Editorial

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.0589
Korean J Radiol 2024;25(11):937-941

Korean Journal of Radiology
KJR

Duration of Response as Clinical Endpoint: A Quick Guide for Clinical Researchers

Seonok Kim[1], Min-Ju Kim[1], Jooae Choe[2]

[1]Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, Seoul, Republic of Korea
[2]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

See the corresponding article "Comparison of Chemoembolization Outcomes Using 70–150 μm and 100–300 μm Drug-Eluting Beads in Treating Small Hepatocellular Carcinoma: A Korean Multicenter Study" at https://doi.org/10.3348/kjr.2024.0231.

Keywords: Survival analysis; Time-to-event analysis; Statistical method; Duration of response; Probability of being in response

A recent study by Lee et al. [1] used 'duration of response (DOR)' as an oncologic outcome parameter. Although uncommon in radiology research studies, DOR is often used as a secondary endpoint in many clinical studies along with the overall response rate (ORR)—the rate of treatment response, such as complete response, partial response, or both depending on the definition used in a study—as a primary endpoint to evaluate the efficacy of treatment [2,3]. Conventional endpoints, including ORR and progression-free survival, have not shown consistent associations with overall survival (OS) benefits in immuno-oncology trials [4]. For example, the biological mechanisms of cytostatic agents can reduce the degree of tumor shrinkage. This can lead to a higher proportion of patients exhibiting only stable disease and consequently lower ORR, despite an achievement of clinically significant improvements in OS. In contrast, the DOR in randomized phase 2 trials may be sensitive and useful for identifying signals of OS [4]. DOR has become a clinically important endpoint for evaluating treatments that offer both immediate and sustained responses, particularly in oncology studies [5,6]. In this article, we discuss the estimation and comparison of DORs between treatment groups.

Received: June 20, 2024 **Revised:** July 31, 2024
Accepted: August 17, 2024
Corresponding author: Seonok Kim, MSc, Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea
• E-mail: seonok@amc.seoul.kr

How to Define and Estimate DOR?

DOR is defined as the duration from the onset of the first response to disease progression or death for any reason. Individuals who do not exhibit any disease progression or death during the follow-up period are censored at the last date of the response assessment.

DOR can differ depending on how intermediate events that may occur during the follow-up period after the first response are handled, and it also has various interpretations [7]. The intermediate events are those that can directly influence the occurrence of an event of interest (i.e., progression or death) and include the following: treatment discontinuation or modification and the start of new therapy. If we are interested in assessing the persistent effect of treatment even after treatment discontinuation and modification, then the intermediate events would not be considered either as the event of interest or as censoring [2]. In contrast, if we are interested in DOR when the patients are on a specific treatment, we could consider the treatment discontinuation as censoring. The handling for the start of a new therapy should be based on the reason why the new therapy was selected. If a physician decides a

new therapy when disease progression is suspected, then the new therapy should be treated as the event of interest. On the other hand, if there is interest in DOR for the patients who have not been administered the new therapy, then those patients who received new therapy are censored at the last response assessment before the start of the new therapy [2,3,8].

The DOR that takes into account the intermediate events mentioned above can be calculated as either a traditional/conditional DOR or an unconditional DOR.

Traditional/Conditional DOR

Traditional/conditional DOR is a conditional estimate calculated only for patients who respond. It is calculated using time-to-event analysis. The common approach is to estimate median DOR, mean DOR, or DOR rate at a specific time using Kaplan–Meier method (Fig. 1). If the censoring rate is high, it may be necessary to determine a truncation time point for estimation. The mean DOR with the truncation time is called the restricted mean DOR [9].

