Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–23

Bayesian Multi-Arm De-Intensification Designs

Steffen Ventz[*] and Lorenzo Trippa[†]

Abstract. In recent years, new cancer treatments have improved survival in multiple histologies. Some of these therapeutics, and in particular treatment combinations, are often associated with severe treatment-related adverse events (AEs). Therefore, It is important to identify alternative de-intensified therapies, such as dose-reduced therapies with reduced AEs and similar efficacy. We introduce a sequential design for multi-arm de-intensification studies. Based on joint modeling of toxicity and efficacy endpoints, the design evaluates multiple de-intensified therapies at different dose levels, one at a time. We study the utility of the design in oropharynx cancer de-intensification studies. We use a joint Bayesian model for efficacy and toxicity outcomes to define decision rules at interim and final analyses. Interim decisions include early termination of the study due to inferior survival of experimental arms compared to a standard of care (SOC), and the transitions from one de-intensified treatment arm to another with a further reduced dose when there is sufficient evidence of non-inferior survival. We evaluate the operating characteristics of the design using data from recent de-intensification studies in oropharynx cancer.

Keywords: Bayesian designs, de-intensification clinical studies, multi-arm clinical trials.

1 Introduction

In the last two decades, several new cancer treatments have improved patient survival (Semenza, 2008). Many new therapies consist of a backbone treatment, often chemotherapy or radiation therapy, combined with an additional drug, for instance, an immune checkpoint inhibitor. Some of these combination therapies improved survival, but they are associated with severe AEs. Intensity-modulated radiotherapy (IMRT) in combination with cisplatin is a SOC in oropharynx cancer (Ang et al., 2010) with three-year survival rates close to 90% (Ang et al., 2010; Gillison et al., 2019). However, adding cisplatin to IMRT is associated with a substantial increase in acute and late AEs compared with IMRT alone (Munker et al., 2001). Similarly, a combination treatment including chemotherapy is the SOC for early-stage HER2-positive breast cancer and has a high rate of treatment-related AEs Mathew and Brufsky (2017).

AEs associated with new treatments are the main motivation for testing if de-intensified therapies maintain efficacy similar to the SOC and reduce AEs. For instance, the recent studies E1308 (Marur et al., 2017), OPTIMA (Seiwert et al., 2019), RTOG1016 (Gillison et al., 2019), DeEscalate (Mehanna et al., 2019), PAMELA (Llombart-Cussac et al., 2017) and KRISTINE (Hurvitz et al., 2018) evaluated de-intensified oropharynx cancer and breast cancer therapies. De-intensification studies

[*] Division of Biostatistics & Health Data Science, University of Minnesota, steffen.ventz.81@gmail.com

[†] Department of Data Science, Dana Farber Cancer Institute

https://doi.org/10.1214/24-BA1417

consider therapies that (i) are dose-reductions of SOC therapies, (ii) replace one component of a SOC combination therapy with a potentially less toxic drug, or (iii) eliminate the backbone treatment from the SOC treatment. In all these cases, the study seeks to demonstrate that the de-intensified treatment has survival outcomes similar to the SOC and reduces AEs.

The design that we introduce is motivated by a study in HPV-associated oropharynx cancer. Several ongoing studies are evaluating de-intensified therapies (Mirghani and Blanchard, 2018). A trial we designed evaluates two de-intensified therapies that differ in the IMRT dose levels. Two large de-intensification studies RTOG1016 (Gillison et al., 2019) and De-ESCALaTE (Mehanna et al., 2019) in HPV-associated oropharynx cancer recently reported strongly inferior survival under the de-intensified therapy compared to the SOC (estimated overall survival (OS) hazard ratios of 1.45 and 5 for RTOG1016 and De-ESCALaTE) without reducing AEs. De-intensified studies tend to use large non-inferiority margins (D'Agostino et al., 2003) for non-inferiority testing to reduce the study's sample size for targeted type I/II error rates. These margins and inadequate interim analyses can lead - as the results of RTOG1016 and De-ESCALaTE suggest - to a large number of patients exposed to treatments with reduced efficacy. This indicates the importance for de-intensification designs to handle trade-offs between power and the number of patients exposed to inferior or toxic treatments using adequate interim analyses (IAs).

We introduce a de-intensification design that allows investigators to test multiple treatments sequentially, for instance, two treatments with 80% and 40% of the original SOC IMRT dose. Using Bayesian modeling for the distribution of survival times and the AEs, we specify sequential decision rules to evaluate treatments, including early futility-stopping rules that are tuned to balance (i) the risk of patients receiving inferior treatments that reduce survival and (ii) the need to identify non-inferior treatments that reduce AEs. The design evaluates de-intensified therapies one at a time, starting from the therapy closest to the SOC. Subsequent arms are only tested if there is evidence of non-inferiority and reductions of AEs for the previous de-intensified treatments. We discuss algorithms to tune early futility-stopping rules according to pre-defined stopping probabilities under the null hypothesis of inferior survival or AEs identical to the SOC. Additionally, the design calibrates the type I error rate to approximately match a targeted α level. We evaluate the operating characteristics of the design using data from recent de-intensification studies in oropharynx cancer.

De-intensification designs use non-inferiority (NI) testing procedures with a predefined NI margin and evaluate if the efficacy of an experimental treatment is comparable to the SOC (D'Agostino et al., 2003). Statistical considerations for NI studies concern the selection of a suitable testing procedure, the specification of the NI margin Δ, and the study design, including early stopping rules and the selection of the sample size (Blackwelder, 1982; Rothmann et al., 2003; Freidlin et al., 2007; Joshua Chen and Chen, 2012; Korn and Freidlin, 2017). Blackwelder (1982) discussed NI tests based on asymptotic techniques and (Farrington and Manning, 1990; Tu, 1998; Laster et al., 2006) focused on the finite-sample operating characteristics of NI tests. Exact NI tests have been discussed in (Chan, 2003; Laster et al., 2006), and extensions to time-to-event outcomes have been proposed in (Rothmann et al., 2003; Freidlin et al., 2007).

Other contributions focused on the selection of suitable NI margins Δ (Snapinn, 2004; Holmgren, 1999) and on the specification of early stopping rules for sequential NI experiments (Freidlin and Korn, 2002; Lachin, 2009; Korn and Freidlin, 2017). Bayesian work on NI experiments includes NI testing methodologies and the use of data from previous clinical studies in the analysis of NI experiments (Simon, 1999; Schmidli et al., 2013). Wellek (2005) and Williamson (2007) discussed NI tests for binary endpoints using beta prior distributions, and Osman and Ghosh (2011) proposed the use of Bernstein priors. Gamalo et al. (2011) used Bayesian modeling to select the NI margin Δ and (Daimon, 2008; Chen et al., 2011) investigated Bayesian sample size calculations for single-stage NI tests.

The main difference between these non-sequential single-stage non-inferiority testing procedures and our work is that we introduce a sequential design for de-intensification studies with efficacy and toxicity co-primary endpoints. The design utilizes a non-parametric Bayesian model to analyze survival data and AEs during the study. Key decisions to pause, stop, or continue evaluating de-intensified treatments are based on data summaries that quantify the trade-off between the risk of exposing patients to inferior treatments and the likelihood of demonstrating relevant reductions of AEs.

After introducing some notation in Section 2.1, we present the de-intensification designs for studies with efficacy primary endpoints (Section 2.2) and efficacy and toxicity (Section 2.3) co-primary endpoints. Section 3.1 compares several de-intensification strategies in HPV-associated oropharynx cancer with efficacy endpoints. Section 3.2 extends this comparison to oropharynx cancer studies with efficacy and toxicity co-primary endpoints.

2 De-intensification design

2.1 Notation and setup

We consider a phase II clinical study with $K \geq 1$ de-intensified treatments. In practice, $K = 2$ or 3, $K = 2$ in our motivating study. We assume the study does not include a SOC (control) arm. Simple modifications of our design allow the conclusion of a SOC arm. The SOC ($k = 0$) survival and toxicity distributions have been estimated previously, and the treatment is associated with a substantial risk of AEs. The de-intensified treatments likely present better AE profiles but may reduce survival. We evaluate if one or multiple de-intensified treatments are non-inferior to the SOC and reduce AEs. The treatments are ranked; for instance, if $k = 1, \cdots, K$ are K de-intensified dose levels, the study starts testing the highest one $k = 1$, followed by the reduced doses $k = 2, \ldots, K$.

A maximum of N patients will be enrolled. For each patient $1 \leq i \leq N$, (C_i, Y_i, X_i) indicates the assignment to treatment $C_i \in \{1, \cdots, K\}$, Y_i and X_i are the efficacy and toxicity outcomes. In our study, Y_i indicates the progression-free survival (PFS) time, and X_i is the time of the first AE (grade $\geq$ 3). We use $n_{t,k}$ and n_t to indicate the number of enrollments to arm k and the total number of enrollments by time t, and Σ_t denotes the data collected until time t since the first enrollment. Table 1 summarizes notation frequently used throughout the manuscript. Supplementary Table S1 (Ventz and Trippa, 2024) contains a complete list of notation.

C_i, Y_i, X_i	treatment assignment, efficacy and toxicity outcome of patient i
Σ_t	data collected until time t since the first enrollment
$n_t, n_{t,k}$	number of enrollments and enrollments to arm k by time t
$m_{\max}$	maximum number of enrollments per treatment arm
m_A, m_0, m_T	minimum number of enrollment to arm k before $\mathcal{H}_{0,k}$ can be rejected (m_A), or arm k can be stopped for inferiority (m_0) or toxicity (m_T)
$S_{k,j}(t)$	efficacy ($j = Y$) and toxicity ($j = X$) survival function for therapy k
$S_{k,Y\|X}(y)$	conditional survival function
θ_k, β_k	efficacy and toxicity summaries for treatment k
$\Delta, \Delta_k, \Delta_L, \Delta_\beta$	margins for testing and IA
$b_j(\cdot), j = 0, T, A$	PFS ($j = 0$) and toxicity ($j = T$) safety boundaries, and non-inferiority ($j = T$) boundary $b_j(\Sigma_t) = 1 - \eta_j \times \max\left[0, \frac{n_{k,t} - m_j}{m_{\max} - m_j}\right]^{\nu_j}$
t_{FU}	time between the last enrollment to treatment k and the final analysis

Table 1: Notation frequently used in the manuscript.

