Journal of Clinical Epidemiology 170 (2024) 111340

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Restricted Net Treatment Benefit in oncology

Max Piffoux[a,b,c,*], Brice Ozenne[d,e], Mickaël De Backer[f], Marc Buyse[f,g], Jean-Christophe Chiem[f], Julien Péron[h,i]

[a]*Medical Oncology, Hospices Civils de Lyon, CITOHL, Lyon, France*
[b]*Direction de la Recherche Clinique et de l'Innovation, Centre Léon Bérard, Lyon, France*
[c]*Laboratoire MSC Matière et Systèmes Complexes, Université de Paris, CNRS UMR 7057, 75006 Paris, France*
[d]*Neurobiology Research Unit and BrainDrugs, Copenhagen University Hospital, Rigshospitalet, 6-8 Inge Lehmanns Vej, 2100 Copenhagen, Denmark*
[e]*Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark*
[f]*International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium*
[g]*I-BioStat, University of Hasselt, Hasselt, Belgium*
[h]*Hospices Civils de Lyon, Oncology Department, Pierre-Bénite, France*
[i]*Université de Lyon, Université Lyon 1, Reshape Laboratory INSERM U1290, Lyon, France*

Accepted 25 March 2024; Published online 1 April 2024

Abstract

Objectives: The restricted Net Treatment Benefit (rNTB) is a clinically meaningful and tractable estimand of the overall treatment effect assessed in randomized trials when at least one survival endpoint with time restriction is used. Its interpretation does not rely on parametric assumptions such as proportional hazards, can be estimated without bias even in the presence of independent right-censoring, and can include a prespecified threshold of minimal clinically relevant difference. To demonstrate that the rNTB, corresponding to the NTB during a predefined time interval, is a meaningful and adaptable measure of treatment effect in clinical trials.

Methods: In this simulation study, we tested the impact on the rNTB value, estimation, and power of several factors including the presence of a delayed treatment effect, minimal clinically relevant difference threshold value, restriction time value, and the inclusion of both efficacy and toxicity in the rNTB definition. The impact of right censoring on rNTB was assessed in terms of bias. rNTB-derived statistical tests and log rank (LR) tests were compared in terms of power.

Results: RNTB estimates are unbiased even in case of right-censoring. rNTB may be used to estimate the benefit/risk ratio of a new treatment, for example, taking into account both survival and toxicity and include several prioritized outcomes. The estimated rNTB is much easier to interpret in this context compared to NTB in the presence of censoring since the latter is intrinsically dependent on the follow-up duration. Including toxicity increases the test power when the experimental treatment is less toxic. rNTB-derived test power increases when the experimental treatment is associated with longer survival and lower toxicity and might increase in the presence of a cure rate or a delayed treatment effect. Case applications on the PRODIGE, Checkmate-066, and Checkmate-067 trials are provided.

Conclusions: RNTB is an interesting alternative to describe and test the treatment's effect in a clear and understandable way in case of restriction, particularly in scenarios with nonproportional hazards or when trying to balance benefit and safety. It can be tuned to take into consideration short- or long-term survival differences and one or more prioritized outcomes.

Keywords: Generalized pairwise comparisons; Restriction; Clinical trial; Nonproportional hazard; Toxicity; Immunotherapy

1. Introduction

Randomized clinical trials (RCTs) in oncology often use survival times as primary endpoints. These time-to-event endpoints are compared and commonly reported using hazard ratios (HRs), representing the relative difference between survival curves. HR only faithfully reflects the effect of the treatment if hazards are proportional over time. However, these conditions are rarely met, especially in immuno-oncology trials. Also, standard analysis strategies

Funding: The sharing of the BMS Checkmate data was done under the consortium agreement of BENEFIT with BMS approval. The manuscript was submitted to BMS for review before publication, without modification of its content. Research partially funded by the Government of Wallonia, BioWin Consortium Agreement No 7979.

* Corresponding author. Medical Oncology, Hospices Civils de Lyon, CITOHL, 165 Chem. du Grand Revoyet, 69495 Oullins-Pierre-Bénite, France.

E-mail address: max.piffoux@cri-paris.org (M. Piffoux).

https://doi.org/10.1016/j.jclinepi.2024.111340

What is new?

Key findings

- Restricted Net Treatment Benefit is unbiased even in case of right-censoring

What this adds to what was known?

- It may be used to estimate the benefit/risk ratio of a treatment by taking into account toxicity
- It is more powerful than log rank in cases when the experimental treatment is less toxic. It is also more powerful when there is an increased cure rate and/or has a delayed treatment effect and when it is tuned to focus on large magnitude survival differences.

What is the implication and what should change now?

- Restricted Net Treatment Benefit is an interesting alternative to describe and test treatment effect in a simple and understandable way, particularly in cases with nonproportional hazards or when trying to balance benefit and safety.

are limited to a single efficacy endpoint and do not include formal quantitative benefit-risk balance analyses while those are necessary for regulatory and clinical decision-making.

