Published in final edited form as:

Stat Med. 2018 February 28; 37(5): 687–709. doi:10.1002/sim.7562.

Relative Efficiency of Precision Medicine Designs for Clinical Trials with Predictive Biomarkers

Weichung Joe Shih* and Yong Lin

Department of Biostatistics, School of Public Health, Rutgers, The State University of New Jersey, 683 Hoes Lane West, Piscataway, NJ 08854, USA and Division of Biometrics, Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08901, USA

Abstract

Prospective randomized clinical trials addressing biomarkers are time-consuming and costly, but are necessary for regulatory agencies to approve new therapies with predictive biomarkers. For this reason, recently there have been many discussions and proposals of various trial designs and comparisons of their efficiency in the literature. We compare statistical efficiencies between the marker-stratified design and the marker-based precision medicine design regarding testing/estimating four hypotheses/parameters of clinical interest, namely: treatment effects in each marker positive and negative cohorts, marker-by-treatment interaction, and the marker's clinical utility. As may be expected, the stratified design is more efficient than the precision medicine design. However, it is perhaps surprising to find out how low the relative efficiency can be for the precision medicine design. We quantify the relative efficiency as a function of design factors including the marker-positive prevalence rate, marker assay and classification sensitivity and specificity, and the treatment randomization ratio. It is interesting to examine the trends of the relative efficiency with these design parameters in testing different hypotheses. We advocate to use the stratified design over the precision medicine design in clinical trials with predictive biomarkers.

Keywords

Efficiency; Biomarker; Clinical trial; Precision medicine; Stratification; Clinical utility

1. Introduction

With the rapid development and use of biomarkers for personalized medicine, it is becoming an increasing trend to consider biomarkers in the study design and analysis for modern clinical trials. Sargent and Allegra,¹ Freidlin et al² and Hayes,³ among others, discussed design issues in clinical trials with biomarkers. Prospective randomized clinical trials addressing biomarkers are time-consuming and costly, but are necessary for regulatory agencies to approve new therapies with predictive biomarkers.³ For this reason, recently

*Correspondence to: Weichung Joe Shih, Department of Biostatistics, School of Public Health, Rutgers, The State University of New Jersey, 683 Hoes Lane West, Room 218, Piscataway, NJ 08854, USA.

there have been many discussions and proposals of various trial designs and comparisons of their efficiency in the literature. Considering a biomarker or a panel of biomarkers that classifies patients as marker positives or negatives, Simon and Maitournam⁴ and Maitournam and Simon⁵ proposed “targeted (enrichment or selected) designs”, which restrict the accrual to only marker positive patients, as opposed to the usual all-comers “untargeted designs”. Sargent et al⁶ proposed the “biomarker-by-treatment interaction design”, which is the “marker-stratified design” that prospectively randomizes patients within each marker-status defined stratum. Stratified randomization designs are classical designs with orthogonality condition in the sense of a two-way ANOVA and are used widely in many clinical trials. When the design stratification is based on biomarker status, however, we need to consider possible misclassifications due to the imperfect assay and classification rule.^{3,7} An example of such a marker-stratified design is the Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010) trial.⁸ The stratification biomarker was TPS (tumour proportion score 50% vs 1 - 49%), which measures the extent of PD-L1 expression. A diagram of the marker-stratified design is presented in Section 3.1.

Considering the case of two treatment groups, test (T) versus control/standard of care (C), Sargent and Allegra,¹ and Mandrekar and Sargent^{9–11} proposed a new family of “modified biomarker-based-strategy designs”, which extended the “usual biomarker-based-strategy design” originally proposed by Hayes.¹² The “modified” version is also called “augmented” marker-based strategy designs.⁷ Young et al¹³ labeled the “usual” and “modified” versions as “biomarker-based strategy design I and II”, respectively. Since all the biomarker clinical trial designs, even for the targeted/enrichment designs, involve some kind of strategy based on biomarkers, Shih and Lin¹⁴ recently unified these multifarious terms and renamed the family of “biomarker-based-strategy designs” as the “precision medicine designs” for better clarity, because this type of designs appeared mostly in oncology journals studying precision medicines. We follow the nomenclature of Shih and Lin¹⁴ in this paper.

The precision medicine design first randomizes patients into a marker-dependent arm and a marker-independent arm; see the diagram in Section 3.2. For the marker-dependent arm, marker-positive patients all receive the test treatment and marker-negative patients all receive the control treatment. For the marker-independent arm, with the “usual” strategy, patients always receive the standard of care C . With the “modified/augmented” strategy, patients follow the untargeted design, i.e., they are further randomized to either treatment group T or C without biomarker information. Sargent and Allegra¹ discussed several advantages of the “modified” strategy over the “usual” strategy. Mandrekar and Sargent¹¹ reported a trial by Potti and Nevins¹⁵ that adopted a modified biomarker-based strategy. Patients were first randomized to the genomics guided arm (marker-dependent) or the unguided arm (marker-independent). In that trial, patients in the marker-independent arm were further randomized to either adriamycin cyclophosphamide or docetaxel cyclophosphamide. These and other designs and generalizations to multiple markers and treatment groups have been reviewed recently by Renfro et al¹⁶ with examples. See more descriptions in later sections.

In consideration of efficiency of these biomarker trial designs,^{4–7,9–11,13} most literature performed relative efficiency either ambiguously without referencing to the hypothesis of interest, or inappropriately with tests for different hypotheses or estimators of different parameters. For example, Maitournam and Simom⁴ and Simom and Maitournam⁵ reported that targeted (or enrichment) designs are more efficient than untargeted (or all-comers) designs. Obviously, this is not an appropriate comparison in efficiency, since the treatment effect in a targeted design is limited to the marker-positive cohort only, while for the untargeted design, the treatment effect refers to the overall unselected population. Comparison of efficiency has to be based on testing the same hypothesis or estimating the same parameter of interest. Moreover, although Maitournam and Simom^{9–11} suggested that the stratified design has greater efficiency than the precision medicine design, their suggestion was based on the sum of the sample sizes required to test the hypotheses of treatment effects in marker-positive and marker-negative cohorts for the stratified design, against the sample size required to compare the marker-independent arm versus the marker-dependent arm for the precision medicine design. These are not compatible hypotheses for a meaningful comparison of efficiency. The relative efficiency issues are complex and not well understood, since there are many factors involved and needed analytic considerations. In this paper, the issue of relative efficiency of the stratified design versus the precision medicine design is analyzed with explicit formulas. We choose these two families of designs because they have been considered in almost all the recent literature in biomarker trial designs. Moreover, Shih and Lin¹⁴ showed that, in their different structures with different emphases, both stratified designs and precision medicine designs are capable in estimating/testing the same set of parameters/hypotheses of interest, including the treatment effects in each marker-defined stratum, marker-by-treatment interaction, and the marker's clinical utility. Therefore, we may (and should) compare the relative efficiency between the stratified design and the precision medicine design with respect to these hypotheses or associated parameters. In this paper, we carry out the efficiency comparisons in the context of mean difference of a continuous endpoint Y for hypotheses or parameters of special clinical interest. Conclusions should be applicable to binary and survival endpoints as well.

The rest of the paper is structured as follows. In Section 2, we layout notation for the design parameters and hypotheses of interest. Sections 3, 4 and 5 address the four hypotheses/parameters of interest: Section 3 for the treatment effects in the marker-status specific sub-cohorts; Section 4 for the biomarker-by-treatment interaction; Section 5 for the marker's clinical utility. The sub-sections are devoted to the two designs: first the stratified design, then the precision medicine design, and subsequently the result of relative efficiency. Section 6 illustrates the methods with two real clinical trials. Section 7 provides summary and more discussions. Detailed derivations and considerations on the robustness issue are given in the appendices.

2. Parameters and Hypotheses

Let D be the true marker status, M be the marker-appeared status, and A be the treatment assignment. Let $\mu_{ij} = E(Y|A = i, D = j)$ and $\sigma_{ij}^2 = \text{Var}(Y|A = i, D = j)$ denote the mean and variance of the response of treatment group i ($i = T$ or C) for patients in marker cohort j ($j =$

0 or 1; $j=0$ being marker-negative and $j=1$ being marker-positive). The marker assay and classification accuracy are measured by sensitivity $\lambda_{sen} \equiv P(M=1|D=1)$ and specificity $\lambda_{spec} \equiv P(M=0|D=0)$. The prevalence rate of true marker-positive is $p \equiv P(D=1)$. The randomization ratio is $r \equiv P(A=T)$ to the test treatment and $1-r$ to the control group for the stratified design in both marker-defined strata. The corresponding randomization ratio is denoted by f : $(1-f)$ for the marker-independent arm of the precision medicine design in general, where treatment randomization is conducted without marker information. We will justify later that it is reasonable to consider $f=r$ for the purpose of assessing relative efficiency of the two designs when $f=0$. Notice that, the precision medicine design is the rename of the “modified/augmented” marker-based strategy design, which includes the so-called “usual” marker-based strategy design as a special case with $f=0$. We now define the parameters and label the corresponding null hypotheses as follows.

The treatment effect for marker-negative stratum is $\delta_0 = \mu_{T0} - \mu_{C0}$, the corresponding null hypothesis is $H_- : \delta_0 = 0$. The treatment effect for marker-positive stratum is $\delta_1 = \mu_{T1} - \mu_{C1}$, the corresponding null hypothesis is $H_+ : \delta_1 = 0$. The marker-by-treatment interaction is $\theta \equiv (\mu_{T1} - \mu_{C1}) - (\mu_{T0} - \mu_{C0})$, the corresponding null hypothesis is $H_I : \theta = 0$. The notion of clinical utility of the biomarker is defined by Hayes³ as “whether the use of the biomarker supports a treatment decision that produces a better outcome for the patient than if the marker results were not available”. The use of biomarker for treatment decision is reflected by the marker-dependent arm. Treatment decision without biomarker is reflected by the marker-independent arm. Thus, the marker's clinical utility is derived directly from the precision medicine design by comparing the mean of marker-dependent arm, μ_m^{DP} , with the mean of the marker independent arm, μ^{IN} , as shown in Shih and Lin¹⁴:

$$U_f(\lambda_{sen}, \lambda_{spec}, p) \equiv \mu_m^{DP} - \mu^{IN} = p(\lambda_{sen} - f)(\mu_{T1} - \mu_{C1}) + (1-p)(1 - \lambda_{spec} - f)(\mu_{T0} - \mu_{C0}).$$

Also see Section 5 later. The corresponding hypothesis is $H_U : U_f(\lambda_{sen}, \lambda_{spec}) = 0$. Shih and Lin¹⁴ highlighted several differences between the marker's clinical utility U_f and the marker-by-treatment interaction θ . In the following, we test each of hypotheses H_- to H_U by the stratified design and the precision medicine design and compare their efficiencies. We quantify the relative efficiencies as functions of design factors including the prevalence rate, sensitivity, specificity, and the treatment randomization ratio, and examine the trends with wide ranges of different scenarios for these design factors.