2.2 De-intensification studies with efficacy primary outcomes

In some cases, AE reductions of de-intensified treatment compared to the SOC can be anticipated or have been demonstrated before the study. In other cases, it is necessary to estimate toxicity and efficacy during the trial. In this subsection, we first introduce a design for studies that utilize efficacy outcomes Y_i for decisions. We then extend it in Section 2.3 to trials with toxicity X_i and efficacy Y_i co-primary endpoints.

The probability model. We use $S_{k,Y}(t) = Pr(Y_i \geq t | C_i = k)$ to indicate the survival function for therapy $k = 1, \ldots, K$. Our design can be combined with any Bayesian model for $S_{k,Y}$. In our motivating study, we considered several parametric models. Based on available prior data, we observed unsatisfactory model fits and used a nonparametric model. $S_{k,Y}$ are random functions with independent prior distributions.

We use a Beta-Stacy (BS) prior distribution (Walker and Muliere, 1997), $S_{k,Y} \sim BS(S_{k,Y} | V_Y, c_Y)$, where $V_Y(t) = E[S_{k,Y}(t)]$ is the prior mean, and the continuous function $c_Y(t) > 0$ controls the prior variability of $S_{k,Y}$ around V_Y. The prior is strictly related to Dirichlet and Beta processes (Ferguson, 1973; Hjort et al., 1990). Under the BS prior, $\{-\log(1 - S_{k,Y}(t))\}_{t \geq 0}$ is a monotone, right-continuous random function with independent increments, and $S_{k,Y}(0) = 0$ and $\lim_{t \to +\infty} - \log(1 - S_{k,Y}(t)) = \infty$ with probability one (Walker and Muliere, 1997). Unlike the Dirichlet process, the BS prior is conjugate with respect to right censored data (Walker and Muliere, 1997). If $Y = \{Y_i\}_{i=1}^n$ is an independent, right-censored sample from a distribution $S_{k,Y}$ and $S_{k,Y} \sim$

$BS(S_{k,Y}|V_Y, c_Y)$, then $p(S_{k,Y}|Y) = BS(S_{k,Y}|V_{Y,n}, c_{Y,n})$ is again a BS distribution with closed from expressions for the posterior mean $V_{Y,n}$ and uncertainty parameter $c_{Y,n}$ (Walker and Muliere, 1997). An advantage of using the BS prior is that conditionally on right censored data, the posterior distribution is available in closed form, and the summaries $\theta_k = \theta(S_k)$ can be easily simulated from the posterior.

Efficacy summaries. For each de-intensified treatment k, efficacy is quantified by a summary $\theta_k = \theta(S_{k,Y}) \in \mathbb{R}$. Large θ_k values correspond to large efficacy. Examples include the median, the mean $\theta_k = E[Y_i|C_i = k]$, or the restricted mean survival time (RMST) $\theta_k = E[\min(Y_i, t_E)|C_i = k]$ at a pre-specified $t_E > 0$.

Hypothesis testing. For each $k = 1, \cdots, K$ the null and alternative hypotheses are

$$\mathcal{H}_{0,k} = \{\theta_k \in \mathbb{R} : \theta_k \leq \theta_0 - \Delta\} \quad \text{and} \quad \mathcal{H}_{A,k} = \{\theta_k \in \mathbb{R} : \theta_k > \theta_0 - \Delta_k\}. \tag{1}$$

Here $\Delta \geq \Delta_k > 0$ are pre-specified margins. Values of θ_k below $\theta_0 - \Delta$ make treatments k inferior, whereas $\theta_k \geq \theta_0 - \Delta_k$ indicates an attractive alternative to the SOC.

Decision rules

We assume that de-intensified treatments can be ordered according to their expected efficacy levels θ_k. For instance, in our motivating study, treatments $k = 1, \cdots, K$ consist of decreasing dose-levels of the backbone treatment, and prior data (Bhide et al., 2012; Jensen et al., 2007; Eisbruch et al., 2004; Levendag et al., 2007) suggest that θ_k is non-increasing in k. The design (Figure 1 for an illustration) evaluates treatments sequentially, one after another, starting with treatment $k = 1$, which is the least likely to be inferior.

A maximum of $m_{\max} \leq N$ patients will be assigned to each treatment, and a minimum of $m_A \leq m_{\max}$ enrollments to the treatment are required before it can be declared non-inferior. At regular time intervals $t = 1, 2, \ldots$ (e.g., monthly, quarterly, etc.) IAs are conducted and the active treatment k may be (Figure 1(B))

(i) declared non-inferior to the SOC, and the trial progresses to evaluate treatment $k + 1$,

(ii) declared inferior to the SOC and the study terminates,

(iii) enrollment to treatment k continues or,

(iv) enrollment is paused for a maximum time t_{FU} after the last enrollment to arm k, to allow the accumulation of sufficient information for testing $\mathcal{H}_{0,k}$.

Let $D_t = (D_{t,1}, D_{t,2}) \in \{\emptyset, 1, \ldots, K\} \times \{d_E, d_P, d_T\}$ indicate the de-intensified treatment $D_{t,1}$ that is evaluated between IA t and $t+1$ ($\emptyset$ if the study is terminated at IA t), and $D_{t,2}$ denotes the status of the study during this time interval, where d_E = "*enrollment is open*", d_P = "*enrollment is paused*", and d_T = "*the study has been terminated*" (Figure 1).

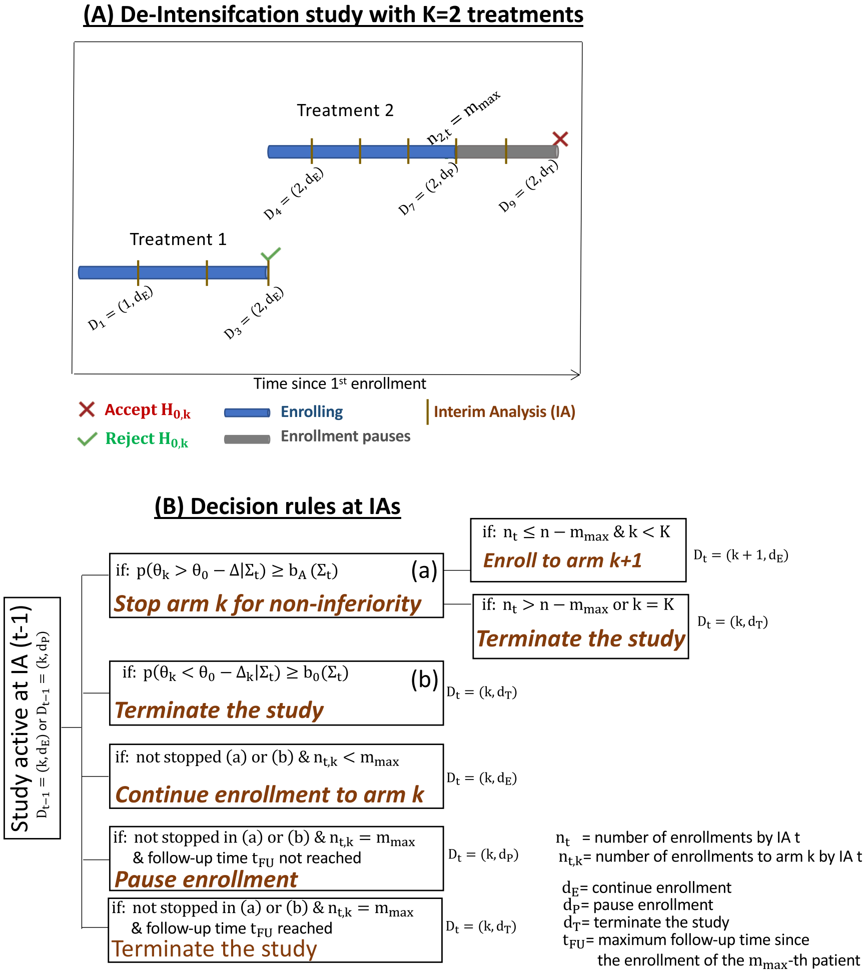
(A) De-Intensifcation study with K=2 treatments
Treatment 2
$n_{2,t} = m_{max}$
$D_4 = (2, d_E)$
$D_7 = (2, d_P)$
$D_9 = (2, d_T)$
Treatment 1
$D_1 = (1, d_E)$
$D_3 = (2, d_E)$
Time since 1st enrollment
Accept $H_{0,k}$
Reject $H_{0,k}$
Enrolling
Enrollment pauses
Interim Analysis (IA)
(B) Decision rules at IAs
Study active at IA (t-1)
$D_{t-1} = (k, d_E)$ or $D_{t-1} = (k, d_P)$
if: $p(\theta_k > \theta_0 - \Delta|\Sigma_t) \geq b_A(\Sigma_t)$ (a)
Stop arm k for non-inferiority
if: $n_t \leq n - m_{max}$ & $k < K$
Enroll to arm k+1
$D_t = (k+1, d_E)$
if: $n_t > n - m_{max}$ or $k = K$
Terminate the study
$D_t = (k, d_T)$
if: $p(\theta_k < \theta_0 - \Delta_k|\Sigma_t) \geq b_0(\Sigma_t)$ (b)
Terminate the study
$D_t = (k, d_T)$
if: not stopped (a) or (b) & $n_{t,k} < m_{max}$
Continue enrollment to arm k
$D_t = (k, d_E)$
if: not stopped in (a) or (b) & $n_{t,k} = m_{max}$
& follow-up time t_{FU} not reached
Pause enrollment
$D_t = (k, d_P)$
if: not stopped in (a) or (b) & $n_{t,k} = m_{max}$
& follow-up time t_{FU} reached
Terminate the study
$D_t = (k, d_T)$
n_t = number of enrollments by IA t
$n_{t,k}$= number of enrollments to arm k by IA t
d_E= continue enrollment
d_P= pause enrollment
d_T= terminate the study
t_{FU}= maximum follow-up time since the enrollment of the m_{max}-th patient

Figure 1: Schematic representation of the trial design with $K = 2$ treatments.

