Dose escalations in phase I studies: feasibility of interpreting blinded pharmacodynamic data

Hassing, G.J.; Esdonk, M.J. van; Westen, G.J.P. van; Cohen, A.F.; Burggraaf, J.; Gal, P.

Received: 2 April 2022 | Revised: 14 July 2022 | Accepted: 18 July 2022

DOI: 10.1111/bcp.15473

ORIGINAL ARTICLE

Dose escalations in phase I studies: Feasibility of interpreting blinded pharmacodynamic data

Gerardus J. Hassing[1,2] | Michiel J. van Esdonk[1] | Gerard J. P. van Westen[3] | Adam F. Cohen[1,2] | Jacobus Burggraaf[1,2,3] | Pim Gal[1,2]

[1]Centre for Human Drug Research, Leiden, the Netherlands

[2]Leiden University Medical Center, Leiden, the Netherlands

[3]Leiden Academic Centre for Drug Research, Leiden, the Netherlands

Correspondence
P. Gal, Centre for Human Drug Research, Zernikedreef 8, 2333 CL Leiden.
Email: info@chdr.nl

Aims: During phase I study conduct, blinded data are reviewed to predict the safety of increasing the dose level. The aim of the present study was to describe the probability that effects are observed in blinded evaluations of data in a simulated phase I study design.

Methods: An application was created to simulate blinded pharmacological response curves over time for 6 common safety/efficacy measurements in phase I studies for 1 or 2 cohorts (6 active, 2 placebo per cohort). Effect sizes between 0 and 3 between-measurement standard deviations (SDs) were simulated. Each set of simulated graphs contained the individual response and mean ± SD over time. Reviewers ($n = 34$) reviewed a median of 100 simulated datasets and indicated whether an effect was present.

Results: Increasing effect sizes resulted in a higher chance of the effect being identified by the blinded reviewer. On average, 6% of effect sizes of 0.5 between-measurement SD were correctly identified, increasing to 72% in 3.0 between-measurement SD effect sizes. In contrast, on average 92–95% of simulations with no effect were correctly identified, with little effect of between-measurement variability in single cohort simulations. Adding a dataset of a second cohort at half the simulated dose did not appear to improve the interpretation.

Conclusion: Our analysis showed that effect sizes <2× the between-measurement SD of the investigated outcome frequently go unnoticed by blinded reviewers, indicating that the weight given to these blinded analyses in current phase I practice is inappropriate and should be re-evaluated.

KEYWORDS
clinical trials, drug development, pharmacodynamics, phase I

1 | INTRODUCTION

During the conduct of clinical studies, scientists evaluate data to judge the safety and futility of continuation of the trial. Responsible scientists and physicians are typically blinded to the study treatment allocation during these early-phase evaluations. These blinded evaluations of data are formalized in the study protocol as dose escalation meetings and are used to project the safety and tolerability profile of the compound in a higher dose. This implies that reviewers must make a subjective interpretation of the

presented, blinded data, on whether a pharmacologically induced (side) effect is present or not. A positive review from the dose escalation review committee is necessary to continue the study, with an increase in the dose of the investigated compound. Data presentation ordinarily does not include statistical analyses and no differentiation as to which subjects were administered active or placebo treatment.

Next to evaluating the safety and tolerability of the increased dose level, intended or unintended pharmacology is evaluated. For example, in case intended pharmacology is observed at a level that is considered therapeutically relevant, additional cohorts with increased doses may not be ethical or rational. Furthermore, if unintended pharmacology is observed, this may raise a safety concern, which may result in a premature termination of the study. Finally, it is not uncommon that follow-up studies are designed in parallel with an ongoing phase I study, using the preliminary blinded data. Thus, the subjective interpretation of blinded data plays a key role in the conduct of these studies and the design of subsequent studies.

However, no data are available on the effect sizes that are required before a pharmacological effect can be observed in these blinded evaluations of data. Therefore, the aim of the present study is to describe the probability that effects are observed in these blinded evaluations of data in a simulated phase I study design and data evaluation.

What is already known about this subject

- In early-phase clinical trials, conventionally reviewers must make subjective interpretations of generally blinded data on whether a pharmacologically induced (side) effect is present or not. However, no data are available on the effect sizes that are required before such a pharmacological effect can be observed in these blinded subjective evaluations of data.

What this study adds

- Effect sizes <2× the between-measurement standard deviation the investigated outcome frequently go unnoticed by blinded reviewers. Although unblinding reviewers resulted in about a 20% increase in correctly identifying a given effect, it also resulted in more false positives when no effect was present.