3. Test Marker-Specific Treatment Effect H_- and H_+ and Relative Efficiency

In the following, we consider the stratified design and the precision medicine design for testing the marker-specific treatment effect hypotheses H_- and H_+ . For both designs, due to the imperfection in assay and classification, the marker-appeared (positive and negative) means are mixtures of the true marker-positive and marker-negative means, as seen in the following expressions (1) and (3). The mixture fractions are the positive predictive and negative predictive values, respectively.

Denote $q \equiv P(M=1) = p\lambda_{sen} + (1-p)(1-\lambda_{spec})$. The positive predictive value (PPV) = $P(D=1|M=1)$ is

$$\tau \equiv \frac{p\lambda_{sen}}{q}.$$

By simple algebra, the mean of marker-appeared positives is

$$\mu_{i*} \equiv E(Y|A=i, M=1) = \tau\mu_{i1} + (1-\tau)\mu_{i0} \quad (1)$$

and the variance is

$$\begin{aligned} \sigma_{i*}^2 \equiv \text{Var}(Y|A=i, M=1) &= \tau(\mu_{i1}^2 + \sigma_{i1}^2) + (1-\tau)(\mu_{i0}^2 + \sigma_{i0}^2) - \mu_{i*}^2 = \tau(1-\tau)(\mu_{i1} - \mu_{i0})^2 \\ &+ \tau\sigma_{i1}^2 + (1-\tau)\sigma_{i0}^2 = \tau(1-\tau)\Delta_i^2 + \tau\sigma_{i1}^2 + (1-\tau)\sigma_{i0}^2 \end{aligned} \quad (2)$$

for $i = T, C$. Note that τ is marker effect in the treatment group T (C).

The negative predictive value (NPV) = $P(D=0|M=0)$ is

$$\eta \equiv \frac{(1-p)\lambda_{spec}}{1-q}.$$

The corresponding mean and variance for the marker-appeared negatives are

$$\mu_{i\phi} \equiv E(Y|A=i, M=0) = (1-\eta)\mu_{i1} + \eta\mu_{i0} \quad (3)$$

and

$$\begin{aligned} \sigma_{i\phi}^2 \equiv \text{Var}(Y|A=i, M=0) &= \eta(\mu_{i0}^2 + \sigma_{i0}^2) + (1-\eta)(\mu_{i1}^2 + \sigma_{i1}^2) - \mu_{i\phi}^2 = \eta(1-\eta)(\mu_{i1} - \mu_{i0})^2 \\ &+ \eta\sigma_{i1}^2 + (1-\eta)\sigma_{i0}^2 = \eta(1-\eta)\Delta_i^2 + \eta\sigma_{i1}^2 + (1-\eta)\sigma_{i0}^2 \end{aligned} \quad (4)$$

for $i = T, C$.

For testing the hypotheses regarding the treatment effects, treatment-by-marker interaction, and marker's clinical utility, we will need to solve Eqs. (1) and (3) for the true means. Notice that in Eqs. (2) and (4), the variances all involve the same quantities $(\Delta_i^2, \sigma_{i1}^2, \sigma_{i0}^2)$, showing that the way they are dependent on treatment groups is the same for both marker cohorts. Moreover, under the conventional assumption of homogeneous variances $\sigma_{i1}^2 = \sigma_{i0}^2 (= \sigma_i^2)$, we can see that in Eqs. (2) and (4), σ_i^2 dominates the magnitude since the first term is at most $\Delta_i^2/4$, which is much smaller.

For both designs, assume the total sample size of N to start the process.

3.1. Stratified Design

A summary of the stratified design is shown in the following layout. We annotate the essential quantities defined previously in the diagram for easy references when reading the formula derivations in the text. Notice that the randomization ratio $r: (1-r)$ to treatment groups is best set according to the variance ratio of the treatment groups to maximize efficiency within each marker cohort. In practice r is usually the same for both marker cohorts. This is justified because, as seen from Eqs. (2) and (4), the way the variances are dependent on treatment groups is the same for both marker cohorts.

With the stratified design, the sample means $\bar{Y}_{ij} = \frac{1}{N_{ij}} \sum_{k: A_k = i, M_k = j} Y_k$ are unbiased estimators of $E(Y|A = i, M = j)$ ($\equiv \mu_{i*}$ for $M = 1$ and $\equiv \mu_{i\phi}$ for $M = 0$). The sample variances $s_{ij}^2 = \frac{1}{N_{ij}-1} \sum_{k: A_k = i, M_k = j} (Y_k - \bar{Y}_{ij})^2$ are unbiased estimators of $Var(Y|A = i, M = j)$ ($\equiv \sigma_{i*}^2$ for $M = 1$ and $\equiv \sigma_{i\phi}^2$ for $M = 0$), where N_{ij} are the sample sizes for the set with $(A = i, M = j)$. Note that $N_{Tj} = rN_j$ where $N_1 \sim \text{Binomial}(N, q)$ is the (random) sample size of the marker-appeared positive stratum. $N_0 = N - N_1$.

Solving from Eqs. (1) and (3), the unbiased estimators of the true means of the marker positive cohort, μ_{i0} , are

$$\hat{\mu}_{i1} = \frac{-\eta}{1-\tau-\eta} \bar{Y}_{i1} + \frac{1-\tau}{1-\tau-\eta} \bar{Y}_{i0}, i = T, C.$$

Hence, the unbiased estimator of the treatment effect in the marker positive cohort is

$$\hat{\delta}_1 = \hat{\mu}_{T1} - \hat{\mu}_{C1} = \frac{-\eta(\bar{Y}_{T1} - \bar{Y}_{C1}) + (1 - \tau)(\bar{Y}_{T0} - \bar{Y}_{C0})}{1 - \tau - \eta} \quad (5)$$

It is shown in Appendix A that the variance of the unbiased estimator is

$$\begin{aligned} \text{Var}(\hat{\delta}_1) &= \text{Var}(\hat{\mu}_{T1}) + \text{Var}(\hat{\mu}_{C1}) \approx \frac{\frac{\eta^2 \sigma_{T*}^2}{rqN} + \frac{(1-r)^2 \sigma_{T\phi}^2}{r(1-q)N} + \frac{\eta^2 \sigma_{C*}^2}{(1-r)qN} + \frac{(1-\tau)^2 \sigma_{C\phi}^2}{(1-r)(1-q)N}}{(1-\tau-\eta)^2} \text{ or} \quad (6) \\ &\approx \frac{\frac{\eta^2 \sigma_{T*}^2}{rN_1} + \frac{(1-r)^2 \sigma_{T\phi}^2}{rN_0} + \frac{\eta^2 \sigma_{C*}^2}{(1-r)N_1} + \frac{(1-\tau)^2 \sigma_{C\phi}^2}{(1-r)N_0}}{(1-\tau-\eta)^2}, \end{aligned}$$

and this variance, with either approximation, can be estimated by

$$\begin{aligned} \widehat{\text{Var}}(\hat{\delta}_1) &= \widehat{\text{Var}}(\hat{\mu}_{T1}) + \widehat{\text{Var}}(\hat{\mu}_{C1}) = \frac{\frac{\eta^2 s_{T1}^2}{N_{T1}} + \frac{(1-r)^2 s_{T0}^2}{N_{T0}} + \frac{\eta^2 s_{C1}^2}{N_{C1}} + \frac{(1-\tau)^2 s_{C0}^2}{N_{C0}}}{(1-\tau-\eta)^2} \quad (7) \\ &= \frac{\eta^2 \left(\frac{s_{T1}^2}{N_{T1}} + \frac{s_{C1}^2}{N_{C1}} \right) + (1-\tau)^2 \left(\frac{s_{T0}^2}{N_{T0}} + \frac{s_{C0}^2}{N_{C0}} \right)}{(1-\tau-\eta)^2}. \end{aligned}$$

Wald test $Z_+^{SF} = \frac{\hat{\delta}_1}{\sqrt{\widehat{\text{Var}}(\hat{\delta}_1)}} \sim N(0, 1)$ under H_+ is readily formed from Eq. (5) and Eq. (7).

For the marker negative cohort, the unbiased estimators of μ_{i0} are

$$\hat{\mu}_{i0} = \frac{1-\eta}{1-\tau-\eta} \bar{Y}_{i1} + \frac{-\tau}{1-\tau-\eta} \bar{Y}_{i0}, \text{ for } i = T, C.$$

Hence, the unbiased estimator of the treatment effect in the marker negative cohort is

$$\hat{\delta}_0 = \hat{\mu}_{T0} - \hat{\mu}_{C0} = \frac{(1-\eta)(\bar{Y}_{T1} - \bar{Y}_{C1}) - \tau(\bar{Y}_{T0} - \bar{Y}_{C0})}{1-\tau-\eta}. \quad (8)$$

It is shown in Appendix A.3 that

$$\begin{aligned}
 \text{Var}(\hat{\delta}_0) &= \text{Var}(\hat{\mu}_{T0}) + \text{Var}(\hat{\mu}_{C0}) \approx \frac{\frac{(1-\eta)^2 \sigma_T^2}{rqN} + \frac{r^2 \sigma_T^2 \phi}{r(1-q)N} + \frac{(1-\eta)^2 \sigma_C^2}{(1-r)qN} + \frac{\tau^2 \sigma_C^2 \phi}{(1-r)(1-q)N}}{(1-\tau-\eta)^2} \text{ or (9)} \\
 &\approx \frac{\frac{(1-\eta)^2 \sigma_T^2}{rN_1} + \frac{r^2 \sigma_T^2 \phi}{rN_0} + \frac{(1-\eta)^2 \sigma_C^2}{(1-r)N_1} + \frac{\tau^2 \sigma_C^2 \phi}{(1-r)N_0}}{(1-\tau-\eta)^2}
 \end{aligned}$$

and the variance can be estimated by

$$\begin{aligned}
 \widehat{\text{Var}}(\hat{\delta}_0) &= \widehat{\text{Var}}(\hat{\mu}_{T0}) + \widehat{\text{Var}}(\hat{\mu}_{C0}) = \frac{\frac{(1-\eta)^2 s_{T1}^2}{N_{T1}} + \frac{\tau^2 s_{T0}^2}{N_{T0}} + \frac{(1-\eta)^2 s_{C1}^2}{N_{C1}} + \frac{\tau^2 s_{C0}^2}{N_{C0}}}{(1-\tau-\eta)^2} \quad (10) \\
 &= \frac{(1-\eta)^2 \left(\frac{s_{T1}^2}{N_{T1}} + \frac{s_{C1}^2}{N_{C1}} \right) + \tau^2 \left(\frac{s_{T0}^2}{N_{T0}} + \frac{s_{C0}^2}{N_{C0}} \right)}{(1-\tau-\eta)^2}.
 \end{aligned}$$

Wald test $Z_-^{SF} = \frac{\hat{\delta}_0}{\sqrt{\widehat{\text{Var}}(\hat{\delta}_0)}} \sim N(0, 1)$ under H_- is readily formed from Eq. (8) and Eq. (10).

3.2. Precision Medicine Design

A summary of the precision medicine design is shown in the following layout. We also annotate the essential quantities in the diagram for easy references when reading the formula derivations in the text.

