Impact of biomarker-based design strategies on the risk of false-positive findings in targeted therapy evaluation

Tat-Thang Vo, Alexandre Vivot, Raphaël Porcher

Impact of biomarker-based design strategies on the risk of false-positive findings in targeted therapy evaluation

Tat-Thang Vo[1,2], Alexandre Vivot[1,3], Raphaël Porcher[1,3]

1. INSERM, UMR1153 Epidemiology and Statistics Sorbonne Paris Cité Research Center (CRESS), METHODS Team, Paris, F-75004, France; Paris Descartes University, France.

2. Department of Applied Mathematics, Computer Science & Statistics, Faculty of Science, Ghent University, Ghent, Belgium

3. Assistance Publique des Hôpitaux de Paris (AP-HP), Hôpital Hôtel Dieu, Centre d'Épidémiologie Clinique, Paris, France.

Address correspondence to: Dr Alexandre Vivot. Centre d'épidémiologie clinique. Hôtel-Dieu de Paris (A2, 1er étage).1 place du parvis Notre-Dame.75181 Paris cedex 04. France.

Tel: +33 1 42 34 78 12 Fax: +33 1 42 34 87 90 Email: alexandre.vivot@aphp.fr

Running Title: Risk of false-positive findings in targeted therapy evaluation

Abbreviations list: BMK, biomarker; CVASD, cross-validated adaptive signature design; pre-BMK, predictive biomarker ; RCT, randomized clinical trial; TIADA, testing-in-all-direction approach

Funding statement: Tat-Thang Vo was supported by the funding from *Conseil Régional, Île-de-France* (Île-de-France Regional Council, Paris, France) within the program *Bourse Master « Île-de-France »* for the 2014/2015 period. The sponsor has no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Disclosure: The authors declare no conflict of interest.

TRANSLATIONAL RELEVANCE

When there is more than one potential predictive biomarker, new targeted agents are often evaluated across several biomarker-defined subpopulations without any correction for multiple testing. This may result in a high risk of false positive findings. In this study, we calibrate the Cross-Validated Adaptive Signature Design (CVASD) and investigate the new design as an alternative to overcome the multiplicity problem. In the modified CVASD, one first evaluates the treatment effect in a sensitive subset of patients identified by a classification algorithm. When there is no effect in this subset, the trialist proceeds to evaluate the treatment effect on the broad population. Type I error is corrected as proposed in the original CVASD. Simulation results show that this slight calibration makes the so-called modified CVASD successfully outweigh the conventional approach, not only in terms of adequately controlling the type I error but also in terms of correctly identifying the predictive biomarker(s).

ABSTRACT

Purpose: When there is more than one potentially predictive biomarker for a new drug, the drug is often evaluated in different subpopulations defined by different biomarkers. We aim to (i) estimate the risk of false-positive findings with this approach and (ii) evaluate the Cross-Validated Adaptive Signature Design (CVASD) as a potential alternative.

Experimental Design: By using numerically simulated data, we compare the current approach and the CVASD across different settings and scenarios. We consider 3 strategies for CVASD. The two first CVASD strategies are different in terms of the partitioning of the overall significance level (between the population test and the subgroup test). In the third CVASD strategy, the order of the two tests is reversed, i.e. the population test is realized when the prioritized subgroup test is not statistically significant.

Results: The current approach results in a high risk of false positive findings, whereas this risk is close to the nominal level of 5% once applying the CVASD, regardless of the strategy. When the treatment is equally effective to all patients, only the CVASD strategies could specify correctly the absence of a sensitive subgroup. When the treatment is only effective for some sensitive responders, the third CVASD strategy stands out by its ability to correctly identify the predictive biomarker(s).

Conclusion: The drug-biomarker co-evaluation based on a series of independent enrichment trials can result in a high risk of false positive findings. CVASD with some appropriate adjustments can be a good alternative to overcome this multiplicity issue.

INTRODUCTION

Precision medicine, also known as stratified or personalized medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person (1–4). One fundamental challenge in precision medicine is to identify a subset of patients with specific biomarkers that will respond adequately to a new targeted agent/regimen. Design strategies have evolved in the past few years to deal with this challenge (4–7). Of all these designs, the biomarker-stratified randomized controlled trial permits to rigorously evaluate the clinical utility of the proposed marker in terms of correctly guiding the treatment selection (8,9). Recent evidence otherwise shows that most trialists are using an enrichment design, in which only biomarker positive patients are eligible and are randomized to receive either the new drug or an appropriate control (7,10).

One common issue of biomarker-stratified and biomarker-enriched designs is that they can only evaluate a single biomarker at a time. In practice, the situation is more complicated. As the development and validation of biomarkers is a complex process that requires considerable time and resources, it often lags behind the therapeutic development of the targeted agent (11). When a treatment is ready to be assessed in clinical trials, early phase data might propose more than one potential predictive biomarker. A recent review finds out that in such a case, the drug is often assessed in a series of independent enrichment trials. For instance, in colon and rectum cancers, panitumumab has been evaluated across 3 biomarker-defined subpopulations, namely BRAF-mutated subpopulation (1 trial), EGFR-positive subpopulation (2 trials) and KRAS wild-type subpopulation (12 trials) (12).

