ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - 2023, Volume 18, Issue 1

Phase III Failures for a Lack of Efficacy can be, in Significant Part, Recovered (Introducing Success Probability Estimation Quantitatively)

Daniele De Martini[(1)]

(1) Università degli Studi di Genova.

CORRESPONDING AUTHOR: Daniele De Martini, daniele.demartini@unige.it, ORCID 0000-0002-6937-5287.

SUMMARY

The rate of phase III trials failures is approximately 42-45%, and most of them are due to a lack of efficacy. Some of the failures for a lack of efficacy are expected, due to type I errors in phase II and type II errors in phase III. However, the rate of these failures is far from saturating the global failure rate due to a lack of efficacy.
In this work, the probability of unexpected failure for a lack of efficacy in phase III trials is estimated to be about 14%, with credibility interval (9%, 18%). These failures can be recovered through an adequate planning/empowering of phase II, and by adopting conservative estimation for the sample size of phase III. The software *SP4CT* (a free web application available at *www.sp4ct.com*) allows these computations. This 14% rate of unexpected failures gives that every year approximately 270,000 patients uselessly undergo a phase III trial with a large damage in individual ethics; moreover, the unavailability of many effective treatments is a considerable damage for collective ethics. The 14% of unexpected failures also produces more than $11bn of pure waste, and generates a much higher lack of revenue given by drugs' marketing.

Keywords: failure reasons; unexpected failures; ethics; waste; conservative sample size; software for sample size.

1. INTRODUCTION

The problem of the high rate of phase III trials that in recent years have failed, approximately 42-45% [1–4], has been widely discussed in our recent paper entltled "Empowering Phase II Clinical Trials to Reduce Phase III Failures" [5], where pros and cons of possible countermeasures have been presented.

In practice, phase III failures are due, for the most part, to a lack of efficacy (approximately 57-66% [1,6,7]). Moreover, given that other failures of this kind are labeled as failures for economic or commercial reasons or failures for safety (we will develop these concepts later), the actual failure rate due to a lack of efficacy is even higher than that reported above.

However, just some of these failures (for a lack of efficacy) are expected: in fact, expected failures of this kind are caused by type I errors committed in phase II and by type II errors of phase III. These statistical errors can not be completely avoided and, given usual settings and data available in the literature (e.g. [8]), they cause the failure of approximately 20-25% of phase III trials, corresponding to about 50-55% of failures.

Then, the global failure rate due to a lack of efficacy results much higher than that of failures due to a lack of efficacy that are expected (i.e. those due to statistical errors).

In the discussion presented in [5], the concept of *failures for a Lack of efficacy that are Not Expected* (viz. *LNE*), that is, failures due to a lack of efficacy minus those due to a lack of efficacy attributable to statistical errors, has been introduced. As far as this concept played a central role within the discussion of the abovementioned paper, the order of magnitude of the rate of *LNE* has been conservatively elicitated, and set at 10% of the trials run (i.e., approximately 22-24% of the failures). This datum supported the conclusion arguing the need of expanding phase II trials to increase phase III success rate.

Therefore, to focus on the rate of *LNE* is: scientifically, ethically, and economically relevant. Through this work, we aim at estimating the rate of *LNE*. Therefore, this paper can be considered a quantitative complement

DOI: 10.54103/2282-0930/20638
Accepted: 31th March 2023

to [5], and an addition to paragraphs 2 and 3 in the Introduction of the book *Success Probability Estimation with Applications to Clinical Trials* [9] (pp. XXIV-XXVI).

Finally, note that *LNE* failures can be recovered through an adequate planning: the conservative estimation of phase III sample size based on phase II data is a useful technique [9–12], and the software *SP4CT* is free web application (*www.sp4ct.com*) that allows these computations.

2. SETS AND PROBABILISTIC EVALUATION

2.1. Defining sets

Consider the following sets: F, representing the failures; L, failures for lack of efficacy; S, failures for safety reasons; C, failures for commercial or economic reasons; O, failures for other reasons; E, failures (lack of statistical significance) due to statistical errors, that is, type I errors in phase II and type II errors in phase III. Thus, we have: $L, S, C, O, E \subset F$. In particular $O = F \setminus L \cup S \cup C$.

