Multistate modeling of clinical hold in randomized clinical trials

Alexandra Nießl1 | Jan Beyersmann1 | Anja Loos2

1Institute of Statistics, Ulm University, Ulm, Germany
2Global Biostatistics, Merck KGaA, Darmstadt, Germany

Correspondence
Alexandra Nießl, Institute of Statistics, Ulm University, Ulm, Germany.
Email: alexandra.niessl@uni-ulm.de

A clinical hold order by the Food and Drug Administration (FDA) to the sponsor of a clinical trial is a measure to delay a proposed or to suspend an ongoing clinical investigation. The phase III clinical trial START serves as motivating data example to explore implications and potential statistical approaches for a trial continuing after a clinical hold is lifted. In spite of a modified intention-to-treat (ITT) analysis introduced to account for the clinical hold by excluding patients potentially affected most by the clinical hold, results of the trial did not show a significant improvement of overall survival duration, and the question remains whether the negative result was an effect of the clinical hold. In this paper, we propose a multistate model incorporating the clinical hold as well as disease progression as intermediate events to investigate the impact of the clinical hold on the treatment effect. Moreover, we consider a simple counterfactual censoring approach as alternative strategy to the modified ITT analysis to deal with a clinical hold. Using a realistic simulation study informed by the START data and with a design based on our multistate model, we show that the modified ITT analysis used in the START trial was reasonable. However, the censoring approach will be shown to have some benefits in terms of power and flexibility.

KEYWORDS
clinical hold, overall survival, progression-free survival, proportional hazards model, time-dependent covariates

1 | INTRODUCTION

A clinical hold (CH) is an order issued by the FDA (US Food and Drug Administration) “to delay a proposed clinical investigation or to suspend an ongoing clinical study.”1 Measures include that subjects must not receive the investigational drug and, when the recruitment period is still ongoing, no new subject may be enrolled. The FDA may place a trial on hold at any time during the clinical development, in particular if it finds that “human subjects are or would be exposed to an unreasonable and significant risk of illness or injury.”1 The investigation may resume after the FDA has notified the sponsor that the CH has been lifted.1 In case of a temporary hold, questions arise whether and which measures should be applied to maintain the potential for a successful trial. Boudes2 analyzed 29 CH orders between August 2008 and August 2015. Twelve development programs were found to be discontinued, and six probably discontinued in consequence of a CH. The author concludes that CH is a significant risk for a subsequent failure to a clinical development program.
underlines the relevance and the need to explore potential implications of a CH and potential remedies from statistical point of view.

The START* trial sponsored by Merck KGaA provides an example for a phase III trial, which was presumed after a CH with exclusion and replacement of patients potentially affected most by the treatment suspension. Results of the trial showed no significant prolongation of overall survival (OS) duration in the modified analysis population used for primary analyses.

Thus, the question arises if the modified intention-to-treat (mITT) analysis, which excluded patients randomized within 6 months preceding the CH, had in principle compensated for potential implications of the CH. Especially, it is of interest (a) to explore to which extent the CH may have caused the failure of the trial, (b) how well the mITT analysis accounted for the CH, and (c) whether other alternatives to the mITT analysis exist.

A model illustrating potential dependencies in a clinical trial with survival endpoints is a useful tool to investigate in simulation studies the impact of the CH on the treatment effect and various methods to account for a CH. For that purpose, we consider a multistate model that can be interpreted as a joint model for the two outcomes progression-free survival (PFS) and OS as well as for the CH as time-dependent covariate. We consider PFS primarily to better understand the impact of the CH. Treatment is stopped after disease progression in the START trial and thus patients who were already progressive at time of the treatment suspension due to the CH are not directly affected by the CH. PFS is defined as the time until disease progression or death, whichever occurs earlier. The simulation framework is closely motivated by the real situation of the START trial. The advantage of the multistate modeling approach is three-fold: Firstly, it naturally models the dependence between PFS and OS, where $\text{PFS} = \text{OS}$ with positive probability, without the need to, eg, resort to copula modeling nor requiring independent latent times to progression, death, and death after progression.

Secondly, a multistate model for jointly modeling PFS and OS is easily extended to additionally account for time-dependent exposures such as ordering and lifting a CH, which is crucial for our investigation. Thirdly, it provides a framework for a simple causal analysis of treatment effects as in Gran et al.

The results of the simulation studies show that the initially proposed mITT analysis was a reasonable proceeding. The mITT approach used prior knowledge on the mode of action of the experimental drug. In this paper, we also develop a more flexible censoring approach that does not require such knowledge and that also provides some improvements in terms of power and accuracy. The censoring approach exploits the fact that occurrence of a CH is independent censoring in a counting process sense. That means that the additional knowledge of a (then hypothetical) CH in a “perfect” data world (where there is no CH) will not alter the intensities of the counting processes at hand. This is a more subtle censoring concept than the common random censoring model, because patients during CH cannot be assumed to have the same OS hazard as patients in the absence of CH. The implication is that random censoring also is independent censoring, but, in general, censoring by CH is not random censoring.

In the next section, the motivating example, the START trial and the original mITT analysis, will be described in more detail, followed by the mathematical framework of survival multistate models in Section 3. In Section 4, we present our multistate model, and in Section 5, we propose alternative analysis strategies to the mITT analysis. Section 4 also suggests a simple multistate summary graphic of the data at hand, generalizing a competing risks tool. The design of the simulation studies and their findings are described in Sections 6 and 7, respectively. Section 8 completes this paper with a critical discussion of the results. Details on the multistate simulation algorithm and additional simulation results are provided in Appendix A.

2 | THE START TRIAL AND A MITT ANALYSIS

The START trial\(^4\) was a phase III, 2:1 randomized, controlled and double-blind group-sequential trial to investigate if the MUC1 antigen-specific cancer immunotherapy tecemotide given as maintenance therapy after chemoradiation improves OS duration in patients with unresectable stage III nonsmall-cell lung cancer. Treatment had to be administered in a six-weekly modality after an initial weekly administration for eight weeks till progression of disease (PD).