With the precision medicine design, patients are first randomized to two arms with $t : (1 - t)$ proportions. The marker-independent arm (sample size Nt) is the untargeted design; patients are further randomized with $f : (1 - f)$ proportions, without marker involvement, to treatment (sample size Ntf) or control (sample size $Nt(1 - f)$) groups. Notice that the randomization ratio $f : (1 - f)$ is best set according to the variance ratio of $v_i^2 (i = T, C)$ in Eq (14). The way that v_i^2 is dependent on the treatment group i in Eq (14) is the same as that in Eqs. (2) and (4) for the randomization ratio r in the stratified design for both marker cohorts. Therefore, it is

justified to set $f = r$ in the sequel. Noticed also that neither r nor f may be 0 in order to estimate the true treatment effects for both marker positive and negative cohorts. The special case of $f = 0$, i.e., the “usual” marker-based strategy, is simply not permissible.

Let Y_{ik}^{IN} be the response of a patient who is randomized to treatment i , $\bar{Y}_i^{IN}$ and $s_{i\#}^2$ be the sample mean and sample variance ($i = T, C$) in the marker-independent arm. Let Y_{ik}^{DP} be the response of a patient in the marker-dependent arm with treatment i , $\bar{Y}_i^{DP}$ and s_{i*}^2 be the sample mean and sample variance ($i = T, C$), where marker appeared-positives receive treatment T and marker appeared-negatives receive treatment C .

It is shown in Appendix B.2 that, for the marker positive cohort, the unbiased estimate of the treatment effect can be obtained by

$$\hat{\delta}_1 = \hat{\mu}_{T1} - \hat{\mu}_{C1} = \frac{(1-p)\bar{Y}_T^{DP} - (1-\tau)\bar{Y}_T^{IN}}{\tau-p} - \frac{(1-p)\bar{Y}_C^{DP} - \eta\bar{Y}_C^{IN}}{1-p-\eta}, \quad (11)$$

and its variance is (with $f = r = 0$)

$$\begin{aligned} \text{Var}(\hat{\delta}_1) &\approx \frac{\frac{(1-p)^2\omega_T^2}{N(1-t)q} + \frac{(1-r)^2v_T^2}{Ntr}}{(\tau-p)^2} + \frac{\frac{(1-p)^2\omega_C^2}{N(1-t)(1-q)} + \frac{\eta^2v_T^2}{Nt(1-r)}}{(\tau-p-\eta)^2} \text{ or } \approx \frac{\frac{(1-p)^2\omega_T^2}{N_{D1}} + \frac{(1-r)^2v_T^2}{Ntr}}{(\tau-p)^2} \\ &+ \frac{\frac{(1-p)^2\omega_C^2}{N_{D0}} + \frac{\eta^2v_T^2}{Nt(1-r)}}{(\tau-p-\eta)^2}, \end{aligned} \quad (12)$$

which, with either approximation, can be estimated by

$$\widehat{\text{Var}}(\hat{\delta}_1) = \frac{\frac{(1-p)^2s_{T*}^2}{N_{D1}} + \frac{(1-r)^2s_{T\#}^2}{Ntr}}{(\tau-p)^2} + \frac{\frac{(1-p)^2s_{C*}^2}{N_{D0}} + \frac{\eta^2s_{C\#}^2}{Nt(1-r)}}{(\tau-p-\eta)^2}, \quad (13)$$

where, for $i = T$ or C ,

$$v_i^2 = p(1-p)\Delta_i^2 + p\sigma_{i1}^2 + (1-p)\sigma_{i0}^2 \quad (14)$$

is the variance of a response in the marker-independent arm with treatment i , and

$$\omega_T^2 = \tau(1-\tau)\Delta_T^2 + \tau\sigma_{T1}^2 + (1-\tau)\sigma_{T0}^2 \quad (15)$$

and

$$\omega_C^2 = \eta(1-\eta)\Delta_C^2 + (1-\eta)\sigma_{C1}^2 + \eta\sigma_{C0}^2 \quad (16)$$

are the variances of a response in the marker-dependent arm with treatments T and C respectively.

Wald test $Z_+^{PM} = \frac{\hat{\delta}_1}{\sqrt{\widehat{Var}(\hat{\delta}_1)}} \sim N(0, 1)$ under H_+ is readily formed from Eq. (11) and Eq. (13).

Similarly, we show in Appendix B.3 for marker negative cohort that the unbiased estimate of the treatment effect is

$$\hat{\delta}_0 = \hat{\mu}_{T0} - \hat{\mu}_{C0} = \frac{-p\bar{Y}_T^{DP} + \tau\bar{Y}_T^{IN}}{\tau - p} - \frac{-p\bar{Y}_C^{DP} + (1-\eta)\bar{Y}_C^{IN}}{1 - p - \eta}, \quad (17)$$

and its variance is (with $f = r = 0$)

$$\begin{aligned} Var(\hat{\delta}_0) &\approx \frac{\frac{p^2\omega_T^2}{N(1-t)q} + \frac{\tau^2v_T^2}{Ntr}}{(\tau - p)^2} + \frac{\frac{p^2\omega_C^2}{N(1-t)(1-q)} + \frac{(1-\eta)^2v_C^2}{Nt(1-r)}}{(1 - p - \eta)^2} \text{ or } \approx \frac{\frac{p^2\omega_T^2}{N_{D1}} + \frac{\tau^2v_T^2}{Ntr}}{(\tau - p)^2} \\ &+ \frac{\frac{p^2\omega_C^2}{N_{D0}} + \frac{(1-\eta)^2v_C^2}{Nt(1-r)}}{(1 - p - \eta)^2}, \end{aligned} \quad (18)$$

which, with either approximation, can be estimated by

$$\widehat{Var}(\hat{\delta}_0) = \frac{\frac{p^2s_{T*}^2}{N_{D1}} + \frac{\tau^2s_{T\#}^2}{Ntr}}{(\tau - p)^2} + \frac{\frac{p^2s_{C*}^2}{N_{D0}} + \frac{(1-\eta)^2s_{C\#}^2}{Nt(1-r)}}{(1 - p - \eta)^2}. \quad (19)$$

Wald test $Z_-^{PM} = \frac{\hat{\delta}_0}{\sqrt{\widehat{Var}(\hat{\delta}_0)}} \sim N(0, 1)$ under H_- is readily formed from Eq. (17) and Eq. (19).

3.3. Relative Efficiency with Respect to Testing H_- , H_+

To assess the relative efficiency (RE) in testing H_- or estimating δ_0 regarding the treatment effect in the marker negative cohort, we compare the variance

$$V_0^{SF} \equiv \text{Var}(\hat{\delta}_0) \approx \frac{\frac{(1-\eta)^2 \sigma_{T*}^2}{rqN} + \frac{\tau^2 \sigma_{T\phi}^2}{r(1-q)N} + \frac{(1-\eta)^2 \sigma_{C*}^2}{(1-r)qN} + \frac{\tau^2 \sigma_{C\phi}^2}{(1-r)(1-q)N}}{(1-\tau-\eta)^2}$$

of the stratified design (c.f., Eq (9)) to the variance

$$V_0^{PM} \equiv \text{Var}(\hat{\delta}_0) \approx \frac{\frac{p^2 \omega_T^2}{N(1-t)q} + \frac{\tau^2 v_T^2}{Ntr}}{(\tau-p)^2} + \frac{\frac{p^2 \omega_C^2}{N(1-t)(1-q)} + \frac{(1-\eta)^2 v_C^2}{Nt(1-r)}}{(1-p-\eta)^2}$$

of the precision medicine design (c.f., Eq (18)).

To assess RE in testing H_+ or estimating δ_1 regarding the treatment effect in the marker positive cohort, we compare the variance

$$V_1^{SF} \equiv \text{Var}(\hat{\delta}_1) \approx \frac{\frac{\eta^2 \sigma_{T*}^2}{rqN} + \frac{(1-\tau)^2 \sigma_{T\phi}^2}{r(1-q)N} + \frac{\eta^2 \sigma_{C*}^2}{(1-r)qN} + \frac{(1-\tau)^2 \sigma_{C\phi}^2}{(1-r)(1-q)N}}{(1-\tau-\eta)^2}$$

of the stratified design (c.f., Eq (6)) to the variance

$$V_1^{PM} \equiv \text{Var}(\hat{\delta}_1) \approx \frac{\frac{(1-p)^2 \omega_T^2}{N(1-t)q} + \frac{(1-\tau)^2 v_T^2}{Ntr}}{(\tau-p)^2} + \frac{\frac{(1-p)^2 \omega_C^2}{N(1-t)(1-q)} + \frac{\eta^2 v_C^2}{Nt(1-r)}}{(1-p-\eta)^2}$$

of the precision medicine design (c.f., Eq(12)).

As shown in the formulas above, RE depends on many factors in a complicated fashion. In the following assessment we set $\sigma_{ij}^2 = 1$ for all $i = T, C$ and $j = 0, 1$. We examine the RE s with respect to a wide range of prevalence rate, proportion randomized to the marker-independent arm for the precision medicine design, sensitivity and specificity, and the marker effects. With homogeneous variances $\sigma_{ij}^2 = 1$ for all $i = T, C$, we examine RE s with $r(=f) = 1/2$.

For testing H_+ , Fig 1A depicts $RE(=V_1^{SF}/V_1^{PM})$ versus marker-positive prevalence rate in the range of 0.1 p 0.9, assuming $\lambda_{sen} = \lambda_{spec} = 0.9$ and $t = 0.5$, where t is the randomization ratio to the marker-independent arm for the precision medicine design. Fig 1B depicts RE

versus $0.1 \leq t \leq 0.9$ when $p = 0.8$. In each figure, we vary the marker effect for the treatment group T , $\tau = \mu_{T1} - \mu_{T0}$, from 0 to 0.75 while set $c = 0$. As seen, $RE < 0.5$ in all cases, indicating that the precision medicine design is much less efficient than the stratified design in testing H_+ . The precision medicine design loses efficiency because it includes patients treated with the same treatment on both marker-independent and marker-dependent arms. It is perhaps expected that this overlap leads to efficiency loss for the precision medicine design; however, the extent to which it loses efficiency is somewhat surprising. Also surprising is that, as seen in Fig 1A, RE increases with increasing marker-positive prevalence rate. Fig 1B shows that RE is quadratic versus t and reaches maximum when t is near 0.55. The marker effect τ does not affect RE much in all cases, as expected since the first term in Eqs. (2), (4) and (14) is dominated by $\sigma_{ij}^2 = 1$. Thus, in Fig 1C we pick $\tau = 0.45$ to depict RE versus $0.5 \leq \lambda_{sen} = \lambda_{spec} \leq 1$, for $p = 0.2$ to 0.8 and $t = 0.5$. Interestingly, the RE vs sensitivity trend is increasing for prevalence $p > 0.4$, but decreasing for prevalence $p < 0.4$. Thus, whether a better marker assay helps the stratified design more (or less) than the precision medicine design depends on the marker-positive prevalence rate.

For testing H_- , contrasting to testing H_+ , in Fig 1D, the trend is reversed for $RE (= V_0^{SF}/V_0^{PM})$ versus marker-positive prevalence rate p . The pattern in Fig 1E (RE versus randomization ratio t) is the same as above, but the RE magnitude is much lower. The pattern in Fig 1F (RE versus sensitivity/specificity $\lambda_{sen} = \lambda_{spec}$) is also reversed from that above; the RE vs sensitivity trend is positive for prevalence $p < 0.4$, but turns negative for prevalence $p > 0.4$.

