Project 1 Paper Adam

Adam Brand
Erin Gabriel

November 2020

1 Trial designs

Enrichment: standard enrichment, patients within a specific subgroup are randomized to A or B (in lit, cite)

Biomarker-stratified: randomized trial stratified by biomarker status (in lit, cite)

Augmented biomarker-stratified: get physician choice of treatment prior to randomizing. Allows ability to estimate proportion of patients who would have received the same therapy regardless of biomarker assessment. Also may allow estimation of clinical utility without a biomarker-strategy design. (develop)

Biomarker-strategy: standard biomarker strategy design, patients randomized to either biomarker-directed therapy or physician choice (without knowledge of biomarker status) (in lit, cite)

Modified biomarker-strategy: patients randomized to either biomarker directed therapy or to a randomized treatment arm. This removes the ability to assess clinical utility. This design is mainly used for co-development for assessing the efficacy of new treatment(s) while also evaluating a biomarker-driven treatment strategy against randomization (not against SOC) (in lit, cite)

Augmented modified biomarker-strategy: get physician recommended treatment for patients on the randomized arm prior to randomizing. This may allow for assessment of clinical utility (develop)

2 Comparing Approved Treatments (A vs B)

Three estimands of interest:

Optimal treatment for a subgroup:

$$E[Y_B|M = m] - E[Y_A|M = m]$$

Or,

$$\frac{E[Y_B|M = m]}{E[Y_A|M = m]}$$

Differential treatment response between subgroups:

$$(E[Y_B|M = pos] - E[Y_A|M = pos]) - (E[Y_B|M = neg] - E[Y_A|M = neg])$$

Or,

$$\frac{E[Y_B|M = pos]}{E[Y_A|M = pos]} \Big/ \frac{E[Y_B|M = neg]}{E[Y_A|M = neg]} = \frac{E[Y_B|M = pos] * E[Y_A|M = neg]}{E[Y_A|M = pos] * E[Y_B|M = neg]}$$

Clinical utility of biomarker treatment strategy:

$$E[Y_{\text{biomarker-directed}}] - E[Y_{\text{physician-directed}}]$$

Or,

$$\frac{E[Y_{\text{biomarker-directed}}]}{E[Y_{\text{physician-directed}}]}$$

2.1 Enrichment Design

- **Optimal treatment for a subgroup:**

$$E[Y_B|M = m] - E[Y_A|M = m]$$

Or,

$$\frac{E[Y_B|M = m]}{E[Y_A|M = m]}$$

In an enrichment design with no interference between units, randomization ensures that the following equalities hold:

$$E[Y_B|M = pos] = E[Y|TRT = B, M = pos]$$
$$E[Y_A|M = pos] = E[Y|TRT = A, M = pos]$$

Therefore, the estimand for treatment effect in the marker-positive subgroup can be estimated by:

$$E[Y|TRT = B, M = pos] - E[Y|TRT = A, M = pos]$$

Or,

$$\frac{E[Y|TRT = B, M = pos]}{E[Y|TRT = A, M = pos]}$$

  - Difference in restricted means

- **Differential treatment response between subgroups:**

$$(E[Y_B|M = pos] - E[Y_A|M = pos]) - (E[Y_B|M = neg] - E[Y_A|M = neg])$$

Or,

$$\frac{E[Y_B|M = pos]}{E[Y_A|M = pos]} \Big/ \frac{E[Y_B|M = neg]}{E[Y_A|M = neg]} = \frac{E[Y_B|M = pos] * E[Y_A|M = neg]}{E[Y_A|M = pos] * E[Y_B|M = neg]}$$

An enrichment design only provides information on the marker-positive subgroup, therefore cannot estimate differential treatment effect between subgroups.

- **Clinical utility of biomarker treatment strategy:**

$$E[Y_{\text{biomarker-directed}}] - E[Y_{\text{physician-directed}}]$$

Or,

$$\frac{E[Y_{\text{biomarker-directed}}]}{E[Y_{\text{physician-directed}}]}$$

An enrichment design only provides information on the marker-positive subgroup, therefore cannot estimate clinical utility of a marker-guided treatment strategy.

2.2 Biomarker Stratified Design

In a biomarker-stratified design with no interference between units, the following four equalities hold due to randomization.

$$E[Y_B|M = pos] = E[Y|TRT = B, M = pos] \tag{1}$$

$$E[Y_A|M = pos] = E[Y|TRT = A, M = pos] \tag{2}$$

$$E[Y_B|M = neg] = E[Y|TRT = B, M = neg] \tag{3}$$

$$E[Y_A|M = neg] = E[Y|TRT = A, M = neg] \tag{4}$$

The design can be powered for only one of the subgroups or both; which could depend on the prevalence of marker positive group and the treatment effects within each group. If powering for both subgroups, may stop enrolling one marker group when sample size reached which can reduce the number enrolled, but not the number screened. This design is basically parallel enrichment designs, and must be powered according to estimand(s) of interest.

- **Optimal treatment for a subgroup:**

$$E[Y_B|M = m] - E[Y_A|M = m]$$

Or,

$$\frac{E[Y_B|M = m]}{E[Y_A|M = m]}$$

Due to the equalities in 1-4, the optimal treatment assignment for each subgroup can be estimated by:

$$E[Y|TRT = B, M = m] - E[Y|TRT = A, M = m]$$

Or,

$$\frac{E[Y|TRT = B, M = m]}{E[Y|TRT = A, M = m]}$$

- **Differential treatment response between subgroups:**

$$(E[Y_B|M = pos] - E[Y_A|M = pos]) - (E[Y_B|M = neg] - E[Y_A|M = neg])$$

Or,

$$\frac{E[Y_B|M = pos]}{E[Y_A|M = pos]} \Big/ \frac{E[Y_B|M = neg]}{E[Y_A|M = neg]} = \frac{E[Y_B|M = pos] * E[Y_A|M = neg]}{E[Y_A|M = pos] * E[Y_B|M = neg]}$$

Also due to the equalities in 1-4, the differential treatment response can be estimatd by:

$$(E[Y|TRT = B, M = pos] - E[Y|TRT = A, M = pos]) - (E[Y|TRT = B, M = neg] - E[Y|TRT = A, M = neg])$$

Or,

$$\frac{E[Y|TRT = B, M = pos]}{E[Y|TRT = A, M = pos]} \Big/ \frac{E[Y|TRT = B, M = neg]}{E[Y|TRT = A, M = neg]} = \frac{E[Y|TRT = B, M = pos] * E[Y|TRT = A, M = neg]}{E[Y|M = pos] * E[Y|TRT = B, M = neg]}$$

- **Clinical utility of biomarker treatment strategy:**

$$E[Y_{\text{biomarker-directed}}] - E[Y_{\text{physician-directed}}]$$

Or,

$$\frac{E[Y_{\text{biomarker-directed}}]}{E[Y_{\text{physician-directed}}]}$$

Augmenting the biomarker-stratified design to obtain the physician's choice of treatment prior to marker testing and independent of randomized treatment assignment allows us to identify the clinical utility of applying the marker-guided treatment strategy, assuming no interference between units.

