Table 2, assuming a Type B error rate of 10%. The percentages of patients here were derived by assuming, for example, that if we expect 80% non DLTs and of those, 90% are truly non drug related, then, 72% (=0.9*0.8) of patients will be correctly classified as non-DLTs. However, 8% of patients will be misclassified as having DLTs when in truth the toxicities they experienced were not drug related. Since the frequency of attribution errors in practice is unknown, we need to evaluate the proposed approach under different error incidences, and under various true dose toxicity rates and score distributions. Attribution error rates of 5, 10, or 15% and mean personalized scores of 50% (high uncertainty) and 75% (low uncertainty) were assessed. Variability for inter patient scores was also assessed by allowing scores to vary from patient to patient around the mean score plus or minus 10%, 20%, or 30%. Two simulation settings were evaluated: one where scores vary by dose level with the assumption that higher scores should be given to patients treated at higher dose levels; and the second setting where dose does not inform the choice of the scores.

Results

We expect that overestimation of DLTs (Type B error) would lead us to a sub-therapeutic, low dose.[13] Thus, we compare how often the designs underestimate the MTD in the presence of Type B error and to what degree the proposed method can compensate for attribution errors.
Figure 2 shows that almost all of the information needed to find the MTD can be recovered, since the MTD will be underestimated under errors as shown by the shift to falsely recommending lower levels. The proposed method recommends the right MTD 88% of the time, as opposed to 11% under the current method when the error rate is 15%, and it is only 20% less accurate than the dose that would be recommended if all attributions were correct; while unrealistic this scenario provides a benchmark to measure how well the method performs. In this simulation setting, we allow investigators to assign different scores for each patient and at each dose level (mean scores are centered at 10, 20%, 30%, 40%, 50 and 70% at each dose level respectively +/- 20% in either direction). These scores vary depending on the dose the patient was treated at, with lower scores be given to lower doses, and higher scores be given to higher dose levels. Under the assumption that we do not know how far we are experimenting from the final MTD, having scores independent of dose is also realistic. In the next simulation setting, we assumed that patients should be assigned scores centered around the same value regardless of the dose they were treated. Figure 3 shows that the percent of times we select the true MTD increases from 43 to 64% compared to the dose we would have selected in the presence of attribution errors. In this scenario the scores were centered around 75% regardless of dose level, and scores varied from patient to patient with a variability of +/-20% around 75%. We repeated all simulations with no or small variability around each score (10 -20%) and whether the patient specific score was centered around the mean or the score had some additional variability around its mean did not affect the results.

We also assigned scores that were systematically high at all levels and systematically low on all levels. As expected, if scored systematically very low, as though less toxicity had been seen in practice, since less toxicity will lead to higher levels, we then tend to overdose. On the other
hand, systematically high scores might lead to treating patients with subtherapeutic dose levels, since we indicate potential DLTs when in fact they are not, shifting experimentation to lower levels (Figure 4). However, in the absence of systematic bias, we found no cases where the proposed method of personalized scores performs worse than the current practice of possibly erroneously attributing AE to the drug. Generally the new method leads to considerable improvement in the accuracy of determining the MTD.

Discussion

An important aspect of Phase I trials in oncology is that the patient population has advanced disease. This makes it difficult for clinicians to distinguish between drug related toxicity and disease deterioration or other non drug related toxicities. In the interest of patient safety, it is not uncommon for investigators to attribute an AE to the study drug when the cause is unclear [14]. Thus, errors in toxicity attribution made by overestimating the number of DLTs will usually lead to a sub-therapeutic MTD, which in turn might rule out a promising agent in subsequent testing.[13] In practice, investigators sometimes reassign toxicity attribution retrospectively based on the cumulated safety profile of the agent and the entire trial history. Our proposed design provides a conceptual link between the clinical challenges of Phase I trials and the statistical complexity of recent model based dose escalation algorithms. These algorithms allow for dynamic learning as the trial progresses and attributions become clearer, and incorporate knowledge from earlier studies together with the clinical investigators’ expertise. In some cases, the investigators may be close to certain on correct attribution. In other cases, they might be less certain and our purpose is to explicitly address this fact. The investigators’ input
tailored to the individual patients can enable more accurate dose finding and greater flexibility in escalation while still maintaining a rigorous adhesion to the protocol so that, all adverse events are recorded as such.

Phase I designs should incorporate a level of uncertainty in toxicity attribution, since in the majority of these trials the therapeutic agent’s mechanism of action is not fully understood. In the proposed framework, the protocol team can assign a score representing the likelihood of an event being drug related. The scores can describe numerically existing verbal descriptors such as possibly, probably or unlikely. These descriptors will map onto separate scores or a group of scores. This mechanism will allow for individual, patient-specific scores as well as dynamic scores that could change as the trial progresses. For example, knowledge based on late-onset toxicities observed in the post-DLT monitoring period can be incorporated when assigning scores, and hence the dose escalation algorithm will reflect accumulating information on the drug's complete toxicity profile. Patients can be assigned a variety of scores: scores that vary by dose level or patient inclusion number so that they reflect accumulated information throughout the trial. For example, in the first 10 patients the scores could be the same or similar given the limited safety information available, while the score can vary for the next 10 patients depending on the dose level the patient received and how far we are experimenting from the starting level.