In order to define unexpected failures for lack of efficacy, *LNE*, recall that it is given by failures for lack of efficacy minus failures for statistical errors belonging to L.

We remark that L is corrected by enlarging it, since the rates of L reported in the literature can be considered underestimated. This is due to the following facts:

a. failures reported as C are often function of L and S;
b. failures for both L and S are usually reported as S, since S is undoubtedly more serious. Consequently, some failures should be reallocated to L, according to the model adopted. To this aim, three models will be presented in the next section.

To conclude, denoting by L^c the corrected set of failures for a lack of efficacy, we have $LNE = L^c \setminus L^c \cap E$. We are interested in $P(LNE)$.

2.2. Calculating probabilities

$P(LNE)$ is given by $P(LNE|F) \times P(F)$, and $P(LNE|F) = P(L^c|F) - P(L^c \cap E|F)$. Then, $P(L^c|F)$ and $P(L^c \cap E|F)$ are computed according to different models (that follow in next section), where it is assumed that $P(F)$, $P(L|F)$, $P(S|F)$, and $P(C|F)$ are given. Moreover, $P(E)$ can be computed given the phase II false discovery rate $FDRII$, the phase III nominal power π, and the phase III type I error probability α. In particular, consider that the probability of running a phase III under the null coincides with the phase II false discovery rate (i.e. $P(H0) = FDRII$). Then, $P(E)$ can be obtained through the Total Probability theorem (i.e. $P(E)$ is the weighted sum of the probabilities of the expexted failures for effective and non effective treatments). Thus, we obtain: $P(E) = P(E|H0)\,P(H0) + P(E|H1)\,P(H1) = (1-\alpha/2)\times FDRII + (1-\pi)\times(1-FDRII)$.

3. MODELS

3.1. Model 1: reallocating a part of C

It is a fact that some failures labeled C are actually a function of safety and efficacy measures [13,14]. Thus, a subset of C, i.e. CR, has to be reallocated to either L or S. In practice, CR is divided into CRL and CRS, to be added to L and S, respectively. Then $CRL \cap CRS = \emptyset$, $CRL \cup CRS = CR \subset C$, and $L^c = L \cup CRL$.

Assume that $P(CR|C)$ is given, and consider $P(CRL|CR)$ and $P(CRS|CR)$ to be proportional to $P(L|F)$ and $P(S|F)$, respectively. In other words, the amount of CR reallocated to L is proportional to the amplitude of L. Given the assumptions above, we obtain that $P(CRL|F) = P(C|F)\,P(CR|C)\,P(L|F) \,/\, (P(L|F)+P(S|F))$, and consequently $P(L^c|F) = P(L|F) + P(CRL|F)$. Analogously, $P(S^c|F)$ is computed.

Now, the point is how to compute $P(L^c \cap E|F)$, which, under independence between E and the failure reasons (viz. L, S, C), would result $P(L^c|F) \times P(E|F)$. However, logic gives that $C \setminus CR$ does not contain parts of E (i.e. $C \setminus CR \cap E = \emptyset$), and therefore we can not exploit independence. Then, we assume E equally distributed over the sets that can contain it: L^c, S^c, O. Therefore we obtain $P(L^c \cap E|F) = (P(L^c|F) \times P(E|F)) \,/\, (P(L^c|F) + P(S^c|F) + P(O|F))$.

For example, if we set $P(F) = 0.43$, $P(L|F) = 0.6$, $P(S|F) = 0.15$, $P(C|F) = 0.2$, $P(CR|C) = 0.7$, $FDRII = 0.1$, $\pi = 0.85$, $\alpha = 0.05$ (the latter three giving $P(E) = 0.2325$), then we obtain $P(LNE) = 0.1301$.