This *"testing-in-all-direction"* approach (TIADA) has several limitations. First, it reduces the chance that a patient could participate in trials, as biomarker-negative patients (who are thus not

eligible for one enriched RCT) are usually not simultaneously evaluated for eligibility for another trial. Second, this approach may result in an inflation of the type I error. When there is no treatment effect in the whole population and one biomarker-enriched trial is performed, the false positive rate is well controlled at a level of 5%. However, when several biomarkers are independently evaluated in several studies, the chance of incorrectly stating the treatment effect under the null hypothesis can be much higher than this conventional threshold of 5%. New designs such as umbrella and basket trials have been recently proposed to overcome the first challenge (10,13–16). Nonetheless, the multiplicity issue still remains, as strata in an umbrella trial are often analyzed separately without much consideration to the overall risk of false positive findings. In fact, multiplicity is less serious when early-phase trials are just for explanatory purposes, as in such a case the findings do not entirely rely on statistical testing. However, recent evidence shows that licensing decisions of regulatory authorities like the FDA are often based on statistical inference carried out in early-phase trials (17–20). As a consequence, the type I error inflation must be taken into account to prevent the risk of restricting the drug indications to an inappropriate sub-population.

The Cross-Validated Adaptive Signature Design (CVASD) has the potential to overcome the issue of multiplicity. Such a design was first proposed by Freidlin et al. to detect signatures in some large multi-dimensional genetic datasets (e.g. more than 10.000 genes) (21). In this setting, CVASD increases the empirical power compared to the traditional broad eligibility approach of RCTs (21). However, it is unclear whether CVASD can also be useful in the context of latter-phase drug-biomarker co-evaluation, especially when the number of biomarker candidates is not considerably large. In other words, it is questionable whether CVASD can be superior to a series

of independent enrichment trials, in terms of controlling the type I error without substantially deteriorating the power of correctly identifying the right predictive biomarker(s).

This study aims to (i) estimate the risk of false-positive findings of the current approach of cancer drug development (TIADA) and (ii) investigate statistical properties of the Cross-Validated Adaptive Signature Design (CVASD) as a potential alternative to the current approach.

METHODS

Data Generation Process

Numerical simulations are conducted to emulate phase III oncology trials evaluating the impact of a new targeted agent versus a standard therapy with respect to a time to event outcome. In practice, this outcome can be either the progression-free survival or the overall survival. In the first scenario, we assume that early-phase data identifies three biomarker candidates that can potentially characterize the drug responders. This scenario mimics the real situation of panitumumab in colon and rectum cancer found previously (12). However, some other biomarkers might have been assessed after the publication of our previous work. Furthermore, non-genetic factors such as gender or age group (e.g. more than 60 years of age) can also be taken into account. Because of this, a second scenario (scenario 2) with up to six biomarker candidates is also considered.

In both scenarios, the status of the binary biomarker i is denoted by Z_i ($Z_i = 0$ or 1), where $i = 1$ to 3 for scenario 1 and $i = 1$ to 6 for scenario 2. Z_i are supposed to be mutually correlated (Appendix 1). Apart from Z_i's, the survival time is also influenced by a continuous non-predictive factor L. The effects of treatment T, of biomarkers Z_i's and of the prognostic factor L are simulated by using a Cox proportional hazard model: where is the linear predictor of the Cox model.

In scenario 1, data are generated such that all three biomarkers are prognostic. By contrast, in scenario 2, only the first four candidates (i.e. Z_1 to Z_4) are prognostic. To generate the predictive effect for a biomarker, a two-way interaction term between treatment T and this biomarker is added into the model (1). For instance, when Z_1 and Z_2 are predictive, the data-generating mechanism proposed for scenario 1 is: and for scenario 2 is:

A biomarker Z_i is considered as strongly, moderately and weakly predictive when the hazard ratio equals to 0.35, 0.5 and 0.65, respectively. In both scenarios, the survival time for a patient profile with = 0 is simulated by using a Weibull distribution with a shape parameter of 2 and a median survival of 14. For generating the censoring time, we applied a Weibull distribution with a shape parameter of 2 and a scale parameter of 30. This results in a censoring rate of 25 – 35% across the settings.

We consider 3 settings in each scenario. In setting 1, there is no predictive biomarker due to no treatment effect in the whole population (Table 1 - 1.1), or because the treatment works equally well for all patients (Table 1 - 1.2). Setting 2 and 3 consider the situation where predictive biomarker(s) is/are present. The new targeted agent is more effective than the standard care for sensitive patients (i.e. those having at least one predictive biomarker positive). In contrast, the two treatments are equally effective for non-sensitive patients. In setting 2, the sensitive subgroup is characterized by one predictive biomarker (Z_1). The predictive value of this biomarker () decreases gradually from sub-setting 2.1 to sub-setting 2.3 (Table 1). In setting 3, the sensitive subgroup is characterized by two predictive biomarkers (Z_1 and Z_2). In sub-setting 3.1, both of them are moderately predictive. In sub-setting 3.2, the first predictive biomarker (Z_1) is moderately predictive and the second one (Z_2) is weakly predictive (and).

The proportion of patients being positive to each biomarker at the population level is given in Table 1. Across three settings, this is fixed at 30% for every non-predictive biomarker. For predictive biomarkers, data is generated such that their positive status will become less frequent when they are more predictive.

Strategy A: performing a series of biomarker-enriched RCTs.

In each simulation, the biomarker-enriched RCTs are independently generated with N patient profiles screened for eligibility per trial. Treatments are then randomized for those who are biomarker-positive. We consider 2 values for N, i.e. N = 500 and N = 1000. The randomization ratio in all enrichment trials is 1:1. The treatment and control groups are compared by a log-rank test at a significance level of 5%. No adjustment for multiplicity is considered since trials are conducted and analyzed separately.