2 | METHODS

2.1 | Overall design

A custom browser-based application was developed that facilitated the review of simulated blinded evaluations of data of the datasets reviewed during routine phase I clinical studies conducted at the Centre for Human Drug Research in Leiden, The Netherlands (CHDR). Various effect sizes, expressed as multiples of the between measurement standard deviation (SD) were simulated for 6 different parameters. Scientific staff members ($n = 34$) each reviewed a median of 100 simulated datasets.

2.2 | Data simulation

2.2.1 | Pharmacokinetics

A 1-compartment pharmacokinetic model simulated a pharmacokinetic profile of an orally administered fictional compound, with the following parameters: volume of distribution = 5 L/h, clearance = 2 L/h, absorption rate constant = 0.8/h. Doses of 5 and/or 10 mg were simulated, resulting in a maximum plasma concentration (C_{max}) of 0.5 and 1.0 ng/mL respectively, with a time to C_{max} (t_{max}) of 1.7 hours. A variance of 0.01 was introduced on the volume of distribution and clearance parameter, resulting in a low level of interindividual variability in the pharmacokinetic profile. This variance was not altered between simulations as the focus of this analysis was on the variability in the pharmacodynamic response.

2.3 | Simulation of pharmacodynamic response

The response over time was simulated using a linear effect model based on the estimated parameters (see Parameter description section):

$$\text{Slope} = \text{Between measurement variability (SD)} \times \text{Effect size}$$
$$\text{Response} = (\text{Baseline} + \text{Slope} \times \text{Concentration}) \times \left(1 + \text{N}\left[0, \sigma^2\right]\right)$$

where *Between measurement variability (SD)* is the estimated standard deviation of the between-measurement variability for each parameter, multiplied by an *Effect size*, ranging between −3 and 3. The occurrence of a no effect scenario in the simulation (*Effect size* = 0) was higher to account for potential bias due to chance. This results in slopes being simulated that show no effect (effect size = 0) up to slopes that show an effect of 3× the SD of the between-measurement variability. *Baseline* is the typical value of a parameter in a population, simulated with the estimated lognormal distribution of the interindividual variability. *Concentration* is the simulated pharmacokinetic concentration (ng/mL) of an individual. Furthermore, a proportional residual error term was added on the response which was identical to the variance of the between-measurement variance, sampled from a normal distribution with mean 0 and variance σ^2.

For example. as the C_{max} of the 10-mg cohort was 1 ng/mL, a *Between measurement variability SD* of 10% and an *Effect size* of 1 would result in a pharmacodynamic response being simulated that has 10% effect (equal to 1 SD) at the t_{max}, with an identical between-measurement variability (residual error) being implemented. Dependent on the simulated effect size, this would result in effect sizes of up to 3× (30%) the between-measurement variability being simulated at the t_{max}.

Sampling of pharmacokinetic and the measurements of the response were performed at timepoint 0 hours (predose), 1, 2, 4, 6, 8 and 12 hours postdose. Parametrical diurnal effects were left out of the simulation to prevent clouding of the evaluation.

2.4 | Graphical user interface description

An internal browser-based R Shiny application was created to simulate pharmacological dose–response curves that mimicked the typical phase I dose escalation profiles frequently encountered.[1] Each set of simulated graphs contained 5 graphs of a single parameter, in which the data of 8 subjects, 6 who received the active compound and 2 who received a placebo treatment, were simulated. The following graphs were generated and displayed:

- Individual datapoints simulated over time
- Individual change from baseline data over time