3.2. Model 2: reallocating a part of S

This model develops point b) in section 2.1, arguing that failures for safety hide a relevant part of those for lack of efficacy, because the former are more serious than the latter. In particular, we assumed that if a treatment fails for both causes (i.e. S and L), then it is labeled S. This implies that a part of S has to be reallocated to F.

Consequently a certain amount of $P(S|F)$ has to be moved to $P(L|F)$. In detail $P(L^c|F) = P(L^c \cap S|F) + P(L^c \cap S^-|F) = P(L^c|S|F)P(S|F) + P(L^c|S^-|F)P(S^-|F)$. Note that $P(L^c \cap S^-|F)$ is in fact $P(L|F)$, giving $P(L^c|S^-|F) = P(L|F) \,/\, P(S^-|F)$. Now, assume that L^c failures have the same probability, in occurence with S failures or under different failures (this is absolutely reliable), giving $P(L^c|S) = P(L^c|S^-|F) = P(L \cap S^-|F)$. Finally, $P(L^c|F) = P(L|F)P(S|F)/P(S^-|F) + P(L|F)$.

To compute $P(L^c \cap E|F)$, independence cannot be advocated since the remaining part of S does not contain parts of E. Then, E is considered equally distributed over the sets that can contain it: L^c, C, O, and we obtain $P(L^c \cap E|F) = (P(L^c|F)P(E|F)) \,/\, (P(L^c|F) + P(C|F) + P(O|F))$.

3.3. Model 3: reallocating parts of *C* and *S*

In this final model, we mix the two reallocation criteria of the above paragraphs. Consequently, the probability of the corrected version of *L* is given by the original one plus that reallocated from *C* and that reallocated from *S*, according to the formulas given above. Thus, we obtain: $P(L^c|F) = P(L|F) + P(CRL|F) + P(L|F)P(S|F)/P(S^-|F)$.

To compute $P(L^c \cap E|F)$, once again independence cannot be applied. *E* is considered equally distributed over the sets that still can contain it, that now are just L^c and *O*. Then, we have $P(L^c \cap E|F) = (P(L^c|F)P(E|F))/(P(L^c|F) + P(O|F))$.

4. STATISTICAL COMPUTATION

4.1. Distribution assumptions

An exact computation of the estimate $P(LNE)$ is not possible, because: a) there is a certain variability among data concerning the estimates of failures due to different causes; b) the type II errors adopted for planning phase III trials and the rate of false positive findings in phase II are not precisely known.

Estimates of the probability of failure due to lack of efficacy (i.e. $P(L|F)$) found in the literature are 57% and 66% [1,6,7]. Assuming these two estimates as equally likely, and considering also likely the values within their range, $P(L|F)$ has been considered uniformly distributed in the range of the estimates, that is $P(L|F) \sim U(.57, .66)$.

Analogously, the probability of failure due to safety concerns ($P(S|F)$) and the probability of failure for economic or commercial reasons ($P(C|F)$) have been considered uniformly distributed in the range of their respective minimum and maximun estimates found in the literature [1,6,7], that is $P(S|F) \sim U(.09, .21)$ and $P(C|F) \sim U(.18, .22)$. Estimates of phase III failures go from 42% to 45% [1–4], so we set $P(F) \sim U(.42, .45)$.

Since failures for economic or commercial reasons are often based on utility functions depending on safety and efficacy measures [13,14], in practice a relevant part of them (i.e. *CR*) is reallocated to failures for safety or lack of efficacy (i.e. *S* or *L*).

The literature does not report estimates of the probability of *CR*. We discussed the problem with some authoritative colleagues, and we elicited $P(CR|C) \sim U(.5, .75)$.

The probability of launching a phase II trial when the treatment is ineffective has been set $P(FDRII) \sim U(.05, .14)$, because: a) in some phase II trials the launching rule is based on statistical significance with threshold 5% or higher; b) it has been estimated that the FDR in top medical literature is 14% [8], where phase II clinical trials represent an even higher class of experiments, so that *FDRII* has be assumed to be at most 14%. As it concerns type I and type II errors, since the power thresholds usually adopted in phase III trials are 80-90% we set $\pi \sim U(.8, .9)$, and $\alpha = 0.05$ according to the requirement of major national and transnational agencies.