HR is a parameter that can be difficult to interpret by clinicians and patients. Some parameters are easier to interpret and communicate including the comparison of survival probability at specific time point, the comparison of median survival, or restricted mean survival time (RMST). It is recognized that median survival and survival probabilities at specific time points are only partial descriptions of survival differences. RMST was proposed to tackle this limitation, but it does not allow the simultaneous analysis of several endpoints.

The *Net Treatment Benefit* (NTB) has been proposed as a possible solution to the previous shortcomings [1–3] and provides a quantitative answer to patients asking "What are my chances of surviving longer with treatment than without?". NTB can be used and interpreted even when hazards are not proportional. When event times are subject to right-censoring, that is, patients are only known to be event-free until a certain date, NTB estimation is known to be biased toward zero. Various corrections of the NTB were proposed in order to mitigate this bias but none of them managed to avoid the dependency between the NTB value and the proportion of censored data [4–7].

In this article, we propose to use a *time-restricted* version of the NTB, called *restricted* NTB (rNTB), which is defined as the probability for a random patient receiving the experimental treatment to have a better outcome (eg, better survival or less toxicity for a similar survival) compared to a random patient in the control group during the first t_r years of treatment. This method requires the choice of a restriction time (eg, t_r = 3 years after the inclusion) [8] and optionally also a threshold of minimal clinically relevant difference in survival (eg, m = 3 months).

We simulated 3 simulated typical scenarios of treatment effects in metastatic setting (chemotherapy vs chemotherapy, immunotherapy vs chemotherapy, and immunotherapy vs immunotherapy), in order to describe the use of rNTB, the power properties of its associated test compared to standard statistical tests, and to provide practical guidance for its use in clinical trials (choice of threshold, restriction time, and inclusion of toxicity or not). We finally illustrate its use by re-analyzing real randomized controlled trial datasets (PRODIGE 24, CHECKMATE 066 and 067).

2. Methods

2.1. Net Treatment Benefit and restricted Net Treatment Benefit

Typical RCTs compare an experimental arm to a control arm to quantify a drug effect with respect to possibly several outcomes (say p) and corresponding thresholds of clinical relevance. These outcomes are denoted by $X=(X_1, ..., X_p)$ in the experimental arm and by $Y=(Y_1, ..., Y_p)$ in the control arm while the thresholds are denoted by $=(m_1, ..., m_p)$. With a single outcome, say survival, the NTB is the probability for a random patient in the experimental arm to survive m months longer than a random patient in the control arm minus the probability of the opposite situation (1):

$$NTB(m) = P[X \geq Y + m] - P[Y \geq X + m]$$

In presence of right-censoring or administrative censoring, it is typically not possible to estimate this estimand nonparametrically as no observation is available at late time points. A more tractable estimand is the rNTB:

$$rNTB(t_r, m) = P[X \wedge t_r \geq Y \wedge t_r + m] - P[Y \wedge t_r \geq X \wedge t_r + m]$$

where $x \wedge t_r$ denotes the minimum between x and the restriction time t_r. Both rNTB and NTB are equal to zero if experimental does not differ from the control, it is positive (up to 100%) if experimental is better than the control and negative (down to −100%) if the control group is superior.

Multiple outcomes are handled by deciding upon a hierarchy and analyzing later outcomes when no difference is found with respect to earlier outcomes [9]. For instance, to balance the benefits and risks of a new treatment, the NTB can be used with survival as a first priority endpoint with a threshold

$m_1 = 2$ *months* and toxicity as a second endpoint (eg, presence or absence of a grade 3 toxicity, m_2 infinitesimal):

$$\Delta = P[X_1 \geq Y_1 + m_1] + P[X_2 \geq Y_2 + m_2, \ |Y_1 - X_1| < m_1] - (P[Y_1 \geq X_1 + m_1] + P[Y_2 \geq X_2 + m_2, |Y_1 - X_1| < \tau_1])$$

Toxicity would then be considered whenever survival was similar.

2.2. Estimation in absence of censoring

Assuming that observations are independent and identically distributed (iid) within each arm and independence between the two arms, a sample of *m* patients in the experimental group and *n* patients in the control arm can be used to estimate Δ and Δ_r through generalized pairwise comparisons (GPC). GPC considers all possible pairs of patients, one from each arm. The pair is a 'win' if the outcome of the patient in the experimental group is better (optionally by a certain threshold) than the outcome of the patient in the control group, and a 'loss' if the outcome is worse. The pair is 'neutral' when the two observations are equal, or when the difference of outcomes does not reach the prespecified threshold of clinical relevance. The estimated NTB is then the difference in proportion of wins vs. proportion of losses. The rNTB only differs in that the outcomes are restricted to be at most t_r so win or losses that happening after t_r may be classified as neutral.

