Regulatory Toxicology and Pharmacology 72 (2015) 231–243

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Improving the power of long term rodent carcinogenicity bioassays by adjusting the experimental design

Matthew T. Jackson

Amgen Inc, 1120 Veterans Blvd, South San Francisco, CA 94080, USA

ARTICLE INFO

Article history:
Received 8 October 2014
Available online 20 April 2015

Keywords:
Carcinogenicity
Power
Experimental design
Rodents
Type 1 error

ABSTRACT

Since long term rodent carcinogenicity studies are used to test a very large number of potential tumor endpoints, finding a balance between the control of Type 1 and Type 2 error is challenging. As a result, these studies can suffer from very low power to detect effects of regulatory significance.

In the present paper, a new design is proposed in order address this problem. This design is a simple modification of the existing standard designs and uses the same number of animals. Where it differs from the currently used designs is that it uses just three treatment groups rather than four, with the animals concentrated in the control and high dose groups, rather than being equally distributed among the groups.

This new design is tested, in a pair of simulation studies over a range of scenarios, against two currently used designs, and against a maximally powerful two group design. It consistently performs at levels close to the optimal design, and except in the case of relatively modest effects and very rare tumors, is found to increase power by 10%–20% over the current designs while maintaining or reducing the Type 1 error rate.

1. Introduction

A long-term carcinogenicity study of rodents is a standard component of a drug approval application (ICH, 1995; McCormick, 2012; US Food and Drug Administration, 2001). Over time, the design and analysis of such studies has become largely standardized (ICH, 1997; Rahman and Lin, 2009; US Food and Drug Administration, 2001). A New Drug Application (NDA) typically includes four separate experiments: on male mice, on female mice, on male rats, and on female rats. Each experiment runs for two years[1]; after this time, all surviving animals are sacrificed. All animals (regardless of cause of death) undergo a complete necroscopy to identify signs of tumors. The data thus collected are then analyzed to detect any dose-related tumorigenic effects of the test article, and tumor types for which a statistically significant dose effect is found are flagged.

The results of these statistical tests are then reviewed by the FDA's Executive Carcinogenicity Assessment Committee (ECAC). The ECAC and the reviewing toxicologist consider any flagged tumor types, and make a determination of which effects need be reported as clinically relevant. Occasionally, generally for extremely rare tumor types, they will report a tumor finding that has not been flagged as statistically significant. Nonetheless, it is appropriate to think of the statistical analysis as the first step in a screening process. In short, a positive statistical flag is an almost necessary but not sufficient condition for a positive carcinogenicity or tumorigenicity finding.

A major concern with this design is that the large number of endpoints under consideration may inflate the false positive rate (Rahman and Lin, 2008). In general, one is not interested in the endpoint *a tumor developed*; the expectation is rather that a drug with genuine tumorigenic properties will increase the incidence of specific tumors or groups of tumors but have no effect on the vast majority of tumor types. Accordingly, each such tumor or tumor group is analyzed independently, with the result that the number of endpoints explicitly tested in a single experiment can be well over fifty. (Including the total number of endpoints that were only implicitly tested – tumor types which were not observed in any experimental animals but which would have been tested had they been found – is likely to raise the total well above 100 (Lin et al., 2015; Giknis and Clifford, 2004; Giknis and Clifford, 2005).) Since a positive finding for a single endpoint is sufficient to allow a conclusion of carcinogenicity[2] (in rodents that is; extrapolation to a corresponding effect in humans is by no means

E-mail address: majackso@amgen.com

[1] Some studies substitute experiments on transgenic mice for the usual mouse experiments (McCormick, 2012; US Food and Drug Administration, 2001). Such experiments have the same basic design as the twenty-four month experiments described above, but are conducted over just 26 weeks, and with smaller sample sizes – typically twenty-five animals per group. This paper addresses only the design of twenty-four month experiments, and not transgenic mouse experiments.

[2] In the language of Huque and Röhmel (2010), each endpoint is a *primary* endpoint.

http://dx.doi.org/10.1016/j.yrtph.2015.04.011
0273-2300/Published by Elsevier Inc.

automatic), the abundance of endpoints has the potential to greatly inflate the study-wise false positive rate. The FDA's response has been to impose fairly stringent requirements for individual tests to be deemed statistically significant. Under such circumstances, the corresponding costs to power may be substantial, meaning that genuine tumorigenic effects, even those of regulatory significance, may remain undetected.

In this paper, two commonly used designs (called D1 and D2) are tested under a variety of simulation models. It is found that in many cases their power to detect even substantial tumorigenic effects is very low, especially when the underlying incidence rate is rare. For example, their power to detect an increase in incidence from 2% in the control group to 16.9% (an odds ratio of 10) is less than 50%. A new design (D3) is also tested in the same scenarios. Compared with D1 and D2, design D3 provides a considerable improvement, yielding improved power (increases of between 10% and 20%) while simultaneously maintaining or decreasing the Type 1 error rate.

In addition, a fourth design (called D4), is tested in the same scenarios. Design D4 uses only two dose groups, which can be reasonably expected to cause a lack of robustness in the case of unexpectedly high dose-related mortality. However, it can be expected to be optimal in terms of power, and therefore serves as a useful reference. Across the scenarios tested, the power of the D3 is consistently found to be only slightly less than D4.