Strategies B1, B2 and B3: applying the CVASD with different partitioning of type I error risk

We investigate 3 different strategies for CVASD. The first two ones (i.e. B1 and B2) apply the original CVASD proposed by Freidlin et al, in which the final analysis begins with an overall comparison between two arms using the data from all patients. If the comparison is statistically significant at a pre-specified significance level α_1 ($\alpha_1 < \alpha$), the new treatment is considered beneficial to the whole population. Otherwise, the design proceeds to the signature development – validation stage to identify a subgroup that is potentially sensitive. The statistical test for the identified subset is carried out at a significance level $\alpha - \alpha_{1\,(21,22)}$. We consider $\alpha_1 = 0.04$ for strategy B1 and $\alpha_1 = 0.01$ for strategy B2. Strategy B1 therefore prioritizes the overall comparison, whereas strategy B2 prioritizes the subgroup one.

In strategy B3, we modify the original CVASD by reversing the order of the two testing levels. The subgroup comparison is performed first, at a pre-specified significance level of α_2 ($\alpha_2 < \alpha$).

The overall comparison in the broad population is only performed (at a significance level of $\alpha - \alpha_2$) when the subgroup test is not statistically significant. We evaluate this strategy when $\alpha_2 = 0.01$.

To ensure that the two approaches are compared on a fair basis, a sample of N profiles is simulated for one CVASD trial. To detect the sensitive responders, we apply the same identification algorithm as previously used by Friedlin et al (Appendix 2) (21). Due to this algorithm, the true predictive biomarker(s) will over-represent in the detected subgroup, in the sense that most of the sensitive patients will possess a positive status of the predictive biomarker(s). Based on this, we propose a classification rule that helps to identify the biomarker characterizing the sensitive responders. For each Z_i, if the proportion of Z_i-positive patients in the identified subgroup $Pr(Z_i = 1|$ sensitive) is maximal among all candidates, Z_i is considered as the biomarker that characterizes the sensitive responders.

Main outcomes

We first focus on the risk of a false positive finding of the four strategies when there is no treatment effect in the whole population. This false positive risk can be estimated in sub-setting 1.1, by calculating the proportion of simulations that show statistical significance.

When the treatment is equally effective for all patients in the population (sub-setting 1.2), the chance of correctly identifying the absence of predictive biomarkers is the main outcome of interest. For strategy A, this requires all enrichment trials showing no statistical significance. For the CVASD strategies, the population test must show statistical significance.

For setting 2 and 3 (i.e. the treatment is only effective for some patients in the population), we compare the four strategies with respect to the chance of identifying a correct sensitive subgroup. In setting 2 (i.e. one predictive biomarker – Z_1), a correct sensitive subgroup is found by strategy

A if the trial enriched on the predictive biomarker Z_1 is the only one showing statistical significance. In contrast, a correct sensitive subgroup is found by the CVASD if the subgroup test is statistically significant and the identified subgroup is characterized by the biomarker Z_1.

In setting 3 (i.e. two predictive biomarkers – Z_1 and Z_2), a correct sensitive subgroup is found by strategy A if at least one out of two predictive biomarker-enriched trials show statistical significance, whereas all trials enriched on a non-predictive biomarker do not. For the CVASD strategies, the subgroup test must be statistically significant and the identified subgroup is characterized by either Z_1 or Z_2.

Ethical Statement

This is a numerical simulation study. No humans nor animals were involved in this study.

Thus there were no ethical guidelines applicable to this study and it did not need institutional review board (IRB) nor written consent.

RESULTS

We first discuss the results when the sample size is N = 1000 patients.

Setting 1.1 – No treatment effect in the whole population

When there is no treatment effect in the whole population, the false positive risk of the current approach (strategy A - series of enrichment trials) inflates up to 12.4% in scenario 1 (i.e. 3 candidates) and 20.0% in scenario 2 (i.e. 6 candidates). By contrast, this risk is close to the nominal level of 5% when applying the CVASD strategies, regardless of the scenario (figure 1 and 2).

Setting 1.2 – Treatment is equally effective for all patients

The current approach (strategy A) has a modest chance to correctly specify the absence of a sensitive subgroup. In scenario 1 (3 candidates), only 51.9% of replicates consist of all enrichment trials showing statistical significance. This proportion in scenario 2 (6 candidates) is 36.4%.

For the original CVASD, an incorrect sensitive subgroup is found in a minor percentage of runs. Instead, the design often comes up with a population-level finding, even when the subgroup test is prioritized (strategy B2). In both scenarios, the population test of strategy B2 is statistically significant in about 99% of the replicates.

When the subgroup test is performed before the population test (strategy B3 – modified CVASD), the percentage of correct population findings decreases but still lies in an acceptable range, e.g. 82% in scenario 2 (6 candidates).

Setting 2 – One sensitive subgroup characterized by one predictive biomarker (Z1)

Simulation results show that the original CVASD will perform better when most of the type I error is dedicated to the subgroup test (strategy B2 vs. B1). However, this is not enough for CVASD to outperform the current approach (strategy A - series of enrichment trials). For instance, in sub-setting 2.2 of scenario 1 (i.e. one moderately predictive biomarker out of 3 candidates), the percentage of picking up the true predictive biomarker is 24.7% for strategy A, but only 3.5% for strategy B1 (original CVASD favoring the population test) and 9.2% for strategy B2 (original CVASD favoring the subgroup test). Meanwhile, the modified CVASD (strategy B3) stands out by its high performance. In the same sub-setting 2.2 (scenario 1), the proportion of correct subgroup findings for B3 is 47.9%, twice and four times higher than for strategy A and B2, respectively.