The resulting estimator can be seen as a two-sample U-statistic and thus shown to be asymptotically normally distributed. A consistent estimator of the variance can be derived from the H-decomposition of the U-statistic—see reference [10] for details. Wald tests can then be used to assess null hypotheses such as $\Delta = 0$ or $\Delta_r = 0$. The software implementation (R package BuyseTest, CRAN) used in the simulation studies applies an inverse tangent hyperbolic transformation for computing *P* values and confidence intervals to improve coverage and type 1 error control in small samples.

2.3. Estimation in presence of censoring

In trials, limited follow-up time and patient drop-out prevent the observation of the outcome for some patients. The former is also referred to as administrative censoring and the latter as random right-censoring. To be able to estimate the (r)NTB, we will assume that the outcome distribution is independent of the censoring given the group. This does not require the censoring distribution to be the same in both groups but implies that patients who drop out in a given group are not more sick or healthy than those who stay in the trial.

For estimation, we considered the approach proposed by Péron et al. [6,7] and referred to as the Péron estimator. The objective of the Péron estimator was to reduce the bias of the NTB estimator observed with the previous Gehan estimator [11]. In this estimator, pairs involving one or two censored observations are no longer classified as uninformative. Instead, their probabilities to be a win, a loss, or neutral are calculated based on the information contained in the observed censoring times and the Kaplan-Meier estimate of the survival function in each arm.

The resulting Péron estimator is a U-statistic involving estimated parameters. It can be shown to be asymptotically normally distributed and its variance can be quantified via two terms: the H-decomposition of the U-statistic had the parameters be known and an asymptotic expansion of the estimated parameters combined with the gradient of the Péron estimator with respect to these parameters [10]. Hypothesis testing, *P* value, and confidence intervals were derived from the estimate and standard error as in the uncensored case.

2.4. Simulation of randomized trial data sets

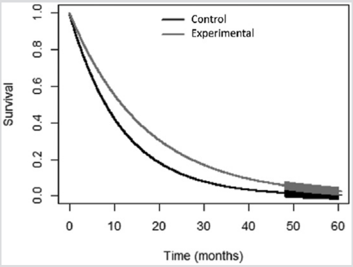
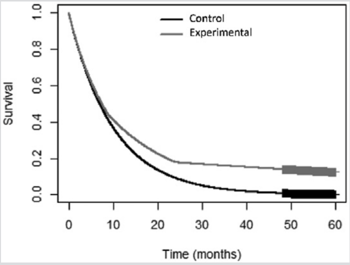
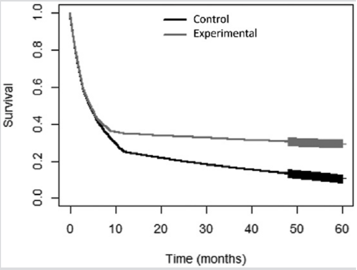
In order to illustrate the features of rNTB, we simulated 3 typical scenarios of survival differences (Table 1, supplementary data 1). In the ''Chemotherapy vs Chemotherapy'' (CvC), the hazards were proportional between the 2 treatment groups, mimicking a trial comparing two palliative chemotherapy regimens. In the other scenarios, the hazards were nonproportionals. The ''Immunotherapy vs Chemotherapy'' scenario (IvC) represents a typical delayed treatment effect as seen in trials comparing an immunotherapy vs a palliative chemotherapy. The ''*Immunotherapy vs Immunotherapy*'' scenario (IvI) represents a comparison between two treatments with delayed effect and different cure rates. For each scenario, three toxicity scenarios were generated: (i) no toxicity, (ii) *equal* 30% toxicity rate, and (iii) *unequal* toxicity favoring the experimental treatment with a 20% toxicity rate in the experimental group and a 30% toxicity rate in the control group. There was no correlation between efficacy and toxicity. Each group was composed of 200 patients included uniformly over a 12-month period. For each dataset, the NTB and rNTB were calculated for various values of follow-up times, thresholds *m,* and restriction times (t_r) (Table 1). The true NTB value was calculated by simulating data with infinite follow-up and not censoring.