2. The design of rodent carcinogenicity bioassays

2.1. Typical current design

Under the current designs (as outlined in McCormick (2012)), within each experiment, three or four equal sized dose groups are treated with different levels of the test article. The highest dose group receives a dose that is close to the estimated maximum tolerated dose (MTD) of the article[3] and the other groups are treated with dose levels spaced by ratios of one third to one half. A typical design with three treated groups might therefore expose the low and mid dose groups to dose levels of one sixth and one half (respectively) of that received by the animals in the high dose group.

Each endpoint, whether a single organ/tumor combination (*carcinoma of the pars distalis*, for example) or a class of related tumor types (*all lipomas, regardless of location*, for example), is then assessed independently, with each animal being categorized as either a tumor bearing animal (TBA) or a non-tumor bearing animal (NTBA). The resulting data are analyzed for a dose-related trend. In addition, each treated group is compared with the control group(s) in a pairwise test. The specific statistical tests used vary, but the FDA currently prefers the exact poly-*k* test (Bailer and Portier, 1988; Bieler and Williams, 1993; Rahman et al., 2014; Lin et al., 2015). In particular, the poly-3 test is preferred for 24 month studies, although no preference for a particular value of *k* has yet been expressed for 26 week transgenic mouse studies.[4]

2.2. Decision rules

The draft guidance (US Food and Drug Administration, 2001) recommends that in most cases the decision to reject the null hypothesis of no tumorigenic effect for a specific endpoint be made on the basis of the trend test alone. In those cases where the trend test is considered inappropriate, the pairwise test between the highest appropriate treated group and the vehicle control group should be used instead. In order to control the study-wise false positive rate (the probability that at least one tumorigenic effect is falsely found when in fact none exist), the significance thresholds for these tests are set quite low, at levels which have been found (Lin and Rahman, 1998; Rahman and Lin, 2008) to yield an approximately 10% study-wise false positive rate.[5]

In practice, however this decision rule has not been adopted. As discussed in Lin et al. (2015), a finding of statistical significance currently requires both the trend test *and* the pairwise comparison between the control and high dose groups yield *p*-values below the thresholds recommended in US Food and Drug Administration (2001). For example, see the discussion of *osteomas and osteosarcomas in female mice* in Center for Drug Evaluation and Research (2014). The use of this rule is motivated by a concern that the decision rule in the draft guidance is too liberal (in the sense of being too quick to reject the null hypothesis of no tumor effect), and yields an unacceptably high false positive rate. We will refer to this decision rule as the Currently Used Joint Test (CUJT).

This *ad hoc* test is necessarily more conservative than the (statistically validated) decision rules recommended in US Food and Drug Administration (2001). Responding to the fact that this rule is nonetheless used, an alternative joint test is proposed in Lin et al. (2015). This new test also requires significant results from both the trend test and pairwise test, and is still more conservative than the decision rule recommended in US Food and Drug Administration (2001); the significance thresholds for the trend test are the same as for those recommended for the trend test alone in US Food and Drug Administration (2001), but the thresholds for the pairwise test are higher than those recommended in US Food and Drug Administration (2001) when the trend test is used alone. We will refer to this rule as the Proposed Joint Test (PJT). The significance thresholds for this test, and those for the CUJT, are summarized in Table 1.

The motivating idea for the PJT is that under the CUJT, the limiting factor was most often the pairwise test. By making this part of the joint test less conservative, it is hoped that the test could be made more liberal overall (compared to the CUJT) – approximately as liberal as the rule recommended in US Food and Drug Administration (2001) – while still satisfying those who feel that some sort of joint test is needed.

It was found in Lin et al. (2015) that for a design with 50 animals per group, the PJT has a study-wise false positive rate of approximately 10%.

It should be noted that for both of these decision rules, the ultimate study-wise false positive rate is likely to be lower than 10%. Since these studies are used as screening tests, with statistically significant findings subject to additional (non-statistical) scrutiny by a committee of experienced pharmacologists and toxicologists, only findings which this committee deems clinically relevant are considered for inclusion in drug labeling. A high proportion of false positives are likely filtered out at this stage in the review process.

2.3. Determination of rarity

Both decision rules draw a distinction between *rare* tumors (those with underlying incidence rates below 1%) and *common* tumors (with underlying incidence rates above 1%). The process by which the determination of whether a tumor should be considered

[3] Alternatives to the estimated MTD are also sometimes used. If the animals can tolerate a dose level that results in a blood serum concentration well above that associated with the anticipated maximum human therapeutic dose, then that dose level is likely to be adequate for use as a high dose level. Conversely, it may be physically unfeasible to administer dose levels above a certain level, in which case, the high dose animals will be treated with the *maximum feasible dose*.

[4] Mortality rates for transgenic mouse studies are typically so low that the results are largely insensitive to the choice of *k*.

[5] It should be noted that Westfall and Soper (1998) found that the study-wise false positive rate was quite sensitive to increases in sample size, and that the 10% error rate found for designs with 50 animals in each group should not be assumed to apply to studies with, say, 75 animals in each group.