When there are two predictive biomarkers among the candidates (setting 3), the original CVASD hardly detect well at least one predictive biomarker. This is worse when the subgroup test is not prioritized (strategy B1 vs. B2). By contrast, the modified CVASD (strategy B3) still behaves properly and outperforms the other strategies. Consider for example the sub-setting 3.2 (i.e. one moderate and one weak predictive biomarker). In this sub-setting, the rate of correct subgroup findings of four strategies is 11.5% (A), 3.2% (B1), 9.0% (B2) and 90.7% (B3), respectively.

Reasons of incorrect findings across the settings 2 and 3

When a sensitive subgroup exists (setting 2 and 3), the most frequent reason for a wrong finding of strategy A (series of enrichment trials) is that it fails to identify an adequate sensitive subgroup (i.e. trials enriched on a non-predictive biomarker also show statistical significance). In contrast, the CVASD strategies often show no findings when coming up with a wrong conclusion.

Impact of candidate number on the performance of different strategies

The CVASD's performance remains stable when the number of biomarker candidates increases. As can be seen from figure 1 & 2, the percentage of each type of findings for the CVASD strategies only varies slightly when passing from scenario 1 (3 biomarker candidates) to scenario 2 (6 biomarker candidates). By contrast, results of strategy A change substantially when there are more biomarkers: the percentage of incorrect subgroup findings increases greatly, whereas the percentage of incorrect population findings decreased quite remarkably (setting 2 and 3).

Impact of sample size on the performance of different strategies

We compare the performance of different strategies when the sample size increases from 500 to 1000 (Figure 1 and 2). Strategy A (series of enrichment trials) does not perform more effectively:

the chance of correctly specifying the predictive biomarker(s) among the candidates decreases, but the chance of an incorrect finding (due to either picking up the incorrect predictive biomarker(s) or showing statistical significance on the population level) increases considerably. This can be seen in both of the two settings 2 and 3. For the original CVASD (strategy B1 and B2), the population test performed in advance will largely take advantage of the increased sample size. As a result, the correct subgroup findings proportion decreases. In contrast, the modified CVASD (strategy B3) is remarkably more effective when the sample size is larger, not only in the settings 2 and 3 (predictive biomarker(s) present) but also in the setting 1.2 (treatment equally effective to the broad population).

DISCUSSION

The drug-biomarker co-evaluation based on a testing-in-all-direction approach has several shortcomings. First, using this approach inflates considerably the risk of finding a false positive result due to the fact that no adjustment for multiplicity issue is realized. The more biomarkers are evaluated and tested in the independent studies, the higher and more serious the risk of false positive findings can be. This approach, however, is common in practice. A new targeted agent can be evaluated across different biomarker-defined subpopulations in several studies addressing one type of cancer, or for the same biomarker in different cancer types (12,23). While the public health community implicitly accepts multiplicity inflation due to independent phase III testing of a new anticancer agent in different stages of the same disease, independent testing of a new agent in multiple biomarker-defined subgroups of the same clinical setting is apparently problematic and should be adjusted for.

Second, if the treatment works well in the whole population and there is no requirement for a guide of treatment selection, performing a series of enrichment trials hardly indicates the absence

of a sensitive subset due to no comparison on a population level. This shortcoming results from the well-known disadvantage of enrichment designs. As the new agent is only evaluated in the biomarker-positive subpopulation, part of the picture regarding the treatment effect in the biomarker-negative subgroup is concealed. Hence, evidence to evaluate the predictivity of a candidate becomes inadequate and negative patients that also gain benefice from the new treatment will apparently be undertreated.

Third, the testing-in-all-direction approach has a quite modest ability to correctly pick up the predictive biomarker (among the candidates) when this presents. In such a situation, the approach often shows either a broad population finding or a wrong subgroup finding. These wrong findings are more apparent when the number of biomarker candidates is high. This is due to the fact that biomarker candidates can be strongly correlated. When the study is enriched on a non-predictive biomarker that is correlated with a predictive one, a remarkable proportion of the participants will be positive to both biomarkers and will respond to the new treatment, since they are actually sensitive responders. As a result, the trial will have a high chance to show statistical significance but leads to a potential misunderstanding that the non-predictive biomarker is actually predictive.

The aforementioned shortcomings of the current approach call for a more appropriate method to evaluate several biomarkers at a time. In this study, we find out that the Cross-Validated Adaptive Signature Design controls well the Family-Wise Type I Error in the weak sense and could be a solution to overcome the multiplicity issue. CVASD behaves stably when the number of biomarker candidates increases. Besides, as the subgroup identification procedure of CVASD has a relatively good specificity, this design guarantees that when no sensitive subgroup exists, the risk of inadequately restricting the drug indications to a subset of patients is minimized.

However, the performance of the original CVASD in terms of identifying the true predictive marker if this presents is quite modest. In such a situation, CVASD often comes up with a conclusion of a broad treatment effect although the targeted agent is only beneficial for certain patients. This result, however, is not surprising. The population test of CVASD actually evaluates the treatment by averaging its effect over the whole population. When the treatment is effective for some but not for others, there is indeed an effect on average. This average effect can be even considerable if the treatment is strongly effective in the sensitive subgroup. Considering this, one might argue about the necessity of the population test. The sensitive patients will be more easily detected when all study power is dedicated for the subgroup identification. However, one can hardly expect the trialists not to carry out a population test but only a subgroup-level test, given that patients are broadly recruited and randomized. Besides, the population test is a gate-keeper which prevents any inadequate findings when there is no predictive biomarker. Keeping the population test is hence necessary, but apparently leads to an important risk of overtreating the patients who do not benefit. This still happens when a large part of the type I error risk is dedicated to the subgroup-level test. In view of this problem, we consider a recalibration for the original CVASD. Simulation results show that by simply changing the orders of the two tests, one can minimize effectively the probability of recommending treatment to the overall population when it is only effective in a subset. Further, this simple calibration has a minor impact on the ability of the design to correctly specify the absence of the predictive biomarker if this is the case, and hence minimize the chance of undertreating any patient subgroups.