Unconditional DOR Using PBIR

Unconditional DOR addresses the question of "what is the DOR for the treatment?", when we do not know if a patient will be a responder. Huang et al. [10,11] introduced the concept of unconditional DOR, integrating the notions of ORR and DOR. To estimate this, they proposed the probability of being in response (PBIR), which is defined as the proportion of patients who have shown a response and remain in response at present. Every patient has one of the four states at a given time after a zero date (e.g., the start

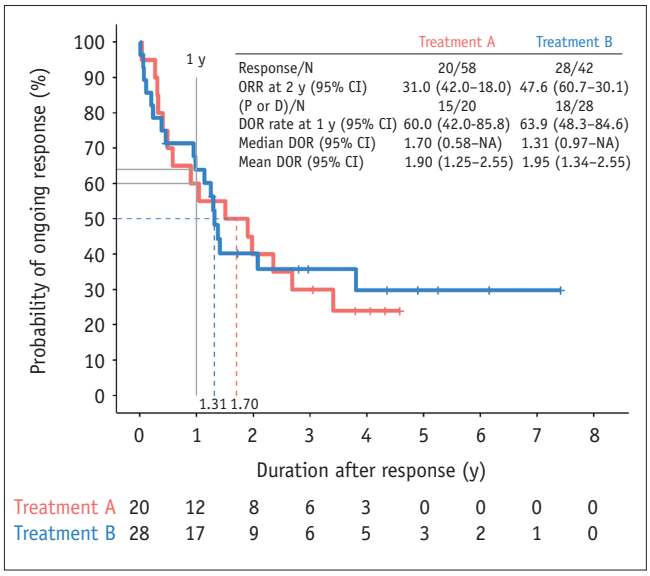

Fig. 1. Example Kaplan–Meier curves with traditional/conditional DOR measures. Two Kaplan–Meier curves for event (P or D)-free survival within responders, one each for treatment A (red) and treatment B (blue). The median DOR, mean DOR, and DOR rate at 1 year of each treatment are presented in the top right portion of the figure. A separate analysis comparing restricted mean DOR up to 4 years between groups showed no statistically significant difference (P = 0.923). DOR = duration of response, P = progression, D = death, ORR = overall response rate, CI = confidence interval, y = years, NA = not available

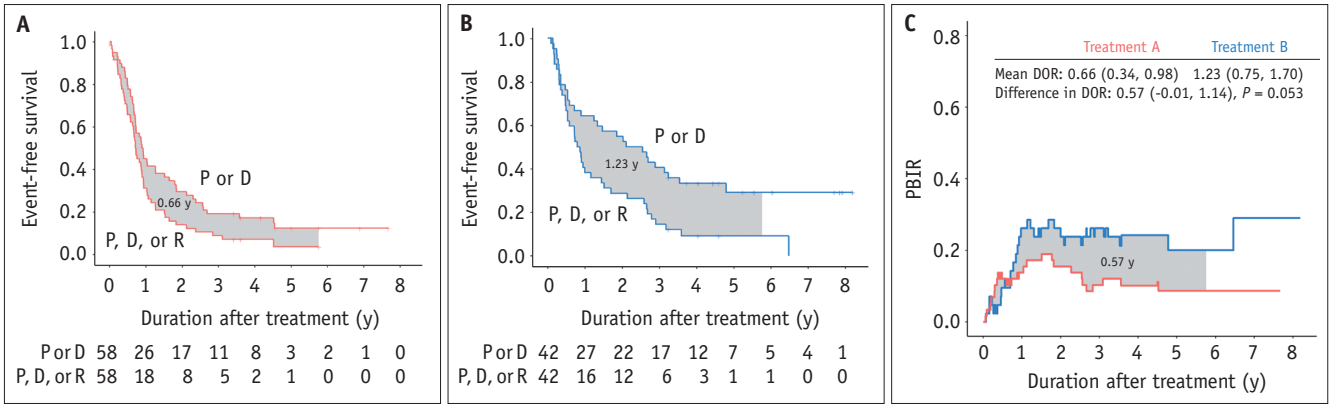

Fig. 2. Unconditional mean DOR using PBIR curve. **A:** Kaplan–Meier curves for event (P, D, or R)-free survival and event (P or D)-free survival in treatment A. The red PBIR curve in **(C)** is the difference between the two Kaplan–Meier curves. The area (gray) between the two Kaplan–Meier curves, which is the same as the area under corresponding PBIR curve indicates the mean DOR. The restricted mean DOR up to 5.75 y is 0.66 y. The truncation point is chosen as the shorter of the last censoring times among the two treatments. **B:** Kaplan–Meier curves for event (P, D, or R)-free survival and event (P or D)-free survival in treatment B. The blue PBIR curve in **(C)** is the difference between the two Kaplan–Meier curves. The restricted mean DOR up to 5.75 y (gray) is 1.23 y. **C:** Difference in mean DORs presented as difference in the areas under the PBIR curves. The gray area is the difference in mean DORs between two treatments, and it is not statistically significant (0.57 y, P = 0.053). DOR = duration of response, PBIR = probability of being in response, P = progression, D = death, R = response, y = years