Finally, the distributions introduced in this section, from that of $P(L|F)$ to that of π, have been considered independent.

4.2. Simulation

The aim of statistical computation is to obtain the distribution of $P(LNE)$. Then, a simulation has been performed, on the basis of distributional assumptions of section 4.1 and probabilistic calculation of section 3.2.

To approximate the distribution of $P(LNE)$ we started from simulating data from the distributions defined in section 4.1. In particular, 10^6 raw data have been generated from the joint distribution of $(P(L|F), P(S|F), P(C|F), P(F), P(CR|C), FDRII, \pi)$. If $P(L|F)+P(S|F)+P(C|F) > 1$, then these summands were rescaled to obtain sum 1 (e.g. $P(L|F)$ became $P(L|F)/(P(L|F)+P(S|F)+P(C|F))$). When $P(L|F)+P(S|F)+P(C|F) < 1$ there was no problem, since some (few) other failure causes are allowed in the model, according to related literature [1,6,7]. In practice, $P(L|F)$ and related probabilities have been rescaled 23% of the simulated raws (i.e. $P(P(L|F) + P(S|F) + P(C|F) > 1) \approx 0.23$); although this correction looks quite frequent and might look as a signal of model inadequacy, note that the extra probability generated by the simulation is quite small, since $P(P(L|F) + P(S|F) + P(C|F) > 1.05) \approx 0.03$.

Finally, $P(LNE)$ has been computed for each raw data, according to probabilistic calculation of section 3.2.

4.3. Results

Adopting Model 1 we obtained that the average of $P(LNE)$ was 13.4%, with the central 90% of data resulting in (8.8%, 18.0%) (this can be viewed as a credibility interval). Model 2 gave the average of $P(LNE)$ equal to 14.2%, with the central 90% of data resulting in (9.8%, 18.7%). The average of $P(LNE)$ given by Model 3 was 13.9%, with the central 90% of data resulting in (8.4%, 19.6%). Note that the latter results lie between those obtained with the two previous Models, since this latter approach is a mix of them, whereas the variability increases a bit.

5. SENSITIVITY ANALYSIS

In this section we modify some hypotheses or relax some assumptions made in the above sections, in

order to evaluate how the results on $P(LNE)$ change.

First, we introduce a different lower bound for $P(C|F)$. Although [1] did not report the rate of failures for commercial or economic reasons, an estimate of $P(C|F)$ can be obtained. Indeed, since the estimates of $P(L|F)$ and $P(S|F)$ where 0.66 and 0.21, respectively, it follows that $P(C|F) \leq 0.13$ (C may be not the only other cause of failure). Given that estimates from other sources were higher (i.e. 0.18, 0.22), we adopted 0.13 as the lower bound, and $P(C|F) \sim U(.13, .22)$. Under this different setting results vary just a bit.

Second, the distribution assumptions of section 4.1 are changed: Gaussian distributions has been used instead of the seven Uniform distributions previously adopted. In particular, $N(\mu, \sigma^2)$ substituted $U(a, b)$, where $\mu = (a + b)/2$ and $\sigma = (b - a)/4$ (i.e. $(\mu - 2\sigma, \mu + 2\sigma) \approx (a, b)$). With these settings, the distribution of $P(LNE)$ was a little tighter.

Table 1. Statistics of the distribution of P (LNE) under different models and settings.

	Mod1	Mod2	Mod3
Basic setting Mean	13.4	14.2	13.9
5th p-tile	8.8	9.8	8.4
95th p-tile	18.0	18.7	19.6
$P(C\|F) \sim U(.13, .22)$ Mean	13.4	14.2	14.0
5th p-tile	9.0	9.9	8.7
95th p-tile	17.9	19.1	19.6
Gaussian priors Mean	13.4	14.2	14.0
5th p-tile	9.6	10.4	9.4
95th p-tile	17.3	18.5	18.8

6. CONCLUSIONS

In a recent paper [5] we discussed the problem of the high rate of phase III failures, and presented pros and cons of possible countermeasures. In that paper, a central role was played by the rate of failure for a lack of efficacy not expected, *LNE*, which has been elicitated to be 10%. Here, we estimated this rate in more depth, through three different models.