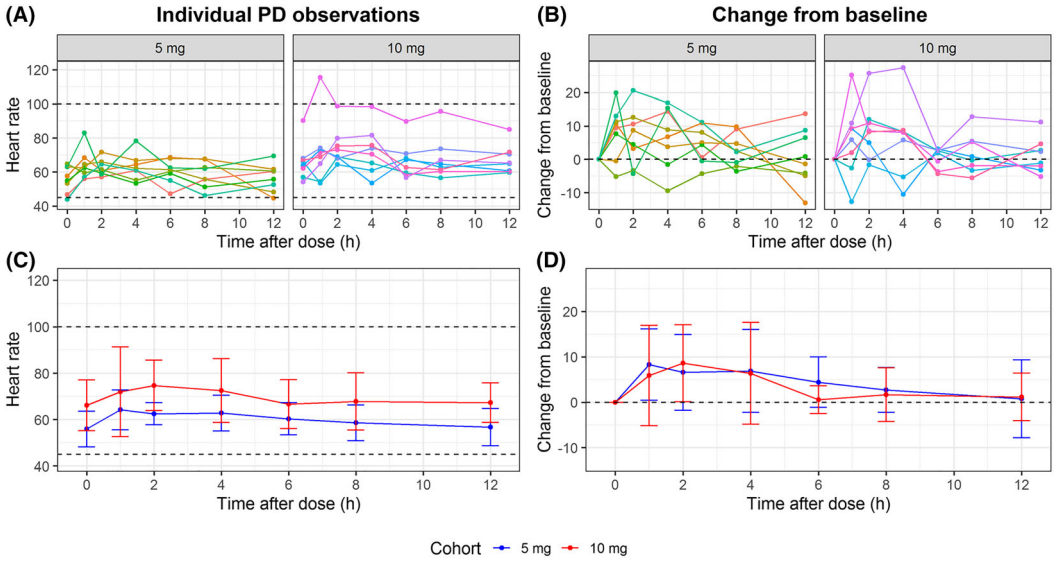

FIGURE 1 Stochastic simulation of blinded response data for 2 cohorts with 6 active and 2 placebo subjects, presented as individual measurements (A), change from baseline (B) and summarized as mean ± standard deviation (C, D) over time. Slope of 2× between-measurement variability of the heart rate parameter was simulated in this scenario. Dashed horizontal lines in Figure A and C present normal range based on literature. Dashed horizontal lines in Figure B and D depict reference line for no change from baseline.

TABLE 1 Dataset information and model estimates for each parameter

Parameter	Number of individuals with 1/2 measurements in dataset	Baseline parameter	IIV CV % (ω^2)	BMV % (σ^2)	BMV group	Simulated effect profiles
Alanine aminotransferase (U/L)	84/711	20.76	39% (0.14)	24.7% (0.06086)	High	None/positive effect
γ-glutamyl transferase (U/L)	154/625	19.95	52% (0.24)	17.6% (0.03081)	High	None/positive effect
Heart rate (beats/min)	261/495	60.44	12% (0.015)	10.1% (0.01027)	Medium	None/negative/positive effect
Diastolic blood pressure (mmHg)	256/495	71.99	10% (0.010)	7.9% (0.0062)	Medium	None/negative/positive effect
Systolic blood pressure (mmHg)	153/636	123	8.9% (0.0079)	5.8% (0.003398)	Low	None/negative/positive effect
QTcF (ms)	170/626	409.7	4.0% (0.0016)	2.2% (0.000503)	Low	None/negative/positive effect

Each parameter with corresponding levels of interindividual (IIV) and between measurement (BMV) variability. CV%, coefficient of variation.

- Mean data over time, including whiskers displaying the SD, of each cohort
- Mean ± SD change from baseline data over time, of each cohort
- The mean ± SD of the pharmacokinetic profile, of each cohort

The model either simulated a single 10-mg dose of the fictional compound or simulated both a single 5- and 10-mg dose of the fictional compound. In the latter case, all graphs were generated for both the 5- and 10-mg data. An example of a simulated response profile over time for the heart rate parameter, based on data of 2 cohorts, is presented in Figure 1. The pharmacokinetic profile, which was also presented in the graphical user interface, is shown in Figure S1.

2.5 | Parameter description

For 18 frequently recorded parameters, the baseline response, the between-measurement and the interindividual variability were calculated based on previously collected data during mandatory medical screening in healthy volunteers. Out of these 18 parameters, 6 were identified that covered a 10-fold range in coefficients of variation for the between-measurement variability, ranging from 2.2% for the Fridericia corrected QT interval to 24.7% for alanine aminotransferase (ALAT). The selected variables are shown in Table 1. Each variable was simulated with either a negative, positive or no effect following dosage except for ALAT and γ-glutamyl transferase, which were simulated as either a positive effect or no effect.