Other concerns could be raised over the fact that CVASD includes biomarker-negative patients, which might be unethical in practice. In fact, the question of whether we need to include or not

biomarker-negative patients in targeted therapy evaluation is a complex and debated question (24,25). This depends on the confidence in the absence of effect in the biomarker-negative patients based on biological rationale, knowledge of the drug's mechanism, preclinical data, the seriousness of the disease treated (i.e. delaying approval for biomarker-positive patients is often considered as not acceptable), etc. (26). For many indications of targeted therapies (e.g., vemurafenib in melanoma), it would be unethical to include "biomarker-negative" patients (in this example, patients with BRAF-wild type tumors) in a randomized clinical trial. However, it could still be possible to include patients with BRAF-mutated tumors in a CVASD to search for one or some additional predictive biomarkers beyond BRAF. On the other hand, there are several drugs for which the relevant predictive biomarker is less straightforward, and hence several trials with different biomarkers evaluated have been conducted (12). In these cases, our key message is that conducting an all-comer design like the (modified) CVASD would be wiser and more appropriate.

This study suffers from some limitations. First, the data generating mechanism is probably oversimplifying the real-life situation. For instance, the simulated biomarkers are all binary, although in practice some markers might classify patients into more than two subgroups (e.g. low-, intermediate- or high-risk subgroup). Besides, we only evaluate in this study one fixed correlation structure among the biomarkers, whereas this can be an important factor that affects the strategies' performance. Future frameworks should therefore address these aspects to develop insight into how different strategies behave in more complicated settings. Second, the performance of the subgroup identification algorithm in CVASD might be suboptimal, due to the fact that the best set of tuning parameters for each development cohort in the main cross validation is not chosen by the leave-one-out cross-validation method recommended by Freidlin

et al (21). In the context of a simulation study, this approach prolongs considerably the overall simulation time and hence, becomes practically infeasible. Our approach is to choose for each sub-setting only one set of parameters that can maximize the empirical power of the algorithm. This set is chosen via an extra simulation of 2000 runs (Appendix 2). Such an approach might be less effective but it limits the simulation time in an acceptable duration. Finally, this paper only deals with the clinical utility of the potential predictive biomarkers, assuming that the other dimensions of the biomarkers' evidence (i.e. the analytic and clinical validity of the test, the ethical, legal and social implications of the use of the biomarkers (27)) are fulfilled. This assumption may not always be the case in practice.

Several propositions could also be considered to further improve the modified CVASD. First, a large variety of methods to identify sensitive patients have been recently suggested, such as the SIDES algorithm (28,29) or other approaches for individualized treatment rules (30–33). These methods should be evaluated to ascertain whether they can help to further increase the modified CVASD performance. Second, this study only focused on randomized trials and compared different design strategies that involve treatment randomization. Further simulation studies should also be conducted to evaluate whether the modified CVASD can assist in the situation where only observational data (i.e. no treatment randomization) is available. Third, one can also think about the application of the cross-validation approach in the context of multistate adaptive enrichment design. In such a design, an intermediate analysis takes place based on first-stage subjects to decide whether the second stage should be enriched on a biomarker (34). This biomarker needs to be pre-specified at the beginning of the trial. If several biomarkers are proposed as in our context, the CVASD can be nested in the first stage and one biomarker that forms the sensitive subset is chosen for the second stage. However, the type I error in such a

design is controlled by using the closure principle rather than splitting the significance level as in the original CVASD (35,36).

CONCLUSION

When several biomarkers are proposed for a new targeted therapy, the current approach of evaluating a drug in a series of independent biomarker-enriched trials can yield a high risk of false-positive findings. CVASD with an appropriate split of type I error risk and a simple recalibration is a good alternative to overcome the problem of multiplicity in several settings.

ACKNOWLEDGEMENT

Tat-Thang Vo was supported by the funding from *Conseil Régional, Île-de-France* (Île-de-France Regional Council, Paris, France) within the program *Bourse Master « Île-de-France »* for the 2014/2015 period. The sponsor has no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

We would like to thank the three “anonymous” reviewers for their insightful comments on an earlier version of the manuscript. Besides, our sincere thanks to Clément Gauvain, Thomas Davergne, Tania Martin, Alice Biggane, Linda Nyanchoka, Justine Jacot, and Thu Van Nguyen for their outstanding emotional support and their diligent English proofreading of this paper.

REFERENCES

1. Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Pers Med. 2010;7:33–47.

2. Simon R. The Use of Genomics in Clinical Trial Design. Clin Cancer Res. 2008;14:5984–93.

3. Simon N. Adaptive enrichment designs: applications and challenges. Clin Investig. 2015;5:383–91.

4. Mandrekar S, Dahlberg SE, Richard S. Improving Clinical Trial Efficiency: Thinking outside the Box. 2015 ASCO Educ Book [Internet]. 2015; Available from: https://meetinglibrary.asco.org/record/104032/edbook

5. Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. Nat Rev Clin Oncol. 2014;11:81–90.