Assuming, without loss of generality, that the marker-guided treatment strategy is to treat all marker-positives with treatment B and all marker-negative patients with treatment A, we have:

$$E[Y_{\text{biomarker-directed}}] = E[Y|(TRT = B, M = pos) \cup (TRT = A, M = neg)]$$

which can be estimated by pooling (due to randomization) the M-positive subjects who received treatment B with the M-negative subjects who received treatment A.

For the physician-directed arm,

$$E[Y_{\text{physician-directed}}] = E[Y|physician - directed = TRT]$$

which is true due to randomization. The sub-sample of patients matching the physician's choice of treatment is a random sample of the trial population, stratified by marker status. So then,

$$E[Y_B|physician - directed = B] = E[Y|TRT = B, physician - directed = B]$$

$$E[Y_A|physician - directed = A] = E[Y|TRT = A, physician - directed = A]$$

The augmented aspect of the design provides us with a physician's choice of treatment for every subject, independent of marker status and randomized assignment. Therefore, we can estimate the outcome of applying physician's choice of treatment by pooling all subjects who received, through independent randomization, the treatment the physician chose. It is important to obtain the physician's choice of treatment without knowledge (e.g., before) of the randomization assignment. This way, physician's choice and random assignment remain independent, and conditioning on physician's choice does not destroy randomization in the restricted sample.

Some patients may be included in both analysis arms. Every subject is included only once per arm, no correlation adjustments are needed when estimating the expected outcome in each analysis arm separately. However, when estimating a contrast between arms, some adjustments may be needed to account for the correlation.

Powering such a study will be difficult as it is not known ahead of time the proportion of subjects who will have physician choice agree with the randomized treatment assignment. This can be estimated through observational data, if available. Randomization probabilities can also be chosen in each stratified arm to attempt to increase or decrease the proportion of patients whose assignment agrees with physician choice based on available data.

2.3 Biomarker Strategy Design

- **Optimal treatment for a subgroup:**

$$E[Y_B|M = m] - E[Y_A|M = m]$$

Or,

$$\frac{E[Y_B|M = m]}{E[Y_A|M = m]}$$

This design cannot identify optimal treatment for any given subgroup. Due to randomization, we can estimate:

$$E[Y_B|M = pos] = E[Y|M = pos, Arm = marker - directed]$$
$$E[Y_A|M = neg] = E[Y|M = neg, Arm = marker - directed]$$

We can also estimate the left hand of the below equations, but the equalities do not hold because physicians choice of treatment could be confounded.

$$E[Y|M = pos, Arm = physician - choice, TRT = A] \neq E[Y_A|M = pos]$$
$$E[Y|M = neg, Arm = physician - choice, TRT = B] \neq E[Y_B|M = neg]$$

- **Differential treatment response between subgroups:**

$$(E[Y_B|M = pos] - E[Y_A|M = pos]) - (E[Y_B|M = neg] - E[Y_A|M = neg])$$

Or,

$$\frac{E[Y_B|M = pos]}{E[Y_A|M = pos]} \Big/ \frac{E[Y_B|M = neg]}{E[Y_A|M = neg]} = \frac{E[Y_B|M = pos] * E[Y_A|M = neg]}{E[Y_A|M = pos] * E[Y_B|M = neg]}$$

This design cannot identify differential treatment between subgroups for the same reasons it cannot identify optimal treatment for either subgroup.

- **Clinical utility of biomarker treatment strategy:**

$$E[Y_{\text{biomarker-directed}}] - E[Y_{\text{physician-directed}}]$$

Or,

$$\frac{E[Y_{\text{biomarker-directed}}]}{E[Y_{\text{physician-directed}}]}$$

This design provides the most direct assessment of clinical utility of applying the marker-guided treatment strategy to the patient population by directly comparing the current standard of care, i.e., physician-directed therapy to the marker-guided strategy. Clinical utility is estimated by:

$$E[Y|Arm = marker - directed] - E[Y|Arm = physician - choice]$$

Or,

$$\frac{E[Y|Arm = marker - directed]}{E[Y|Arm = physician - choice]}$$

Because randomization ensures that:

$$E[Y_{\text{biomarker-directed}}] = E[Y|Arm = marker - directed]$$
$$E[Y_{\text{physician-choice}}] = E[Y|Arm = physician - choice]$$

2.4 Modified Biomarker Strategy Design

The modified biomarker strategy design compares a marker-guided treatment arm to a fully randomized arm, with or without stratification by marker status. It is a combination of a biomarker stratified design and a biomarker strategy design.

- **Optimal treatment for a subgroup:**

$$E[Y_B|M = m] - E[Y_A|M = m]$$

Or,

$$\frac{E[Y_B|M = m]}{E[Y_A|M = m]}$$

This design can identify optimal treatment for each subgroup only if marker testing is performed in the randomized arm. If it is, and there is no interference between units,

$$E[Y_B|M = pos] = E[Y|M = pos, TRT = B]$$
$$E[Y_A|M = pos] = E[Y|M = pos, TRT = A]$$
$$E[Y_A|M = neg] = E[Y|M = neg, TRT = A]$$
$$E[Y_B|M = neg] = E[Y|M = neg, TRT = B]$$

Notice the right side of the above equations involves pooling subjects across the marker-directed and randomized arms which will increase power by utilizing all of the information. The equalities still hold when pooling, because of the double round of randomization. And we can estimate optimal treatment in each subgroup using the same estimates as for the marker-stratified design.

$$E[Y|TRT = B, M = m] - E[Y|TRT = A, M = m]$$

Or,

$$\frac{E[Y|TRT = B, M = m]}{E[Y|TRT = A, M = m]}$$

The design will have different power/sample size implications than the biomarker stratified design, and the differences will be dependent on the proportion of marker positive subjects, the treatment effect sizes within each subgroup and the randomization probabilities.