Due to a case of encephalitis occurring in a phase II trial of tecemotide, the FDA put the START trial on hold for enrollment and treatment in March 2010. At that time, the trial was with 1182 (787 Tecemotide vs 395 Placebo) subjects out of 1322 planned close to accrual completion. This CH was lifted in June 2010, and the treatment restarted once regulatory approval of the amended clinical trial protocol had been obtained after a median suspension of 135 days. Of 531 (372

\(^4\) Stimulating Targeted Antigenic Response To nonsmall-cell cancer
Tecemotide vs 159 Placebo) patients receiving study treatment at the time of the hold, 180 (133 Tecemotide vs 47 Placebo) patients did not resume treatment.

The sponsor decided to continue the trial after enlarging the overall trial population and to exclude those patients in an mITT analysis believed to be most affected by suspension of treatment. Under the assumption that 6 months of treatment are at least necessary to induce an immuno-therapeutic effect on survival, all 274 patients (177 Tecemotide vs 97 Placebo) who were randomly assigned within the 6 months preceding the CH were excluded from primary analysis. This exclusion of patients is illustrated in Figure 1. The sample size calculation of the trial protocol was adjusted to cover for the patients who were randomly assigned within the 6 months preceding the CH were excluded from primary analysis. Overall, from 22 February 2007, until 15 November 2011, 1513 patients were enrolled, of whom 1006 were randomly assigned to tecemotide and 507 to placebo. In the mITT subset 1239 patients remained, 829 receiving tecemotide, and 410 receiving placebo.

3 | SURVIVAL MULTISTATE MODELS

Multistate models generalize the standard survival model, where an individual can only move from the initial state “alive” to the state “dead,” and are a flexible approach to handle complex survival scenarios. A multistate model allows the analysis of any finite number of states and any transition between these states. States out of which no transitions are modeled are called absorbing states, whereas states with possible outgoing transitions are called transient states.

Multistate models complying with the Markov property and allowing the transition hazards to vary over time are called time-inhomogeneous Markov models. The Markov property states that the future course of an individual depends only on the current time and the state currently occupied. In terms of the multistate process $(X_t)_{t \geq 0}$, where $X_t$ denotes the state where an individual is in at time $t$, the transition hazards $a_{ij}(t)$ from state $i$ to state $j$ of a time-inhomogeneous Markov model are defined via:

$$a_{ij}(t)dt = P(X_{t+dt} = j | X_t = i), \quad l \neq j.$$  

(1)

Thus, the transition hazard $a_{ij}(t)$ times $dt$ in Equation (1) can be interpreted as the conditional probability of making an $l \rightarrow j$ transition in the very small interval $[t, t + dt)$. We assume a finite state space $\{0, 1, 2, \ldots, J\}$.

Now we consider $n$ replicates assumed to be independent conditional on the initial states with individual process $(X^{(i)}_t)_{t \geq 0}$, $l = 1, 2, \ldots, n$. Let

$$Y^{(i)}_l := \mathbb{1}(X^{(i)}_{C_l} = l, t \leq C_l)$$

(2)

denote whether the individual $i$ is in state $l$ and under observation just before time $t$. $C_l$ is defined as the right-censoring time of individual $i$. The number of $i$’s observed direct $l \rightarrow j$ transitions in $[0, t]$ is denoted by the counting process $N^{(i)}_{lj}(t)$.

Here, ‘direct $l \rightarrow j$ transition’ means a transition from $l$ to $j$ without visiting another state in between.

Then, the Nelson-Aalen estimator of the cumulative transition hazard $\hat{A}_{ij}(t) = \int_0^t a_{ij}(u)du$ is given by:

$$\hat{A}_{ij}(t) := \sum_{s \leq t} \frac{\Delta N_{ij}(s)}{Y_l(s)}, \quad l \neq j,$$

(3)

where the sum is over all observed, unique event times $s$, and

$$Y_l(t) := \sum_{i=1}^n Y^{(i)}_l(t) \quad \text{and} \quad N_{ij}(t) := \sum_{i=1}^n N^{(i)}_{ij}(t), \quad l \neq j,$$

(4)
and we write $\Delta N_l(s)$ for the increments between time $s$ and the previous time point of a jump of $N_l(s)$.

To relate the transition specific hazards to one or more covariates Cox models can be used. A proportional hazard model for the $l \rightarrow j$ transition hazard can be formulated as:

$$ a_{lj}^{(l)}(t; Z_i) = a_{lj,0}(t) \cdot \exp(\beta_{lj} \cdot Z_i), \quad l, j \in \{0, 1, 2, \ldots, J\}, \quad l \neq j, \quad i = 1, \ldots, n, $$ (5)

where $\beta_{lj}$ is the $1 \times p$ vector of regression coefficients and $Z_i$ the $p \times 1$ vector of the time-fixed baseline covariates for individual $i$. In principle, different transition hazards may be related to different covariate vectors. Moreover, $a_{lj,0}(t)$ denotes an unspecified, non-negative baseline hazard function. It is important to note that the model (5) reflects the Markov assumption. That means, the hazards of the transient states do not depend on the time at which an individual enters that transient state but only on the time elapsed since time origin, the state currently occupied and on the baseline covariates.7

A complete picture is to be obtained only when analyzing every transition hazard in turn.

### Time-Dependent Covariates and Multistate Models

One distinguishes between two categories of time-dependent covariates: External covariates are identified as covariates whose existence does not depend on the individual under study. In contrast, internal covariates are determined at each time point depending on the life experience of the individual up to this time point.11 CH is an example of an external covariate, while progression status is an example of an internal covariate.