Stopping rules. Let $b_A(\cdot)$ be a predefined non-inferiority stopping boundary with $0 \leq b_A(\Sigma_t) \leq 1$ (see Section 2.2 for details), and $D_{t-1} = (k, d_E)$ between IA $t-1$ and t. Arm k is stopped for non-inferiority and $\mathcal{H}_{0,k}$ is rejected at IA t, if the probability of non-inferiority crosses the boundary, i.e., if

$$p(\theta_k > \theta_0 - \Delta|\Sigma_t) \geq b_A(\Sigma_t). \tag{2}$$

Conditionally on the availability of sufficient sample size, i.e., $n_t \leq n - m_{\max}$, the study then proceeds to evaluate the next treatment $k+1$, i.e., $D_t = (k+1, d_E)$. Whereas if $n_t > N - m_{\max}$, the study is terminated, $D_t = (\emptyset, d_T)$.

If the null hypothesis was not rejected and there is evidence of lack of non-inferiority, i.e., if

$$p(\theta_k \leq \theta_0 - \Delta_k | \Sigma_t) \geq b_0(\Sigma_t), \tag{3}$$

then treatment k is stopped early for inferiority and the study terminates. Here $0 \leq b_0(\Sigma_t) \leq 1$ is a pre-specified futility safety stopping boundary.

Pause enrollment. If the probabilities in (2) and (3) don't cross the stopping boundaries, then enrollment to arm k continues until the maximum enrollment per arm $m_{\max}$ is reached. When $n_{t,k} = m_{\max}$, enrollment will pause until either (i) treatment k is stopped for safety (inferiority) or non-inferior according to (3) and (2) at later times, or (ii) the maximum follow-up time t_{FU} since the last enrollment to treatment k is reached. In the latter case, the first (last) patient enrollment to arm k was followed for $t_{FU} + t_{enr}$ (t_{FU}) months or until the time of progression, whichever comes first, where $t_{enr,k}$ to indicate the time between the 1st and the $m_{\max}$-th enrollment to arm k. If the probabilities in (2) don't cross the boundaries at the FA, the study is closed and $\mathcal{H}_{0,k}$ is not rejected.

Calibration of the design thresholds

We use stopping boundaries of the form

$$b_j(\Sigma_t) = 1 - \eta_j \times \max\left[0, \frac{n_{t,k} - m_j}{m_{\max} - m_j}\right]^{\nu_j} \text{ for } t \geq 1 \text{ and } j = 0, A. \tag{4}$$

The parameter $\nu_j \geq 0$ determines the shape of $b_j(\cdot)$, which is decreasing from 1 to $1 - \eta_j \in [0, 1]$ when $\nu_j > 0$, and is constant across IAs t (such that $n_{t,l} \geq m_j$) when $\nu_j = 0$. Here $m_j, j = 0, A$, indicate the minimum number of enrollments $n_{k,t}$ to a treatment k necessary before $\mathcal{H}_{0,k}$ can be rejected ($j = A$) and before the treatment can be stopped early for inferiority ($j = 0$).

Different summary information could be utilized as an input for $b_j(\Sigma_t)$, including the cumulative number of follow-up times or the number of observed PFS events for arm k up to time t. For instance, Figure 2 summarizes simulations to compare the boundary (4) and an alternative boundary $b'_j(\cdot)$ that uses the number of PFS events by time t to summarize information by IA t in simulation tailored to our motivating study (see Section 3.1). In these simulations, we observed nearly identical type I/II error rates for both boundaries.

We fix $\{\nu_A, m_A, \eta_0, \nu_0, m_0\}$, see Supplementary Section S1 for a discussion on the selection of these parameters. We then calibrate the parameter η_A of the boundary $b_A(\cdot)$ to bound the type I error rate at a desired α level across a set $\mathcal{S}_0$ of survival distributions that satisfy $\theta(S) = \theta_0 - \Delta$ for each $S \in \mathcal{S}_0$. We use the historical control $S_{0,Y}$ (published PFS Kaplan-Meier estimator of the SOC) and select a set of transformations $S = g(S_0)$ (proportional hazards, accelerated failure time, proportional odds, etc.) such that $\theta(S) = \theta_0 - \Delta$ for each S in $\mathcal{S}_0$.

Controlling the type I error rate. For each $S \in \mathcal{S}_0$, we determine the largest value $\eta_A(S)$ that bounds the type I error rate of the design for arm 1 at level α when $S_{1,Y} = S$.

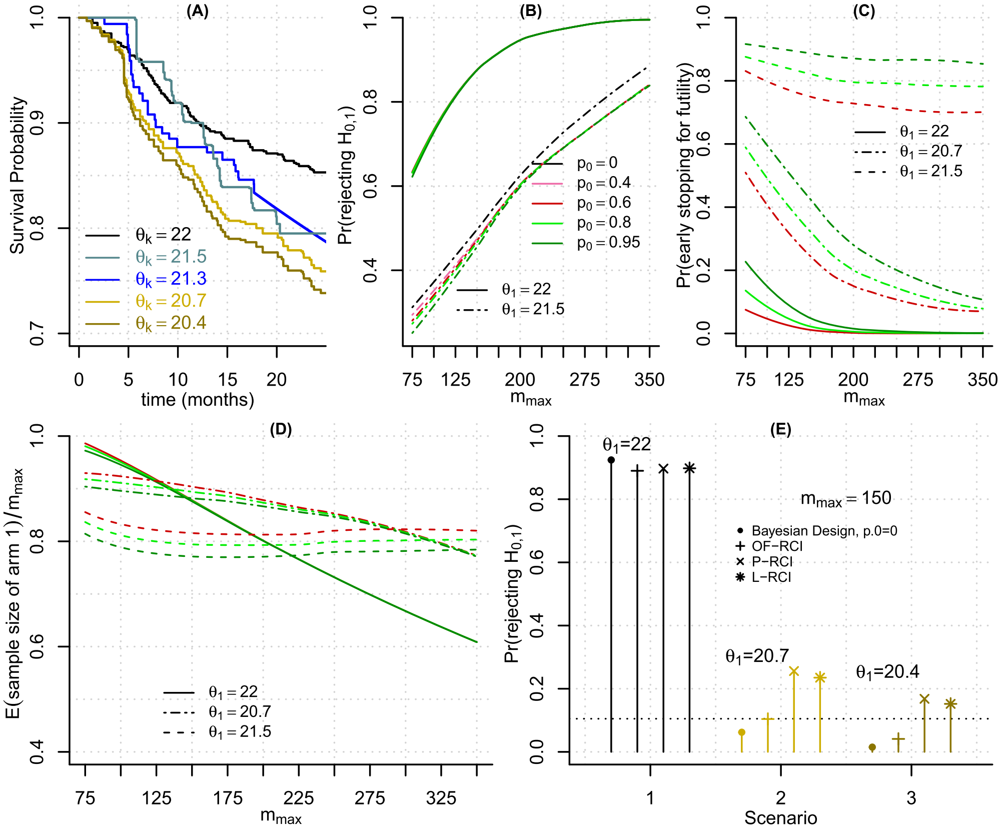

Figure 2: Panel (A) shows PFS Kaplan-Meier curves extracted from recent de-intensification studies (RTOG 1016, DeEscalate, Optima, and E1308) which we use to generate outcomes. Panels (B) to (D) show selected operating characteristics of treatment $k = 1$ for a maximum sample sizes $m_{\max} = 75, \cdots, 350$. Panel (E) shows, for $m_{\max} = 150, p_0 = 0, \nu_1 = 6, m_1 = 50$, power when $\theta_1 = 22$ (black bars), $\theta_k = 20.7$ (yellow bars) and $\theta_k = 22$ (brown bars) for the proposed Bayesian design and the three RCI methods with O'Brien-Fleming (O'Brien and Fleming, 1979), Pocock (Pocock, 1977) and linear spending functions (Reboussin et al., 2000) (OF-RCI, P-RCI and L-RCI).

We then set $\eta_A = \min_{S \in \mathcal{S}_0} \eta_A(S)$. Relevant operating characteristics, such as power and the average study duration, depends on the selected $\{\nu_A, m_A, \eta_0, \nu_0, m_0\}$. We discuss the selection of these parameters In Supplementary Section S1.

Calibration. We estimate $\eta_A(S)$ using a Monte-Carlo (MC) procedure, by simulating C trials (we use $C = 2000$ in Section 3) with individual outcomes generated from S and random enrollment times of $m_{\max}$ patients with a fixed enrollment rate. For each simulation $c = 1, \ldots, C$, we compute the number of enrollments $n_t^{(c)}$ by IA $t \geq 1$ and the posterior probabilities $p_{0,t}^{(c)}$ in (3) and $p_{A,t}^{(c)}$ in (2). Since $p_{A,t}^{(c)} \geq b_A(n_t^{(c)})$ in (2) is equivalent to $\eta_{A,F} \leq \eta_t^{(c)}$, where

$$\eta_t^{(c)} = \left(1 - p_{A,t}^{(c)}\right) \Big/ \max\left[0, \frac{n_t^{(c)} - m_{\min}}{m_{\max} - m_{\min}}\right]^{\nu_A}, \tag{5}$$

the simulated trial c does not reject the null hypothesis $\mathcal{H}_{0,1}$ at time t, or at any other interim analysis, if $\eta_A(S)$ is larger than $\eta^{(c)} = \min \eta_t^{(c)}$ where the minimum is over all t such that $p_{0,t'}^{(c)} < b_0(n_{t'}^{(c)})$ for $t' \leq t$. We then estimate $\eta_A(S)$ as the α-percentile of $\{\eta_A^{(c)}\}_{c=1}^{C}$.