In addition, we also show supplementary graphs in Appendix C for the robustness of the variance to design parameter values. Figures 3A to 3F indicate that, in testing H_- or H_+ , the variances of the tests are more stable/robust for the stratified design (solid lines) than the precision medicine design (dashed lines) with respect to the varying marker-positive prevalence rate (Figs 3A and 3D), marker's sensitivity (Figs 3B and 3E) and specificity (Figs 3C and 3F).

4. Test Marker-by-Treatment Interaction H_I and Relative Efficiency

The marker-by-treatment interaction is $\theta = (\mu_{T1} - \mu_{C1}) - (\mu_{T0} - \mu_{C0}) = \delta_1 - \delta_0$, the corresponding null hypothesis is $H_I: \theta = 0$. It indicates how the marker status modifies the treatment effect. Derivations follow straightforwardly from the previous section. Again, as for testing H_- and H_+ , $f = r = 0$ in testing H_I .

4.1. Stratified Design

For the stratified design, from Eqs. (5) and (8), unbiased estimator of θ can be obtained by

$$\hat{\theta}^{SF} = \hat{\delta}_1 - \hat{\delta}_0 = \frac{-\eta(\bar{Y}_{T1} - \bar{Y}_{C1}) + (1 - \tau)(\bar{Y}_{T0} - \bar{Y}_{C0})}{1 - \tau - \eta} - \frac{(1 - \eta)(\bar{Y}_{T1} - \bar{Y}_{C1}) - \tau(\bar{Y}_{T0} - \bar{Y}_{C0})}{1 - \tau - \eta} = \frac{(\bar{Y}_{T0} - \bar{Y}_{C0}) - (\bar{Y}_{T1} - \bar{Y}_{C1})}{1 - \tau - \eta}, \quad (20)$$

with its variance

$$\begin{aligned} \text{Var}(\hat{\theta}^{SF}) &\approx \frac{\frac{\sigma_{T*}^2}{rqN} + \frac{\sigma_{T\phi}^2}{r(1-q)N} + \frac{\sigma_{C*}^2}{(1-r)qN} + \frac{\sigma_{C\phi}^2}{(1-r)(1-q)N}}{(1-\tau-\eta)^2} \text{ or } \\ &\approx \frac{\frac{\sigma_{T*}^2}{rN_1} + \frac{\sigma_{T\phi}^2}{rN_0} + \frac{\sigma_{C*}^2}{(1-r)N_1} + \frac{\sigma_{C\phi}^2}{(1-r)N_0}}{(1-\tau-\eta)^2} \end{aligned} \quad (21)$$

and variance estimator

$$\widehat{\text{Var}}(\hat{\theta}^{SF}) = \frac{\frac{s_{T1}^2}{N_{T1}} + \frac{s_{T0}^2}{N_{T0}} + \frac{s_{C1}^2}{N_{C1}} + \frac{s_{C0}^2}{N_{C0}}}{(1-\tau-\eta)^2}. \quad (22)$$

Wald test $Z_I^{SF} = \frac{\hat{\theta}^{SF}}{\sqrt{\widehat{\text{Var}}(\hat{\theta}^{SF})}} \sim N(0, 1)$ under H_I is readily formed from (20) and (22).

4.2. Precision Medicine Design

For the precision medicine design, from Eqs. (11) and (17), unbiased estimator of θ can be obtained by

$$\begin{aligned} \hat{\theta}^{PM} = \hat{\delta}_1 - \hat{\delta}_0 &= \left[\frac{(1-p)\bar{Y}_T^{DP} - (1-\tau)\bar{Y}_T^{IN}}{\tau-p} - \frac{(1-p)\bar{Y}_C^{DP} - \eta\bar{Y}_C^{IN}}{1-p-\eta} \right] - \\ &\left[\frac{-p\bar{Y}_T^{DP} + \bar{Y}_C^{DP} + (1-\eta)\bar{Y}_C^{IN}}{1-p-\eta} \right] = \frac{\bar{Y}_T^{DP} - \bar{Y}_T^{IN}}{\tau-p} - \frac{\bar{Y}_C^{DP} - \bar{Y}_C^{IN}}{1-p-\eta}, \end{aligned} \quad (23)$$

with its variance

$$\begin{aligned} \text{Var}(\hat{\theta}^{PM}) &\approx \frac{\frac{\omega_T^2}{N(1-t)q} + \frac{v_T^2}{Ntr}}{(\tau-p)^2} + \frac{\frac{\omega_C^2}{N(1-t)(1-q)} + \frac{v_C^2}{Nt(1-r)}}{(1-p-\eta)^2} \text{ or } \approx \frac{\frac{\omega_T^2}{N_{D1}} + \frac{v_T^2}{N_{tr}}}{(\tau-p)^2} + \frac{\frac{\omega_C^2}{N_{D0}} + \frac{v_C^2}{N_{t(1-r)}}}{(1-p-\eta)^2} \\ &\quad (24) \end{aligned}$$

and variance estimator

$$\widehat{Var}(\hat{\theta}^{PM}) = \frac{\frac{s_{T*}^2}{N_1} + \frac{s_{T\#}^2}{N_{tr}}}{(\tau - p)^2} + \frac{\frac{s_{C*}^2}{N(1-t) - N_1} + \frac{s_{C\#}^2}{Nt(1-r)}}{(1 - p - \eta)^2}. \quad (25)$$

Wald test $Z_I^{PM} = \frac{\hat{\theta}^{PM}}{\sqrt{\widehat{Var}(\hat{\theta}^{PM})}} \sim N(0, 1)$ under H_I is readily formed from (23) and (25).

4.3. Relative Efficiency with Respect to Testing H_I

To assess the relative efficiency in testing H_I or estimating θ regarding the treatment by marker interaction, we compare $Var(\hat{\theta}^{SF})$ in Eq. (21) of the stratified design to $Var(\hat{\theta}^{PM})$ in Eq. (24) of the precision medicine design. We use the same setup for the parameters as in Section 3.3. In Fig 2A, setting $\lambda_{sen} = \lambda_{spec} = 0.9$ and $t = 0.5$, we see that the precision medicine design is only 25% or less efficient than the stratified design and that the RE trend is quadratic and reaches maximum when marker-positive prevalence rate $p = 0.5$. Similarly, Fig 2B shows that RE is quadratic versus t and reaches maximum when t is near 0.5. The marker effect τ makes minuscule difference in the RE trends. Thus, in Fig 2C we pick $\tau = 0.45$ to depict RE versus $0.5 - \lambda_{sen} = \lambda_{spec} - 1$, for $p = 0.2$ to 0.8 and $t = 0.5$. As shown, the RE trend is similar for $p = (0.2 \text{ or } 0.8)$, $(0.4 \text{ or } 0.6)$. Different from testing H_+ , a better marker assay always helps the stratified design more than helping the precision medicine design in testing H_I . In addition, we also show supplementary graphs in Appendix C for the robustness of the variance to design parameter values. Figures 4A to 4C indicate that, in testing H_I , the variances of the tests are also more stable/robust for the stratified design (solid lines) than the precision medicine design (dashed lines) with respect to the varying marker-positive prevalence rate (Fig 4A), marker's sensitivity (Fig 4B) and specificity (Fig 4C).

5. Test Marker's Clinical Utility H_U and Relative Efficiency

The notion of marker's clinical utility is not so explicit from the stratified design at the first glance, but can be derived directly from the precision medicine design by comparing the mean of marker-dependent arm, μ_m^{DP} , with the mean of the marker independent arm, μ^{IN} .

$$\begin{aligned} \mu_m^{DP} &= q\mu_{T*} + (1-q)\mu_{C\phi} = q[\tau\mu_{T1} + (1-\tau)\mu_{T0}] + (1-q)[(1-\eta)\mu_{C1} + \eta\mu_{C0}] \\ \mu^{IN} &= f\mu_{T\#} + (1-f)\mu_{C\#} = f[p\mu_{T1} + (1-p)\mu_{T0}] + (1-f)[p\mu_{C1} + (1-p)\mu_{C0}]. \end{aligned}$$

Hence,

$$U_f(\lambda_{sen}, \lambda_{spec}, p) \equiv \mu_m^{DP} - \mu^{IN} = p(\lambda_{sen} - f)(\mu_{T1} - \mu_{C1}) + (1 - p)(1 - \lambda_{spec} - f)(\mu_{T0} - \mu_{C0}) \\ = p(\lambda_{sen} - f)\delta_1 + (1 - p)(1 - \lambda_{spec} - f)\delta_0 = [p\lambda_{sen}\delta_1 + (1 - p)(1 - \lambda_{spec})\delta_0] - f[p\delta_1 + (1 - p)\delta_0].$$

(26)

The corresponding hypothesis is $H_U: U_f(\lambda_{sen}, \lambda_{spec}, p) = 0$. Shih and Lin¹⁴ discussed differences between the marker clinical utility and marker-by-treatment interaction. Both the third and the last equations above show that U_f is a weighted sum of δ_1 and δ_0 , which will be used for the stratified design to estimate the marker's clinical utility by replacing f by r (see next section for justification). The last equation further indicates an interpretation that U_f as a weighted sum of treatment effects using the biomarker to support treatment decision discounting the treatment effects when the marker results were not available and the choice of treatment was made by randomization with ratio f . Different from testing H_- , H_+ , and H_f , f may be 0 in testing H_U . When $f = 0$, marker's clinical utility is the “marker-directed treatment effect” relative to the “always control-treated effect in the absence of biomarker”. For $f \neq 0$, from the third equation above, with corresponding estimators of δ_1 and δ_0 ,

$$\hat{U}_f(\lambda_{sen}, \lambda_{spec}, p) = p(\lambda_{sen} - f)\hat{\delta}_1 + (1 - p)(1 - \lambda_{spec} - f)\hat{\delta}_0. \quad (27)$$