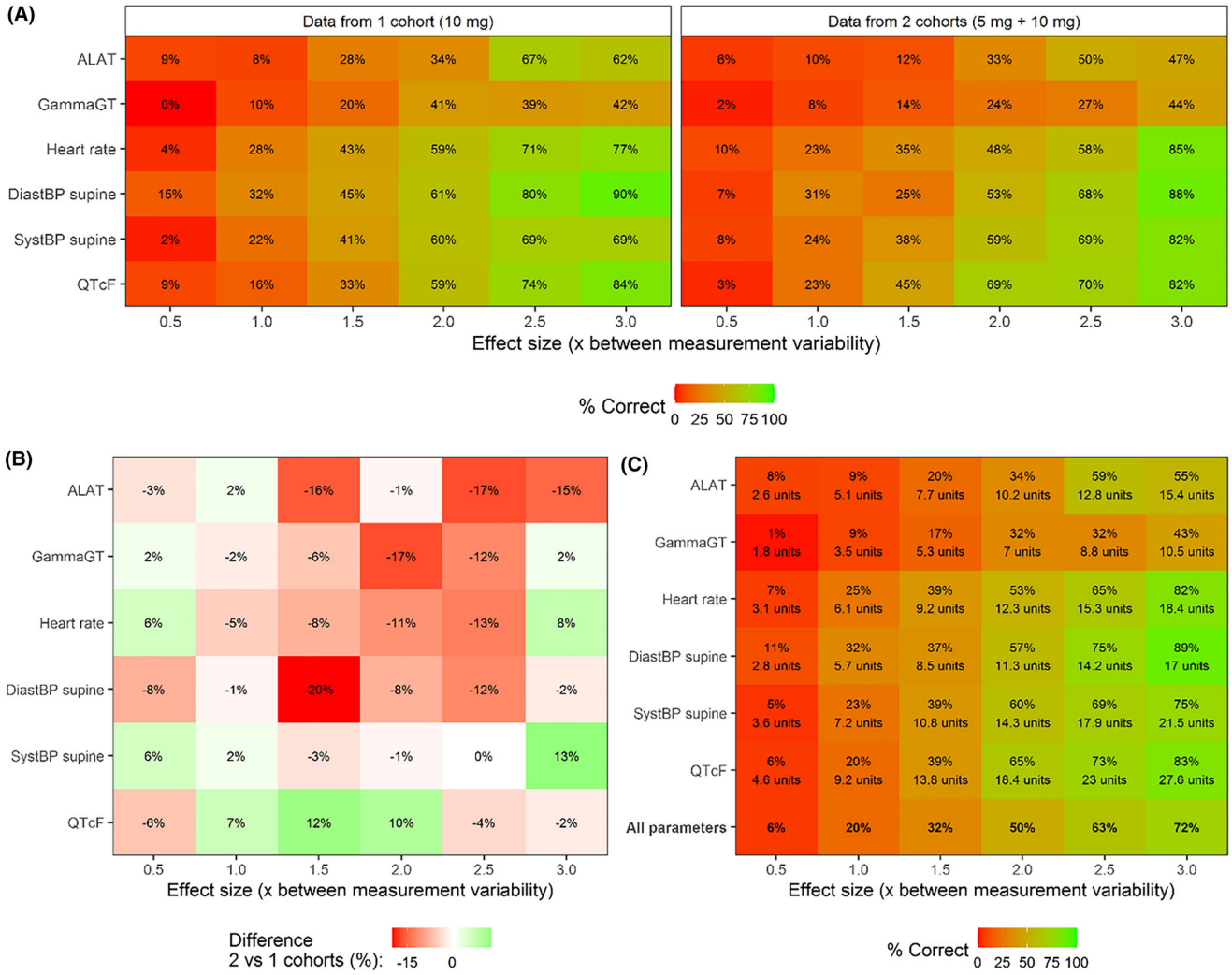

FIGURE 2 Heatmap with % correct decision of each parameter over effect size, based on blinded data of 1 or 2 cohorts (A), the calculated delta (B) and all data combined (C). The % correct and the simulated effect size of a parameter is reported in each cell (C). Parameters are sorted on the level of between measurement variability (high to low). ALAT = alanine aminotransferase; GammaGT = γ-glutamyl transferase; DiastBP supine = diastolic blood pressure in supine position; SystBP supine = systolic blood pressure in supine position; QTcF = Fridericia corrected QT interval.

2.6 | Reviewers

All reviewers ($n = 34$) were part of the scientific staff of the CHDR with ample experience at different research institutions or pharma companies. This included research physicians or clinical scientists ($n = 20$), postdocs ($n = 12$), or professors ($n = 2$) with a mean of 1.7, 4.3 and 15.5 years of clinical experience, respectively. Reviewers were provided with standardized instructions, among which that they could only consider an effect positive or negative in case they would defend this effect against a study sponsor that is developing the fictional compound irrespective of whether the effect would be desirable or not.

2.7 | Unblinded evaluation

To differentiate the effects of blinding on the interpretation of these data, a simulation was performed with the only difference that participants on active treatment were identifiable from participants treated with placebo. The cohort average plots also displayed the averages of the actively treated participants from the placebo-treated participants. The same parameters with the same settings were used. Two CHDR staff members (G.J.H. and P.G.) reviewed a total of 1423 blinded and 1230 unblinded datasets.