Multiplicity Adjustments. There is an ongoing debate in the literature about the need for multiplicity-adjustments in multi-arm clinical studies (Wason et al., 2014; Proschan and Waclawiw, 2000). In our setting, where the θ_k's are assumed to be non-increasing with $k = 1, \ldots, K$, i.e., $\theta_k \geq \theta_{k+1}$, it can be shown that the family-wise type I error rate (FWER) of the design is bounded by $FWER \leq \frac{\alpha(1-\alpha^K)}{1-\alpha} < \frac{\alpha}{1-\alpha}$, where α is the treatment-specific maximum type I error rate. For instance, the FWER ≤ 0.112 (0.0527) for $\alpha = 0.1$ (0.05). In confirmatory trials, where regulatory restrictions may require the control of the FWER at level $\tilde{\alpha}$, one can set the arm-specific type-I error levels α equal to $\alpha = \tilde{\alpha}/(1+\tilde{\alpha})$ to bound the FWER below $\tilde{\alpha}$. In Sections 3.1 and S2, we evaluate our design and alternative designs when treatment-specific (Section 3.1) and family-wise (Section S2) control of the type I error rate is desirable.

2.3 Efficacy and toxicity co-primary outcomes

When little is known about AEs of the de-intensified treatments, it becomes necessary to evaluate them together with efficacy as co-primary endpoints.

Probability Model. Recall that X_i indicates the time (months since enrollment) of the first grade ≥ 3 AE for patient i. The joined probability model for Y_i and X_i given $C_i = k$ is specified via a probability distribution for X_i and for $Y_i|X_i$. The survival distributions $S_{k,X}(\cdot) = P(X_i > \cdot|C_i = k), k = 1, \ldots, K$ follow again independent BS prior probabilities, $S_{k,X} \sim BS(V_X, c_X)$, with mean distribution V_X and precision function c_X. For the conditional survival function of $Y_i|X_i, C_i = k$, we use a proportional hazards model

$$-\log\{S_{k,Y|X=x}(t)\} = \int_0^t h_k(z) \exp\{\gamma_k I(z > x)\} dz, \tag{6}$$

with treatment-specific baseline hazard rate $h_k(\cdot)$ for patients that do not experience an AE, and the effect $\gamma_k \geq 0$ of an AE at time $X_i = x < +\infty$ on the hazard rate. For γ_k we use a truncated normal prior $\gamma_k \sim N(0, \sigma_\gamma^2) I(\gamma_k \geq 0)$. The hazard rate $h_k(t)$ is approximated by a piecewise-constant model, $h_k(z) = \sum_{j=1}^{J} h_{kj} I_{(z_j, z_{j+1})}(t)$, where $z_j < z_{j+1}$, with independent Gamma priors for the hazard rate $h_{kj} \sim Ga(v_Y, v_Y/h_j)$ of interval $(z_j, z_{j+1}]$ with prior mean $E[h_{kj}] = h_j$ and variance h_j^2/v_Y.

Efficacy summaries. We use β_k to indicate a toxicity summary. Small values of β_k indicate high toxicity. Examples for β_k include the median, the mean $\beta_k = E[X_i|C_i = k]$, the RMST $\beta_k = E[\min(X_i, t_T)|C_i = k]$ and the probability $\beta_k = P[X_i > t_T|C_i = k]$ of no AE by time t_T for pre-specified $t_T > 0$. In Section 4.2, we consider the RMST.

We consider the null and alternative hypotheses

$$\begin{aligned}\mathcal{H}_{0,k} &= \{(\theta_k, \beta_k) \in \mathbb{R}^2 : \theta_k \leq \theta_0 - \Delta \text{ or } \beta_k \leq \beta_0\}, \text{ and} \\ \mathcal{H}_{A,k} &= \{(\theta_k, \beta_k) \in \mathbb{R}^2 : \theta_k > \theta_0 - \Delta_k \text{ and } \beta_k > \beta_0 + \Delta_\beta\},\end{aligned} \tag{7}$$

where $\Delta_\beta > 0$, and extend the design in Section 2.2 to include toxicity outcomes.

Stopping rules. At IA $t \geq 1$, treatment k is declared non-inferior and less toxic than the SOC and the null hypothesis is rejected if the posterior probability of $\mathcal{H}_{0,k}$ becomes smaller than a pre-defined threshold $b_A(\cdot)$, i.e. when

$$p\big(\{\theta_k > \theta_0 - \Delta\} \cap \{\beta_k > \beta_0\}|\Sigma_t\big) \geq b_A(\Sigma_t). \tag{8}$$

Here the boundary $b_A(\cdot)$, as well as the safety boundaries $b_T(\cdot)$ and $b_0(\cdot)$ introduced below, belong to the same family as (4). When the null hypothesis is rejected, the trial starts enrolling patients to arm $k+1$ if the sample size $n_t \leq N - m_{\max}$; otherwise, the study terminates.

Toxicity safety rules. If $\mathcal{H}_{0,k}$ was not rejected at IA t, then treatment k can be stopped early due to safety concerns, i.e., insufficient reductions of AEs or inferior survival. Specifically, treatment k will be dropped at IA t due to insufficient early evidence of toxicity reductions if the posterior probability of the event $\{\beta_k \leq \beta_0 + \Delta_\beta\}$ exceeds the toxicity boundary b_T, i.e. $p\big(\beta_k \leq \beta_0 + \Delta_\beta|\Sigma_t\big) \geq b_T(\Sigma_t)$. If, at this point, (i) $n_t \leq n - m_{\max}$ and (ii) treatment k shows sufficiently evidence for non-inferiority, i.e. $p\big(\theta_k \geq \theta_0 + \Delta|\Sigma_t\big) \geq b_C$ for a pre-specified $b_C \in [0,1]$, then the trial starts enrollment to treatment $k+1$. The study then evaluates if a further de-intensification of treatment $k+1$ reduces treatment-related AEs and has non-inferior survival. Otherwise, the study terminates.

Efficacy safety rules. If $\mathcal{H}_{0,k}$ was not rejected, and $p\big(\beta_k \leq \beta_0 + \Delta_\beta|\Sigma_t\big) < b_T(\Sigma_t)$, then therapy k can be stopped for inferiority, and the study terminates, if the posterior probability $p\big(\theta_k \leq \theta_0 - \Delta_k|\Sigma_t\big) \geq b_0(\Sigma_t)$ becomes larger than the boundary $b_0(\Sigma_t)$. The margin $\Delta_k = \Delta_k(\Sigma_t)$ now depends on the current evidence of toxicity reductions and can vary during the study. In particular, with low evidence of toxicity reductions, we use a smaller margin than in presence of strong evidence,

$$\Delta_k(\Sigma_t) = \Delta - (\Delta - \Delta_L) \times \max\left(0, \frac{p\big(\beta_k \leq \beta_0 + \Delta_\beta|\Sigma_t\big) - p_\beta}{1 - p_\beta}\right)^{\nu_\beta}, \tag{9}$$

where $p_\beta \in [0,1]$ and $\Delta_L, \nu_\beta \geq 0$ are fixed design parameters such that $\Delta \geq \Delta_L$. The function $\Delta_k(\Sigma_t) \in [\Delta_L, \Delta]$, is constant and equal to $\Delta_k(\Sigma_t) = \Delta_L$ when $\nu_\beta = 0$. It is equal to $\Delta_k(\Sigma_t) = \Delta$ when $p_\beta = 0, \nu_\beta > 0$. While it is monotone decreasing from Δ to Δ_L otherwise.

Pause enrollment. Similar to Section 2.2, if the posterior probabilities don't cross the stopping boundaries, the study continues enrollment to arm k until a maximum of

$m_{\max}$ enrollments to arm k are reached. If $n_{t,k} = m_{\max}$, enrollment is paused, and IAs are conducted until the maximum follow-up time t_{FU} is reached.

Calibration. Supplementary Section S3.2 outlines a modification of the calibration algorithm detailed in Section 2.2 to tune the rejection boundary $b_A(\cdot)$ in (8) to approximately bound that treatment-arm specific type I error rates of the design at a target level α for a set of reference scenarios. We developed our Bayesian design for applications in non-confirmatory Phase II clinical settings where regulatory agencies typically do not require the control of FWERs. In confirmatory settings, regulatory restrictions may demand the control of the FWER at level $\tilde{\alpha} > 0$. For such settings, we can specify arm-specific type I error levels $\alpha_k \geq 0$ such that $\tilde{\alpha} = \sum_{k=1}^{K} \alpha_k$, and tune individually the parameters $\eta_{A,k}$ of arm-specific rejection boundaries for type I error levels α_k.

3 Application in HPV-associated oropharynx cancer

We evaluate the de-intensification designs with efficacy outcomes (Sections 3.1) and with efficacy and toxicity co-primary outcomes (Sections 3.2) in HPV-associated oropharynx cancer. Section S1 discusses the sensitivity of the operating characteristics of our design for different values of the parameters $\{\nu_A, m_A, \eta_0, \nu_0, m_0\}$ and different prior parameters.

3.1 De-escalation design with efficacy endpoint

Data. We apply the design of Section 2.2 to four studies in HPV-associated oropharynx cancer. We extracted published PFS distributions (Figure 2(A)) from the de-intensification studies RTOG 1016, DeEscalate, Optima and E1308 (Gillison et al., 2019; Mehanna et al., 2019; Seiwert et al., 2019; Marur et al., 2017) using the software DigitizeIt (I. Bormann, 2018). RTOG 1016 is a large randomized phase III study (849 patients); the remaining studies were smaller single-arm studies. The IMRT+*Cisplatin* SOC (black curve) has a 24-month RMST of $\theta_0 = 22$ months.

Simulation Setup. We consider a study with two de-intensified therapies, an average of 5 enrollments per month, monthly IAs, $t_{FU} = 12$ months follow-up time after the last enrollment, and the null hypotheses $\theta_k \leq 20.7$ $(\Delta = 1.3)$ are tested at an 0.1 significance level. The last patients will be enrolled to arm k approximately $t_{enr} = 30$ months after the 1st assignment to arm k, and the final analysis will be conducted approximately $42 = 30 + 12$ months after the 1st enrollment to arm k. At this time point, about 90 of the 150 enrolled patients have been followed for the full 24 months or until disease progression (whichever occurred first). Patients that did not progress within the first $t_E = 24$ months also contribute to the posterior distribution of $\theta_k = \int_0^{t_E} S_{Y,k}(z)dz$ since the posterior of $S_{k,Y}(z)$, for $z \in [0, t_E]$, depends on the data through the number of events at time z and the number of patients that progressed after time z.