5.1. Stratified Design

Eq. (26) reveals that marker's clinical utility for the stratified design is a weighted sum of δ_1 and δ_0 . In the stratified design, the randomization ratio $r: (1 - r)$ is the same (and $r \neq 0$) for each marker cohort regardless of the marker status, i.e., it is marker-independent. Therefore, it matches the meaning of $f: (1 - f)$, the randomization ratio in the marker-independent arm of the precision medicine design. Therefore $r = f \neq 0$ for the stratified design in Eqs. (26) and (27), and the estimated marker's clinical utility is, from Eqs. (5) and (8):

$$\begin{aligned} \hat{U}_r(\lambda_{sen}, \lambda_{spec}, p) &= p(\lambda_{sen} - r) \left[\frac{-\eta(\bar{Y}_{T1} - \bar{Y}_{C1}) + (1 - \tau)(\bar{Y}_{T0} - \bar{Y}_{C0})}{1 - \tau - \eta} \right] + (1 - p)(1 - \lambda_{spec} - r) \\ &\quad \left[\frac{(1 - \eta)(\bar{Y}_{T1} - \bar{Y}_{C1}) - \tau(\bar{Y}_{T0} - \bar{Y}_{C0})}{1 - \tau - \eta} \right] = \frac{1}{1 - \tau - \eta} \left[-\eta p(\lambda_{sen} + \lambda_{spec} - 1) + (1 - p - \eta)(1 \right. \\ &\quad \left. - \lambda_{spec} - r) \right] (\bar{Y}_{T1} - \bar{Y}_{C1}) + \left[-\tau p(\lambda_{sen} + \lambda_{spec} - 1) + p(\lambda_{sen} - r) - \tau(1 - \lambda_{spec} - r) \right] (\bar{Y}_{T0} \\ &\quad - \bar{Y}_{C0}). \end{aligned} \quad (28)$$

with its variance

$$\begin{aligned} V_U^{SF} = \text{Var}(\hat{U}_r(\lambda_{sen}, \lambda_{spec}, p)) &= \frac{c_1 [\text{Var}(\bar{Y}_{T1}) + \text{Var}(\bar{Y}_{C1})] + c_2 [\text{Var}(\bar{Y}_{T0}) + \text{Var}(\bar{Y}_{C0})]}{(1 - \tau - \eta)^2} \quad (29) \\ &\approx \frac{c_1 \left(\frac{\sigma_{T*}^2}{rqN} + \frac{\sigma_{C*}^2}{(1-r)qN} \right) + c_2 \left(\frac{\sigma_{T\phi}^2}{r(1-q)N} + \frac{\sigma_{C\phi}^2}{(1-r)(1-q)N} \right)}{(1 - \tau - \eta)^2} \text{ or} \\ &\approx \frac{c_1 \left(\frac{\sigma_{T*}^2}{rN_1} + \frac{\sigma_{C*}^2}{(1-r)qN} \right) + c_2 \left(\frac{\sigma_{T\phi}^2}{rN_0} + \frac{\sigma_{C\phi}^2}{(1-r)N_0} \right)}{(1 - \tau - \eta)^2}, \end{aligned}$$

where

$$c_1 = \left[-\eta p(\lambda_{sen} + \lambda_{spec} - 1) + (1 - p - \eta)(1 - \lambda_{spec} - r) \right]^2$$

and

$$c_2 = \left[-\tau p(\lambda_{sen} + \lambda_{spec} - 1) + p(\lambda_{sen} - r) - \tau(1 - \lambda_{spec} - r) \right]^2.$$

5.2. Precision Medicine Design

For the marker dependent arm, we have direct unbiased estimator:

$$\hat{\mu}_m^{DP} = q\hat{\mu}_{T*} + (1-q)\hat{\mu}_{C\phi} = q\bar{Y}_T^{DP} + (1-q)\bar{Y}_C^{DP}$$

with its variance

$$\begin{aligned} \text{Var}\left(\hat{\mu}_m^{DP}\right) &= q^2 \text{Var}\left(\bar{Y}_T^{DP}\right) + (1-q)^2 \text{Var}\left(\bar{Y}_C^{DP}\right) \simeq q \frac{\omega_T^2}{N(1-t)} + (1-q) \frac{\omega_C^2}{N(1-t)} \text{ or } \simeq q^2 \frac{\omega_T^2}{N_{D1}} \\ &+ (1-q)^2 \frac{\omega_C^2}{N_{D0}}, \end{aligned}$$

(30)

where

$$\omega_T^2 = \tau(1-\tau)\Delta_T^2 + \tau\sigma_{T1}^2 + (1-\tau)\sigma_{T0}^2$$

from Eq. (2), and

$$\omega_C^2 = \eta(1-\eta)\Delta_C^2 + (1-\eta)\sigma_{C1}^2 + \eta\sigma_{C0}^2$$

from Eq. (4).

For the marker independent arm, we have direct unbiased estimator

$$\hat{\mu}^{IN} = f\hat{\mu}_{T\#} + (1-f)\hat{\mu}_{C\#} = f\bar{Y}_T^{IN} + (1-f)\bar{Y}_C^{IN}$$

with its variance

$$\text{Var}\left(\hat{\mu}^{IN}\right) = f^2 \text{Var}\left(\bar{Y}_T^{IN}\right) + (1-f)^2 \text{Var}\left(\bar{Y}_C^{IN}\right) = f \frac{v_T^2}{Nt} + (1-f) \frac{v_C^2}{Nt}, \quad (31)$$

where

$$v_T^2 = p(1-p)\Delta_T^2 + p\sigma_{T1}^2 + (1-p)\sigma_{T0}^2$$

and

$$v_C^2 = p(1-p)\Delta_C^2 + p\sigma_{C1}^2 + (1-p)\sigma_{C0}^2$$

from Eq. (B.2).

Hence, the estimated marker's clinical utility is

$$\hat{U}_f(\lambda_{sen}, \lambda_{spec}, p) = \hat{\mu}_m^{DP} - \hat{\mu}^{IN} = [q\bar{Y}_T^{DP} + (1-q)\bar{Y}_C^{DP}] - [f\bar{Y}_T^{IN} + (1-f)\bar{Y}_C^{IN}]$$

with its variance

$$V_U^{PM} = \text{Var}\left(\hat{U}_r(\lambda_{sen}, \lambda_{spec}, p)\right) = \text{Var}\left(\hat{\mu}_m^{DP}\right) + \text{Var}\left(\hat{\mu}^{IN}\right) = q\frac{\omega_T^2}{N(1-t)} + (1-q)\frac{\omega_C^2}{N(1-t)} \quad (33) \\ + f\frac{v_T^2}{Nt} + (1-f)\frac{v_C^2}{Nt}.$$

Notice that, unlike the situation of testing H_- , H_+ , and H_f , where it is necessary for $f = 0$, it is permissible in testing H_U that $f = 0$ here. Moreover, as commented previously, since the variance $v_T^2 \approx v_C^2$, V_U^{PM} above is not sensitive to f . This can be seen from the supplementary graphs provided in Appendix D (Fig 5), where in each graph (A to H) all the dash lines ($f = 0$ to 0.8 for precision medicine design) overlap.

5.3. Relative Efficiency with Respect to Testing H_U

To assess the relative efficiency in testing H_U or estimating U_f regarding the marker utility, we compare V_U^{SF} in Eq. (29) of the stratified design versus V_U^{PM} in Eq. (33) of the precision medicine design. We use the same setup for the parameters as in Sections 3.3 and 4.3, in particular, the usual homogeneous variance $\sigma_{ij}^2 = 1$ for $i = T, C; j = 0, 1$. This leads to examining RE s with the optimal $r = 0.5$ for the stratified design. Also, as commented above, V_U^{PM} is insensitive to f . Thus we set $f = 0.5$ as well. We vary other design parameters over wide ranges as before. In Fig 2D, setting $\lambda_{sen} = \lambda_{spec} = 0.9$ and $t = 0.5$, we vary the marker-positive prevalence rate. In Fig 2E, we vary t . In Fig 2F, we vary λ_{sen} .

From Fig 2D we see that the precision medicine design is 60% or less efficient than the stratified design and that the RE trend is a quadratic function of the marker-positive prevalence rate. But different from testing H_f , RE reaches minimum (25%) when marker-positive prevalence rate $p = 0.5$ in testing H_U . This finding shows two important facts: First, it verifies the assertion of Shih and Lin¹⁴ on the differences between biomarker-treatment interaction and biomarker's clinical utility; Second, contrary to the suggestion in previous

literature,^{9–11} it is not true that higher prevalence of the marker-positives always leads more loss of efficiency of the precision medicine design. The relative efficiency trend is mostly hypothesis-specific.

Fig. 2E shows that, similar to testing interaction, RE is quadratic versus t and reaches maximum (around 40%) when t is about 0.5. Again, the marker effect τ makes minuscule difference in the RE trends. Thus, in Fig 2F we pick $\tau = 0.45$ to depict RE versus 0.5 $\lambda_{sen} = \lambda_{spec} - 1$, for $p = 0.2$ to 0.8 and $t = f = 0.5$. As shown, the RE trend with respect to marker's sensitivity is similar for $p = (0.2 \text{ or } 0.8)$, $(0.4 \text{ or } 0.6)$. Different from testing H_I , a better marker assay helps the precision medicine design more than helping the stratified design in testing H_U .

In addition, we also address the question: How the two designs are likely to perform if the true parameter values are different from the assumed values. Toward this end, we examine how the variances V_U^{SF} and V_U^{PM} behave with varying design parameter values in supplementary graphs in Appendix C (Figures 4D–F). It is interesting to see that, in testing H_U , the variance of the test is more stable/robust for the precision medicine design (V_U^{PM} dashed lines) than the stratified design (V_U^{SF} solid lines) with respect to the varying marker-positive prevalence rate (Fig 4D), marker's sensitivity (Fig 4E) and specificity (Fig 4F). This is the opposite to what we see in tests for H_L , H_+ , and H_I but not overly stable or sufficiently robust – this can be seen in all Figs 4D–F that the variances of the precision medicine design are larger than the corresponding variances of the stratified design over the entire design parameter regions. This means that it is not possible that one can have a better efficiency in testing H_U for the precision medicine design due to an assumed parameter value (such as prevalence rate, sensitivity or specificity) being different from the true value or vice versa, even though V_U^{PM} is more stable/robust than V_U^{SF} . For example, looking at Fig 4E to examine V_U^{SF} and V_U^{PM} versus λ_{sen} with other design parameters as specified on the heading, the solid curve labeled (prevalence) $p = 0.8$ is obviously changing from low (≈ 0.0038 at $\lambda_{sen} = 0.6$) to high (≈ 0.006 at $\lambda_{sen} = 0.9$) for V_U^{SF} , but the corresponding dashed line is quite flat (≈ 0.0142) for V_U^{PM} . Suppose one assumes that the marker's sensitivity $\lambda_{sen} = 0.9$, but actually the true sensitivity $\lambda_{sen} = 0.6$, then the true relative efficiency $RE \approx 0.0038/0.0142 = 0.27$ rather than the assumed $RE \approx 0.006/0.0142 = 0.42$. Both are quite low. Similar observations and comments can be made with Fig 4F for RE with respect to the specificity λ_{spec} .

6. Examples

Example 1 is a phase III trial in non-small-cell lung cancer using quantitative excision repair cross-complementing 1 (ERCC1) mRNA expression as a biomarker to customize the treatment of cisplatin.¹⁸ We use this example to illustrate issues in the precision medicine design used in this trial and potential advantages of a stratified design in case it were adopted. Corresponding to the diagram of the precision medicine design in Section 3.2, this trial first randomized patients with 1:2 ratio into the marker-independent arm and marker dependent arm (i.e., $t = 1/3$). In the marker-independent arm, patients always received the

control/standard of care (SOC), docetaxel followed by cisplatin (i.e., $f=0$). In the marker-dependent arm, patients with low levels of ERCC1 mRNA expression were also treated with the SOC, while the high genotypic group always received the test treatment, docetaxel followed by gemcitabine. Since there was not a test treatment group in the marker-independent arm ($f=0$), the true treatment effects in the marker cohorts ($\delta_j, j=0, 1$) and true marker effects in each treatment group ($\gamma_i, i=T, C$) could not be unbiasedly estimated/ tested, neither could be the treatment-by-marker interaction. However, the primary goal of this study to assess the utilization of ERCC1 expression to guide the selection of platinum chemotherapy for lung cancer patients could be achieved as follows. The overall response rate (ORR) was 39.3% (out of 141 subjects) in the marker-independent arm, and 51.2% (out of 225 subjects) in the marker-dependent (genotypic) arm; c.f. Table 2 of Cobo et al.¹⁸ Thus, the marker's utility was estimated to be, in terms of ORR, a gain of 12% with the use of ERCC1 mRNA expression to guide the selection of platinum therapy versus not to use biomarker and always to treat patients with the SOC. The marker-positive (high level of ERCC1 mRNA expression) prevalence was about $p \approx 96/225 = 43\%$.