3. Results

3.1. Restricted Net Treatment Benefit value does not depend on the follow-up duration

The estimated values of rNTB (without threshold of minimal clinically relevant difference *m*) according to the length of follow-up are reported in Figure 1. NTB estimations converged toward the exact NTB with follow-up time but may require a very long time before reaching it. On the contrary, rNTB tended to converge toward its exact value very fast in an unbiased manner. Interestingly, the expected rNTB value *after a particular restriction time* (t_r) may be

Table 1. Summary of simulation plan and scenarios simulated. The thick bar during the last year corresponds to administrative censoring expected due to the progressive inclusion during the first year

Scenario	Chemotherapy *vs* chemotherapy	Immunotherapy *vs* chemotherapy	Immunotherapy *vs* immunotherapy
Survival curve			
Toxicity (2nd priority outcome)	No toxicity, equal or unequal toxicity		
Restriction time (t_r)	From 12 to 60 mo		
Threshold (m)	From 0 to 24 mo		

Hazard ratio over time is depicted in Figure S1.

deduced from the NTB value when follow-up is equal to the restriction time (t_r). NTB and rNTB at a particular time point before restriction times were similar. In the CvC and IvC scenario, NTB and rNTB reached the true NTB value during the first years of follow-up, as most events were observed in at least one of the treatment groups (Table 1). In the IvI scenario, NTB did not reach the true NTB at 5 years, that is, more than a relevant timeline for a clinical trial.

Interestingly, rNTB had the remarkable property to reach a fixed value corresponding to its true value on complete data, and this value did not change when additional follow-up was gathered. Attaining a stable value is an important characteristic for cross-study comparisons and for interim analyses. It is of particular interest in scenarios involving immunotherapies where the estimation of NTB requires extended follow-up to reach its true value on complete data. rNTB-derived tests had the same power as NTB-derived tests when the restriction time was equal to the follow-up time and its power was stable when additional follow-up is gathered (Fig 2). The effect of a 1-year vs 4-year inclusion period had a limited impact in our simulation results (Figure S2).

3.2. Impact of threshold of clinical relevance on rNTB

rNTB can optionally include a threshold of minimal clinically relevant difference in order to only look at differences that are perceived as *meaningful* for patients and clinicians. Threshold increase tended to decrease the rNTB estimations in the CvC scenario whereas it tended to increase it in scenarios including immunotherapies with cure rates (Fig 3). The power of the associated tests followed a similar interesting behavior with a higher power observed with no or small thresholds for CvC scenarios and higher power observed with large thresholds in scenarios involving

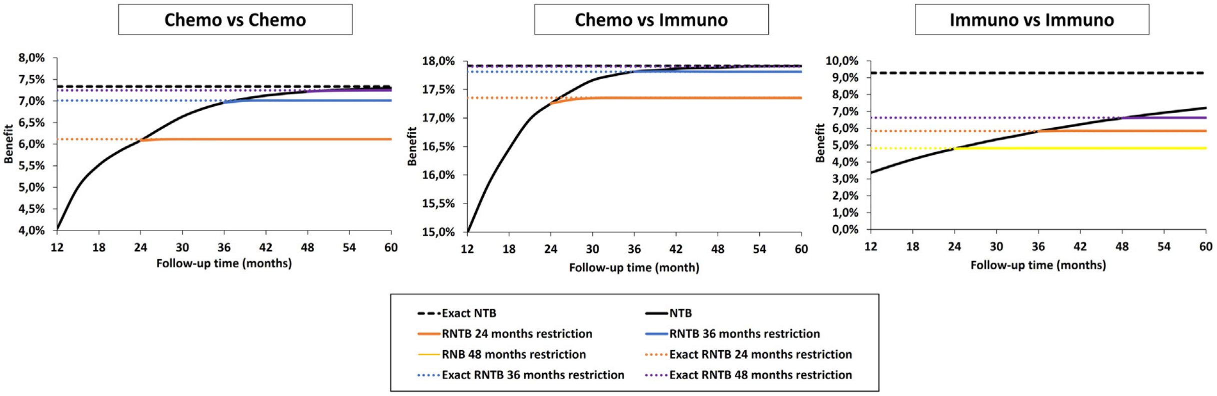


Figure 1. Effect of restriction time (t_r) on rNTB and comparison to NTB. Comparison of true NTB and rNTB with various restriction times depending on time. True NTB and rNTB are the values calculated on complete data without censoring. Of note, NTB and rNTB at a particular time point before restriction are exactly the same. NTB, Net Treatment Benefit; rNTB, restricted Net Treatment Benefit. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