For the Bayesian design we used $(\nu_A, \nu_0, m_A, m_0) = (6, 5, 50, 0)$. In each scenario we consider below, we used a different pair of distributions from Figure 2(A), to sample PFS outcomes for the 1st and 2nd treatment. Blue and black survival functions have RMSTs, which are non-inferior to the SOC. The remaining two survival functions in yellow and brown have inferior RMSTs.

Sample size determination. We initially determined a sample size $m_{\max}$ for the first experimental arm to achieve $\approx$ 90% power when $S_{1,Y} = S_{0,Y}$, see Figure 2(B). For each candidate $m_{\max}$, we tuned the safety stopping boundary such that under $\mathcal{H}_{0,1}$, when $\theta_1 = 20.7$, the study would stop with predefined probability p_0 early due to inferiority. Supplementary Section S3.1 outlines our tuning algorithm. The power shows little sensitivity to the choice of p_0 when $\theta_1 = 22$, whereas for $\theta_1 = 21.5$, the power varies substantially with $p_0 > 0.6$. Based on Figure 2(B), we select $(m_{\max}, p_0) = (150, 0.7)$.

Supplementary Table S3, reports additional simulations for different combinations (t_{FU}, t_E) of the restriction time t_E of the RMST and the time t_{FU} between the last enrollment and final analysis. Follow-up times $t_{FU} > 12$ only slightly increase power; we selected $t_{FU} = 12$ as a compromise between power and the average study duration. For general recommendation on the selection of (t_{FU}, t_E) we refer the reader to (Royston and Parmar, 2013; Nemes et al., 2020; Tian et al., 2020).

Comparator designs. We compare the Bayesian design to alternative de-intensification designs with different combinations of testing and futility-stopping rules (Jennison and Turnbull, 1989; O'Brien and Fleming, 1979; Pocock, 1977; Reboussin et al., 2000). Similar to our Bayesian design, these designs may declare a treatment k non-inferior at an IA or FA, and start evaluating arm $k + 1$, or declare treatment k inferiority and stop the study. Noninferiority IAs are conducted monthly for all designs, starting after $m_{\min} = 50$ enrollments.

Non-inferiority is tested in the comparator designs using the *RCI* method (Jennison and Turnbull, 1989; Gillison et al., 2019) based on O'Brien-Fleming (O'Brien and Fleming, 1979), Pocock (Pocock, 1977) and linear (Reboussin et al., 2000) error-spending functions (OF-RCI, P-RCI, and L-RCI, see Section S4 for details). We consider three frequently used rules for futility IAs (F1, F2 and F3) in the comparator designs. (1) At each IA t we compute a p-value for the ''null" hypothesis $(\theta_k \geq 22)$ using a normal approximation for the distribution of $\widehat{\theta}_k$, and stop the study if the p-value ≤ 0.0025 as suggested in (Freidlin et al., 2010; Gillison et al., 2019). (2) Alternatively, (Lachin, 2009) suggested a p-value ≤ 0.05 cut-off. (3) The last rule Freidlin et al. (2010) stops early if the $(1 - \alpha_t)$ confidence interval $(-\infty, \widehat{U}_t]$ for θ_k doesn't include $\theta_0 = 22$, with overall $\sum_t \alpha_t = 0.025$.

Operating characteristics for the 1st treatment. We first compared type I error rates and the power of the designs for the first de-intensified therapy $(k = 1)$ in three scenarios with $\theta_1 = 20, 20.7, 22$ and $m_{\max} = 150$ patients (Panel E of Figure 2). To simplify the evaluation, we don't consider interim futility analyses in these three scenarios. RCIs with Pocock (Pocock, 1977) and linear (Reboussin et al., 2000) spending functions do not control the type I error rate at the targeted $\alpha = 0.1$ level with empirical type I error rates of 0.26 and 0.23 across 10,000 simulations. O'Brien-Fleming boundaries have type I error rates nearly identical to the nominal α level. Figure S1(A) shows that the normal approximation of the RMST estimates $\widehat{\theta}_k$ in the *RCI* is not accurate for the initial IAs, which leads to these inflated error rates. The approximation becomes better towards the end of the study (Panel B). O'Brien-Fleming boundaries are significantly more conservative during the initial IAs than the linear and Pocock's boundaries. Hence, they are less affected by these approximation errors. We, therefore, use O'Brien-Fleming

RCI boundaries for the remaining comparisons, in combination with each of the three futility-stopping rules described above (RCI-F1, RCI-F2, and RCI-F3).

Tables S4 and S5 report simulations for longer follow-up and restriction times $(t_{FU}, t_E) = (24, 36)$. We observed a similar relative performance of the designs compared to simulations with $(t_E, t_{FU}) = (24, 12)$. The Bayesian and OF design had the highest power (Table S4 scenarios 1 and 2, Bayesian: 83% and 77%, OF-RCI: 81% and 74%), P-RCI and L-RCI had inflated type I error rates.

Operating characteristics for $K = 2$ treatments. Figure 3(A) summarizes the eight scenarios we consider in the two-arm study. For each scenario (x-axis), the two vertical bars indicate the RMSTs θ_1 and θ_2 that we consider with distributions $S_{1,Y}$ and $S_{2,Y}$ selected from Figure 2 (same colors). Treatments $k = 1, 2$ are non-inferior in the first three scenarios, whereas the second de-intensified treatment is inferior in the last four scenarios.

In scenario 1, both de-intensified treatments are non-inferior to the SOC with identical RMSTs $\theta_0 = \theta_1 = \theta_2$. The Bayesian design has 94% and 88% power to declare the two treatments non-inferior, compared to 89% and 80% for the RCI-F1 design, Figure 3(B). The remaining two designs RCI-F2 and RCI-F3 have lower power (74% and 54% for RCI-F2 and 75% and 57% for RCI-F3), respectively. The power in Figure 3(B) for the second experimental arm is defined as the probability that the study starts testing treatment two and rejects $\mathcal{H}_{0,2}$ at final or IAs. Panel C shows, for both treatments $k = 1, 2$, the probability that the study started testing treatment k and stopped treatment k early for futility at IAs (solid vertical bar). For the 2nd treatment, we also show the probability that the study does not start testing the therapy (dashed vertical bar). For instance, for the Bayesian design in scenario 5, the inferior 2nd treatment is not tested or stopped early for futility with a probability of 0.95. Here, the study does not start testing the therapy with a probability 0.40 and is stopped early for futility with a probability 0.45. Panel C shows that the futility-stopping rule of RCI-F1 leads to a low probability of stopping inferior treatments early (scenario 8, $\theta_1 = 20.7$) early for futility (38%) compared to the RCI-F2 (93%) and RCI-F3 (53%) and the proposed Bayesian design (81%). This leads in scenarios 5 and 7, where the first de-intensified treatment is non-inferior, but θ_1 is close to $\theta_0 - \Delta = 20.7$ ($\theta_1 = 21.5$ and 21.3), to a slightly larger power of the RCI-F1 compared to the remaining designs.

Scenarios 4 to 8, where the 2nd treatment is inferior to the SOC, show the benefit of testing experimental arms sequentially. For instance, if the first experimental arm is inferior ($\theta_1 = 20.7$, scenario 8), all designs start testing the inferior 2nd treatment with less than 10% probability (10% for RCI-F1, RCI-F3, and the Bayesian design, and 7% for RCI-F2).

Supplementary Figure S3 reports operating characteristics for two additional scenarios (Scenarios 9 and 10) when both de-intensified treatments are inferior to the SOC. Across all ten simulation scenarios, we observed a maximum FWER of 11%, 10%, 10%, and 6% (in Scenario 9) for RCI-F1, RCI-F3, RCI-F3, and the Bayesian design (recall that both de-intensified treatments are inferior to the SOC in Scenarios 8-10). Supplementary Section reports the results of a simulation study nearly identical to the one