- **Differential treatment response between subgroups:**

$$(E[Y_B|M = pos] - E[Y_A|M = pos]) - (E[Y_B|M = neg] - E[Y_A|M = neg])$$

Or,

$$\frac{E[Y_B|M = pos]}{E[Y_A|M = pos]} \Big/ \frac{E[Y_B|M = neg]}{E[Y_A|M = neg]} = \frac{E[Y_B|M = pos] * E[Y_A|M = neg]}{E[Y_A|M = pos] * E[Y_B|M = neg]}$$

As above, we can identify treatment effect in each subgroup as the below equalities hold due to randomization.

$$E[Y_B|M = pos] = E[Y|TRT = B, M = pos]$$
$$E[Y_A|M = pos] = E[Y|TRT = A, M = pos]$$
$$E[Y_B|M = neg] = E[Y|TRT = B, M = neg]$$
$$E[Y_A|M = neg] = E[Y|TRT = A, M = neg]$$

So then the estimand of differential treatment effect is captured by:

$(E[Y|TRT = B, M = pos] - E[Y|TRT = A, M = pos]) - (E[Y|TRT = B, M = neg] - E[Y|TRT = A, M = neg])$

Or,

$$\frac{E[Y|TRT = B, M = pos]}{E[Y|TRT = A, M = pos]} \Big/ \frac{E[Y|TRT = B, M = neg]}{E[Y|TRT = A, M = neg]} = \frac{E[Y|TRT = B, M = pos] * E[Y|TRT = A, M = neg]}{E[Y|M = pos] * E[Y|TRT = B, M = neg]}$$

which, again involves pooling patients from the marker-directed treatment arm and the randomized arm.

- **Clinical utility of biomarker treatment strategy:**

$$E[Y_{\text{biomarker-directed}}] - E[Y_{\text{physician-directed}}]$$

Or,

$$\frac{E[Y_{\text{biomarker-directed}}]}{E[Y_{\text{physician-directed}}]}$$

The augmented version of this design can identify the clinical utility of using the marker-directed treatment only if marker status is obtained in the randomized arm. As in the augmented-stratified design, we can estimate clinical utility of using the marker-guided treatment, because:

$$E[Y_{\text{biomarker-directed}}] = E[Y|(TRT = B, M = pos) \cup (TRT = A, M = neg)]$$

which can be estimated by pooling (due to randomization) the M-positive subjects who received treatment B with the M-negative subjects who received treatment A.

For the physician-directed arm,

$$E[Y_{\text{physician-directed}}] = E[Y|physician - directed = TRT]$$

which is true due to randomization. The sub-sample of patients matching the physician's choice of treatment is a random sample of the trial population, stratified by marker status. So then,

$$E[Y_B|physician - directed = B] = E[Y|TRT = B, physician - directed = B]$$

$$E[Y_A|physician - directed = A] = E[Y|TRT = A, physician - directed = A]$$

Because randomization ensures that:

$$E[Y_{\text{biomarker-directed}}] = E[Y|(Arm = marker{-}directed) \cup (Arm = randomized, TRT = marker{-}directed)]$$
$$E[Y_{\text{physician-choice}}] = E[Y|TRT = physician - directed]$$

2.5 Analysis Methods

- Absolute differences in survival
  - Difference in restricted means
- Relative differences in survival
  - Cox PH hazard ratios
  - Time point specified comparisons (ratio)
- **CONSIDERATIONS**:
  - how to account for correlation when including a patient in both comparison groups???
  - how to compare 4 groups in come of these estimands??

We compare survival outcomes between groups using four different estimands; hazard ratio, ratio of survival probability up to a pre-specified time point, survival probability ratio at pre-specified time points, and absolute difference in restricted mean survival time (RMST).

Hazard ratio

To estimate hazard ratios between groups, Cox proportional hazards regression is used. The commonly-used Cox model is

$$\lambda(t; z) = \lambda_0(t)\epsilon^{\beta z}$$

where $\lambda(t; z)$ is the hazard function of T evaluated at t, given a vector of covariates $Z = z$, and β is the vector of model coefficients (Cox et al. [1972]). As Martinussen et al. [2020] states, when $exp(\beta)$ is interpreted as the ratio of log survival probabilities, or cumulative hazards, at time t, that is,

$$exp(\beta) = \frac{logP(T_A > t)}{logP(T_B > t)}$$

then $exp(\beta)$ can be interpreted as a causal, relative measure of the log survival probability at an arbitrary time t between those receiving treatment A and those receiving treatment B under two conditions:

1. the proportional hazards assumption holds
2. $T_A \perp\!\!\!\perp A$ and $T_B \perp\!\!\!\perp B$, as is the case in a randomized trial

However, $exp(\beta)$ is more often interpreted as a hazard ratio, that is

$$exp(\beta) = \frac{\lim_{h\to 0} P(t \leq T_A < t + h | T_A \geq t)}{\lim_{h\to 0} P(t \leq T_B < t + h | T_B \geq t)}$$

As identified by Hernán [2010], the causal interpretation of the hazard ratio is invalid even when there is no confounding and conditions 1 and 2 above are met. The hazard rate at time t conditions on surviving to time t, which occurs after randomization, destroying the causal interpretation granted by randomization.

When fitting a Cox model to survival data however, the test for whether β is significantly different from zero is a valid causal test under conditions 1 and 2, and $exp(\beta)$ can be interpreted causally as the relative increase/decrease in log survival probabilities at the selected follow-up time point. Although $exp(\beta)$ cannot be interpreted causally as the hazard ratio as of the follow-up time point, it is most often reported as such and therefore will be reported as such in this paper.