Categorical time-dependent covariates such as a CH can be included in a multistate model through transitions from one transient state to another.12 Then, the multistate model covers both, a time-dependent covariate process and time-to-event endpoints (through the time until the multistate process enters an absorbing state). Thus, the multistate model is interpreted as a joint model for both the time-dependent covariate and the time to event.7

### Simulating Multistate Data

A multistate model can be interpreted as a nested series of competing risks experiments. Therefore, competing risk simulation is the basic element of the algorithm for simulating multistate data. Consider an individual who is in the initial state at time $0$7:

1. The waiting time in the initial state $0$ ($=: t_0$) is determined by the hazard $a_0(t) = \sum_{m=1}^{J} a_{0m}(t)$, $t \geq 0$ and survival function $t \mapsto \exp(-\int_{0}^{t} a_0(u)du)$.
2. Then, the state entered at $t_0$ is determined by a multinomial experiment, which decides with probability $\frac{a_{0m}(t)}{a_{0}(t)}$ on state $m, m \neq 0$. If $m$ is an absorbing state, the algorithm stops. Otherwise,
3. the waiting time $t_1$ in state $m$ is generated with hazard $a_m(t) = \sum_{k=0,k \neq m}^{J} a_{mk}(t)$, $t \geq t_0$ and conditional survival probability $t \mapsto \exp(-\int_{t_0}^{t} a_m(u)du)$.
4. A multinomial experiment decides with probability $\frac{a_{mk}(t_0+t_1)}{a_m(t_0+t_1)}$ on state $k, k \neq m$. If $k$ is an absorbing state, the algorithm stops.

If $k$ is a transient state, the next competing risk experiment (steps 3 and 4) starts at time $t_0+t_1$. The algorithm perspective is a useful tool to understand the structure of the data, as it describes the natural way how multistate data occur over time. In particular, only real life event times such as PFS are modeled, and the algorithm does not require latent failure times such as “time to progression” and “time to death without prior progression”. Advantages of the multistate model approach are discussed in more detail in Meller et al13 for the purpose of jointly modeling PFS and OS and in Bluhmki et al12 with a view toward modeling of time-dependent covariates.

The next section will suggest multistate models for jointly modeling PFS, OS, and CH, and Section 6 will employ the present simulation algorithm for these multistate models. In Section 6, we will also discuss two subtleties in the implementation of the simulation algorithm for CH: Firstly, we will also need to simulate CH times for patients not truly affected by CH. Secondly, we will need to approximate via simulation an (average) OS hazard ratio under model misspecification. These subtleties will be explained below.
In general, treatment with the same drug continues till PD in trials on treatment of solid tumors. In the START trial, treatment changed upon PD for the tecemotide group as well as for the placebo group. As a consequence, we include a state representing the intermediate event disease progression in our model. Since it is assumed that the CH primarily affects the hazards in the treatment group, we do not model the CH in the placebo group (see Figure 2). As PFS is defined as the time until progression or death, whichever occurs earlier, PFS in the placebo group is the waiting time $t_0$ in the initial state which is determined by the all-cause hazard $\alpha_0(t)$. A binomial experiment decides with probability $\frac{\alpha_0(t_0)}{\alpha_0(t_0)}$ on progression (state 2) and with probability $\frac{\alpha_0(t_0)}{\alpha_0(t_0)}$ on death (state 3). If the state “progression” is entered the waiting time till the absorbing state “death” is determined by the hazard $\alpha_{23}(t)$. OS is defined as the time until the absorbing state “death” is reached.

To describe the event of the CH in the treatment group, we add two transient states representing start and end of the CH to our model (see Figure 3). In this model, CH is modeled as an event of a patient’s course of disease and treatment, as are PD and death. A difference between the latter two events and CH is that CH occurs at one point in calendar time for the entire population. This is similar to censoring as a consequence of study closure in calendar time, leading, however, to individual censoring times (and times of CH) as a consequence of staggered study entry.

We have estimated the Nelson-Aalen estimators of all cumulative hazards of our models from the START data including all patients of the ITT population with the following results: The Nelson-Aalen estimates of the death hazards without prior progression are rather low, and no clear difference can be observed between the $0 \rightarrow 3$ hazards when comparing the treatment with the placebo group nor when comparing the $0 \rightarrow 3$, $1 \rightarrow 3$, and $4 \rightarrow 3$ hazards in the treatment group. That means, only a small part of the patients die without being progressive before (ITT START: 30 [3.6%] Tecemotide vs...
The estimated cumulative hazards into the progression state indicate a protective effect of treatment on the time until progression, which can not be observed during treatment suspension due to CH (i.e., we do not see any clear difference between the cumulative estimates of the $1 \rightarrow 2$ hazard of the treatment group and of the $0 \rightarrow 2$ hazard of the placebo group).

The large number of transition hazards and states of the model complicates the interpretation of the Nelson-Aalen estimates. Although it cannot be assumed that all transition hazards of our models are time-homogeneous, i.e., time-constant, a graphical illustration using incidence rates, as described in Grambauer et al. for the simpler competing risks model, enables a quick overview over the situation of the START trial. In order to avoid a too strong influence of late time points, only values till month 30 are included in the estimation and data after month 30 were artificially censored for the graphical presentation in Figures 4 and 5. In Figures 4 and 5, the arrow thickness describes the particular amount of every incidence density and relating one arrow to another has a proportional hazards interpretation.

The interpretation of Figures 4 and 5 is that hardly any patient dies without prior diagnosed progression and that progression increases the death hazard. The thickness of the arrows may also indicate that treatment is protective w.r.t. the progression hazard, while imposing a CH may diminish this protective effect and lifting the CH may recover a protective effect. We stress that these figures must not be over-interpreted but that they can be useful to summarize the complex multistate situation at hand, see also the simulation scenarios studied in Section 6.

5 | ANALYSIS METHODS TO COMPENSATE FOR THE IMPACT OF THE CH

Before we use the multistate model of Section 4 for the simulation of scenarios similar to the START trial, we present in this section alternative methods to the mITT analysis of the START trial. The mITT analysis population of the START trial was purely based on the randomization time and did not take into account whether a patient receives study treatment or not. As we have already noted in Section 4, we assume that the CH primarily affects the treatment group. Hence, we apply the modification used for the START ITT population to the treatment group only with the advantage of a larger sample size. These mITT methods will be labeled II and III, respectively, in the following. (A “naive” ITT method ignoring the difficulty posed by CH will be labeled I.)