2.8 | Statistical analysis

Data are presented as % of correctly identified as an effect in a heatmap, specified for each effect size and parameter, or % of correctly identified as no effect for each parameter as a bar plot. A sigmoidal nonlinear least squares model was fitted in R on blinded data with a simulated effect for each parameter, and for comparison of the blinded and unblinded simulations.[1]

To achieve a resolution of maximum 5% for each parameter, effect size and number of cohorts (72 combinations) it was estimated that 1440 simulated datasets with an effect size higher or lower than 0 needed to be reviewed. A linear mixed effect model to identify a minimal detectable effect size with a statistical power of 80% was used based on 1 or 2 (4 placebo subjects, 6 active per cohort) cohorts for each parameter with a contrast up to 6 hours after dosing. The minimal detectable effect size was based on simulated effect sizes ranging from 2 to 4.5 with 500 iterations for each effect size, parameter and number of cohorts. The statistical power of each scenario was derived, and an overall mean power was calculated.

3 | RESULTS

In total, 34 scientific reviewers evaluated between 3 and 773 simulated datasets (median = 100, mean = 139), a total of 3779 data simulations were evaluated, of which 1945 data simulation with an effect size higher or lower than 0.

3.1 | Effect identification

An increasing effect size resulted in a higher chance of the effect being identified by the reviewer, as can be observed in Figure 2. On average, only 6% of effect sizes of 0.5 SD and 20% of effect sizes of 1× the SD of the between-measurement variability were correctly identified by the reviewers. For example, for systolic blood pressure effects, this resulted in a 5% chance to identify a 3.6-mmHg effect and 23% probability to identify a 7.2-mmHg effect. Also, a lower between-measurement variability resulted in a higher chance of the effect being identified by the reviewer, in particular for ALAT and γ-glutamyl transferase at the largest effect sizes, which have the highest between-measurement variability, as is best observed in Figure S2. The effect of reviewing 2 dose level as opposed to 1 dose level resulted in a similar probability of the effect being identified by the reviewer and no clear improvement was identified, as shown in Figure 2B and S3.

3.2 | No effect identification

On average, the reviewer correctly identified 92–95% of simulations with no effect, with negligible effect of intrasubject variability as illustrated in Figure 3. Also, there appears to be no effect of reviewing data from 1 cohort as compared to 2 cohorts.

3.3 | Unblinded analysis

A substantial improvement in the ability to identify effects was made by unblinding the reviewer to the treatment allocation.

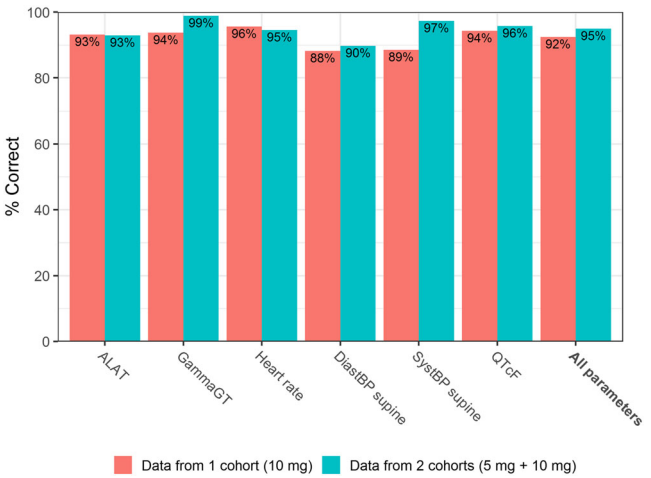

FIGURE 3 Barplot with % correct decision when no effect was present in different simulation scenarios. ALAT = alanine aminotransferase; GammaGT = γ-glutamyl transferase; DiastBP supine = diastolic blood pressure in supine position; SystBP supine = systolic blood pressure in supine position; QTcF = Fridericia corrected QT interval.

Unblinding the reviewer resulted in about a 20% increase in the ability of reviewers to identify the simulated effects correctly. However, this also resulted in an increased incorrect assessment when no effect was simulated, which was on average 13% less as shown in Figures 3 and 4.

3.4 | Minimal detectable effect size

Correct identification of a minimal detectable effect size with an 80% statistical power was not reached with a linear mixed effects model for effect sizes of up to 4.5 SD for a single cohort data simulation, as illustrated in Figure 5. However, for 2 cohort data simulations, the minimal detectable effect size with an 80% power was reduced to 3.5 SD because of the increased number of placebo subjects.