6. Temple R. Enrichment of clinical study populations. Clin Pharmacol Ther. 2010;88:774–8.

7. US Food and Drug Administration. Guidance for Industry. Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products [Internet]. 2012 [cited 2016 Mar 8]. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf

8. Simon R. Biomarker based clinical trial design. Chin Clin Oncol. 2014;3:39.

9. Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: a systematic review and classification. Clin Cancer Res Off J Am Assoc Cancer Res. 2013;19:4578–88.

10. Mandrekar SJ, Richard S. Improving Clinical Trial Efficiency: Thinking outside the Box. J Clin Oncol [Internet]. Available from: http://meetinglibrary.asco.org/content/11500141-156

11. Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: design issues. J Natl Cancer Inst. 2010;102:152–60.

12. Vivot A, Li J, Zeitoun J-D, Mourah S, Crequit P, Ravaud P, et al. Pharmacogenomic biomarkers as inclusion criteria in clinical trials of oncology-targeted drugs: a mapping of ClinicalTrials.gov. Genet Med Off J Am Coll Med Genet. 2016;18:796–805.

13. Menis J, Hasan B, Besse B. New clinical research strategies in thoracic oncology: clinical trial design, adaptive, basket and umbrella trials, new end-points and new evaluations of response. Eur Respir Rev. 2014;23:367–78.

14. Simon R. Genomic Alteration-Driven Clinical Trial Designs in Oncology. Ann Intern Med. 2016;165:270–8.

15. Mullard A. NCI-MATCH trial pushes cancer umbrella trial paradigm. Nat Rev Drug Discov. 2015;14:513–5.

16. West HJ. Novel Precision Medicine Trial Designs: Umbrellas and Baskets. JAMA Oncol. 2017;3:423.

17. Downing NS, Krumholz HM, Ross JS, Shah ND. Regulatory watch: Characterizing the US FDA's approach to promoting transformative innovation [Internet]. Nat. Rev. Drug Discov. 2015 [cited 2018 May 23]. Available from: https://www.nature.com/articles/nrd4734

18. Kesselheim AS, Myers JA, Avorn J. Characteristics of clinical trials to support approval of orphan vs nonorphan drugs for cancer. JAMA. 2011;305:2320–6.

19. Downing NS, Zhang AD, Ross JS. Regulatory Review of New Therapeutic Agents - FDA versus EMA, 2011-2015. N Engl J Med. 2017;376:1386–7.

20. Vivot A, Boutron I, Béraud-Chaulet G, Zeitoun J-D, Ravaud P, Porcher R. Evidence for Treatment-by-Biomarker interaction for FDA-approved Oncology Drugs with Required Pharmacogenomic Biomarker Testing. Sci Rep. 2017;7:6882.

21. Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design. Clin Cancer Res Off J Am Assoc Cancer Res. 2010;16:691–8.

22. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res Off J Am Assoc Cancer Res. 2005;11:7872–8.

23. Vivot A, Boutron I, Ravaud P, Porcher R. Guidance for pharmacogenomic biomarker testing in labels of FDA-approved drugs. Genet Med. 2015;17:733–8.

24. Deverka P, Messner DA, McCormack R, Lyman GH, Piper M, Bradley L, et al. Generating and evaluating evidence of the clinical utility of molecular diagnostic tests in oncology. Genet Med. 2016;18:780–7.

25. Pletcher MJ, McCulloch CE. The Challenges of Generating Evidence to Support Precision Medicine. JAMA Intern Med. 2017;177:561–2.

26. European Medicines Agency. Reflection paper on methodological issues associated with pharmacogenomic biomarkers in relation to clinical development and patient selection. 2011;21.

27. Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group. The EGAPP initiative: lessons learned. Genet Med. 2014;16:217–24.

28. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search--a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med. 2011;30:2601–21.

29. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. Stat Med. 2011;30:2867–80.

30. Shen J, Wang L, Daignault S, Spratt DE, Morgan TM, Taylor JMG. Estimating the Optimal Personalized Treatment Strategy Based on Selected Variables to Prolong Survival via Random Survival Forest with Weighted Bootstrap. J Biopharm Stat. 2018;28:362–81.

31. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012;68:1010–8.

32. Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR. Doubly Robust Learning for Estimating Individualized Treatment with Censored Data. Biometrika. 2015;102:151–68.

33. Zhou X, Mayer-Hamblett N, Khan U, Kosorok MR. Residual Weighted Learning for Estimating Individualized Treatment Rules. J Am Stat Assoc. 2017;112:169–87.

34. Stallard N, Hamborg T, Parsons N, Friede T. Adaptive designs for confirmatory clinical trials with subgroup selection. J Biopharm Stat. 2014;24:168–87.

35. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. Pharm Stat. 2011;10:347–56.

36. Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. Stat Med. 2012;31:4309–20.

TABLES

Table 1 – Definition of different settings and sub-settings

Setting 1 – No predictive biomarker (pre-BMK)	
1.1	No treatment effect in the broad population.
1.2	The targeted agent applies equally to all patients. The treatment effect is weak.
Setting 2 – One pre-BMK (Z_1) & no treatment effect for non-sensitive patients	
2.1	The pre-BMK (Z_1) has a high predictive value and a positive proportion of 25% in the population.
2.2	The pre-BMK (Z_1) has a moderate predictive value and a positive proportion of 35% in the population.
2.3	The pre-BMK (Z_1) has a low predictive value and a positive proportion of 50% in the population.
Setting 3 – Two pre-BMKs (Z_1 and Z_2) & no treatment effect for non-sensitive patients	
3.1	Both pre-BMKs (Z_1 and Z_2) have a moderate predictive value and a positive proportion of 25% in the population.
3.2	One pre-BMK (Z_1) has a low predictive value and the other (Z_2) has a moderate predictive value. The positive proportion in the population is 25% and 35%, respectively.