Alternatively, one may exploit the fact that CH is an external mechanism independent of the individual patient. For this reason, censoring the patients at the beginning of the CH should be expected to give unbiased estimates of OS and PFS hazard ratios. Censoring CH leads to methods IV and V below, either applied to all patients or to the treatment group only. A note on censoring as a consequence of CH is in place, because it is more subtle than the common random censorship model, which assumes stochastically independent death and censoring times. In this model, the hazard of death is the same for patients alive and censored and for patients alive and uncensored. This is clearly not the case for censoring by CH, because CH leads to treatment discontinuation for progression-free patients in the treatment group.

On the other hand, our multistate models contain the relevant information for stopping treatment at the time of CH save for external random variation, see section 9.6.2 in Aalen et al. for such multistate models and Pearl et al. for the role of external random variation in causal models. Aalen et al. then demonstrate that censoring treatment discontinuation (here: by CH) provides valid inference, counter to fact, i.e., in the hypothetical situation of no CH.

More precisely, Aalen et al. use a partial empirical transition matrix, i.e., the usual estimator of the transition probabilities in a multistate model but additionally censoring treatment discontinuation and demonstrate that this coincides with the causal g-computation formula, see also Gran et al. In what follows, we will use a hybrid approach aligned with the mITT analysis. Here, censoring CH is applied to both groups (method IV) and to the treatment group only (method V). In the placebo group, this may be viewed as random censoring because the treatment of the patient is not affected. Moreover, we will censor both PFS and OS times. In the treatment group, this may again be viewed as random censoring of OS times of patients in the progression state.

The following list gives an overview over the considered analysis methods:

1. **naive**: An ITT analysis including all patients (for reasons of comparability).
2. **mITT (placebo and treatment)**: Patients randomized less than six months before the CH are excluded from the ITT analysis. This is the method that has been applied in the START trial, and we will refer to this analysis as original mITT analysis.
3. **mITT (only treatment)**: Only patients randomized to the treatment group within the 6 months prior to the CH are excluded from the ITT analysis.
4. **CH as censoring (placebo and treatment)**: Patients of both groups are censored with start of the CH regardless of progression status.
5. CH as censoring (only treatment): Patients of the treatment group only are censored with start of the CH regardless of progression status.

The mITT analyses differ from the censoring approaches in that patients are excluded, whereas the censoring approaches use the information of these excluded patients that no event is observed until the CH and, hence, use a larger sample. Both approaches may lead to a considerable loss of power because of a reduced number of observed events.

6 | SIMULATION STUDIES

We performed simulation studies to assess the impact of the CH on OS and PFS in scenarios similar to the real study situation of the START trial and to compare the performance of the different analysis methods I to V. We use the hazard-based algorithm presented in Section 3 and apply the multistate model presented in Section 4 with CH as an external time-dependent covariate and PD as an internal one to simulate several scenarios. The cumulative transition hazards are parametrically estimated from the START data assuming piece-wise constant hazards. All these parametric estimates approximate the true cumulative transition hazards precisely enough to use them for our simulation studies.

At first, we simulate a set of scenarios with a considerable treatment effect. Following the results of the non-parametric estimation described in Section 4, we vary only the hazard ratios between treatment and placebo concerning the time to PD. The hazard ratio of the direct progression hazards of treatment and placebo group \((a_t^{(T)}/a_t^{(P)}) = \exp(\beta_0)\) is chosen to be 0.6, where superscripts \((T)\) and \((P)\) refer to the treatment group and to the placebo group, respectively.

The adverse effect of the CH is realized by a hazard ratio of \((1 \pm 0.1)\) between treatment and placebo during the CH \((a_t^{(T)}/a_t^{(P)}) = \exp(\beta_1)\). As we assume that resumption of treatment more or less restores the treatment effect, we consider a hazard ratio of \((0.6 \pm 0.1)\) after resumption of treatment \((a_t^{(T)}/a_t^{(P)}) = \exp(\beta_2)\). Additionally, a second set of scenarios with a treatment effect closer to the estimates received in the START trial is simulated. Table 1 summarizes the simulated scenarios.

Random censoring times are generated from a Gompertz distribution for all patients, which roughly mimicked the empirical censoring distribution in the original data. An iteration number of 1000 with 1500 observations each, assigned in a 2:1 ratio to treatment or placebo group, is chosen.

Our multistate model provides the starting time of the CH only for patients actually affected by the CH. Thus, an additional simulation of the event times of the CH needs to be performed for the placebo group and patients with direct progression to reproduce the mITT analyses and the censoring approaches. The approach is presented in Appendix A. In a nutshell, a binomial experiment decides whether a patient is included in the study before the CH order. Only for these patients additional simulation of CH times is needed, which is informed by the survival distribution of time-to-CH from START estimated through Kaplan-Meier.

To assess how the proposed analysis methods compensate the impact of the CH on OS and PFS, it is necessary to know the treatment effect on OS and PFS, which would be observed if no CH has occurred and if the true hazard ratio between the progression hazards of treatment and placebo is 0.6 or 0.8, respectively. This treatment effect cannot be analytically identified from our multistate model. Thus, we estimate the treatment effect on OS and PFS in a setting without CH by means of simulation for the respective scenarios (see Table 2), see previous studies\(^{15,16}\) for simulation-based numerical approximation of average hazard ratios.