Even though this trial used $f=0$, according to Section 5 and Appendix D Figs 5A to 5H the variance of the marker's utility estimate was insensitive to f for the precision medicine design. The relative efficiency is applicable to designs with $f>0$ (say, $f=0.5$) as well. Moreover, even though there were no explicit information on the sensitivity and specificity of the ERCC1 mRNA assay and classification rule and no estimate of the marginal marker's effect, we may still entertain the following question: If the same study employed a marker-stratified design instead, say with $r=0.5$, how would this study gain in efficiency in testing the hypothesis of marker's clinical utility? Figs 2D-F can shed light on the answer. First, as commented before, Figs. 2D and 2E show that the marker's effects ($\tau - \gamma$) makes minuscule difference in the RE trends. Figs 2E and 2F shows the maximum RE occurs at randomization $t \approx 0.5$ and RE increases with increasing sensitivity and specificity. Giving an optimistic scenario that sensitivity=specificity=0.90, checking the RE for the case of prevalence $p \approx 43\%$, both Figs. 2D and 2F show $RE \approx 0.25$. This implies that if the trial were using a stratified design, the efficiency in testing H_U could increase, or equivalently the sample size could be cut, by four folds. In addition, with a stratified design, the other hypotheses H_+ , H_- , and H_f could also be tested.

Example 2 is the pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (NSCLC) trial.⁸ This trial stratified qualified subjects by biomarker's TPS (tumour proportion score 50% vs 1 - 49%), which measures the extent of PD-L1 expression, then randomized subjects with 1:1:1 ratio to three treatment groups within each high and low TPS stratum. For illustration, also since there was no significant difference between the two test doses of pembrolizumab, we only look at the pembrolizumab 2 mg (Pem) versus docetaxel (Dox), which is the control/SOC. The high TPS ("strongly positive" of PD-L1 marker) stratum (TPS+) had enrolled 442 patients and the low TPS cohort (TPS-) had 592 patients. Thus the prevalence rate of the PD-L1 strongly positives among the PD-L1 positive NSCLC patients was $p \approx 44\%$. According to Merck & Co., Inc.,¹⁹ 66% of the NSCLC patients are PD-L1 positive (TPS+). Thus, the prevalence of the PD-L1 strongly positives is $p \approx 30\%$ among the NSCLC patient population. The PD-L1 IHC 22C3 assay and classification was shown to have high sensitivity and specificity and

was approved by the FDA.²⁰ The overall survival results were the following: In the TPS+ cohort, 58 deaths out of 139 patients in the Pem group (42%), 86 deaths out of 152 patients in the Dox group (57%). In the TPS- cohort, 114 deaths out of 205 patients in the Pem group (56%), 107 deaths out of 191 patients in the Dox group (56%). With $\lambda_{sen} = \lambda_{spec} \approx 0.95$, the positive predictive value $\tau = 0.89$ and the negative predictive value $\eta = 0.98$. From Eq. (5) we obtain $\hat{\delta}_1 = [-0.98(0.42 - 0.57) + (1 - 0.89)(0.56 - 0.56)] / (1 - 0.89 - 0.98) = -0.16897$, and from Eq. (8) we have $\hat{\delta}_0 = [(1 - 0.98)(0.42 - 0.57) - 0.89(0.56 - 0.56)] / (1 - 0.89 - 0.98) = 3.4483 \times 10^{-3}$. The marker's clinical utility with respect to overall survival can be directly calculated by Eq. (27):

$\hat{U}_R(\lambda_{sen}, \lambda_{spec}, p) = p(\lambda_{sen} - r)\hat{\delta}_1 + (1 - p)(1 - \lambda_{spec} - r)\hat{\delta}_0 \approx 0.3(0.95 - 0.5)(-0.16897) + 0.7(1 - 0.95 - 0.5)(3.4483 \times 10^{-3}) = -2.3897 \times 10^{-2}$. This means that, in practice, we estimate to save about additional 2.4% lives by treating advanced NSCLC patients with pembrolizumab or docetaxel using marker guided strategy (i.e., treat with Pem for TPS+ or else treat with Dox) versus equal randomization without marker's information. If one were using the precision medicine design with $t = f = 0.5$ for this study, a 3-fold sample size would be required to show this result since $RE \approx 0.30$ according to Figs 2D or 2E.

7. Discussion

As pointed out by Hayes,³ the past President of American Society of Clinical Oncology (ASCO), prospective randomized clinical trials (PRCT) addressing biomarkers are precious and few because they are time-consuming and costly. But PRCT are necessary for the regulatory agencies to approve new therapies with predictive biomarkers. Therefore, statistical efficiency is a critical issue for considering trial designs. In this paper, we compared the efficiency between two families of designs: the marker-stratified designs and the precision medicine designs. We focused on these two designs because they are popular; almost all literature in biomarker trials discussed them. Stratified design are classical, but with biomarker we need to consider the assay and classification rule's sensitivity and specificity. On the other hand, the precision medicine designs (rename of the so-called "marker-based strategy designs") is a relatively new proposal in the modern precision medicine era. More importantly, these two designs are capable of addressing the same set of hypotheses of interest, namely, H_- (treatment effect in the marker-negative cohort), H_+ (treatment effect in the marker-positive cohort), H_I (the marker-by-treatment interaction), and H_U (the marker clinical utility) as shown in this paper. Of course, the composite hypothesis $H_C = H_- \cap H_+$ (overall treatment effect in the unselected population) is also of interest and testable with either of these two designs. It is sufficient and clearer to consider the relative efficiency in testing the individual hypothesis rather than multiple tests or a composite of these hypotheses. We emphasize that our focus is on the relative efficiency issue of these two designs, not on how to control the type-I error rate in testing a family of hypotheses. For the latter problem, we reference Millen et al.²¹

It is clear that relative efficiency of designs has to be based on testing/estimating the same hypothesis/parameter of interest. We showed that in addressing each H_- , H_+ , H_I , and H_U , the stratified design is more efficient than the precision medicine design. The precision medicine design loses efficiency largely because it includes patients treated with the same treatment

on both marker-independent and marker-dependent arms, as suggested by Mandrekar and Simon.^{9–11} However, unlike previously suggested by Mandrekar and Simon,^{9–11} our results (Fig 1 and Fig 2, Panels A and D) showed that the relative efficiency showed very different trends as a function of the marker-positive prevalence, depending on what hypothesis the two designs we address and compared them to. Fig 1A (with respect to testing H_+) and 1D (with respect to testing H_-) showed linear functions, but opposite (increasing/decreasing) trends. Fig 2A (with respect to testing interaction hypothesis) and 2D (with respect to testing marker utility hypothesis) showed quadratic functions, also with opposite (concave/convex) directions. Moreover, the extent to which the inferior efficiency of the precision medicine design to the stratified design is quite surprising in our investigation. Our evaluations (Figs 1 and 2) showed the patterns of the relative efficiency under wide ranges of design parameter values. The patterns of REs vary and are hypothesis-specific as expected, but the conclusion is consistent: Marker stratified design is much more efficient than the precision medicine design in all the hypotheses of interest. The examples we discussed in Section 6 illustrate how much gain in efficiency for the stratified design compared to the precision medicine design in testing/estimating biomarker's clinical utility.

Finally, based on the relative efficiency considered in this paper, we recommend marker-stratified designs over precision-medicine designs. In practice, when considering a marker-stratified design, the information on design parameters such as sensitivity, specificity, and prevalence of the biomarker-positives need to be assumed, together with the true treatment effect and variance, for sample size calculations. Information on these design parameters should be obtained from other studies. As uncertainty always exists when using information from other studies, a prudent researcher would contemplate various scenarios.

A. Derivations of Equations (6), (7), (9) and (10)

A.1. Approximation method

In the following when calculating

$$Var(\bar{Y}_{T1}) = E_{N_1}[Var(\bar{Y}_{T1}|N_1)] + Var_{N_1}[E(\bar{Y}_{T1}|N_1)] = E_{N_1}\left[Var\left(\frac{1}{rN_1}\sum_{\{k:A_k=T, M_k=1\}}Y_k|N_1\right)\right] = \frac{\sigma_T^2}{r}E\left(\frac{1}{N_1}\right)$$

we need to consider $E\left(\frac{1}{N_1}\right)$. Since sample size needs to be at least 2 to calculate variance, in practice we consider $E\left(\frac{1}{N_1}|N_1 \geq 2\right)$ and have

$$E\left(\frac{1}{N_1} | N_1 \geq 2\right) = \frac{\sum_{k=2}^{N-2} \frac{1}{k} (N-2k) q^k (1-q)^{N-2-k}}{\sum_{k=2}^{N-2} (N-2k) q^k (1-q)^{N-2-k}}.$$

Liu et al¹⁷ approximated N_1 by its expected value, $E(N_1) = Nq$, and replaced

$Var(\bar{Y}_{T1}) \approx \frac{\sigma_{T*}^2}{Nqr}$. Or we may treat N_1 as given, then

$$Var(\bar{Y}_{T1}) \approx \frac{\sigma_{T*}^2}{rN_1}.$$

A.2. Derivations of (6) and (7)

Using the method shown in Section A.1, we have

$$\begin{aligned} Var(\hat{\mu}_{T1}) &= \left(\frac{-\eta}{1-\tau-\eta}\right)^2 Var(\bar{Y}_{T1}) + \left(\frac{1-\tau}{1-\tau-\eta}\right)^2 Var(\bar{Y}_{T0}) \quad (A.1) \\ &\approx \frac{\frac{\eta^2 \sigma_{T*}^2}{rqN} + \frac{(1-\tau)^2 \sigma_{T\phi}^2}{r(1-q)N}}{(1-\tau-\eta)^2} \text{ or } \frac{\frac{\eta^2 \sigma_{T*}^2}{rN_1} + \frac{(1-\tau)^2 \sigma_{T\phi}^2}{rN_0}}{(1-\tau-\eta)^2}. \end{aligned}$$

Since our goal here is to compare the relative efficiency between designs, the approximation is consistently applied to both designs. Similarly,

$$\begin{aligned} Var(\hat{\mu}_{C1}) &= \left(\frac{-\eta}{1-\tau-\eta}\right)^2 Var(\bar{Y}_{C1}) + \left(\frac{1-\tau}{1-\tau-\eta}\right)^2 Var(\bar{Y}_{C0}) \quad (A.2) \\ &\approx \frac{\frac{\eta^2 \sigma_{C*}^2}{(1-r)qN} + \frac{(1-\tau)^2 \sigma_{C\phi}^2}{(1-r)(1-q)N}}{(1-\tau-\eta)^2} \text{ or } \frac{\frac{\eta^2 \sigma_{C*}^2}{(1-r)N_1} + \frac{(1-\tau)^2 \sigma_{C\phi}^2}{(1-r)N_0}}{(1-\tau-\eta)^2}. \end{aligned}$$

Equation (6) is the sum of Equations (A.1) and (A.2).