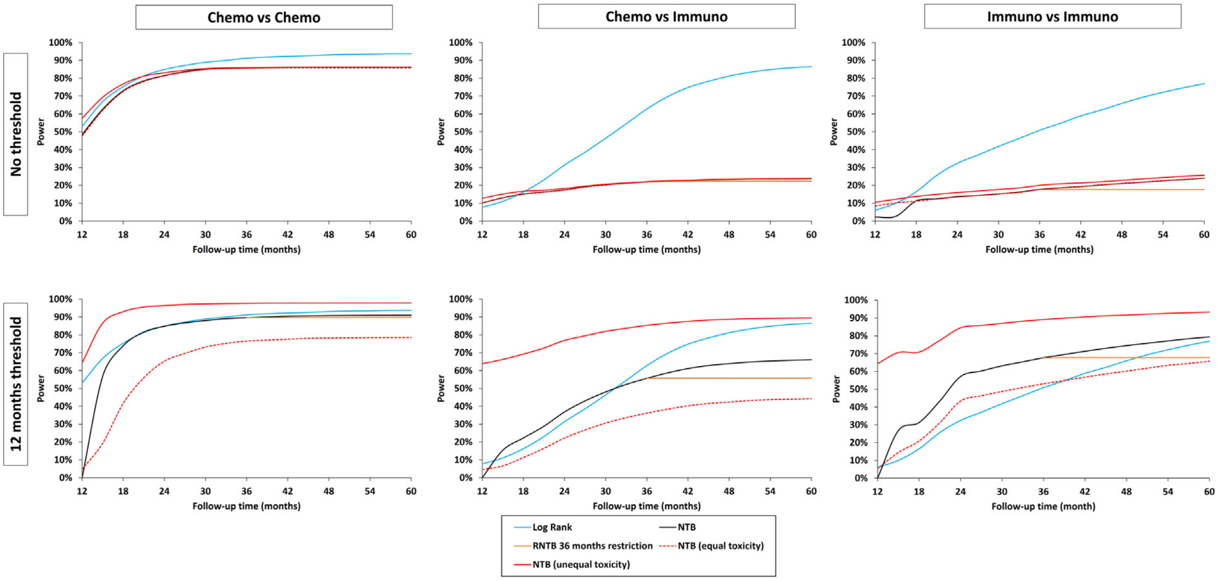


Figure 2. Power comparison. NTB and rNTB with or without toxicity and with or without threshold compared to log rank (LR) in terms of power. NTB, Net Treatment Benefit; rNTB, restricted Net Treatment Benefit. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

immunotherapy, which is explained by the long-term survival benefit observed in these cases (Fig 3).

3.3. *Use of rNTB including toxicity to assess the benefit/risk ratio*

rNTB may be used to evaluate a benefit/risk ratio when toxicity is defined as a second priority criteria. In a scenario of equal toxicity (30%), adding toxicity as a second priority endpoint did not change rNTB estimations (Fig 3), whereas it largely increased in the case of an unequal toxicity that favors the experimental group. This large increase in (r)NTB estimations led to a major power improvement (Fig 2).

3.4. *Comparison with log rank*

Power of rNTB-derived tests was compared with the power of the log rank (LR) test and NTB. Power was compared with or without toxicity, threshold and restriction time in each scenario (Fig 2 and S3). In the CvC scenario with proportional hazards, the power of tests associated with rNTB and NTB were uniformly below the LR test power. It was expected as LR is known to have an optimal power in this setting. When toxicity was included as a second priority endpoint, and in the scenario of unequal toxicity favoring the experimental groups, the power of NTB- and rNTB-derived tests was greatly improved, exceeding the power of the LR test.

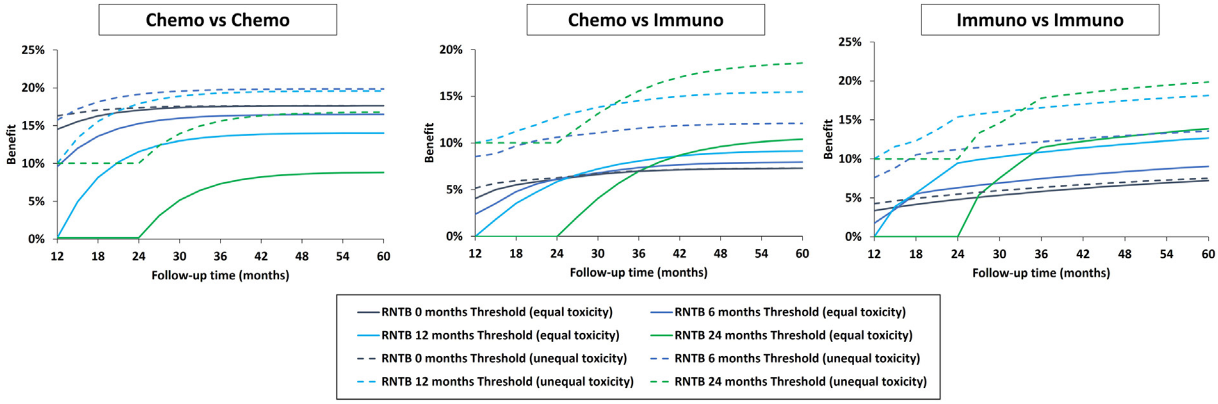


Figure 3. Effect of varying the threshold (m) and the inclusion of toxicities on rNTB in each scenario. Effect of threshold variation with or without inclusion of an equal (30%) or unequal toxicity (20 vs 30%). Of note, NTB and rNTB at a particular time point before restriction are exactly the same. NTB, Net Treatment Benefit; rNTB, restricted Net Treatment Benefit. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

In the IvC and IvI scenarios, NTB- and rNTB-associated tests had similar or lower power compared to LR when thresholds of minimal clinically relevant difference were small (0 to 12 months), while they achieved better power when thresholds were large enough (18 or 24 months, Figure S1). Long thresholds are particularly of interest in cases where a long-term control is expected, for example, in adjuvant or immunotherapy trials where time restriction is particularly of interest. In case of a "crossing curve" scenario, rNTB result highly depends on the restriction period used and reaches a better power than LR if restriction is sufficiently long to capture long-term treatment benefit (Figure S4).