7 | RESULTS

In the following section, the results of our simulation studies are provided. We analyzed the simulated scenarios according to methods I to V presented in Section 5. In order to assess how well the considered analysis methods compensate potential implications of the CH, we compare the mean hazard ratio of OS and PFS over 1000 simulations with the “target” hazard ratio (see Table 2). Additionally, we consider the mean-squared error (MSE) of the averaged hazard ratios, as the MSE incorporates both the variance of the estimated hazard ratios and its bias. In the context of a clinical study, it is crucial, whether a true treatment effect could have been shown. Thus, we consider the empirical power resulting from the different analysis methods as well.
### Results for OS

Table 3 lists the averaged hazard ratios for the first set of scenarios together with their MSE resulting from the analysis methods I to V. The corresponding empirical confidence intervals are shown in Appendix A. Since without occurrence of a CH, a hazard ratio of 0.707 would be expected (cf Table 2), a negative impact of the CH in the naive analysis (column I) can be observed for all scenarios in that the estimated average hazard ratio is biased upwards. However, the amount of that impact is rather moderate and depends on the specific scenario. The original mITT analysis as well as the censoring approaches provide estimates close to the comparison value of 0.707. The censoring approach concerning the treatment group only (method V) produces the best results. Comparing the MSEs across the scenarios, we find that censoring method IV produces MSEs that are typically larger than in the other scenarios, while the smallest MSEs are found for the naive analysis I, the original mITT approach II and censoring method V. Figure 6 illustrates the distribution of the hazard ratios of 1000 simulations via box plots for scenario 1 (\(\exp(\beta_{12}) = 1\) and \(\exp(\beta_{42}) = 0\)). The empirical power keeps a sufficient level for each method, but the loss of power is less for method V compared with the other methods (results are provided in Appendix A).

### TABLE 3 Results of the simulation studies for overall survival—first set of scenarios

| Scenario ID | \(\exp(\beta_{12})\) | \(\exp(\beta_{42})\) | HR (MSE) | I (MSE) | II (MSE) | III (MSE) | IV (MSE) | V (MSE) |
|-------------|----------------------|----------------------|----------|---------|---------|---------|---------|---------|
| I           | 0.6                  | 0.6                  | 0.738 (0.0037) | 0.715 (0.0037) | 0.724 (0.0042) | 0.704 (0.0050) | 0.712 (0.0040) |
| II          | 1.1                  | 0.6                  | 0.740 (0.0034) | 0.714 (0.0035) | 0.726 (0.0039) | 0.699 (0.0048) | 0.705 (0.0039) |
| III         | 0.9                  | 0.6                  | 0.729 (0.0046) | 0.710 (0.0040) | 0.714 (0.0052) | 0.701 (0.0052) | 0.709 (0.0041) |
| IV          | 1                    | 0.5                  | 0.727 (0.0045) | 0.706 (0.0039) | 0.712 (0.0051) | 0.703 (0.0052) | 0.710 (0.0039) |
| V           | 1.1                  | 0.5                  | 0.732 (0.0043) | 0.708 (0.0041) | 0.718 (0.0050) | 0.700 (0.0054) | 0.709 (0.0042) |
| 6           | 0.9                  | 0.5                  | 0.719 (0.0054) | 0.705 (0.0042) | 0.705 (0.0059) | 0.701 (0.0052) | 0.710 (0.0042) |
| 7           | 1                    | 0.7                  | 0.747 (0.0033) | 0.718 (0.0037) | 0.731 (0.0037) | 0.702 (0.0053) | 0.708 (0.0042) |
| 8           | 1.1                  | 0.7                  | 0.751 (0.0033) | 0.718 (0.0039) | 0.736 (0.0038) | 0.699 (0.0054) | 0.707 (0.0043) |
| 9           | 0.9                  | 0.7                  | 0.739 (0.0038) | 0.715 (0.0039) | 0.725 (0.0044) | 0.701 (0.0053) | 0.707 (0.0041) |

Note. Target value: 0.707 (see Table 2). First set of scenarios: \(\exp(\beta_{12}) = 0.6\). I: naive, II: mITT (placebo and treatment), III: mITT (only treatment), IV: CH as censoring (placebo and treatment), V: CH as censoring (only treatment).
TABLE 4  Results of the simulation studies for progression-free survival—second set of scenarios

| Scenario ID | exp(β_{12}) | exp(β_{42}) | HR (MSE)   |
|-------------|-------------|-------------|------------|
|             |             |             | I           | II          | III          | IV           | V           |
| 1           | 1.0         | 0.8         | 0.838 (0.0033) | 0.820 (0.0037) | 0.839 (0.0039) | 0.817 (0.0036) | 0.816 (0.0032) |
| 2           | 1.1         | 0.8         | 0.846 (0.0036) | 0.818 (0.0036) | 0.842 (0.0036) | 0.819 (0.0038) | 0.818 (0.0032) |
| 3           | 0.9         | 0.8         | 0.827 (0.0032) | 0.814 (0.0032) | 0.826 (0.0037) | 0.814 (0.0035) | 0.815 (0.0030) |
| 4           | 1.0         | 0.7         | 0.823 (0.0027) | 0.816 (0.0033) | 0.818 (0.0032) | 0.815 (0.0036) | 0.816 (0.0031) |
| 5           | 1.1         | 0.7         | 0.835 (0.0031) | 0.816 (0.0039) | 0.833 (0.0037) | 0.819 (0.0040) | 0.818 (0.0035) |
| 6           | 0.9         | 0.7         | 0.816 (0.0030) | 0.811 (0.0036) | 0.818 (0.0036) | 0.818 (0.0039) | 0.818 (0.0031) |
| 7           | 1.0         | 0.9         | 0.852 (0.0042) | 0.829 (0.0037) | 0.853 (0.0047) | 0.818 (0.0037) | 0.818 (0.0032) |
| 8           | 1.1         | 0.9         | 0.860 (0.0032) | 0.829 (0.0036) | 0.856 (0.0038) | 0.819 (0.0039) | 0.818 (0.0032) |
| 9           | 0.9         | 0.9         | 0.840 (0.0045) | 0.821 (0.0036) | 0.838 (0.0048) | 0.815 (0.0036) | 0.815 (0.0031) |

Note. Target value: 0.819 (see Table 2). Second set of scenarios: exp(β_{12}) = 0.8. I: naive, II: mITT (placebo and treatment), III: mITT (only treatment), IV: CH as censoring (placebo and treatment), V: CH as censoring (only treatment).

For the second set of scenarios, similar results are observed but with a generally diminished negative impact of the CH (results are provided in Appendix A).