When comparing survival between groups via the Cox proportional hazards model, the following Cox models will be fit

$$\lambda(t; trt) = \lambda_0(t)\epsilon^{\beta_0 * trt} \tag{5}$$

$$\lambda(t; trt, marker) = \lambda_0(t)\epsilon^{\beta_0 * trt + \beta_1 * marker + \beta_2 * trt * marker} \tag{6}$$

$$\lambda(t; arm) = \lambda_0(t)\epsilon^{\beta_0 * arm} \tag{7}$$

where (5) will test optimal treatment for a subgroup with $H_0 : \beta_0 = 0$, (6) will test differential response to treatment between subgroups with $H_0 : \beta_2 = 0$, and (7) will test the clinical utility of marker-directed treatment with $H_0 : \beta_0 = 0$, and where arm defines the marker-directed versus physician-directed arm. The models will be fit in R using the `coxph` package, and adjustments for correlation due to subject pooling in the analysis of some designs will be based on the counting process formulation of Andersen and Gill [1982].

Ratio of survival probability up to a pre-specified time point

The ratio of survival probability up to a pre-specified time point will be compared using a pseudo-observation technique developed by Andersen et al. [2003] which uses non-parametric, Kaplan-Meier based modelling of right-censored survival data while allowing for the estimation of covariate effects, competing risks, and correlated observations (Klein et al. [2008]). Klein et al. [2007] applied this technique to the comparison of entire survival functions up to a specified time point and survival probabilities at fixed time points, and showed that this technique works well when incorporating covariates and when the proportional hazards assumption is violated. Overgaard et al. [2017] presents the asymptotic theory of the pseudo-observation technique and proves that the estimating procedure used by Klein et al. [2007] to compare survival probabilities is reasonable under a condition of completely independent censoring, defined as that censoring is independent of event time, event type and covariates. Overgaard et al. [2017] also presents asymptotic theory results proving that estimating the restricted mean survival time using pseudo observations is reasonable when completely independent censoring is met. In a randomized trial setting, completely independent censoring is a reasonable assumption, so we will use the pseudo-observation technique to compare survival probabilities and restricted mean survival. The asymptotic variances of the estimates provided by using the pseudo-observation technique are complex and not readily available in software, so we will use the jackknife variance estimates available in the R package `geese`.

The general pseudo-observation technique is as follows. Let X_i, i from $1, ..., n$, be independent and identically distributed random variables, vectors or processes, and let θ be the expected value of some function of X_i, that is $\theta = E[f(X_i)]$ where θ may be multivariate. Also assume that there is an unbiased estimator of θ, $\hat{\theta}$, and suppose there are measured covariates $\mathbf{Z}_i$. Then the conditional expectation of $f(X_i)$ given $\mathbf{Z}_i$ is defined by

$$\theta_i = E[f(X_i)|\mathbf{Z}_i]$$

The ith pseudo-observation is defined by

$$\hat{\theta}_i = n \cdot \theta - (n-1)\widehat{\theta^{-i}}$$

where $\widehat{\theta^{-i}}$ is the "leave-one-out" estimator for θ. Regressing on θ now corresponds to specifying the relationship between θ_i and $\mathbf{Z}_i$ using a generalized linear model with link function $g(\cdot)$,

$$g(\theta_i) = \beta^T \mathbf{Z}_i$$

Estimates of the β's, which are based on unbiased estimating equations, are shown to be asymptotically normal by Liang and Zeger [1986], and Klein et al. [2008] claim that a sandwich estimator for the variance of $\hat{\beta}$ converges in probability to the true variance. Once the pseudo-observations are computed, estimates of β can be obtained by using standard GEE or GLM methods widely available in statistical software, regressing the outcome on the pseudo-observations as well as the covariates of interest.

In the context of comparing entire survival curves up to a specified time point, pseudo-observations are computed for a grid of time points up to the specified ending time point, that is $\theta = (\theta_1, ..., \theta_M)$ where $S(t_i) = P[T > t_i]$ and $\theta_i = S(t_i)$ is estimated using the Kaplan-Meier estimator. When $M = 1$, the regression model compares the survival probability at a single time point. When fitting the regression model to an entire curve, Andersen et al. [2004] and Klein and Andersen [2005] found that five to ten time points spaced equally by number of events works well (Klein et al. [2008]). A detailed explanation of available R and SAS functions for calculating pseudo-observations and incorporating them into standard regression methods using right-censored survival data is available in Klein et al. [2008].

When comparing entire survival curves up to a pre-specified time point, pseudo observations will be computed at ten time points, spaced roughly equally based on number of events, for each subject. The pseudo-observations will be regressed according to the following models where $g(\cdot)$ is the complimentary log-log link function, $ln[-lnS(t)]$,

$$g(\theta_i) = \beta_0 + \beta_1 \cdot trt + \beta_{tpseudo} \tag{8}$$

$$g(\theta_i) = \beta_0 + \beta_1 \cdot trt + \beta_2 \cdot marker + \beta_3 \cdot trt \cdot marker + \beta_{tpseudo} \tag{9}$$

$$g(\theta_i) = \beta_0 + \beta_1 \cdot arm + \beta_{tpseudo} \tag{10}$$

where (8) will test optimal treatment for a subgroup with $H_0 : \beta_1 = 0$, (9) will test differential response to treatment between subgroups with $H_0 : \beta_3 = 0$, and (10) will test the clinical utility of marker-directed treatment with $H_0 : \beta_1 = 0$, and where arm defines the marker-directed versus physician-directed arm. $\beta_{tpseudo}$ represents the β's for each of the ten time points used to calculate the pseudo-observations. Approximate jackknife variance estimates for the β's from Yan and Fine [2004] will be used following the recommendation by Klein et al. [2008].

Pseudo-observations will be computed in R using the `pseudo` package. GEE models accounting for correlation through an id variable and using the independent correlation variance structure will be fit in R using the `geese` package. Using GEE to estimate parameters accounts for the correlation when including subjects in more than one comparison group as well as for the multiple pseudo-observations per subject.

Survival probability ratio at a time point

Survival probability ratio at pre-specified time points will also be compared using the pseudo-observation technique according to Klein et al. [2007]. Estimation will follow the same process as when comparing entire survival curves up to a time point except pseudo-observations will only be computed at a single time point. GEE still accounts for the correlation when including subjects in more than one arm even though only one pseudo-observation per subject will be available.