BMK: biomarker, pre-BMK: predictive biomarker

Table 2 – Parameter setup for the outcome generating mechanism (1) in the different sub-settings

and are the positive proportions of the two biomarkers Z_1 and Z_2, respectively. The proportion of patients being positive to each non-sensitive biomarker is fixed at 30% in all sub-settings.

Sub-setting				
1.1	1	1	1	0.85
		0.3	0.3	
1.2	0.65	1	1	0.85
		0.3	0.3	
2.1	1	0.35	1	0.85
		0.25	0.3	
2.2	1	0.5	1	0.85
		0.35	0.3	
2.3	1	0.65	1	0.85
		0.5	0.3	
3.1	1	0.5	0.5	0.85
		0.25	0.25	
3.2	1	0.5	0.65	0.85
		0.25	0.35	

Figures Legends

Figure 1

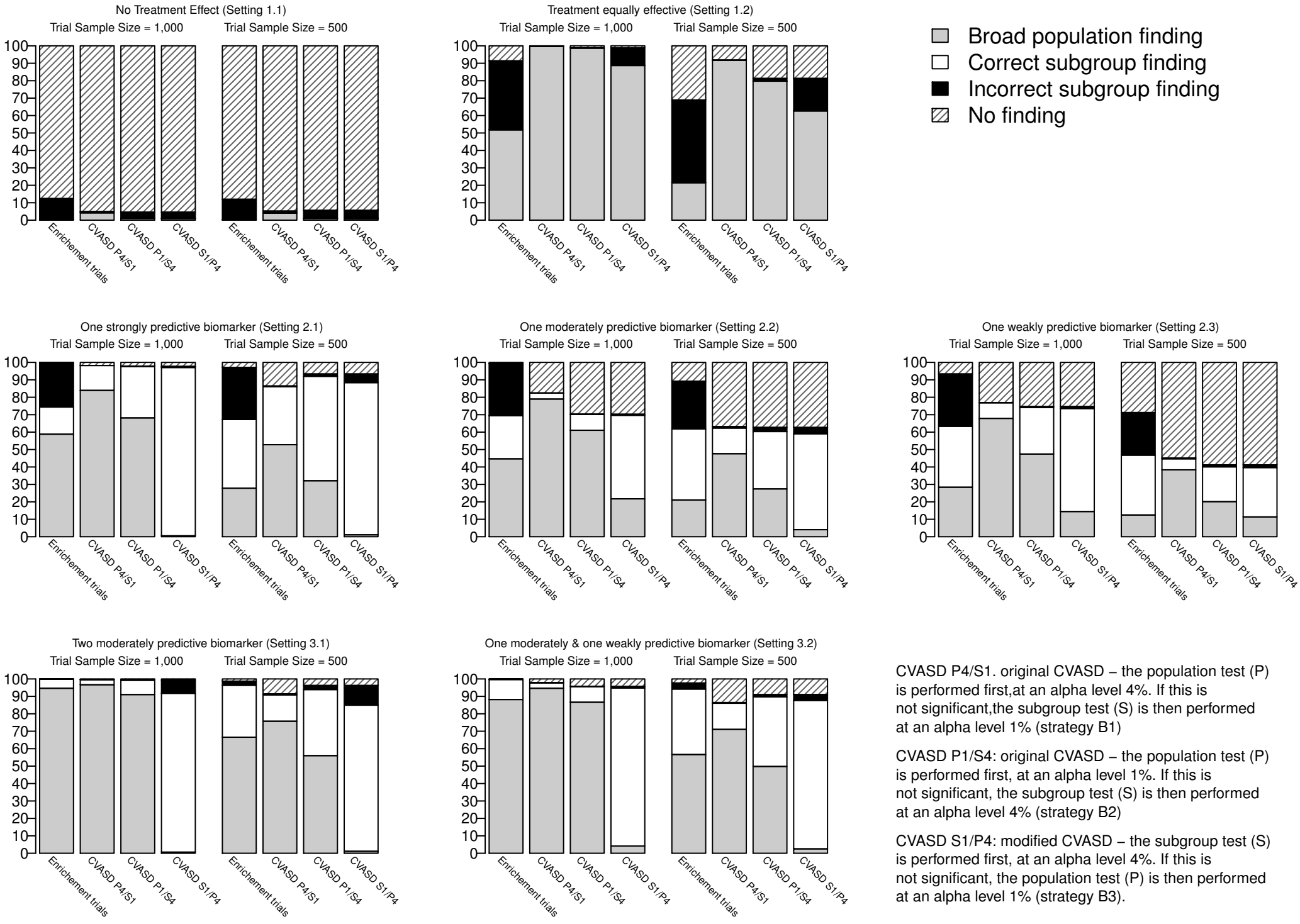
Overall comparison of four strategies: **A** (series of enrichment trials), **B1** (CVASD P4/S1 – original CVASD favoring the population test P), **B2** (CVASD P1/S4 – original CVASD favoring the subgroup test S) and **B3** (CVASD S4/P1 – modified CVASD: the subgroup test S is performed before the population test P) across three settings: 1 (no sensitive subgroup), 2 (one predictive biomarker) and 3 (two predictive biomarkers) in scenario 1 (three biomarker candidates). Note that the correct findings are represented in stripe for sub-setting 1.1, in grey for sub-setting 1.2 and in white for setting 2 and 3.