Results for PFS

Table 4 shows the results for the scenarios with a direct progression hazard ratio of 0.8. The first scenario of the table is illustrated in Figure 7 via box plots. Unlike the results for OS, there is barely any difference between the results of the censoring approaches (methods IV and V) regarding the averaged hazard ratios. However, the MSEs of these two methods indicate less variability of the estimates obtained by method V. The results of the original mITT analysis (II) depend on the scenario, but a moderate negative impact of the CH in the naive analysis, ie, an estimated average hazard ratio that is biased upwards, is compensated well. Moreover, the original mITT analysis compensates the impact of the CH better than method III where patients of the treatment group only have been excluded.

Method V provides benefits in terms of power compared with the other methods (cf Table 5).

FIGURE 6  Comparing the analysis methods by box plots of the hazard ratios for overall survival (OS). The target value 0.707 (cf Table 2) is represented by the grey line. Scenario 1 of first set of scenarios: exp(β_{02}) = 0.6, exp(β_{12}) = 1, exp(β_{42}) = 0.6. I: naive, II: modified intention-to-treat (mITT) (placebo and treatment), III: mITT (only treatment), IV: clinical hold (CH) as censoring (placebo and treatment), V: CH as censoring (only treatment)
FIGURE 7  Comparing the box plots of the hazard ratios for progression-free survival (PFS) to the target value 0.819 (grey line). Scenario: \( \exp(\beta_{12}) = 0.8, \exp(\beta_{42}) = 1, \exp(\beta_{42}) = 0.8 \). I: naive, II: mITT (placebo and treatment), III: mITT (only treatment), IV: CH as censoring (placebo and treatment), V: CH as censoring (only treatment).

TABLE 5  Results of the simulation studies for PFS regarding the empirical power—second set of scenarios

| Scenario | Empirical Power |
|----------|-----------------|
| ID       | \( \exp(\beta_{12}) \) | \( \exp(\beta_{12}) \) | \( \exp(\beta_{42}) \) | I | II | III | IV | V |
| 1        | 0.8             | 1               | 0.8            | 0.800 | 0.787 | 0.725 | 0.793 | 0.833 |
| 2        | 0.8             | 1.1             | 0.8            | 0.741 | 0.781 | 0.709 | 0.770 | 0.823 |
| 3        | 0.8             | 0.9             | 0.8            | 0.846 | 0.821 | 0.788 | 0.799 | 0.848 |
| 4        | 0.8             | 1               | 0.7            | 0.853 | 0.826 | 0.795 | 0.786 | 0.836 |
| 5        | 0.8             | 1.1             | 0.7            | 0.795 | 0.815 | 0.752 | 0.761 | 0.807 |
| 6        | 0.8             | 0.9             | 0.7            | 0.907 | 0.820 | 0.823 | 0.788 | 0.837 |
| 7        | 0.8             | 1               | 0.9            | 0.705 | 0.733 | 0.627 | 0.766 | 0.820 |
| 8        | 0.8             | 1.1             | 0.9            | 0.668 | 0.759 | 0.644 | 0.765 | 0.822 |
| 9        | 0.8             | 0.9             | 0.9            | 0.794 | 0.788 | 0.730 | 0.793 | 0.843 |

Note. Second set of scenarios: \( \exp(\beta_{12}) = 0.8 \). I: naive, II: mITT (placebo and treatment), III: mITT (only treatment), IV: CH as censoring (placebo and treatment), V: CH as censoring (only treatment).

8 | DISCUSSION

In this paper, we established a multistate model enabling the investigation of the impact of a CH on the treatment effect and the comparison of different analysis methods to deal with a CH. We proposed a censoring CH approach as alternative to the mITT analysis used in the START trial.

It can be assumed that the performance of the mITT analysis mainly depends on the adequate choice of the exclusion window, which requires (potentially arbitrary) assumptions of the influence of the CH on the treatment effect on the primary endpoint. Butts at al\(^3\) performed sensitivity analyses using larger exclusion windows than 6 months with the result that there might be a negative impact on the treatment effect in terms of OS that extended beyond the selected 6 months. In contrast, the censoring CH approach is expected to provide unbiased results independent of the duration of the CH and its impact on the treatment effect. It should be noted, that, for reason of comparison, in the simulation study, all analysis methods were applied to the same number of randomized patients inspired by the number of randomized patients of the START trial after the extended recruitment period. A sample size calculation adjusting for the censoring approach may result in another, probably larger, number of required patients as compared with the initial planning of the study.

As stated earlier, the CH is an external mechanism and fulfills the independent (but not: random) censoring assumption in that the additional knowledge of a CH order in a complete data world does not change the intensities of the counting processes. Then, standard survival analyses aim to make statements that would be valid in the absence of censoring. That means, censoring induces a counterfactual thinking.\(^9\) With regard to the START trial, censoring patients at the beginning of the CH estimates the treatment effect in the ITT analysis if the CH would not have occurred. Therefore, our censoring
CH approach has a causal interpretation for the intention of the initial treatment and, in this sense, could be viewed as a hypothetical estimand\(^\text{17}\).

In conclusion, the simulation results showed that the mITT analysis applied to both groups is an appropriate approach to deal with a CH if sufficient information about the mode of action of the treatment exists to determine an exclusion window. However, the censoring approach is more flexible and can generally be recommended as analysis method to account for a CH. This is particularly attractive in situations where, unlike in START, there is less a priori knowledge about a “therapeutic window” around the CH. Applying the censoring only to the treatment group keeps the loss of power in limits.

Typically, strategic decisions on further conduct and analysis of a running clinical trial, especially in the situation of a CH have to be done without knowing the results of the trial. Our simulation study showed that the framework of multistate models is able to support discussions around appropriate analyses and with this decision making for trials in similar situation.

A line of future research would be to only censor progression-free patients in the treatment group at the time of CH, estimate the distribution of the multistate model in the absence of CH using the partial empirical transition matrix as described in Section 5 and subsequent OS analysis using the simulation approach of Sections 3 and 6. However, such investigations go beyond the scope of the present work whose primary aims were to investigate the merits of the original mITT analysis and to provide for a simple alternative without the need of a priori knowledge about a therapeutic window.