Restricted mean survival time (RMST)

The pseudo-observation technique, extended to estimate restricted mean survival time (RMST) by Andersen et al. [2004], will also be used to estimate RMST. Restricted mean survival time, first used by Irwin [1949], is the area under the survival curve up to time t, that is

$$RMST = \int_0^t S(u)du$$

and the estimate of RMST is

$$\hat{\theta} = \int_0^t \hat{S}(u)du$$

Here, unlike when comparing survival curves across time points, θ is a one-dimensional summary of the data, so instead of computing pseudo-observations, a pseudo-mean is computed estimating RMST. This estimated mean is regressed on the covariates to provide an estimate of covariate effects. The following regression models are fit using GEE which accounts for correlation from including subjects in more than one group.

$$\hat{\theta} = \beta_0 + \beta_1 \cdot trt \tag{11}$$

$$\hat{\theta} = \beta_0 + \beta_1 \cdot trt + \beta_2 \cdot marker + \beta_3 \cdot trt \cdot marker \tag{12}$$

$$\hat{\theta} = \beta_0 + \beta_1 \cdot arm \tag{13}$$

where (11) will test optimal treatment for a subgroup with $H_0 : \beta_1 = 0$, (12) will test differential response to treatment between subgroups with $H_0 : \beta_3 = 0$, and (13) will test the clinical utility of marker-directed treatment with $H_0 : \beta_1 = 0$, and where arm defines the marker-directed versus physician-directed arm.

Andersen et al. [2004] showed that the pseudo-observation technique for estimating RMST works well when incorporating covariate information and when the proportional hazards assumption does not hold.

3 Comparing Experimental Treatment B with Single-option Standard of Care A

The estimands of interest, as stated above, remain the same. However, this scenario is comparing an experimental, unapproved treatment which may or may not improve outcomes in both or either of the marker-defined subgroups. There are some nuances in this scenario which change the estimands available to each trial design. One nuance is that physician's choice in this scenario is known for every subject, because only treatment A is currently approved for use. Without information on the efficacy/safety of the experimental drug, a physician cannot assign any patient to treatment B. Therefore in this scenario, treatment B only becomes a treatment option for a particular subgroup if it is shown to be non-inferior to treatment A for that subgroup.

Because we need treatment B to be shown as non-inferior to treatment A for a particular subgroup before it can used as a treatment option for that subgroup, declaration of clinical utility requires at a minimum that the treatment effect of treatment B is beneficial in one subgroup and inferior in the other subgroup. If treatment B is inferior in both subgroups, obviously it is not a treatment option over treatment A. If B is superior in one subgroup and non-inferior in the other subgroup, then there is no clinical utility of applying the marker-guided treatment strategy, because treating every subject with treatment B regardless of marker status would provide the same result.

In this scenario we will hypothesize that treatment B is superior to treatment A in the marker-positive subgroup while it is inferior in the marker-negative subgroup.

- **Optimal treatment for a subgroup:**

$$E[Y_B|M = m] - E[Y_A|M = m]$$

Or,

$$\frac{E[Y_B|M = m]}{E[Y_A|M = m]}$$

Enrichment design: Same as above

Biomarker stratified design (no augmentation needed): Same as above

Biomarker strategy design: Previously, this design could not identify the optimal treatment for either subgroup because the following equality did not hold due to confounding of physician's choice:

$$E[Y|M = pos, Arm = physician - choice] \neq E[Y_A|M = pos]$$
$$E[Y|M = neg, Arm = physician - choice] \neq E[Y_B|M = neg]$$

However, now physician's choice is known to be treatment A for all subjects. Therefore,

$$E[Y|M = pos, Arm = physician - choice] = E[Y_A|M = pos]$$

And as above, we know that,

$$E[Y_B|M = pos] = E[Y|M = pos, Arm = marker - directed]$$
$$E[Y_A|M = neg] = E[Y|M = neg, Arm = marker - directed]$$

So we can now estimate the optimal treatment for the marker-positive subgroup with:

$$E[Y|M = pos, Arm = marker - directed] - E[Y|M = pos, Arm = physician - choice]$$

Or,

$$\frac{E[Y|M = pos, Arm = marker - directed]}{E[Y|M = pos, Arm = physician - choice]}$$

However, we still cannot estimate the optimal treatment for the marker-negative subgroup, because we have no information on marker-negative subjects receiving treatment B.

Modified biomarker strategy design: same as above

- **Differential treatment response between subgroups:**

$$(E[Y_B|M = pos] - E[Y_A|M = pos]) - (E[Y_B|M = neg] - E[Y_A|M = neg])$$

Or,

$$\frac{E[Y_B|M = pos]}{E[Y_A|M = pos]} \Big/ \frac{E[Y_B|M = neg]}{E[Y_A|M = neg]} = \frac{E[Y_B|M = pos] * E[Y_A|M = neg]}{E[Y_A|M = pos] * E[Y_B|M = neg]}$$

Enrichment design: Same as above

Biomarker stratified design (no augmentation needed): Same as above

Biomarker strategy design: Even though we can identify the optimal treatment for the marker-positive subgroup we still cannot identify the treatment effect in the marker-negative subgroup. Therefore, any differential treatment effect cannot be identified with this design.

Modified biomarker strategy design: same as above

- **Clinical utility of biomarker treatment strategy:**

$$E[Y_{\text{biomarker-directed}}] - E[Y_{\text{physician-directed}}]$$

Or,

$$\frac{E[Y_{\text{biomarker-directed}}]}{E[Y_{\text{physician-directed}}]}$$

Enrichment design: Same as above

Biomarker stratified design (no augmentation needed): Same as above except physician's choice is now known to be treatment A for all subjects. Therefore, no augmentation is needed, and the estimate of clinical utility remains the same as above, imputing treatment A for physician's choice. In this scenario, randomization probabilities can be manipulated to create the desired proportion of subjects agreeing with physician's choice.

Biomarker strategy design: Same as above although everyone on the physician's choice arm will receive treatment A.

Modified biomarker strategy design (no augmentation needed): Same as above except no augmentation is needed, because we know physician's choice is treatment A.

3.1 Analysis Methods

- **Testing**

  When testing for superior/inferior treatment within any subgroup, we will use the most common test statistic for survival outcomes in confirmatory clinical trials, the logrank test. It should be noted however, that the logrank statistic remains invalid whenever survival curves cross, so the p-value from the logrank test should always be accompanied by a Kaplan-Meier curves of the groups being compared.