Results were very close among the models: estimates of the rate of *LNE* were ap- proximately 14%, with 90% credibility interval approximately (9%, 18%). Thus, the elicitation has been confirmed by technical results. Moreover, a sensitivity analysis supported the estimates obtained.

Given that every year approximately 3,800 phase III trials are run with, on average, 500 patients each, the estimated 14% of *LNE* translates into an individual ethical loss [15] of 266,000 patients uselessly undergoing a phase III trial, annually. Moreover, the damage for collective ethics [15] is the unavailability of many effective treatments. Since the cost of each patient enrolled in a phase III is, on average, $42,000, the 14% of unexpected failures also produces more than $11bn of pure waste, and the loss of revenue given by drugs' marketing.

In fact, it is worth noting that *failures for a Lack of efficacy that are Not Expected* can be recovered through adequate planning: the above numbers argue for the need of empowering phase II trials, and for that of adopting conservative strategies for phase III sample size computation, in order to reduce phase III failures.

The software *SP4CT* allows conservative sample size estimation for phase III trials, and can help in determining phase II sample size on the basis of the overall probability of success of phase II *and* phase III. *SP4CT* is a free web application that can be run at *www.sp4ct.com*. Moreover, *SP4CT* performs profit computations [16] and is a useful tool for portfolio strategic planning.

The problem still open is whether enlarging phase II is worth it or not, given that resources are limited and that enlarging some phase II trials might imply that some other phase II would be not launched.

REFERENCES

1. Arrowsmith J. Phase III and submission failures: 2007-2010. Nature Reviews Drug Discovery 2011; 10(2): 1-1.
2. DiMasi JA, Reichert JM, Feldman L, Malins A. Clinical Approval Success Rates for Investigational Cancer Drugs. Clinical Pharmacology & Therapeutics 2013; 94: 329-335.
3. Thomas DW, Burns J, Audette J, Carroll A, Dow-Hygelund C, Hay M. Clinical Development Success Rates 2006-2015. BIO Industry Analysis, 2016.
4. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics 2018; 20 (2): 273-286.
5. De Martini D. Empowering Phase II Clinical Trials to Reduce Phase III Failures. Pharmaceutical Statistics, DOI:10.1002/pst.1980, 2019.
6. Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA* Internal Medicine 2016; 176(12): 1826-1833.
7. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nature Biotechnology 2014; 32(1): 40-51.

8. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 2014; 15(1): 1-12
9. De Martini D. Success Probability Estimation with Applications to Clinical Trials. Wiley and Sons, Hoboken, 2013.
10. Chuang-Stein C. Sample Size and the Probability of a Successful Trial. Pharmaceutical Statistics 2006; 5: 305–309.
11. Wang SJ, Hung HMJ, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. Pharmaceutical Statistics 2006; 5: 85–97.
12. De Martini D. Adapting by calibration the sample size of a phase III trial on the basis of phase II data. Pharmaceutical Statistics 2011, 10(2): 89-95.
13. Patel N, Bolognese J, Chuang-Stein C, Hewitt D, Gammaitoni A, Pinheiro J. Designing PhII trials based on program-level considerations: a case study for neuropathic pain. Drug Information Journal 2012; 46: 439-454.
14. Antonijevic Z, Kimber M, Manner D, Burman C-F, Pinheiro J, Bergenheim K. Optimizing drug development programs: type 2 diabetes case study. Therapeutic Innovation and Regulatory Science 2013; 47: 363-374.
15. Lellouch J, Schwartz D. L'essai therapeutique: ethique individuelle ou ethique collective? Revue de l'Institut International de Statistique 1971; 39: 127-36.
16. De Martini D. Profit Evaluations When Adaptation by Design Is Applied. Therapeutic Innovation and Regulatory Science 2016; 50(2): 213-220.