Another topic of future research would be the implication of the proportional hazard assumption, which has been checked for OS and PFS in the original START data. However, technically, Cox models for transition hazards as used in our simulations will imply non-proportional hazards of composites such as OS and PFS. We assume a minor impact, because our simulations nicely mimicked the real START data. We also refer to Klein\(^\text{18}\) who has discussed the issue in the simpler competing risks model.

ACKNOWLEDGEMENTS

Jan Beyersmann acknowledges support by the German Research Foundation (BE 4500/3-1).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from Merck KGaA. Data sharing is not applicable to this article, as the START data include human subject data. The data were used under license for this study.

ORCID

Alexandra Nießl  \[https://orcid.org/0000-0002-8582-9664\]

REFERENCES

1. U.S. Food and Drug Administration. IND Application Procedures: Clinical Hold. 2013. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/InvestigationalNewDrugINDApplication/ucm362971.htm.
2. Boudes PF. An Analysis of US Food and Drug Administration Clinical Hold Orders for Drugs and Biologics: A Prospective Study Between 2008 and 2014. *Pharmaceutical Medicine*. 2015;29(4):203-209.
3. Butts C. others. Tecemotide (L-BLP25) versus placebo after chemoradiotherapy for stage III non-small-cell lung cancer (START): a randomised, double-blind, phase 3 trial. *Lancet Oncology*. 2014;15(1):59-68.
4. Fu H, Wang Y, Liu J, Kulkarni PM, Melemed AS. Joint modeling of progression-free survival and overall survival by a Bayesian normal induced copula estimation model. *Statistics in Medicine*. 2013;32(2):240-254.
5. Fleischer F, Gaschler-Markefski B, Bluhmki E. A statistical model for the dependence between progression-free survival and overall survival. *Statistics in Medicine*. 2009;28(21):2669-2686.
6. Li Y, Zhang Q. A Weibull multi-state model for the dependence of progression-free survival and overall survival. *Statistics in Medicine*. 2015;34(17):2497-2513.
7. Beyersmann J, Allignol A. *Schumacher M*. Competing risks and multistate models with R. Springer; 2012.
8. Gran JM, Lie SA, Oyeflaten I, Borgen Ø, Aalen OO. Causal inference in multi-state models–sickness absence and work for 1145 participants after work rehabilitation. *BMC Public Health*. 2015;15(1):1082. http://doi.org/10.1186/s12889-015-2408-8
9. Aalen OO, Borgan Ø, Gjessing HK. *Survival and event history analysis: a process point of view*. Springer. 2008.
APPENDIX A: SIMULATING ADDITIONAL EVENT TIMES FOR THE CLINICAL HOLD

Two groups of patients are considered: On the one hand, patients randomized before the start of the clinical hold and potentially affected by it; and on the other hand, patients entering the study after the end of the clinical hold and definitely not affected by it. Hence, the first step of simulating additional event times for the clinical hold is to decide by a binomial experiment to which group a patient belongs. Then, the event times are drawn from a distribution whose assumption is informed by survival distribution from START estimated through Kaplan-Meier considering clinical hold as event and death as censoring. This procedure is described more precisely in the following:

- Simulation of event times of the clinical hold for the placebo group by applying the following steps:
  
  There are two subgroups of the placebo group:
  
  1. Patients of the placebo group that are potentially affected by the clinical hold
  2. Patients of the placebo group that are not potentially affected by the clinical hold because of their randomization date after the clinical hold

  A binomial experiment decides with probability

  \[ P(\text{potentially affected}) = \frac{\# \, \text{Placebo subgroup 1 of START}}{\# \, \text{Placebo group of START}} \]  

  on “potentially affected.”

  Event times are drawn from the empirical distribution (estimated through Kaplan-Meier) in the subgroup 1 considering clinical hold as event and death as censoring

- Simulation of the event times of the clinical hold for the patients with direct progression in the treatment group by applying the following steps:
  
  There are two subgroups:
  
  1. Patients of the treatment group with direct progression that are potentially affected by the clinical hold
  2. Patients of the treatment group with direct progression that are not potentially affected by the clinical hold because of their randomization date after the clinical hold
### TABLE A1  Results of the simulation studies for overall survival including empirical CIs—first set of scenarios

| Scenario | ID | exp(\(\beta_{12}\)) | exp(\(\beta_{42}\)) | Lower CI | HR | Upper CI |
|----------|----|----------------------|----------------------|---------|----|----------|
| 1        | 1  | 0.6                  | 0.657                | 0.663   | 0.738 | 0.822    |
| 2        | 1.1| 0.6                  | 0.657                | 0.663   | 0.738 | 0.822    |
| 3        | 0.9| 0.6                  | 0.657                | 0.663   | 0.738 | 0.822    |
| 4        | 1  | 0.6                  | 0.657                | 0.663   | 0.738 | 0.822    |
| 5        | 1  | 0.6                  | 0.657                | 0.663   | 0.738 | 0.822    |
| 6        | 0.9| 0.5                  | 0.657                | 0.663   | 0.738 | 0.822    |
| 7        | 1  | 0.7                  | 0.657                | 0.663   | 0.738 | 0.822    |
| 8        | 1  | 0.7                  | 0.657                | 0.663   | 0.738 | 0.822    |
| 9        | 0.9| 0.7                  | 0.657                | 0.663   | 0.738 | 0.822    |

Note. Target value: 0.707 (see Table 2) First set of scenarios: \(\text{exp}(\beta_{02}) = 0.6\). I: naive, II: mITT (placebo and treatment), III: mITT (only treatment), IV: CH as censoring (placebo and treatment), V: CH as censoring (only treatment).