Since the variances in (A.1) and (A.2) can be estimated by

$$\widehat{Var}(\hat{\mu}_{T1}) = \frac{\frac{\eta^2 s_{T1}^2}{N_{T1}} + \frac{(1-\tau)^2 s_{T0}^2}{N_{T0}}}{(1-\tau-\eta)^2} \text{ and } \widehat{Var}(\hat{\mu}_{C1}) = \frac{\frac{\eta^2 s_{C1}^2}{N_{C1}} + \frac{(1-\tau)^2 s_{C0}^2}{N_{C0}}}{(1-\tau-\eta)^2}.$$

Equation (7) is the sum of these two estimated variances.

A.3. Derivations of (9) and (10)

Since

$$\begin{aligned} Var(\hat{\mu}_{T0}) &= \left(\frac{1-\eta}{1-\tau-\eta} \right)^2 Var(\bar{Y}_{T1}) + \left(\frac{-\tau}{1-\tau-\eta} \right)^2 Var(\bar{Y}_{T0}) \quad (A.3) \\ &\approx \frac{\frac{(1-\eta)^2 \sigma_{T*}^2}{rqN} + \frac{\tau^2 \sigma_{T\phi}^2}{r(1-q)N}}{(1-\tau-\eta)^2} \text{ or } \frac{\frac{(1-\eta)^2 \sigma_{T*}^2}{rN_1} + \frac{\tau^2 \sigma_{T\phi}^2}{rN_0}}{(1-\tau-\eta)^2}, \end{aligned}$$

and

$$\begin{aligned} Var(\hat{\mu}_{C0}) &= \left(\frac{1-\eta}{1-\tau-\eta} \right)^2 Var(\bar{Y}_{C1}) + \left(\frac{-\tau}{1-\tau-\eta} \right)^2 Var(\bar{Y}_{C0}) \quad (A.4) \\ &\approx \frac{\frac{(1-\eta)^2 \sigma_{C*}^2}{(1-r)qN} + \frac{\tau^2 \sigma_{C\phi}^2}{(1-r)(1-q)N}}{(1-\tau-\eta)^2} \text{ or } \frac{\frac{(1-\eta)^2 \sigma_{C*}^2}{(1-r)N_1} + \frac{\tau^2 \sigma_{C\phi}^2}{(1-r)N_0}}{(1-\tau-\eta)^2}, \end{aligned}$$

Equation (9) is the sum of Equations (A.3) and (A.4).

Since the variances in (A.3) and (A.4) can be estimated by

$$\widehat{Var}(\hat{\mu}_{T0}) = \frac{\frac{(1-\eta)^2 s_{T1}^2}{N_{T1}} + \frac{\tau^2 s_{T0}^2}{N_{T0}}}{(1-\tau-\eta)^2} \text{ and } \widehat{Var}(\hat{\mu}_{C0}) = \frac{\frac{(1-\eta)^2 s_{C1}^2}{N_{C1}} + \frac{\tau^2 s_{C0}^2}{N_{C0}}}{(1-\tau-\eta)^2}.$$

Equation (10) is the sum of these two estimated variances.

B. Derivations of Equations (11) to (19)

B.1. Unbiased estimators of μ_{ij}

For marker-independent arm, the means and variances for the treatment group $i (= T \text{ or } C)$ are

$$\mu_{i\#} \equiv E(Y_{ik}^{IN}) = p\mu_{i1} + (1-p)\mu_{i0} \quad (\text{B.1})$$

and

$$\begin{aligned} v_i^2 &\equiv \text{Var}(Y_{ik}^{IN}) = \text{Var}(E(Y_{ik}^{IN} | D)) + E_D \text{Var}(Y_{ik}^{IN} | D) = E\mu_{iD}^2 - (E\mu_{iD})^2 + p\sigma_{i1}^2 + (1-p)\sigma_{i0}^2 \\ &= p\mu_{i1}^2 + (1-p)\mu_{i0}^2 - (p\mu_{i1} + (1-p)\mu_{i0})^2 + p\sigma_{i1}^2 + (1-p)\sigma_{i0}^2 = p(1-p)\Delta_i^2 + p\sigma_{i1}^2 + (1-p)\sigma_{i0}^2 \end{aligned} \quad (\text{B.2})$$

respectively.

Unbiased estimators of $\mu_{i\#}$ can be obtained for $i = T, C$ respectively by the sample means:

$$\hat{\mu}_{T\#} = \frac{1}{Ntr} \sum_{k=1}^{Ntr} Y_{Tk}^{IN} = \bar{Y}_T^{IN}$$

and

$$\hat{\mu}_{C\#} = \frac{1}{Nt(1-r)} \sum_{k=1}^{Nt(1-r)} Y_{Ck}^{IN} = \bar{Y}_C^{IN}.$$

The variances of the estimators are

$$\text{Var}(\hat{\mu}_{T\#}) = \frac{v_T^2}{Ntr} = \frac{p(1-p)\Delta_T^2 + p\sigma_{T1}^2 + (1-p)\sigma_{T0}^2}{Ntr} \quad (\text{B.3})$$

and

$$\text{Var}(\hat{\mu}_{C\#}) = \frac{v_C^2}{Nt(1-r)} = \frac{p(1-p)\Delta_C^2 + p\sigma_{C1}^2 + (1-p)\sigma_{C0}^2}{Nt(1-r)}, \quad (\text{B.4})$$

where v_T^2, v_C^2 can be estimated by the sample variances $s_{T\#}^2 = \frac{1}{Ntr-1} \sum_{k=1}^{Ntr} (Y_{Tk}^{IN} - \bar{Y}_T^{IN})^2$ and $s_{C\#}^2 = \frac{1}{Nt(1-r)-1} \sum_{k=1}^{Nt(1-r)} (Y_{Ck}^{IN} - \bar{Y}_C^{IN})^2$, respectively.

For Marker-dependent arm, the means for the two treatment groups are

$$E(Y_{Tk}^{DP} | M = 1) \equiv \mu_{T*} = \tau\mu_{T1} + (1-\tau)\mu_{T0} \quad (\text{B.5})$$

and

$$E(Y_{Ck}^{DP} | M = 0) \equiv \mu_{C\phi} = (1-\eta)\mu_{C1} + \eta\mu_{C0} \quad (\text{B.6})$$

The variances for the two treatment groups are

$$\text{Var}(Y_{Tk}^{DP} | M = 1) \equiv \omega_T^2 = \tau(\mu_{T1}^2 + \sigma_{T1}^2) + (1-\tau)(\mu_{T0}^2 + \sigma_{T0}^2) - \mu_{T*}^2 = \tau(1-\tau)\Delta_T^2 + \tau\sigma_{T1}^2 + (1-\tau)\sigma_{T0}^2$$

and

$$\text{Var}(Y_{Ck}^{DP} | M = 0) \equiv \omega_C^2 = \eta(\mu_{C0}^2 + \sigma_{C0}^2) + (1-\eta)(\mu_{C1}^2 + \sigma_{C1}^2) - \mu_{C\phi}^2 = \eta(1-\eta)\Delta_C^2 + (1-\eta)\sigma_{C1}^2 + \eta\sigma_{C0}^2.$$

Let N_{Di} be the sample size of the marker-appeared positive group ($M=1$) with treatment i ($= T$ or C). Note that $N_{D1} \sim \text{Binomial}(N(1-t), q)$ and $N_{D0} = N(1-t) - N_{D1}$. The unbiased estimators of means in this arm can be obtained by

$$\hat{\mu}_{T*} = \frac{1}{N_{D1}} \sum_{k=1}^{N_{D1}} Y_{Tk}^{DP} = \bar{Y}_T^{DP}$$

and

$$\hat{\mu}_{C\phi} = \frac{1}{N_{D0}} \sum_{k=1}^{N_{D0}} Y_{Ck}^{DP} = \bar{Y}_C^{DP},$$

respectively.

Using the expected sample size $E(N_{D1}) = N(1 - t)q$ to approximate N_1 , or treating N_{D1} as given, the variances of the estimators are, respectively,

$$\text{Var}(\hat{\mu}_{T*}) \approx \frac{\omega_T^2}{N(1-t)q} \text{ or } \frac{\omega_T^2}{N_{D1}} \quad (\text{B.7})$$

and

$$\text{Var}(\hat{\mu}_{C\phi}) \approx \frac{\omega_C^2}{N(1-t)(1-q)} \text{ or } \frac{\omega_C^2}{N_{D0}}, \quad (\text{B.8})$$

where ω_T^2 and ω_C^2 can be estimated by the sample variances

$s_{T*}^2 = \frac{1}{N_{D1}-1} \sum_{k=1}^{N_{D1}} (Y_{Tk}^{DP} - \bar{Y}_T^{DP})^2$ and $s_{C*}^2 = \frac{1}{N_{D0}-1} \sum_{k=1}^{N_{D0}} (Y_{Ck}^{DP} - \bar{Y}_C^{DP})^2$, respectively. Note that $\sigma_{T*}^2 = \omega_T^2$ and $\sigma_{C\phi}^2 = \omega_C^2$ by definition.