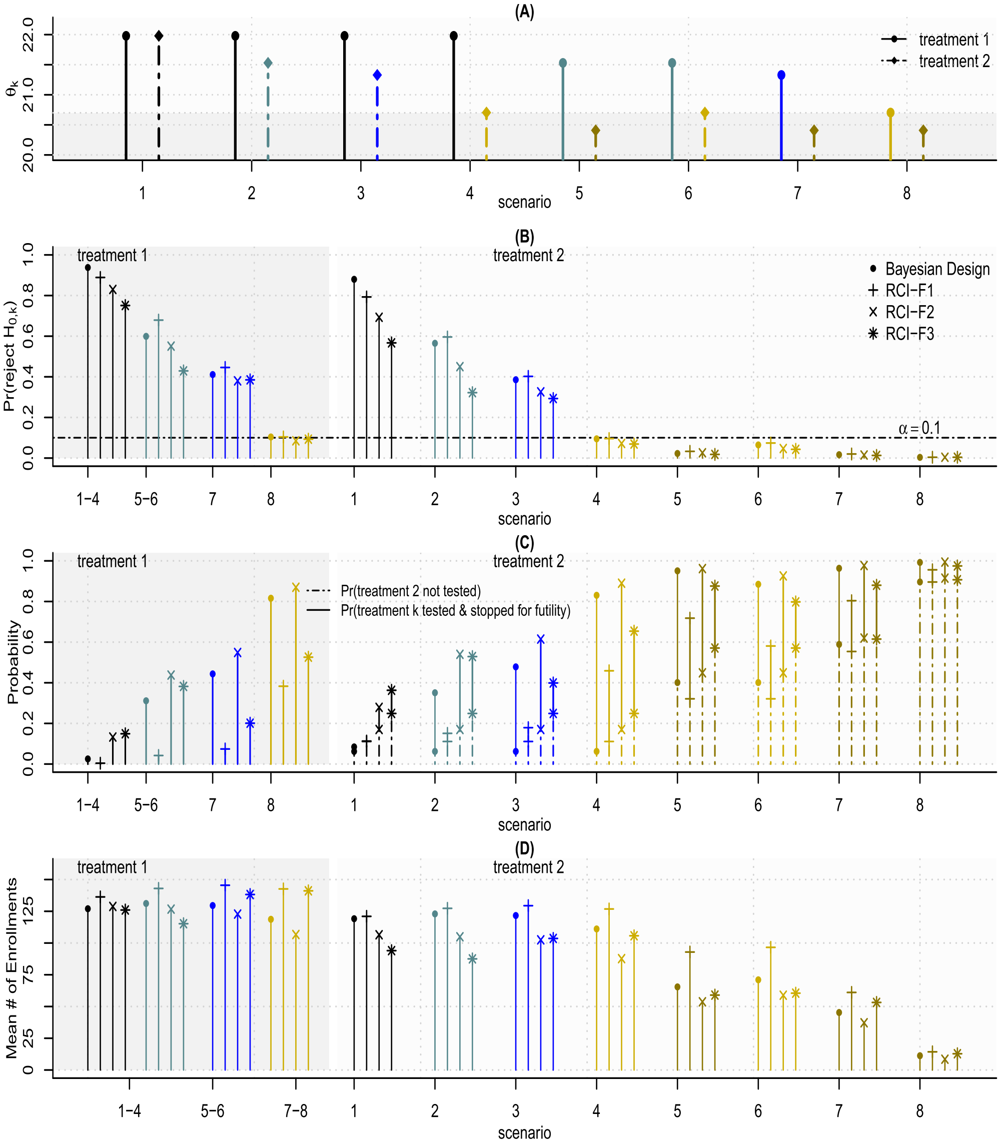

Figure 3: Operating characteristics of the de-intensification designs for a two-arm study with efficacy primary endpoint and a maximum of $m_{\max} = 150$ patients for each treatment. Panel A summarizes the 24-month RMSTs θ_1 and θ_2 (y-axis) in each of the eight scenarios (x-axis) that we consider. Panel B shows the power of the Bayesian design and the three alternative designs (RCI-F1, RCI-F3, and RCI-F3). Panel C shows the probability of stopping treatments 1 and 2 at an IA for futility (solid vertical line), and the probability that the 2nd treatment is not tested due to early termination of the study (dashed vertical line). Panel D shows the average enrollment of the two-arms trial on the 1st and 2nd de-intensified treatments.

presented here for settings where regulatory restrictions require the control of the FWER level $\tilde{\alpha} = 0.1$, using arm-specific type I error levels $\alpha = 0.1/(1 + 0.1)$ for testing. We observed a similar relative performance for our design compared to alternative designs.

3.2 Using efficacy and toxicity co-primary outcomes

Testing. We consider testing non-inferior survival and toxicity reductions. We assume the same enrollment rate (5/month), $t_{FU} = 12$ months, 24-month RMST of $\theta_0 = 22$ for the historical SOC, and a 24-month RMST for the 1st grade ≥ 3 AE of $\beta_0 = 12.5$ months (estimated from data of RTOG-1016). We are testing $\mathcal{H}_{0,k} = \{(\theta_k, \beta_k) \in \mathbb{R}^2 : \theta_k \leq 20.7 \text{ or } \beta_k \leq 12.5\}, k = 1, 2$ at levels $\alpha_1 = \alpha_2 = 0.05$ to control the FWER at level $\tilde{\alpha} = 0.1$.

Prior distribution. The BS-prior for $S_{k,X}$ was centered at an exponential model with a 24-month RMST of 12.5. For the conditional efficacy survival distribution, we used a piecewise constant hazard model with $J = 5$ intervals, $(z_0, \ldots, z_5) = (0, 6, 12, 18, 24, 80)$, and parameters $\sigma_\lambda = v_Y = 10$ for the prior precision of the baseline-hazard (v_Y) and the prior standard deviation (σ_λ) of for the effect of toxicity on survival. See Supplementary Section S1.1 for recommendations on selecting these prior parameters.

Outcome scenarios and design parameters. We consider scenarios with efficacy distributions $S_{k,Y}$ (and θ_k) identical to Kaplan-Meier curves Figure 2(A) and exponential distributions $S_{k,X}$ for the time X_i until the 1st grade ≥ 3 AE (jointly generated from a gaussian copulas (Nelsen, 2006) with correlation 0.5) with 24-month RMST equal to $\beta_k = 12.5$ or 14.5 months. We used $(\Delta, \Delta_L, \Delta_\beta) = (2, 1, 0), \nu_\beta = 2, p_\beta = 0$ in (9), and $\nu_A = \nu_0 = \nu_T = 6$ for the shape parameters of the boundaries b_A, b_0, b_T in (4) and required 75 assignments before applying (toxicity, futility, and non-inferiority) early stopping rules. We tuned the parameters η_j for b_0, b_T in (4) so that with probability 0.5 inferior treatments or treatments that do not reduce toxicities are stopped early at IAs.

Sample size determination. We determined for the 1st arm the power of the Bayesian design with a maximum arm-specific sample size $m_{\max}$ (Figure 4(A)) when $\theta_1 = \theta_0$ and $\beta_1 = \beta_0 + 2$. With a $\alpha = 0.05$ significance level, the design requires ≈ 250 patients to achieve 90% power, respectively. We then considered a two-arm de-intensification study ($K = 2$) with a maximum overall sample size per arm of $m_{\max} = 250$ patients. We evaluated the operating characteristics of the proposed Bayesian design in 8 scenarios that are summarized in Figure 4(C). PFS parameters θ_k are represented by vertical bars in Panel C (the bars and colors for θ_1 and θ_2 are consistent with $S_{k,Y}$'s in Figure 2(A)). Toxicity parameters are indicated by green triangles ($\beta_k = 14.5$) and red stars ($\beta_k = 12.5$) on top of the vertical bars.

Results. Figure 4(B) shows the benefit of interim monitoring of efficacy and toxicity endpoints. The figure shows, for the first treatment, the cumulative probability of stopping the therapy for futility (y-axis), i.e. for inferiority or toxicity, by time t since the first enrollment (x-axis) for four scenarios (combinations of $\theta_1 = \theta_0 - \Delta, \theta_0$ and $\beta_1 = \beta_0, \beta_0 + 2$). For instance, if the treatment is inferior with $\theta_1 = \theta_0 - \Delta = 20.7$ but reduces toxicities ($\beta_1 = \beta_0 + 2$), then 56% of all simulated de-intensification trials are stopped early for futility at IAs (scenario 8, dashed golden curve). In comparison, if

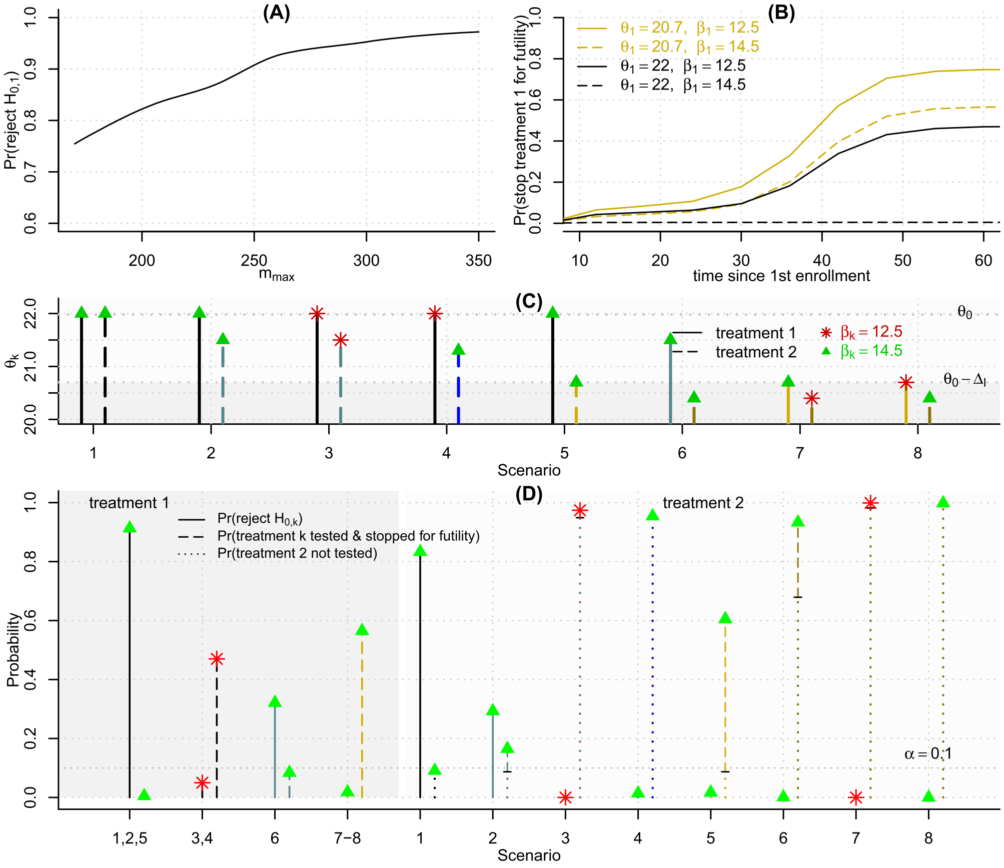

Figure 4: Operating characteristics of Bayesian de-intensification design for a study with efficacy and toxicity co-primary endpoints. Panel A shows the power for treatment $k = 1$ when $\theta_1 = \theta_0$ and $\beta_1 = \beta_0 + 2$ with maximum sample size $m_{\max}$ between 170 and 350 patients. Panel B shows the probability of stopping treatment $k = 1$ for futility (either for inferior survival or low evidence of reduced toxicities) when $\theta_1 = \theta_0$ (black curves) and $\beta_1 = \beta_0, \beta_0 + 2$ (red cross and green triangles) and when $\theta_1 = \theta_0 - \Delta = 20.7$ (yellow curves) and $\beta_1 = \beta_0, \beta_0 + 2$ (red cross and green triangles) for a study with maximum sample size $m_{\max} = 250$ patients. Panel C summarizes the 24-month RMSTs $(\theta_k, \beta_k), k = 1, 2$ in each of the eight scenarios (x-axis) that we consider. The vertical bars (y-axis) indicate θ_k, whereas green arrows ($\beta_k = 14.5$) and red stars ($\beta_k = 12.5$) on top of the vertical bars indicate toxicity parameters. Panel D shows, for both treatments, the power (solid bar), the probability of stopping treatment evaluation early for futility at IAs (dashed bars), and the probability that the 2nd treatment is not tested due to early termination of the study (dotted bars).

the treatment fails to reduce toxicities ($\theta_1 = 20.7$ and $\beta_1 = \beta_0$), then the treatment is stopped early for futility in 75% of the simulations (scenarios 7, solid golden curve).