### TABLE A2  Results of the simulation studies for overall survival—second set of scenarios

| Scenario | ID | exp(\(\beta_{12}\)) | exp(\(\beta_{42}\)) | Lower CI | HR | Upper CI |
|----------|----|----------------------|----------------------|---------|----|----------|
| 1        | 1  | 0.8                  | 0.771                | 0.750   | 0.862 | 0.990    |
| 2        | 1.1| 0.8                  | 0.771                | 0.750   | 0.862 | 0.990    |
| 3        | 0.9| 0.8                  | 0.771                | 0.750   | 0.862 | 0.990    |
| 4        | 1  | 0.7                  | 0.771                | 0.750   | 0.862 | 0.990    |
| 5        | 1  | 0.7                  | 0.771                | 0.750   | 0.862 | 0.990    |
| 6        | 0.9| 0.7                  | 0.771                | 0.750   | 0.862 | 0.990    |
| 7        | 1  | 0.9                  | 0.771                | 0.750   | 0.862 | 0.990    |
| 8        | 1  | 0.9                  | 0.771                | 0.750   | 0.862 | 0.990    |
| 9        | 0.9| 0.9                  | 0.771                | 0.750   | 0.862 | 0.990    |

Note. Second set of scenarios: \(\text{exp}(\beta_{02}) = 0.8\). I: naive, II: mITT (placebo and treatment), III: mITT (only treatment), IV: CH as censoring (placebo and treatment), V: CH as censoring (only treatment).

A binomial experiment decides with probability

\[
P(\text{potentially affected}) = \frac{\text{#Treatment subgroup 1 of START}}{\text{#Treatment group with direct progression of START}} \quad (A2)
\]

on “potentially affected.”

Event times of the clinical hold are drawn from the distribution whose assumption is informed by survival distribution from START estimated through Kaplan-Meier of the subgroup 1 of the START trial that considers clinical hold as event and death as censoring. The time scale “time since progression” and not “time since randomization” is used to avoid left-truncated data. The simulated event times are translated to the randomization time scale afterwards.

### A.1 Additional simulation results

The present subsection of the appendix provides additional simulation results following the rationale of Section 6 of the main paper. These additional results confirm the interpretations of the main paper, with a preference for method V. The empirical confidence intervals are displayed as indices, following a suggestion in Louis and Zeger.19
### TABLE A3 Results of the simulation studies for OS regarding the empirical power—second set of scenarios

| Scenario ID | \(\exp(\beta_{02})\) | \(\exp(\beta_{12})\) | Empirical Power - \(\exp(\beta_{42})\) |
|-------------|-----------------|-----------------|-----------------|
| I           | 0.8             | 1               | 0.8             |
| II          | 0.8             | 1.1             | 0.8             |
| III         | 0.8             | 0.9             | 0.8             |
| IV          | 0.8             | 1               | 0.7             |
| V           | 0.8             | 1.1             | 0.7             |
| 1           | 0.8             | 1               | 0.8             |
| 2           | 0.8             | 1.1             | 0.8             |
| 3           | 0.8             | 0.9             | 0.8             |
| 4           | 0.8             | 1               | 0.7             |
| 5           | 0.8             | 1.1             | 0.7             |
| 6           | 0.8             | 0.9             | 0.7             |
| 7           | 0.8             | 1.1             | 0.9             |
| 8           | 0.8             | 0.9             | 0.9             |
| 9           | 0.8             | 0.9             | 0.9             |

**Note:** I: naive, II: mITT (placebo and treatment), III: mITT (only treatment), IV: CH as censoring (placebo and treatment), V: CH as censoring (only treatment).

### TABLE A4 Results of the simulation studies for progression-free survival—first set of scenarios

| Scenario ID | \(\exp(\beta_{12})\) | \(\exp(\beta_{42})\) | Lower CIHR | Upper CI |
|-------------|-----------------|-----------------|------------|----------|
| I           | 0.6             | 0.617            | 0.753      |
| II          | 0.6             | 0.623            | 0.680      |
| III         | 0.6             | 0.608            | 0.664      |
| IV          | 0.6             | 0.605            | 0.660      |
| V           | 0.6             | 0.610            | 0.666      |
| 1           | 0.6             | 0.617            | 0.753      |
| 2           | 0.6             | 0.623            | 0.680      |
| 3           | 0.6             | 0.608            | 0.664      |
| 4           | 0.6             | 0.605            | 0.660      |
| 5           | 0.6             | 0.610            | 0.666      |
| 6           | 0.6             | 0.595            | 0.650      |
| 7           | 0.6             | 0.626            | 0.683      |
| 8           | 0.6             | 0.633            | 0.691      |
| 9           | 0.6             | 0.621            | 0.678      |

**Note:** First set of scenarios: \(\exp(\beta_{02}) = 0.6\). I: naive, II: mITT (placebo and treatment), III: mITT (only treatment), IV: CH as censoring (placebo and treatment), V: CH as censoring (only treatment).

### TABLE A5 Results of the simulation studies for progression-free survival including empirical CIs—second set of scenarios

| Scenario ID | \(\exp(\beta_{12})\) | \(\exp(\beta_{42})\) | Lower CIHR | Upper CI |
|-------------|-----------------|-----------------|------------|----------|
| I           | 0.8             | 0.753            | 0.838      |
| II          | 0.8             | 0.760            | 0.846      |
| III         | 0.8             | 0.743            | 0.827      |
| IV          | 0.8             | 0.739            | 0.823      |
| V           | 0.8             | 0.750            | 0.835      |
| 1           | 0.8             | 0.753            | 0.838      |
| 2           | 0.8             | 0.760            | 0.846      |
| 3           | 0.8             | 0.743            | 0.827      |
| 4           | 0.8             | 0.739            | 0.823      |
| 5           | 0.8             | 0.750            | 0.835      |
| 6           | 0.8             | 0.733            | 0.816      |
| 7           | 0.8             | 0.765            | 0.852      |
| 8           | 0.8             | 0.772            | 0.860      |
| 9           | 0.8             | 0.754            | 0.840      |

**Note:** Target value: 0.819 (see Table 2) Second set of scenarios: \(\exp(\beta_{02}) = 0.8\). I: naive, II: mITT (placebo and treatment), III: mITT (only treatment), IV: CH as censoring (placebo and treatment), V: CH as censoring (only treatment).