3.5. Application of rNTB to real RCTs

3.5.1. The PRODIGE trial

The *PRODIGE* trial [12] compared the FOLFIRINOX regimen to gemcitabine in metastatic pancreatic ductal adenocarcinoma. It is similar to the simulated CvC scenario, with nearly proportional hazards (Fig 4). At the time of trial design, it could have been anticipated that FOLFIRINOX would be more toxic than gemcitabine. An a priori determination of parameters necessary to plan the analysis could have been done based on prior knowledge like median survival expected in the control arm, threshold of minimal clinical relevance in this pathology, expectation for long-term responses and expectation of larger toxicity. It may a priori have been set to m = 3 months, t_r = 18 months, and no inclusion of toxicity data in the definition of rNTB (Table 2).

We performed a reanalysis of the PRODIGE trial data using rNTB and explored the effect of varying thresholds (m), restriction times (t_r), as well as the inclusion of toxicity as a second priority endpoint. In order to fairly compare rNTB to LR, we calculated the LR *P* value using data with a follow-up equal to the restriction time. Median survival was largely increased from 6.8 months in the gemcitabine arm to 11.1 months in the FOLFIRINOX arm. As expected, the occurrence of clinically relevant toxicities was more frequent in the FOLFIRINOX arm (69.0 vs 59.6%).

In our main rNTB analysis, rNTB (t_r = 18 months, m = 3 months) was equal to 23.2% (95% CI 12.1−33.7) with an associated *P* value = 5.2×10^{-5} (9). This means that patients had a 23.2% higher chance to survive at least 3 months longer in the FOLFIRINOX arm than in the gemcitabine arm during the first 18 months. Of note, in the *quasi*-proportional hazard context in the treatment arm, LR and rNTB *P* value tended to be similar (5.1×10^{-5} at 18 months follow-up for LR) but LR test associated *P* value was the lowest.

Including toxicity as a second priority criteria led to lower rNTB values and higher *P* values for rNTB-associated statistical tests (Fig 4 and S5). As most events occurred before 18 months of follow-up, increasing restriction time above 18 months was of limited interest in terms