  Include any other tests for sensitivity analyses?

- **Estimation**

  - Absolute difference in survival ($E[Y_B] - E[Y_A]$):
    - Restricted means - choice of ending timepoint affects estimate
    - additive hazards model? Assumptions?
    - anything else?
  - Relative difference in survival ($\frac{E[Y_B]}{E[Y_A]}$)
    - Cox ph - proportional hazards assumption, choice of ending follow-up time affects estimate
    - accelerated failure time model - parametric model

Table 1: Estimands Available when Comparing Approved Treatments (A vs B) Among Biomarker-defined Subgroups

	Optimal treatment for a subgroup	Differential treatment response between subgroups	Clinical utility of biomarker treatment strategy
Enrichment	X		
Biomarker-stratified	X	X	
Augmented biomarker-stratified	X	X	X
Biomarker-strategy			X
Modified biomarker-strategy	X	X	
Augmented modified biomarker-strategy	X	X	X

4 Introduction

4.1 Questions

How to arrange the diagrams into a 2x2 figure?

Thoughts on my tables? Explanations for the estimands and trial designs are below table 1. Explanations for why things changed in the 2nd table are below table 2.

Were you thinking of putting the tables in the introduction?

Table 2: Estimands Available when Comparing Experimental Therapy B vs Single-option Standard of Care A

	Optimal treatment for a subgroup	Differential treatment response between subgroups	Clinical utility of biomarker treatment strategy
Enrichment	X		
Biomarker-stratified	X	X	X
Augmented biomarker-stratified	X	X	X
Biomarker-strategy	X		X
Modified biomarker-strategy	X	X	X
Augmented modified biomarker-strategy	X	X	X

4.2 Outline and ideas

- some distinction of number screened and sample size enrolled for comparing power. Distinction importance depends on rarity of biomarker-subgroup, expense/invasiveness of test, etc... - for time to event endpoints, look how others compare methods in simulations (logrank test?). Also look at real trials for how they test interaction in a biomarker-stratified design (Cox interaction term?) - compare designs using time to event endpoint with interim futility/efficacy analyses - look in literature for how people have compared power between strategy and interaction designs in the past (apples to apples?; estimand they use even estimable per design?) - some discussion on comparing biomarker-stratified designs to independent enrichment trials (shared .05 vs .05 for each subgroup; fair comparison?) - diagrams for each of the 3 designs with optional box included for augmentation

Motivate the problem/what is the problem?

Biomarker-strategy design should be used more, specifically when lots of biomarker-treatment strata and when obtaining the biomarker is expensive or causes risk, harm, and/or discomfort for the patient. Is the biomarker useful? Even if biomarker defines subgroups benefiting differentially from treatments, physicians may already be prescribing treatment along the lines of the biomarker (they're already doing a good job prescribing for the subgroups without aid of a biomarker).

Literature

Cite research papers developing the prediction-driven trial types and their recommended use. Cite specific trials, and state how the biomarker-interaction and enrichment designs are highly favored over the biomarker-strategy designs, mostly due to sample size differences. Cite any literature stating our believed reasons for conducting a biomarker-strategy design, i.e., finding out if the biomarker is useful and how much instating a rule defined by the biomarker would benefit patients. Cite any other sources agreeing that this design has been largely ignored.

Must incorporate ASD designs relevant to Phase 3 confirmatory trials. Motivating examples use response rates more suited to Phase 2 trials, but look for use in Phase 3 with time to event.

Read Renfro paper, 2020 about two-stage biomarker-driven designs using time to event data.

Sachs et al. [2020] et al. motivates our research by stating that any biomarker-directed treatment strategy should be evaluated against SOC in a randomized trial. Good place to bring up that physicians may already be prescribing close to the recommended treatment rule. So then even if the treatment rule is the optimum treatment, it may not be worth the obtainment of potentially risky, uncomfortable, and/or costly biomarker status. Is the biomarker treatment rule useful even when optimal treatment?

Introduce our research (ideas for research)

See about extending the adaptive signature design (k-fold version) to the biomarker-strategy design. Explore ways of decreasing necessary sample size for biomarker-strategy design: related to the proportion of patients whose standard of care agrees with the rule (potentially do a interaction design in Phase 2 to identify a rule, including estimating the proportion of subjects who would've been prescribed the rule regardless of biomarker assessment). Possibly re-weight analysis based on agreement with standard of care, e.g., give patients that disagree with SOC higher weight, or exclude all patients who agree with SOC? Design

a multi-tiered study where biomarker-interaction design used to identify treatment rule and strategy design to test if the rule is useful. Compare this to the extended version of the ASD for the strategy design.

Idea for analyzing different utilities for marker-strategy designs. Potential for marker-strategy and SOC to perform similarly on average, but must explore what happens to the patients on SOC who don't receive strategy-ordained treatment. For example, say 90% on SOC am agree with strategy treatment, so on average we detect no difference. But what if those 10% different patients perform much less well than the strategy arm. Or, more generally, find a way to compare the distributions of outcomes between the marker-strategy arm and the SOC arm patients who did not receive the strategy ordained treatment.

5 Writing

As scientific advances lead to increased understanding of disease pathology, researchers are able to further classify disease sub types based on measurable biologic features called biomarkers. Biomarkers are sometimes able to increase prognostic accuracy (prognostic biomarkers) and/or predict response to treatment (predictive biomarkers). Therapeutics developed specifically for treating diseases with a particular biomarker signature, comprised of one or more biomarkers, have shown to be highly effective in treating, for example, chronic myelogenous leukemia (imatinib) (Cohen et al. [2002]) and Her-2 overexpressing breast cancer (trastuzumab) (Albanell and Baselga [1999]), among others. Development of biomarkers to better prognose outcomes and predict treatment response can increase the efficiency and effectiveness of therapeutic product development (Woosley and Cossman [2007]).