Abbreviations: CVASD, cross-validated adaptive signature design;

Figure 2

Overall comparison of four strategies: **A** (series of enrichment trials), **B1** (CVASD P4/S1 – original CVASD favoring the population test P), **B2** (CVASD P1/S4 – original CVASD favoring the subgroup test S) and **B3** (CVASD S4/P1 – modified CVASD: the subgroup test S is performed before the population test P) across three settings: 1 (no sensitive subgroup), 2 (one predictive biomarker) and 3 (two predictive biomarkers) in scenario 2 (six biomarker candidates). Note that the correct findings are represented in stripe for sub-setting 1.1, in grey for sub-setting 1.2 and in white for setting 2 and 3.

Abbreviations: CVASD, cross-validated adaptive signature design;

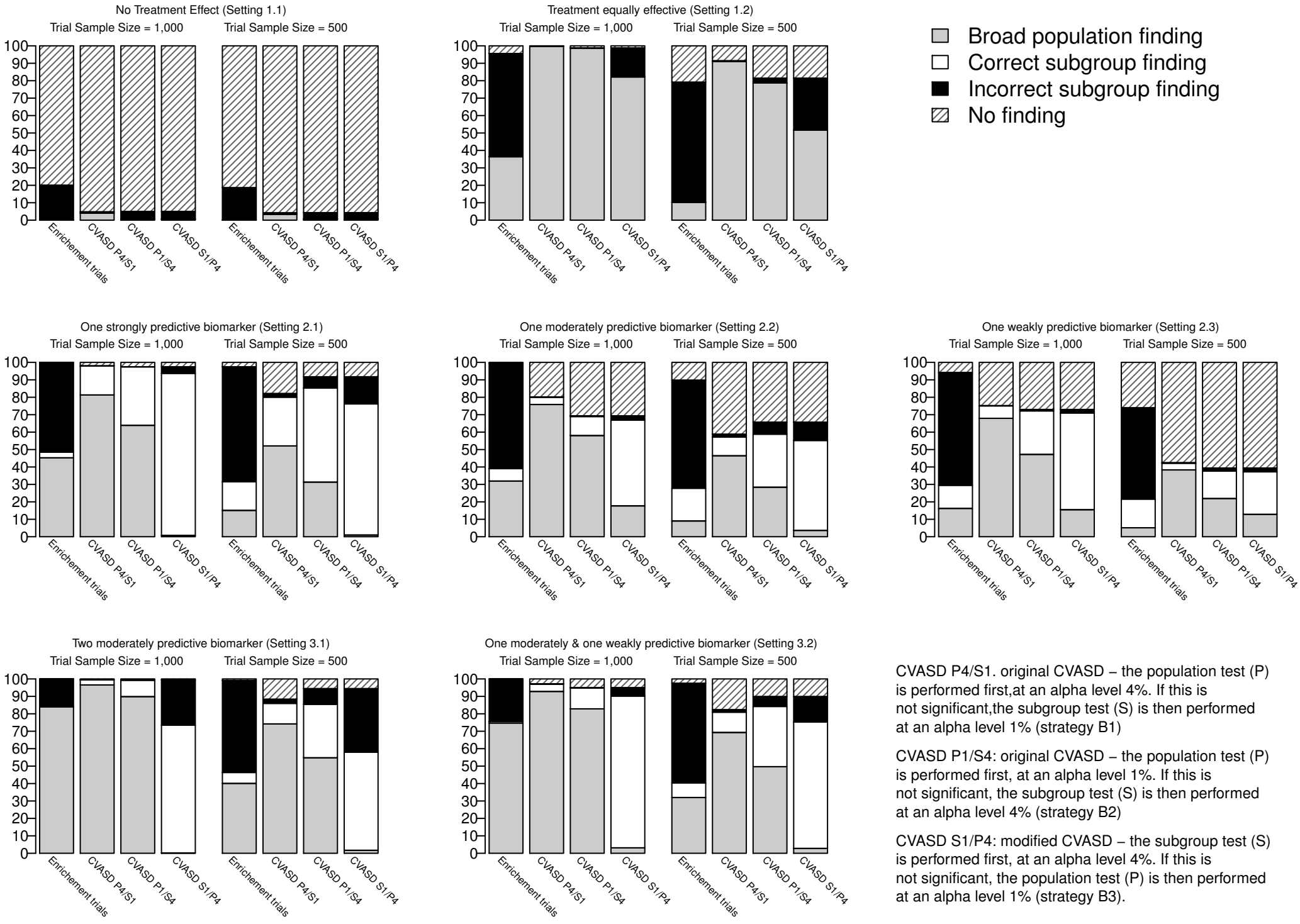


CVASD P4/S1. original CVASD – the population test (P) is performed first,at an alpha level 4%. If this is not significant,the subgroup test (S) is then performed at an alpha level 1% (strategy B1)

CVASD P1/S4: original CVASD – the population test (P) is performed first, at an alpha level 1%. If this is not significant, the subgroup test (S) is then performed at an alpha level 4% (strategy B2)

CVASD S1/P4: modified CVASD – the subgroup test (S) is performed first, at an alpha level 4%. If this is not significant, the population test (P) is then performed at an alpha level 1% (strategy B3).

Figure 1



CVASD P4/S1. original CVASD – the population test (P) is performed first,at an alpha level 4%. If this is not significant,the subgroup test (S) is then performed at an alpha level 1% (strategy B1)

CVASD P1/S4: original CVASD – the population test (P) is performed first, at an alpha level 1%. If this is not significant, the subgroup test (S) is then performed at an alpha level 4% (strategy B2)

CVASD S1/P4: modified CVASD – the subgroup test (S) is performed first, at an alpha level 4%. If this is not significant, the population test (P) is then performed at an alpha level 1% (strategy B3).

Figure 2