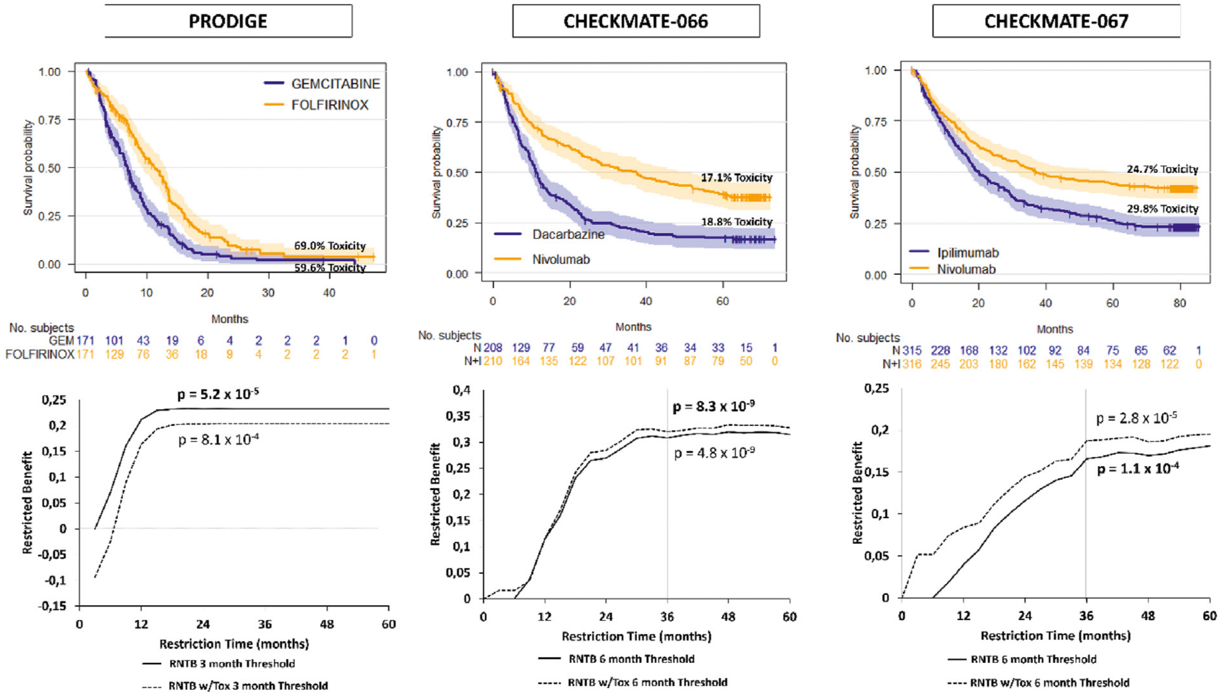


Figure 4. Application of rNTB in randomized controlled trials. The PRODIGE trial, CHECKMATE-066, and CHECKMATE-067 trial (nivolumab arm vs ipilimumab arm) results with or without inclusion of toxicity were used to exemplify the use of rNTB in oncology. Survival curves are displayed on the upper panel, estimation of restricted Net Treatment Benefit depending on inclusion or not of toxicity as secondary prioritized criteria are depicted in the lower panel (with a follow-up at least equal to the restriction time). Detailed *P* values and choice of other thresholds are available in Figure S5. rNTB, restricted Net Treatment Benefit. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2. Parameters that may have been used a priori for rNTB in the 3 clinical trials of interest

Trial	Restriction time (t_r)	Threshold (m)	Inclusion of toxicity as a 2nd criterion
PRODIGE	18 mo	3 mo	No
Checkmate-066	3 y	6 mo	Yes
Checkmate-067	3 y	6 mo	No

rNTB, restricted Net Treatment Benefit.

of power and clinical pertinence. The effect of changing the threshold from 3 months to 0 or 6 months had a limited effect on *P* value and rNTB estimate (Figure S5).

3.6. The Checkmate 066 trial

The *Checkmate 066* trial [13] was designed to compare nivolumab and dacarbazine in metastatic melanoma without BRAF mutation. It lies in between the IvC and IvI scenario (Fig 4) as many patients in the dacarbazine group crossed arm and had immunotherapy in subsequent lines. In the context of an expected cure rate in the nivolumab arm (immunotherapy), a focus on long-term improvement of survival may have been preferred by choosing a quite long minimal clinically relevant benefit in survival (threshold m) and a quite high restriction time to ensure a reasonable follow-up. It may a priori have been set to m = 6 months and t_r = 3 years (Table 2).

Occurrence of clinically relevant toxicities was less frequent in the nivolumab arm (17.1% vs 18.6%, Fig 4). Median survival was largely increased, from 11.2 months in the dacarbazine arm to 37.3 months in the nivolumab arm. In the final analysis, rNTB(t_r = 3 years, m = 6 months, with toxicity) was equal to 32.0% (95% CI 21.7—41.6) with an associated *P* value = 4.8×10^{-9}. Patients had a 32.0% higher chance to either live 6 months longer or to have less toxicity in the nivolumab arm than in the dacarbazine arm during the first 3 years. Of note, LR *P* value with 3 years of follow-up (2×10^{-10}) tends to reach a quite similar value.

A posteriori, including toxicity as a second priority criteria led to a small increase in terms of rNTB value. The rNTB-associated *P* value was not substantially modified by the inclusion of toxicity as a second priority endpoint when a short restriction time is considered and decreased with a longer threshold (Figure S5). As most events occurred at 3 years (both curves achieved their cure rate), increasing restriction time above 3 years had limited interest both in terms of power and clinical pertinence. The effect of changing the threshold from 6 to 12 months had an interesting effect on *P* value and rNTB estimate. A posteriori analysis may have proposed rNTB(t_r = 3 years, m = 12 months, with toxicity) as an even more interesting choice for final analysis.

3.7. The Checkmate 067 trial

The *Checkmate 067* trial [14] was designed to compare nivolumab, ipilimumab, and their combination in metastatic melanoma. We chose to focus the comparison on the nivolumab vs ipilimumab combination in this 3-arm trial. The Kaplan-Meir curves were similar to the IvI scenario (Fig 4).

No major hypothesis regarding a potential increase in toxicity in one group vs another emerged at the time of the trial design, and therefore estimating a benefit-risk ratio (by including toxicity as a second priority endpoint) may have been perceived as an unreasonable risk. The rNTB was then estimated based on survival *only* with a typical restriction time (t_r) set at 3 years and with a 6-month threshold in order to focus on long-term survivors (Table 2).

Occurrence of toxicities was less frequent in the nivolumab arm (24.7% vs 29.8%, Fig 4). Median survival was better in the nivolumab arm at 36.9 months vs 19.9% in the ipilimumab arm. In the final analysis, rNTB(t_r = 3 years, m = 6 months) was equal to 16.6% (95% CI 8.2—24.7) with an associated *P* value = 1.13×10^{-4}. This means that patients had a net 16.6% increase in chance to survive at least 6 months longer in the nivolumab arm than in the ipilimumab arm during the first 3 years.