As with any therapeutic development, evidence of a desirable benefit-risk profile must be proven through one or more confirmatory clinical trials after gathering evidence from exploratory trials. The development of biomarkers to direct treatment decisions undergoes similar development. First, the biomarker(s) must be analytically validated, i.e. it must be possible to accurately and reliably measure the biomarker(s) (Omenn et al. [2012], Simon [2010]). Next, the biomarker signature, comprised of one or more analytically validated biomarkers, must show some capability of defining subgroups with differential treatment effect. This is called clinical validation and usually includes information from exploratory trials where cut points for continuous biomarkers are established and the biomarker signature is refined to optimize negative and positive predictive values (Simon [2010]). Lastly, a biomarker signature-directed therapy, known as a prediction-based decision rule, must show clinical utility, defined as the difference in expected outcome of using the prediction-based decision rule versus not (Sachs et al. [2020], Simon [2010], Sargent et al. [2005], Freidlin et al. [2010]).

It is this last step of evaluating clinical utility that is often ignored [cite?] even though evaluating clinical utility is crucial to effective and efficient treatment decisions [cite]. When testing new treatments, the common goal is to evaluate efficacy. In the context of prediction-driven trials, often the goal is to evaluate efficacy within one or more subgroups either with or without an overall population analysis. Depending on the design however, the clinical utility of the prediction-based decision rule cannot be estimated. The enrichment design is such a design, and is frequently used due to its advantage of superior sample size efficiency over other prediction-driven designs [cite!!!]. Enrolling only patients from the subgroup hypothesized to benefit from treatment, the enrichment design cannot evaluate clinical utility, because there is no evaluation of the treatment effect in other subgroups. Its possible that multiple or all subgroups can benefit, albeit differentially, meaning there may be no clinical utility of the decision rule even though the biomarker signature is predictive of treatment outcome. By not evaluating the clinical utility of the decision rule, clinical benefit may be missed.

Another frequently used prediction-driven trial design is the biomarker-stratified design where all patients are randomized, stratified by biomarker signature strata [cite initial work]. This design can also be thought of as a collection of enrichment trials that includes all strata. Unlike a single enrichment trial, the biomarker-stratified design can estimate treatment effect in all subgroups as well as in the entire population. This design also allows for formal testing of differential treatment effects between strata. However, this design can only evaluate clinical utility in settings where there is only one option for standard of care. When there are multiple options for standard of care, randomizing between those standards of care may not perform as well as a physician's choice of treatment. It is possible that physicians are already treating similarly to the prediction-driven decision rule without knowledge of the biomarker signature, meaning that the biomarker signature accurately predicts the optimal treatment per strata while achieving no clinical utility.

The biomarker-stratified design is also limited by the number of treatment-biomarker strata it can handle efficiently.

Biomarkers can be expensive and invasive for patients, causing discomfort and risking harm. If a biomarker signature provides no meaningful clinical utility, it should not be used even if it accurately predicts the optimal treatment per subgroup.

The prediction-driven trial design that directly evaluates clinical utility is the biomarker-strategy design, which randomizes between the prediction-based decision rule and physician's choice of care [cite]. Although this design directly evaluates clinical utility it is not often used due to requiring higher (sometimes much higher) sample sizes than the enrichment or biomarker-stratified designs [cite!!!].

Clinical utility of a prediction-based decision rule is often overlooked in favor of evaluating treatment efficacy due to requiring higher sample sizes. However, evaluating clinical utility is crucial to ensure that no benefit to subgroups is missed and unnecessary biomarker tests are not conducted. The optimal clinical trial design for evaluating the clinical utility of prediction-based decision rules depends on a number of factors: whether there exists a well-defined decision rule or the decision rule is simultaneously developed and evaluated, the number of options for standard of care, the number of experimental treatments under study, the number of subgroups defined by the biomarker signature, and existing medical knowledge (Sargent et al. [2005], Freidlin and Korn [2010], Hu and Dignam [2019], Renfro et al. [2016]).

In this paper, we will compare the sample size and power of confirmatory trial designs for evaluating the clinical utility of a well-defined prediction-based decision rule in various settings. We will also propose new designs to improve on sample size efficiency when evaluating clinical utility, and provide recommendations for when to use which design(s).

In Section blah, blah [will fill in this compulsory paragraph when appropriate]

References

Joan Albanell and José Baselga. Trastuzumab, a humanized anti-her2 monoclonal antibody, for the treatment of breast cancer. *Drugs Today (Barc)*, 35(12):931–946, 1999.

Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

Per Kragh Andersen, John P Klein, and Susanne Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

Per Kragh Andersen, Mette Gerster Hansen, and John P Klein. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime data analysis*, 10(4):335–350, 2004.

Martin H Cohen, Grant Williams, John R Johnson, John Duan, Jogarao Gobburu, Atiqur Rahman, Kimberly Benson, John Leighton, Sung K Kim, Rebecca Wood, et al. Approval summary for imatinib mesylate capsules in the treatment of chronic myelogenous leukemia. *Clinical cancer research*, 8(5):935–942, 2002.

David R Cox et al. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

Boris Freidlin and Edward L Korn. Biomarker-adaptive clinical trial designs. *Pharmacogenomics*, 11(12): 1679–1682, 2010.

Boris Freidlin, Lisa M McShane, and Edward L Korn. Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*, 102(3):152–160, 2010.

Miguel A Hernán. The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13, 2010.

Chen Hu and James J Dignam. Biomarker-driven oncology clinical trials: Key design elements, types, features, and practical considerations. *JCO Precision Oncology*, 1:1–12, 2019.

JO Irwin. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Epidemiology & Infection*, 47(2):188–189, 1949.

John P Klein and Per Kragh Andersen. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61(1):223–229, 2005.

John P Klein, Brent Logan, Mette Harhoff, and Per Kragh Andersen. Analyzing survival curves at a fixed point in time. *Statistics in medicine*, 26(24):4505–4519, 2007.

John P Klein, Mette Gerster, Per Kragh Andersen, Sergey Tarima, and Maja Pohar Perme. Sas and r functions to compute pseudo-values for censored data regression. *Computer methods and programs in biomedicine*, 89(3):289–300, 2008.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Torben Martinussen, Stijn Vansteelandt, and Per Kragh Andersen. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*, 26(4):833–855, 2020.

Gilbert S Omenn, Sharly J Nass, Christine M Micheel, et al. *Evolution of translational omics: lessons learned and the path forward*. National Academies Press, 2012.

Morten Overgaard, Erik Thorlund Parner, and Jan Pedersen. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics*, 45(5):1988–2015, 2017.