4 | DISCUSSION

This analysis found that effect sizes <2× the SD of the between-measurement variability are frequently not observed during the blinded reviewers subjective interpretation. This probability does not seem to improve when multiple dose levels were modelled. In contrast, when no effect is present, this is usually correctly observed. Unblinding the data resulted in some improvement, but effect sizes <1× the SD of between-measurement variability frequently remain unobserved and a false positive effect is more common in this scenario. Moreover, to detect a minimal effect with an 80% statistical power in linear mixed effect model required effect sizes of 3.5 or higher in simulations with 2 cohorts while single cohort simulations did not even reach those levels at effect sizes of 4.5. These results indicate that the weight given to these blinded analyses for predicting

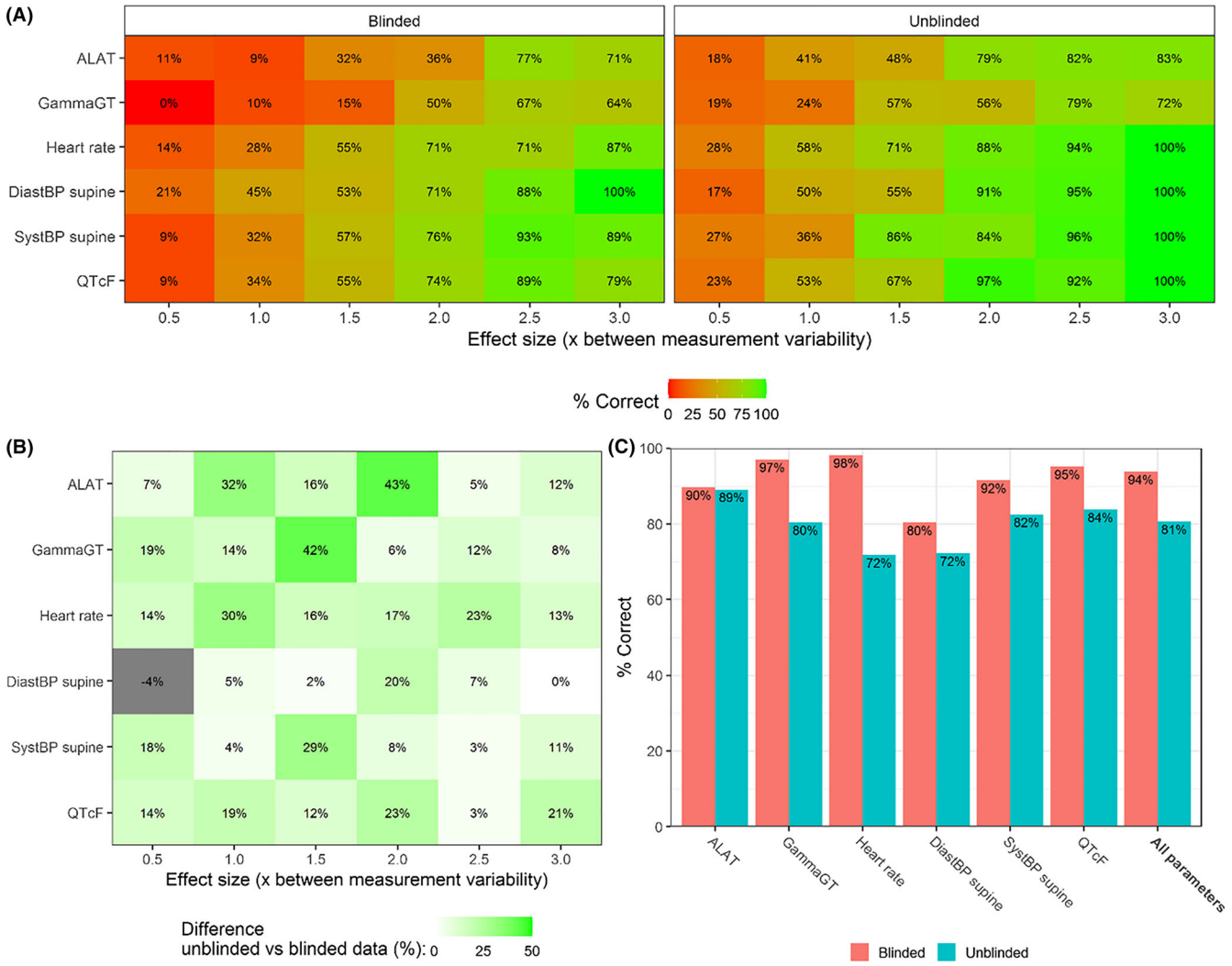

FIGURE 4 Heatmap with % correct decision of each parameter over effect size, based on blinded or unblinded data (A), the calculated delta (B) and all data combined (C). Data are presented of 2 individuals who were shown both blinded or unblinded data. The % correct when no effect was simulated is presented in C. Parameters are sorted on the level of between measurement variability (high to low). ALAT = alanine aminotransferase; GammaGT = γ-glutamyl transferase; DiastBP supine = diastolic blood pressure in supine position; SystBP supine = systolic blood pressure in supine position; QTcF = Fridericia corrected QT interval.