We showed that a method that uses a personalized score instead of a binary DLT outcome can accurately lead to the MTD; if the scores are well calibrated, on average, we expect patients with high probabilities of DLTs to correlate with an actual presence of DLT and similarly low scores to correlate with absence of DLT. For example, if the clinicians' scores have no systematic
variability and always agree with the presence of DLTs, then the proposed method is more accurate than using a binary outcome with misclassification errors. However, if the scores do not correlate with the underlying true DLTs, then we do not expect improved accuracy. As long as clinicians' estimates are not systematically biased, i.e., generally too low or too high, then a range of widely varying scores will lead to improved accuracy. Thus, the scores can be inexact for every patient but the method’s accuracy will not be unduly affected as long as the scores are well calibrated and not systematically biased. This novel conceptual framework incorporates the subjectivity involved in toxicity attribution into Phase I designs, such that the final recommended dose reflects this uncertainty. It also allows the investigators to strictly adhere to the protocol by recording all DLTs and, yet, still allow expert opinion to prevent AE that are most likely not drug related from seriously compromising the exercise of identifying the correct MTD.

Although the statistical literature has examined the effect of different types and grades of toxicity, current designs do not address the lack of certainty in toxicity attribution. Ordinal responses that represents the toxicity grade on a 0-5 scale [15, 16], or a toxicity burden that summarizes different types and grades of toxicities [17, 18] have been proposed. The concept of personalized Phase I design has been considered by Piantadosi and Liu [19] in the context of pharmacodynamics/pharmacokinetics (PK/PD), continuous doses or patient specific doses based on a protein level for example. Our proposed framework could potentially work in conjunction with a number of other advances. PK/PD modeling and other data can inform the choice of these scores. For example, a PK model that gives predicted DLT rates at each dose level can be used to inform these scores. The modeling and score assignment can change over time. Moreover, the chance of an AE being a DLT might depend on the type of toxicity we review, e.g. AE such as
rash or fatigue might get lower scores, while neuropathy, diarrhea, or other AE known to be associated with the investigational drug could be assigned high scores.

We consider the clinician’s assessment of toxicity attribution to be the gold standard, as this is typical in the Phase I setting. Previous work investigated the agreement between the clinician’s assessment of the severity (grade) of AE with patient reported outcomes (PRO).[20] This is a potentially promising area for improving our proposed patient-specific scores for AE attribution, and more research is needed in the Phase I setting before we can incorporate PRO in toxicity attribution to guide dose escalation.[20-22] The basic assumptions of Phase I designs thus far have been that a toxicity is either drug related or not, and that the DLT assignment can be made without error [23, 24]. We suggest a new paradigm for Phase I trials by proposing model based designs that can enable investigators to combine all the knowledge they have regarding the investigational agent’s safety profile, information from PK/PD studies, patient disposition and individualized patient risk at baseline into a consensus score that can represent a new expanded Phase I endpoint that guides dose escalation.
Figure/Tables Legends:

**Table 1:** Illustrative example of scores for drug related toxicities.

**Table 2:** Phase I toxicity outcomes with misclassification error of 10%

**Figure 1:** Dose escalation algorithm: using a personalized score instead of dose limiting toxicity as the Phase I endpoint

**Figure 2:** Percent of trials out of 1000 simulated trials (y axis) that select each dose level among 6 levels (x axis). True DLT probabilities are 1, 5, 7, 11, 20, 50% at each dose level and scores vary depending on dose level (scores are 10, 20%, 30%, 40%, 50 and 70% at each dose level respectively with +/-20% variability from patient to patient). Error rates vary from 0-15%.

**Figure 3:** Percent of trials out of 1000 simulated trials (y axis) that select each dose level among 6 levels (x axis). True DLT probabilities are 1, 5, 7, 11, 20, 50% at each dose level and scores are constant across dose level but vary per patient (scores are 75% with +/-20%). Error rates vary from 0-15%.

**Figure 4:** Percent of trials out of 1000 simulated trials (y axis) that select each dose level among 6 levels (x axis). True DLT probabilities are 1, 5, 7, 11, 20, 50% at each dose level and scores vary depending on dose level but they are systematically overestimating DLT risk, i.e. scores induce bias by systematically assigning higher scores compared to true rates.
Table 1: Illustrative example of scores for drug related toxicities 0: non DLT; 1: DLT

<table>
<thead>
<tr>
<th>Attribution</th>
<th>Hypothetical Score %</th>
<th>Current designs: DLT yes/no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated</td>
<td>0</td>
<td>No (0)</td>
</tr>
<tr>
<td>Unlikely</td>
<td>0 - &lt; 30</td>
<td>No (0)</td>
</tr>
<tr>
<td>Possibly</td>
<td>30 - &lt; 65</td>
<td>Yes (1)</td>
</tr>
<tr>
<td>Probably</td>
<td>65 - &lt; 100</td>
<td>Yes (1)</td>
</tr>
<tr>
<td>Definitely</td>
<td>100</td>
<td>Yes (1)</td>
</tr>
</tbody>
</table>

Table 2: Phase I toxicity outcomes with misclassification error of 10%

<table>
<thead>
<tr>
<th>Classified DLT</th>
<th>True drug related DLT</th>
<th>Percentage Of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>20%</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>8%</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>72%</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>0%</td>
</tr>
</tbody>
</table>
REFERENCES:


Treat a patient at dose level $d_i$

**DLT:** assign each patient a score to reflect the chance of this toxicity being a DLT. If it is a certain DLT, assign 100%; if it is a certain non-DLT, assign 0. If attribution is uncertain assign a score from 0 to 100%

Re-evaluation of the toxicity probability associated with each dose level

The dose level to be administered to the next patient is that associated with the estimated toxicity probability closest to the toxicity target (e.g. 30%) while not allowing to skip more than one level.

Accrue 20-25 subjects, and at the end of the trial estimate the MTD based on the updated curve that uses all patients’ responses.