Figure 4(D) shows for both treatments the probability of rejecting $\mathcal{H}_{0,k}$ (power, solid vertical bars), and the probability that the study evaluates treatment k and stops this arm early for futility (inferior survival or insufficient reduction of toxicities) at IAs (dashed vertical bars). As before, the power for the 2nd treatment is defined as the probability that the study starts testing treatment 2 and rejects $\mathcal{H}_{0,2}$. For the 2nd treatment, Panel D also shows the probability that the 2nd treatment is not tested due to the early termination of the study (dotted vertical bars). If both treatments ($k = 1, 2$) improve the 24-month RMST β_k by two months ($\beta_1 = \beta_2 = 14.5$) compared to the SOC and have identical survival outcomes ($\theta_1 = \theta_2 = 22$) the 1st and 2nd de-intensified treatment have 91% and 83% power (scenario 1). In comparison, with a moderate non-inferior treatment effect of $\theta_k = 21.5$ in scenario six, the power for the 1st treatment decreases to 36%. Similar to the Bayesian design with efficacy primary endpoint of Section 2.2, Scenarios 7 and 8 indicate the advantage of testing the 1st and 2nd treatment sequentially one after the other. If the first treatment is inferior, the second treatment $k = 2$ is tested in 2% (scenario 7, $\beta_1 = 14.5$) or in less than 1% (scenario 8, $\beta_1 = 12.5$) of all simulations.

4 Discussion

There has been a recent interest in developing de-intensified treatments with similar survival rates and reduced AEs as the current SOC. (Elrefaey et al., 2014; Mirghani and Blanchard, 2018) identified 12 de-intensification studies in oropharyngeal cancer that are currently ongoing or recently reported results.

Compared to traditional superiority trials, which test the superiority of experimental treatments compared to the SOC, demonstrating similarity in survival between de-intensified treatments and the SOC and reductions in AEs require large sample sizes. Investigators often select large NI margins to reduce the size of the study (Ventz et al., 2019). Recent results in oropharyngeal cancer showed that many de-intensification treatments fail to reduce toxicities and have inferior survival compared to the SOC (Gillison et al., 2019; Mehanna et al., 2019). As discusses in Ventz et al. (2019), many of these studies (i) evaluate only survival or toxicity, (ii) do not have explicit futility early stopping rules for survival and toxicity endpoints and (iii) tend to use conservative early stopping rules to avoid power reductions.

Motivated by an oropharyngeal cancer study that tests two dose-reduced treatments, we proposed a Bayesian design for multi-arm de-intensification trials. Using a Bayesian semi-parametric model, our design sequentially tests non-inferior survival and toxicity reductions. We proposed early safety and non-inferiority decision rules to monitor both endpoints. The design parameters can be tuned to calibrate trade-offs between power and the probabilities of stopping treatments early due to inferior survival or insufficient evidence of toxicity reductions. We defined decreasing non-inferiority stopping boundaries that encourage the rejection of the null hypothesis $\mathcal{H}_{0,k}$ for a non-inferior experimental treatment k towards the end of the study when the data for arm k are mature.

Since de-intensified treatments are typically dose-reduced versions of SOC treatments, there is a high likelihood that these treatments might be inferior to the SOC. Examples include the recent RTOG1016 (Gillison et al., 2019) and De-ESCALaTE (Mehanna et al., 2019) studies in oropharynx cancer, which tested treatments that were strongly inferior to the SOC and showed no evidence of reductions in toxicities. We, therefore, defined stringent early safety-stopping rules to limit the risk of exposing patients to potentially toxic or ineffective treatments when there is sufficient early evidence for inferiority. In oropharynx cancer, where survival rates five years after *IMRT+cisplatin* treatment are $> 90\%$, the number of OS and PFS events during the trial are typically small. Standard frequentist methods based on large-sample normal approximations can perform poorly in this setting, and the Bayesian approach is an attractive alternative.

Our design tests de-intensified treatments one at a time, starting with the treatment with the highest dose level. This controls the number of patients exposed to inferior treatments. The aim of our motivating de-intensification study was to recommend all treatments that show strong evidence of non-inferior survival and reduce the risk of AEs. Oncologists can then further personalize patient care and select one of the recommended treatments based on additional patient characteristics such as performance status, smoking history, additional markers, and patient preferences. Consequently, our design recommends a set of $0 \leq K_{NI} \leq K$ treatments as an alternative to the SOC at the end of the study. If only a single treatment should be recommended, one can specify a treatment selection rule (see for instance (Domenicano et al., 2019; Houede et al., 2010; Lee et al., 2015; Lin et al., 2021; Ventz et al., 2017)) to select a final de-intensified treatment based on the available data Σ at the end of the study. Moreover, we focused on non-controlled phase II de-intensification studies, but the design could be modified to include a concurrent control arm. If the SOC is associated with severe toxicities, continuing the assignment of patients to the SOC may be unethical after the null hypothesis $\mathcal{H}_{0,1}$ has been rejected. In this case, the 1st de-intensified treatment could be utilized as the *new control arm* for testing the 2nd de-intensified experimental treatment.

Acknowledgments

The authors would like to thank the anonymous referees, the Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

SV was supported by the National Cancer Institutes (5P30CA077598-23), a DSI Grant of the Minnesota Supercomputing Institute, a Medtronic Faculty Fellowship, and partial support by the National Institutes of Health (1R01LM013352-01A1).

LT has been supported by the National Institutes of Health (NIH Grant 1R01LM013352-01A1).

Supplementary Material

Supplementary Material for “Bayesian Multi-Arm De-Intensification Designs” (DOI: 10.1214/24-BA1417SUPP; .pdf). The file contains additional simulation results and descriptions of the calibration algorithms that we used for tuning of the trial design parameters.

References

Ang, K. K., Harris, J., Wheeler, R., Weber, R., Rosenthal, D. I., Nguyen-Tan, P. F., Westra, W. H., Chung, C. H., Jordan, R. C., Lu, C., et al. (2010). "Human papillomavirus and survival of patients with oropharyngeal cancer." *New England Journal of Medicine*, 363(1): 24–35. 1

Bhide, S., Newbold, K., Harrington, K., and Nutting, C. (2012). "Clinical evaluation of intensity-modulated radiotherapy for head and neck cancers." *The British journal of radiology*, 85(1013): 487–494. 5

Blackwelder, W. C. (1982). "Proving the null hypothesis in clinical trials." *Cont clinical trials*, 3(4): 345–353. 2

Chan, I. S. (2003). "Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods." *Stat Med in Med Research*, 12(1): 37–58. MR1977234. doi: `https://doi.org/10.1191/0962280203sm314ra`. 2

Chen, M.-H., Ibrahim, J. G., Lam, P., Yu, A., and Zhang, Y. (2011). "Bayesian design of noninferiority trials for medical devices using historical data." *Biometrics*, 67(3): 1163–1170. MR2829252. doi: `https://doi.org/10.1111/j.1541-0420.2011.01561.x`. 3

D'Agostino, R. B., Massaro, J. M., and Sullivan, L. M. (2003). "Non-inferiority trials: design concepts and issues–the encounters of academic consultants in statistics." *Stat Med*, 22(2): 169–186. 2

Daimon, T. (2008). "Bayesian sample size calculations for a non-inferiority test of two proportions in clinical trials." *Cont clinical trials*, 29(4): 507–516. 3

Domenicano, I., Ventz, S., Cellamare, M., Mak, R., and Trippa, L. (2019). "Bayesian uncertainty-directed dose finding designs." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5): 1393–1410. MR4022818. doi: `https://doi.org/10.1111/rssc.12355`. 18

Eisbruch, A., Schwartz, M., Rasch, C., Vineberg, K., Damen, E., Van As, C. J., Marsh, R., Pameijer, F. A., and Balm, A. J. (2004). "Dysphagia and aspiration after chemoradiotherapy for head-and-neck cancer: which anatomic structures are affected and can they be spared by IMRT?" *International Journal of Radiation Oncology* Biology* Physics*, 60(5): 1425–1439. 5

Elrefaey, S., Massaro, M., Chiocca, S., Chiesa, F., and Ansarin, M. (2014). "HPV in oropharyngeal cancer: the basics to know in clinical practice." *Acta Otorh Italica*, 34(5): 299. 17

Farrington, C. P. and Manning, G. (1990). "Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk." *Stat Med*, 9(12): 1447–1454. 2

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The annals of statistics*, 209–230. MR0350949. 4

Freidlin, B. and Korn, E. L. (2002). "A comment on futility monitoring." *Cont Clinical Trials*, 23(4): 355–366. 3

Freidlin, B., Korn, E. L., George, S. L., and Gray, R. (2007). "Randomized clinical trial design for assessing noninferiority when superiority is expected." *JCO*, 25(31): 5019–5023. 2

Freidlin, B., Korn, E. L., and Gray, R. (2010). "A general inefficacy interim monitoring rule for randomized clinical trials." *Clinical Trials*, 7(3): 197–208. 12

Gamalo, M. A., Wu, R., and Tiwari, R. C. (2011). "Bayesian approach to noninferiority trials for proportions." *J of Biopharm Stat*, 21(5): 902–919. MR2823357. doi: https://doi.org/10.1080/10543406.2011.589646. 3

Gillison, M. L., Trotti, A. M., Harris, J., Eisbruch, A., Harari, P. M., Adelstein, D. J., Sturgis, E. M., Burtness, B., Ridge, J. A., Ringash, J., et al. (2019). "Radiotherapy plus cetuximab or cisplatin in human papillomavirus-positive oropharyngeal cancer (NRG Oncology RTOG 1016): a randomised, multicentre, non-inferiority trial." *The Lancet*, 393(10166): 40–50. 1, 2, 11, 12, 17, 18

Hjort, N. L. et al. (1990). "Nonparametric Bayes estimators based on beta processes in models for life history data." *The Annals of Statistics*, 18(3): 1259–1294. MR1062708. doi: https://doi.org/10.1214/aos/1176347749. 4