Lindsay A Renfro, Himel Mallick, Ming-Wen An, Daniel J Sargent, and Sumithra J Mandrekar. Clinical trial designs incorporating predictive biomarkers. *Cancer treatment reviews*, 43:74–82, 2016.

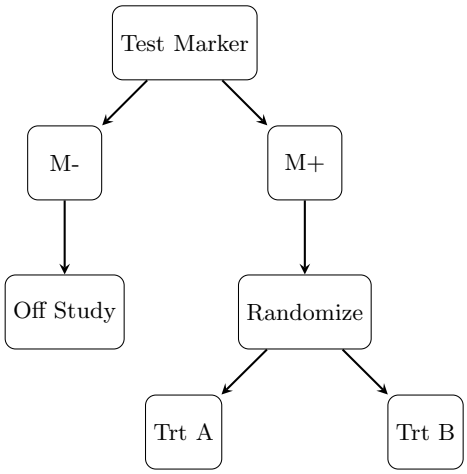
Michael C Sachs, Arvid Sjölander, and Erin E Gabriel. Aim for clinical utility, not just predictive accuracy. *Epidemiology (Cambridge, Mass.)*, 31(3):359, 2020.

Daniel J Sargent, Barbara A Conley, Carmen Allegra, and Laurence Collette. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*, 23(9):2020–2027, 2005.

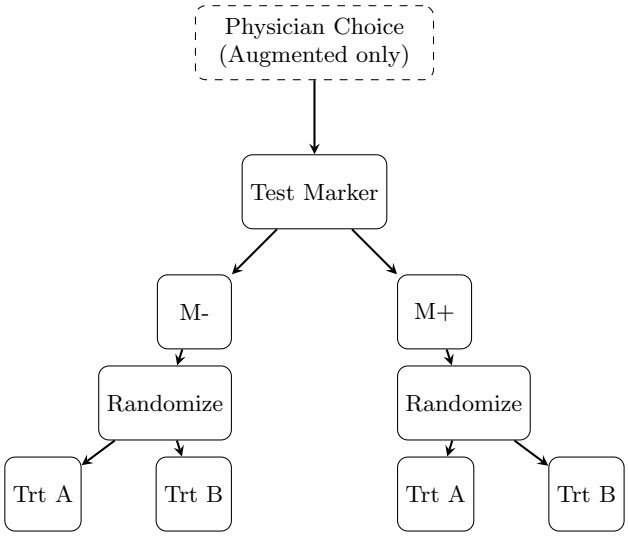
Richard Simon. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized medicine*, 7(1):33–47, 2010.

RL Woosley and J Cossman. Drug development and the fda's critical path initiative. *Clinical Pharmacology & Therapeutics*, 81(1):129–133, 2007.

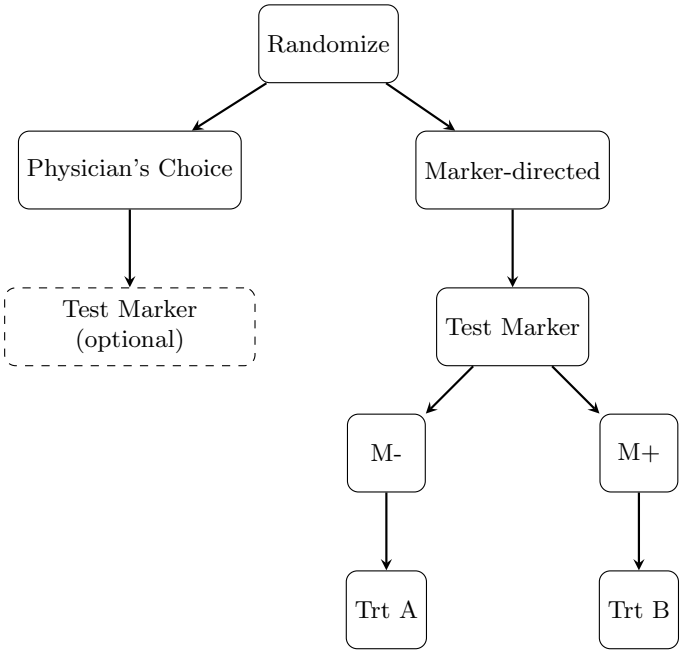
Jun Yan and Jason Fine. Estimating equations for association structures. *Statistics in medicine*, 23(6): 859–874, 2004.



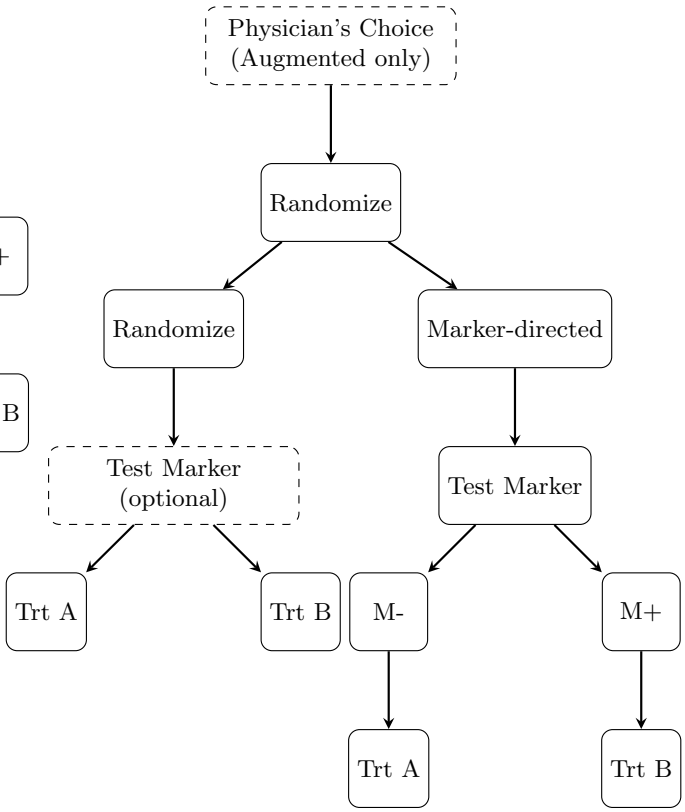
(a) Enrichment Design



(b) Potentially Augmented Biomarker Stratified Design



(c) Biomarker-strategy Design



(d) Potentially Augmented Modified Biomarker-strategy Design