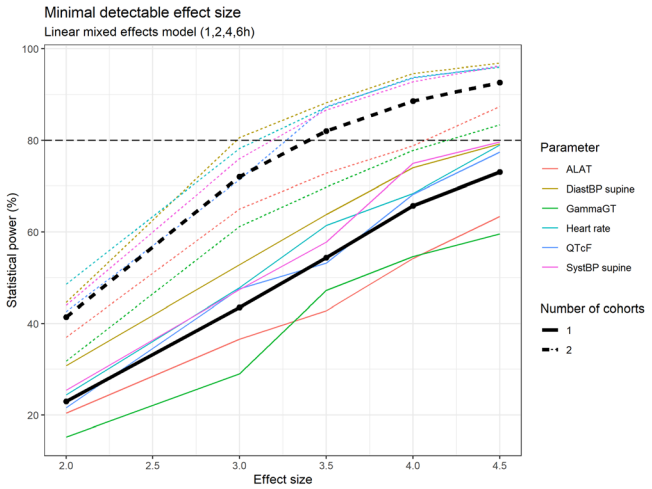

FIGURE 5 Statistical power for each parameter with an effect size between 2.0 and 4.5 standard deviation effects as calculated with a linear mixed effect model for both single or double drug cohort simulations. Dotted horizontal line highlights the statistical power to detect a minimal detectable effect size with 80% certainty. ALAT = alanine aminotransferase; GammaGT = γ-glutamyl transferase; DiastBP supine = diastolic blood pressure in supine position; SystBP supine = systolic blood pressure in supine position; QTcF = Fridericia corrected QT interval.

the safety and efficacy of an increasing dose level, as is common in current phase I practice, may be too big and should be re-evaluated.

Although the mainstay of phase I dose escalation studies is the ability to review safety, pharmacokinetic and intended or unintended pharmacology, there are surprisingly few data available to support the rationale for this approach. Most effect sizes in phase I studies, particularly in otherwise healthy volunteers, are relatively small in comparison to the variability. For example, the maximum pharmacodynamic effect of anti-hypertensive drugs is usually about 10 mmHg, which is about 1 SD between-measurement variability of systolic blood pressure.[2] The reviewers of our data detected this only 23% of the time. Another example is QTc prolongation, in which some noncardiovascular drugs with a mean increase in the QT interval of 5–10 ms have been withdrawn from the market because of an increased risk of torsade de pointes after coadministration with a strong CYP inhibitor.[3] Most notable is terfenadine, which averaged a QT interval prolongation of approximately 6 ms in healthy individuals which was about 0.5 SD intrasubject variability in our data set.[3] The reviewers of our data detected this only 6% of the time. Unblinded review of the same data resulted in a higher possibility of correctly identifying smaller effect sizes for all parameters including systolic blood pressure and QTc. Therefore, our data support unblinded review of systolic blood pressure and QTc in phase I. This should result in a higher chance of detection of effect sizes <1× the SD of intrasubject measurement variability at the cost of an increased probability of incorrect identification of an effect.

Conversely, incorrect identification of an effect may have unwanted consequences. First, overlooking a pharmacodynamic effect may lead to inaccurate prediction of the pharmacodynamic effects of the next dosing step. This leads to misinterpretation of safety risks. Also, our analysis shows that an investigator's capabilities to identify the maximum tolerable dose are limited. We have previously published that identifying a maximum tolerated dose should never be the purpose of phase I trials.[4] Moreover, when desired pharmacodynamic effects are overlooked this may lead to irrational continuation of the study to a dose level that is excessive. This may even lead to discontinuation of development when desired pharmacodynamic effects are not observed especially when further dose escalation is unsafe based on preclinical studies. Such a false-negative finding may also affect the design of follow-up studies.