A posteriori, including toxicity as a second priority criteria would have led to an overall increase in terms of rNTB estimate and a decrease in *P* value (Figure S5). A posteriori analysis may have proposed rNTB(t_r = 3 years, m = 12 months, with toxicity) as an even more interesting choice for final analysis.

4. Discussion

We describe the use of rNTB as an interesting statistical method to describe the treatment's effect on patients in a meaningful and understandable way, based on one or multiple endpoints analyzed simultaneously and including optionally thresholds of minimal clinically relevant differences.

Compared to NTB, the rNTB can be estimated without bias even when the follow-up time is short and when the final form of the survival curves is still unknown. This estimator could then be particularly useful when long-term survival is anticipated.

Just like NTB, rNTB may be used to evaluate a benefit/risk balance, for example, when taking into account both survival and toxicity or quality of life in oncology trials. As expected, including toxicity in the analysis allows to increase power in case of unequal toxicity in favor of the experimental group.

In terms of power, NTB- and rNTB-associated tests are more powerful than the standard LR test in case of unequal toxicity favoring the experimental group, and in scenarios with long-term survival differences when high thresholds of minimal clinically relevant difference are chosen (2). RMST is an interesting alternative to rNTB but it is less versatile as it does not allow the simultaneous analysis of multiple endpoints or the use of thresholds of minimal clinically relevant difference. In oncology, as most relevant toxicities will occur within the first weeks, we did not explore the impact of late adverse events that may induce an imbalance censoring in long-term adverse events data.

5. Conclusion

rNTB is particularly interesting as an alternative to standard statistical tests in nonproportional hazards or in cases with expected unequal toxicities in favor of the experimental arm where it leads to a substantial increase in power compared to LR. In these situations, we recommend the use of rNTB with *Peron* estimator and with use of a threshold of minimal clinically relevant difference (m). rNTB also has the interest to be interpretable by patients and physicians.

CRediT authorship contribution statement

Max Piffoux: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Brice Ozenne:** Writing – review & editing, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Mickaël De Backer:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Marc Buyse:** Writing – review & editing, Validation, Supervision, Methodology. **Jean-Christophe Chiem:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Julien Péron:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Data availability

We will share the code (already online).

Declaration of competing interest

None of the authors declare conflicts of interest regarding this work.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111340.

References

[1] Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Stat Med 2010;29:3245–57.

[2] Peron J, Roy P, Ozenne B, Roche L, Buyse M. The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials. JAMA Oncol 2016;2:901–5.

[3] Péron J, Lambert A, Munier S, Ozenne B, Giai J, Roy P, et al. Assessing long-term survival benefits of immune checkpoint inhibitors using the net survival benefit. JNCI 2019;111:1186–91.

[4] Deltuvaite-Thomas V, Verbeeck J, Burzykowski T, Buyse M, Tournigand C, Molenberghs G, et al. Generalized pairwise comparisons for censored data: an overview. Biom J 2023;65:e2100354.

[5] De Backer M, Legrand C, Péron J, Lambert A, Buyse M. On the use of extreme value tail modeling for generalized pairwise comparisons with censored outcomes. Pharm Stat 2023;22:284–99.

[6] Péron J, Idlhaj M, Maucort-Boulch D, Giai J, Roy P, Collette L, et al. Correcting the bias of the net benefit estimator due to right-censored observations. Biom J 2021;63:893–906.

[7] Péron J, Buyse M, Ozenne B, Roche L, Roy P. An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. Stat Methods Med Res 2018;27:1230–9.

[8] Zhang S, LeBlanc ML, Zhao YQ. Restricted survival benefit with right-censored data. Biom J 2022;64:696–713.

[9] Bebu I, Lachin JM. Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. Biostatistics 2016;17:178–87.

[10] Ozenne B, Budtz-Jørgensen E, Péron J. The asymptotic distribution of the Net Benefit estimator in presence of right-censoring. Stat Methods Med Res 2021;30:2399–412.

[11] Gehan EA. A generalized two-sample Wilcoxon test for doubly censored data. Biometrika 1965;52:650–3.

[12] Conroy T, Desseigne F, Ychou M, Bouché O, Guimbaud R, Bécouarn Y, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. N Engl J Med 2011;364:1817–25.

[13] Robert C, Long GV, Brady B, Dutriaux C, Maio M, Mortier L, et al. Nivolumab in previously untreated melanoma without BRAF mutation. N Engl J Med 2015;372:320–30.

[14] Wolchok JD, Chiarion-Sileni V, Gonzalez R, Rutkowski P, Grob J-J, Cowey CL, et al. Overall survival with combined nivolumab and ipilimumab in advanced melanoma. N Engl J Med 2017;377:2503–4.