The unbiased estimators of μ_{ij} can be derived by solving the above Eqs. (B.1), (B.5) and (B.6), we have

$$\hat{\mu}_{T1} = \frac{1-p}{\tau-p} \hat{\mu}_{T*} - \frac{1-\tau}{\tau-p} \hat{\mu}_{T\#} \quad (\text{B.9a})$$

$$\hat{\mu}_{T0} = \frac{-p}{\tau-p} \hat{\mu}_{T*} + \frac{\tau}{\tau-p} \hat{\mu}_{T\#} \quad (\text{B.9b})$$

$$\hat{\mu}_{C1} = \frac{1-p}{1-p-\eta} \hat{\mu}_{C\phi} - \frac{\eta}{1-p-\eta} \hat{\mu}_{C\#} \quad (\text{B.9c})$$

$$\hat{\mu}_{C0} = \frac{-p}{1-p-\eta} \hat{\mu}_{C\phi} + \frac{1-\eta}{1-p-\eta} \hat{\mu}_{C\#}. \quad (\text{B.9d})$$

B.2. Derivations of (11), (12) and (13)

From Equations B.9a to B.9d, for the marker positive cohort, the unbiased estimate of the treatment effect, i.e., Equation (11), can be obtained by

$$\hat{\delta}_1 = \hat{\mu}_{T1} - \hat{\mu}_{C1} = \frac{(1-p)\hat{\mu}_{T*} - (1-\tau)\hat{\mu}_{T\#}}{\tau-p} - \frac{(1-p)\hat{\mu}_{C\phi} - \eta\hat{\mu}_{C\#}}{1-p-\eta} = \frac{(1-p)\bar{Y}_T^{DP} - (1-\tau)\bar{Y}_T^{IN}}{\tau-p} - \frac{(1-p)\bar{Y}_C^{DP} - \eta\bar{Y}_C^{IN}}{1-p-\eta}.$$

From Equations B.3, B.4, B.7 and B.8 we have (Equation (12))

$$\begin{aligned} \text{Var}(\hat{\delta}_1) \approx & \frac{(1-p)^2\omega_T^2}{N(1-t)q} + \frac{(1-\tau)^2v_T^2}{Ntr} + \frac{(1-p)^2\omega_C^2}{N(1-t)(1-q)} + \frac{\eta^2v_C^2}{Nt(1-r)} \text{ or } \approx \frac{(1-p)^2\omega_T^2}{N_{D1}} + \frac{(1-\tau)^2v_T^2}{Ntr} \\ & + \frac{(1-p)^2\omega_C^2}{N_{D0}} + \frac{\eta^2v_C^2}{Nt(1-r)}, \end{aligned}$$

which, with either approximation, can be estimated by (Equation (13))

$$\widehat{\text{Var}}(\hat{\delta}_1) = \frac{(1-p)^2s_{T*}^2}{N_{D1}} + \frac{(1-\tau)^2s_{T\#}^2}{Ntr} + \frac{(1-p)^2s_{C*}^2}{N_{D0}} + \frac{\eta^2s_{C\#}^2}{Nt(1-r)},$$

B.3. Derivations of (17), (18) and (19)

Similarly from Equations B.9a to B.9d, for the marker negative cohort, the unbiased estimate of the treatment effect, i.e., Equation (17), can be obtained by

$$\hat{\delta}_0 = \hat{\mu}_{T0} - \hat{\mu}_{C0} = \frac{-p\hat{\mu}_{T*} + \tau\hat{\mu}_{T\#}}{\tau-p} - \frac{-p\hat{\mu}_{C\phi} + (1-\eta)\hat{\mu}_{C\#}}{1-p-\eta} = \frac{-p\bar{Y}_T^{DP} + \tau\bar{Y}_T^{IN}}{\tau-p} - \frac{-p\bar{Y}_C^{DP} + (1-\eta)\bar{Y}_C^{IN}}{1-p-\eta}.$$

From Equations B.3, B.4, B.7 and B.8, we have (Equation (18))

$$\text{Var}(\hat{\delta}_0) \approx \frac{p^2\omega_T^2}{N(1-t)q} + \frac{\tau^2v_T^2}{Ntr} + \frac{p^2\omega_C^2}{N(1-t)(1-q)} + \frac{(1-\eta)^2v_C^2}{Nt(1-r)} \text{ or } \approx \frac{p^2\omega_T^2}{N_{D1}} + \frac{\tau^2v_T^2}{Ntr} + \frac{p^2\omega_C^2}{N_{D0}} + \frac{(1-\eta)^2v_C^2}{Nt(1-r)},$$

which, with either approximation, can be estimated by (Equation (19))

$$\widehat{Var}(\hat{\delta}_0) = \frac{\frac{p^2 s_{T*}^2}{N_{D1}} + \frac{\tau^2 s_{T\#}^2}{N_{tr}}}{(\tau - p)^2} + \frac{\frac{p^2 s_{C*}^2}{N_{D0}} + \frac{(1-\eta)^2 s_{C\#}^2}{N_{t(1-r)}}}{(1 - p - \eta)^2}.$$

C. Robustness of Variances to Design Parameters

In this appendix, we address the following question (raised by the Associate Editor): Is it possible that, (regarding the efficiency), one design may be overly sensitive (or not sufficiently robust against) potential deviations from the assumed parameter values. Toward this end, we plot the variances of the tests for each of the hypotheses (H_- , H_+ , H_I , H_U) versus the design parameters (marker-positive prevalence rate, sensitivity and specificity).

In testing H_- we plot V_0^{SF} of the stratified design (c.f., Eq (9)) together with V_0^{PM} of the precision medicine design (c.f., Eq (18)). In testing H_+ we plot V_1^{SF} of the stratified design (c.f., Eq (6)) together with V_1^{PM} of the precision medicine design (c.f., Eq (12)). In testing H_- or H_+ , the variances of the tests are more stable/robust for the stratified design (solid lines) than the precision medicine design (dashed lines) with respect to the varying marker-positive prevalence rate (Figs 3A and 3D), marker's sensitivity (Figs 3B and 3E), and specificity (Figs 3C and 3F).

Similarly in testing H_I we plot the test variances $Var(\hat{\theta}^{SF})$ (c.f. Eq. (21)) of the stratified design and $Var(\hat{\theta}^{PM})$ (c.f. Eq. (24)) of the precision medicine design. It is seen that the variance of the test for H_I is also more stable/robust for the stratified design (solid lines) than the precision medicine design (dashed lines) with respect to the varying marker-positive prevalence rate (Fig 4A), marker's sensitivity (Fig 4B) and specificity (Fig 4C).

In testing H_U , we also examine how the variances V_U^{SF} (c.f. Eq. (29)) of the stratified design versus V_U^{PM} (c.f. Eq. (33)) of the precision medicine design behave with varying design parameters. It is interesting to see that the variance of the test is more stable/robust for the precision medicine design (V_U^{PM} dashed lines) than the stratified design (V_U^{SF} solid lines) with respect to the varying marker-positive prevalence rate (Fig 4D), marker's sensitivity (Fig 4E) and specificity (Fig 4F). This is the opposite to what we see in tests for H_- , H_+ , and H_I . Nevertheless, in all graphs, the variances of the precision medicine design are larger, indicating lower efficiency, than the corresponding variances of the stratified medicine.

D. Sensitivity to f in Testing Marker's Clinical Utility

In this appendix, we demonstrate that under the conventional setting of homogeneous variances, $\sigma_{ij}^2 = 1 (i = T, C; j = 0, 1)$, the variance V_U^{PM} (c.f. Eq. (33)) of the precision medicine design in testing H_U is insensitive to the parameter f , the randomization ratio in the marker-independent arm. Since f only applies to the precision medicine design (dash lines), not the stratified design (solid lines), we see several dash lines for varying f (0 to 0.8) but only one

solid line (for the stratified design, $r = 0.5$) in all the graphs in Fig 5. From all the graphs, we see that all the dashed line are almost indistinguishable, meaning that the variance V_U^{PM} , as functions of other parameters such as marker-positive prevalence rate, proportion of marker independent arm, sensitivity and specificity, is insensitive to f . For this reason, it is justified to set $f = 0.5$ without loss of generality in Section 5.3.

References

1. Sargent D, Allegra C. Issues in clinical trial designs for tumor marker studies. *Seminars in Oncology*. 2002; 29:222–230. [PubMed: 12063675]
2. Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: design issues. *J Natl Cancer Inst*. 2010; 102:152–160. [PubMed: 20075367]
3. Hayes, DF. A bad tumor marker test is as bad a bad drug: The case for more consistent regulation of cancer diagnostics. *ASC Connection*. Dec 22. 2016 <http://connection.asco.org/blogs>
4. Simon R, Maitournam A. Evaluating the Efficiency of Targeted Designs for Randomized Clinical Trials. *Clinical Cancer Research*. 2004; 10:6759–6763. [PubMed: 15501951]
5. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine*. 2005; 24:329–339. [PubMed: 15551403]
6. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol*. 2005; 23:2020–27. [PubMed: 15774793]
7. Hoering A, LeBlanc M, Crowley JJ. Randomized phase III clinical trial designs for targeted agents. *Clin Cancer Res*. 2008; 14:4358–4367. [PubMed: 18628448]
8. Herbst, RS., Baas, P., Kim, DW., et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet*. Published Online December 19, 2015 [http://dx.doi.org/10.1016/S0140-6736\(15\)01281-7](http://dx.doi.org/10.1016/S0140-6736(15)01281-7)
9. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: One size does not fit all. *J Biopharm Stat*. 2009; 19:530–542. [PubMed: 19384694]
10. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J Clin Oncol*. 2009; 27:4027–34. [PubMed: 19597023]
11. Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: lessons from real trials. *Clinical Trials*. 2010; 7:567–573. [PubMed: 20392785]
12. Hayes DF, Trock B, Harris AL. Assessing the clinical impact of prognostic factors: when is statistically significant clinically useful? *Breast Cancer Res Treat*. 1998; 52:305–319. [PubMed: 10066089]
13. Young KY, Laird A, Zhou XH. The efficiency of clinical trial designs for predictive biomarker validation. *Clinical Trials*. 2010; 7:557–566. [PubMed: 20571132]
14. Shih WJ, Lin Y. On study designs and hypotheses for clinical trials with predictive biomarkers. *Contemporary Clinical Trials*. 2017; 62:140–145. [PubMed: 28838813]
15. Potti A, Nevins JR. Utilization of genomic signatures to direct use of primary chemotherapy. *Curr Opin Genet Develop*. 2008 Feb; 18(1):62–67.
16. Renfro LA, Mallick H, An MW, Sargent DJ, Mandrekar SJ. Clinical trial designs incorporating predictive biomarkers. *Cancer Treatment Reviews*. 2016; 43:74–82. [PubMed: 26827695]
17. Liu C, Liu A, Hu J, Yuan V, Halabi S. Adjusting for misclassification in stratified biomarker clinical trial. *Statistics in Medicine*. 2014; 33:3100–3113. [PubMed: 24733510]
18. Cobo M, Isla D, Massuti B, et al. Customizing cisplatin based on quantitative excision repai cross complementing 1 mRNA expression: a phase III trial in non-small-cell lung cancer. *J Clin Oncol*. 2007; 25:2747–2754. [PubMed: 17602080]
19. Merck & Co., Inc. KEYTRUDA: Start Informed With PD-L1 Expression in mNSCLC. <https://www.keytruda.com/static/pdf/keytruda-pd-l1-expression-testing-guide.pdf>. ONCO-1212896-0000 05/17

20. Agilent Pathology Solutions. PD-L1 IHC 22C3 pharmDx is FDA-approved For In Vitro Diagnostic Use. https://www.agilent.com/cs/library/usermanuals/public/29158_pd-l1-ihc-22C3-pharmdx-nslcl-interpretation-manual.pdf
21. Millen BA, Dmitrienko A, Ruberg S, Shen L. A statistical framework for decision making in confirmatory multipopulation tailoring clinical trials. *Drug Information Journal*. 2012; 46:647–656.

Figure 1. Relative Efficiency of Test for Treatment Effect on Marker Positive and Negative Cohorts

Figure 2. Relative Efficiency for Tests of Interaction and Marker Utility

Figure 3. Variances in Testing for Treatment Effect on Marker Positive and Negative Cohorts (solid line for Stratified design, and dashed line for Precision Medicine Design)

Figure 4. Variances in Tests of Interaction and Marker Utility (solid line for Stratified design, and dashed line for Precision Medicine Design)

Figure 5. Variances in Tests of Marker Utility for varying f (solid line for Stratified design with $r = 0.5$, and dashed line for Precision Medicine Design with $t = 0.5$ (except Figure B))