Holmgren, E. B. (1999). "Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained." *J of Biopharm Stat*, 9(4): 651–659. 3

Houede, N., Thall, P. F., Nguyen, H., Paoletti, X., and Kramar, A. (2010). "Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials." *Biometrics*, 66(2): 532–540. MR2758833. doi: https://doi.org/10.1111/j.1541-0420.2009.01302.x. 18

Hurvitz, S. A., Martin, M., Symmans, W. F., Jung, K. H., Huang, C.-S., Thompson, A. M., Harbeck, N., Valero, V., Stroyakovskiy, D., Wildiers, H., et al. (2018). "Neoadjuvant trastuzumab, pertuzumab, and chemotherapy versus trastuzumab emtansine plus pertuzumab in patients with HER2-positive breast cancer (KRISTINE): a randomised, open-label, multicentre, phase 3 trial." *Lancet Onc*, 19(1): 115–126. 1

I. Bormann (2018). "DigitizeIt." URL https://www.digitizeit.de 11

Jennison, C. and Turnbull, B. W. (1989). "Interim analyses: the repeated confidence interval approach." *JRSS-B*, 51(3): 305–334. MR1017201. 12

Jensen, K., Overgaard, M., and Grau, C. (2007). "Morbidity after ipsilateral radiotherapy for oropharyngeal cancer." *Radiotherapy and Oncology*, 85(1): 90–97. 5

Joshua Chen, Y. and Chen, C. (2012). "Testing superiority at interim analyses in a non-inferiority trial." *Stat Med*, 31(15): 1531–1542. MR2947525. doi: https://doi.org/10.1002/sim.5312. 2

Korn, E. and Freidlin, B. (2017). "Interim monitoring for non-inferiority trials: minimizing patient exposure to inferior therapies." *Ann of Oncology*, 29(3): 573–577. 2, 3

Lachin, J. M. (2009). "Futility interim monitoring with control of type I and II error probabilities using the interim Z-value or confidence limit." *Clinical Trials*, 6(6): 565–573. 3, 12

Laster, L., Johnson, M. F., and Kotler, M. (2006). "Non-inferiority trials." *Stat Med*, 25(7): 1115–1130. MR2225582. doi: https://doi.org/10.1002/sim.2476. 2

Lee, J., Thall, P. F., Ji, Y., and Müller, P. (2015). "Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity." *Journal of the American Statistical Association*, 110(510): 711–722. MR3367259. doi: https://doi.org/10.1080/01621459.2014.926815. 18

Levendag, P. C., Teguh, D. N., Voet, P., van der Est, H., Noever, I., de Kruijf, W. J., Kolkman-Deurloo, I.-K., Prevost, J.-B., Poll, J., Schmitz, P. I., et al. (2007). "Dysphagia disorders in patients with cancer of the oropharynx are significantly affected by the radiation therapy dose to the superior and middle constrictor muscle: a dose-effect relationship." *Radiotherapy and Oncology*, 85(1): 64–73. 5

Lin, R., Thall, P. F., and Yuan, Y. (2021). "A phase I–II basket trial design to optimize dose-schedule regimes based on delayed outcomes." *Bayesian analysis*, 16(1): 179–202. MR4194278. doi: https://doi.org/10.1214/20-BA1205. 18

Llombart-Cussac, A., Cortés, J., Paré, L., Galván, P., Bermejo, B., Martínez, N., Vidal, M., Pernas, S., López, R., Muñoz, M., et al. (2017). "HER2-enriched subtype as a predictor of pathological complete response following trastuzumab and lapatinib without chemotherapy in early-stage HER2-positive breast cancer (PAMELA): an open-label, single-group, multicentre, phase 2 trial." *Lancet Onc*, 18(4): 545–554. 1

Marur, S., Li, S., Cmelak, A. J., Gillison, M. L., Zhao, W. J., Ferris, R. L., Westra, W. H., Gilbert, J., Bauman, J. E., Wagner, L. I., et al. (2017). "E1308: phase II trial of induction chemotherapy followed by reduced-dose radiation and weekly cetuximab in patients with HPV-associated resectable squamous cell carcinoma of the oropharynx." *JCO*, 35(5): 490. 1, 11

Mathew, A. and Brufsky, A. (2017). "Less is more? De-intensification of therapy for early-stage HER2-positive breast cancer." *Lancet Onc*, 18(4): 428–429. 1

Mehanna, H., Robinson, M., Hartley, A., Kong, A., Foran, B., Fulton-Lieuw, T., Dalby, M., Mistry, P., Sen, M., O'Toole, L., et al. (2019). "Radiotherapy plus cisplatin or cetuximab in low-risk human papillomavirus-positive oropharyngeal cancer: an open-label randomized controlled phase 3 trial." *The Lancet*, 393(10166): 51–60. 1, 2, 11, 17, 18

Mirghani, H. and Blanchard, P. (2018). "Treatment de-escalation for HPV-driven oropharyngeal cancer: Where do we stand?" *Clinical and translational radiation oncology*, 8: 4–11. 2, 17

Munker, R., Purmale, L., Aydemir, Ü., Reitmeier, M., Pohlmann, H., Schorer, H., and Hartenstein, R. (2001). “Advanced head and neck cancer: long-term results of chemoradiotherapy, complications and induction of second malignancies.” *Oncology Research and Treatment*, 24(6): 553–558. 1

Nelsen, R. B. (2006). *An introduction to copulas*. Springer. MR2197664. doi: https://doi.org/10.1007/s11229-005-3715-x. 15

Nemes, S., Bülow, E., and Gustavsson, A. (2020). “A brief overview of restricted mean survival time estimators and associated variances.” *Stats*, 3(2): 107–119. 12

O’Brien, P. C. and Fleming, T. R. (1979). “A multiple testing procedure for clinical trials.” *Biometrics*, 549–556. 8, 12

Osman, M. and Ghosh, S. K. (2011). “Semiparametric Bayesian testing procedure for noninferiority trials with binary endpoints.” *J of Biopharm Stat*, 21(5): 920–937. MR2823358. doi: https://doi.org/10.1080/10543406.2010.544526. 3

Pocock, S. J. (1977). “Group sequential methods in the design and analysis of clinical trials.” *Biometrika*, 64(2): 191–199. 8, 12

Proschan, M. A. and Waclawiw, M. A. (2000). “Practical guidelines for multiplicity adjustment in clinical trials.” *Controlled clinical trials*, 21(6): 527–539. 9

Reboussin, D. M., DeMets, D. L., Kim, K., and Lan, K. G. (2000). “Computations for group sequential boundaries using the Lan-DeMets spending function method.” *Controlled clinical trials*, 21(3): 190–207. 8, 12

Rothmann, M., Li, N., Chen, G., Chi, G. Y., Temple, R., and Tsou, H.-H. (2003). “Design and analysis of non-inferiority mortality trials in oncology.” *Stat Med*, 22(2): 239–264. 2

Royston, P. and Parmar, M. K. (2013). “Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome.” *BMC medical research methodology*, 13(1): 1–15. 12

Schmidli, H., Wandel, S., and Neuenschwander, B. (2013). “The network meta-analytic-predictive approach to non-inferiority trials.” *Stat Med in Med Research*, 22(2): 219–240. MR3190655. doi: https://doi.org/10.1177/0962280211432512. 3

Seiwert, T., Foster, C., Blair, E., Karrison, T., Agrawal, N., Melotek, J., Portugal, L., Brisson, R., Dekker, A., Kochanny, S., et al. (2019b). “OPTIMA: a phase II dose and volume de-escalation trial for human papillomavirus-positive oropharyngeal cancer.” *Annals of Oncology*, 30(2): 297–302. 1, 11

Semenza, G. L. (2008). “A new weapon for attacking tumor blood vessels.” *New England Journal of Medicine*, 358(19): 2066–2067. 1

Simon, R. (1999). “Bayesian design and analysis of active control clinical trials.” *Biometrics*, 55(2): 484–487. 3

Snapinn, S. M. (2004). “Alternatives for discounting in the analysis of noninferiority trials." *J of Biopharm Stat*, 14(2): 263–273. MR2190480. doi: https://doi.org/10.1081/BIP-120037178. 3

Tian, L., Jin, H., Uno, H., Lu, Y., Huang, B., Anderson, K. M., and Wei, L. (2020). "On the empirical choice of the time window for restricted mean survival time." *Biometrics*, 76(4): 1157–1166. MR4186832. doi: https://doi.org/10.1111/biom.13237. 12

Tu, D. (1998). "On the use of the ratio or the odds ratio of cure rates in therapeutic equivalence clinical trials with binary endpoints." *J of Biopharm Stat*, 8(2): 263–282. 2

Ventz, S., Barry, W. T., Parmigiani, G., and Trippa, L. (2017). "Bayesian response-adaptive designs for basket trials." *Biometrics*, 73(3): 905–915. MR3713124. doi: https://doi.org/10.1111/biom.12668. 18

Ventz, S., Trippa, L., and Schoenfeld, J. D. (2019). "Lessons Learned from De-escalation trials in favorable risk HPV-associated SCHN Cancer." *Clinical Cancer Research*, 25(24): 7281–7286. 17

Ventz, S. and Trippa, L. (2024). "Supplementary Material for "Bayesian Multi-Arm De-Intensification Designs"", *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1417SUPP. 4

Walker, S. and Muliere, P. (1997). "Beta-Stacy processes and a generalization of the Pólya-urn scheme." *The Annals of Statistics*, 1762–1780. MR1463574. doi: https://doi.org/10.1214/aos/1031594741. 4, 5

Wason, J., Stecher, L., and Mander, A. P. (2014). "Correcting for multiple-testing in multi-arm trials: is it necessary and is it done?" *Trials*, 15(1): 1–7. 9

Wellek, S. (2005). "Statistical methods for the analysis of two-arm non-inferiority trials with binary outcomes." *Biom Journal*, 47(1): 48–61. 3

Williamson, P. P. (2007). "Bayesian equivalence testing for binomial random variables." *Journal of Statistical Computation and Simulation*, 77(9): 739–755. MR2409860. doi: https://doi.org/10.1080/10629360600643496. 3