Secondly, our results suggest that investigators should design their study keeping in mind the observable effect sizes with their respective statistical power in blinded interim analyses given the study population. For pharmacodynamic readouts that are implemented to evaluate desired pharmacodynamic effects, the effect size should exceed at least 1 SD, but preferably 2× the SD of the intrasubject measurement variability. Obviously, increasing the number of study participants will lead to smaller effect size to be detectable and an adequate power analysis must be performed. However, from a safety perspective, increasing the entire study population is not considered feasible. Moreover, our data show that having intermediate dosing levels up until the effect size that was simulated does not improve the interpretation of the data. Rather, it is recommended to include sensitive methodologies in phase I studies with a lower intrasubject variability to allow more accurate identification of pharmacologically induced effects.

A workable solution for key variables would be to conduct grouped or unblinded analyses to allow more accurate evaluation of pharmacodynamic and pharmacokinetic effects. Nevertheless, it must be noted that it may still not be possible to detect small effects considering the limited group size typically used in phase I studies. In addition, it is not unimaginable that the traditional phase III studies may be abolished and replaced with intensive measurements with sensitive methodologies in phase I and II studies in the future. This will allow a more accurate assessment of the chance that a compound will demonstrate favourable effects in clinical practice. Lastly, compound or even disease specific model-informed drug development as proposed by the Food and Drug Administration may ultimately improve decision-making through integration of data from each new (pre)clinical study into biological and statistical models during development.[5]

4.1 | Limitations

The initial training simulations with medical screening data provided a fit for purpose selection of parameters with a broad intrasubject distribution and intersubject variability specific for the selected dataset. However, through implementation of additional data per subjects, or potential circadian rhythm in an outcome, may change the level of inter- and intrasubject variability in each parameter. Moreover, effect

sizes are also variable between individuals in real-life and variability in the pharmacokinetics is more present, which may further affect the total level of variability in the data, although both effects were not accounted for in the present simulations. Finally, our results were solely based on the evaluation by reviewers of a single institute while perhaps variability exists between research centres. To minimize this potential bias, we included reviewers with a wide range of years of clinical experience of up to 15 years with work experience at various different institutes.

5 | CONCLUSION

Our analysis showed that effect sizes <2× the between-measurement SD of the investigated outcome frequently go unnoticed by blinded reviewers even when multiple dose levels were modelled. Unblinding the data resulted in some improvement, but effect sizes <1× the between-measurement SD frequently remain unobserved and a false positive effect is more common in this scenario. These results indicate that the weight given to these blinded analyses for predicting the safety and efficacy of an increasing dose level, as is common in current phase I practice, may be too large and should be re-evaluated.

ACKNOWLEDGEMENTS

None.

CONTRIBUTORS

G.H., M.E., G.W., A.C., J.B. and P.G., wrote the manuscript, G.H., M.E., A.C., J.B. and P.G. designed the research, G.H., M.E. and P.G. performed the research, G.H., M.E. and P.G. analysed the data.

COMPETING INTERESTS

The authors declare that there are no conflicts of interest regarding this analysis.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Gerardus J. Hassing https://orcid.org/0000-0002-7979-3546
Michiel J. van Esdonk https://orcid.org/0000-0001-8159-0273
Adam F. Cohen https://orcid.org/0000-0001-7452-6222
Jacobus Burggraaf https://orcid.org/0000-0001-7912-0918
Pim Gal https://orcid.org/0000-0001-7622-9029

REFERENCES

1. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. Shiny: web application framework for R. *R Package Version*. 2017;1.
2. Bramlage P, Hasford J. Blood pressure reduction, persistence and costs in the evaluation of antihypertensive drug treatment – a review. *Cardiovasc Diabetol*. 2009;8(1):1-13. doi:10.1186/1475-2840-8-18
3. Roden DM. Drug-Induced Prolongation of the QT Interval. *N Engl J Med*. 2004;350(10):1013-1022. doi:10.1056/NEJMra032426
4. Cohen A. Should we tolerate tolerability as an objective in early drug development? *Br J Clin Pharmacol*. 2007;64(3):249-252. doi:10.1111/j.1365-2125.2007.03023.x
5. Wang Y, Zhu H, Madabushi R, Liu Q, Huang S-M, Zineh I. Model-Informed Drug Development: Current US Regulatory Practice and Future Considerations. *Clin Pharm Therap*. 2019;105(4):899-911. doi:10.1002/cpt.1363

SUPPORTING INFORMATION

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

How to cite this article: Hassing GJ, van Esdonk MJ, van Westen GJP, Cohen AF, Burggraaf J, Gal P. Dose escalations in phase I studies: Feasibility of interpreting blinded pharmacodynamic data. *Br J Clin Pharmacol*. 2022;1-8. doi:10.1111/bcp.15473