https://doi.org/10.3348/kjr.2024.0589

Korean Journal of Radiology
KJR

of treatment): currently responding [R], disease progression [P], death [D], or not yet in the state of P, D, or R. PBIR is defined as the difference between the probability of P, D, or R event and the probability of P or D event at a given time, which can be calculated by Kaplan–Meier estimates (Fig. 2). The area under PBIR curve during follow-up period corresponds to the unconditional mean DOR.

How to Compare DORs Between Treatment Groups?

In a comparative study, it is essential to appropriately compare estimated DORs. Most studies traditionally used the comparison of estimated DORs observed only in responders. Calculating DOR for only responders and excluding the non-responders cannot determine whether one treatment is better than another. For example, consider an ineffective treatment that achieves a response only in patients who have a low disease burden at baseline and are less likely to experience disease progression [11]. For this ineffective treatment, the DOR in responders alone may appear deceptively large [11]. If two treatments have the same ORR, traditional/conditional DOR can be used to determine the more effective treatment; however, if one treatment has a lower ORR, a longer traditional/conditional DOR does not necessarily indicate better efficacy. In these situations, a comparison of traditional/conditional DOR between treatment groups may lead to a biased result. Therefore, to reduce the bias caused by differences in ORR, we propose comparing DORs using PBIR or within a subset where the probability of response between two groups is similar.

Comparison of Unconditional DORs Using PBIR

The comparison of DORs using PBIR involves an unconditional comparison of the entire population instead of limiting it to the responder population, thereby avoiding the selection bias (Fig. 2C).

One key consideration for comparing DORs using PBIR is the choice of truncation point i.e., comparison of restricted mean DORs. The principle of selecting a truncation point is to enable valid inferences about the probability of P, D, or R event and the probability of P or D event [9,12]. This is connected to the censoring rates at each time point. If a sufficient period of follow-up has been conducted to ensure that all patients experience P or D events, then the selection of the truncation point becomes irrelevant. If there are only sparse censoring cases before and after the selected truncation point, we need to be cautious in making inference for DOR up to the truncation point. In these scenarios, advancing a time point to where some patients are still at risk is more reliable [12]. R package 'PBIR' are available for analysis [13].

Comparison of Conditional DORs Within a Subset Where the Probability of Response is Similar Between the Two Groups

Korn et al. [14] proposed that reconstructing the subset of responders should be considered to reduce the bias in estimating treatment effects to compare the traditional/conditional DOR between treatment groups. Specifically, the subset was constructed by removing patients in the experimental group who had the lowest likelihood of response when administered control treatment (e.g., those with the shortest survival) and conversely by adding non-responders in the control group who had the highest likelihood of response when administered experimental treatment (e.g., those with the most tumor-burden reduction). Given that DOR of the added non-responders in the control group was not observed, it was assumed as the disease-free period from the zero date (e.g., the start of treatment). The size of the subset was set to be the same as the overall allocation ratio. This approach requires the validity of the measures used to estimate the likelihood of response in the hypothetical group that did not actually occur. Even if the measures are valid, the meaning of the subsets in this approach remains ambiguous. It is doubtful whether it can effectively prove scientific hypotheses.

Matsuyama & Morita [15] proposed applying Frangakis & Rubin's method [16] for estimating the average causal effects within a subset of patients who are likely to respond in both treatment groups. This method is similar to propensity score analysis for causal inference [17,18], and the proposed approach consists of three steps.

1) Modelling: Fit separate logistic regression models with several covariates to predict the probability of response in each treatment group.

2) Prediction (the generation of propensity score): Predict the probability of response if each patient were administered the alternative treatment by applying the regression parameters of the other group estimated in step 1.

3) Weighting: Calculate the weighted mean DOR for each group by applying the probability of response generated in step 2 as individual weights.

In step 3, the potential outcomes are compared among

 https://doi.org/10.3348/kjr.2024.0589

patients who are responders, specifically within the principal stratum. The efficiency of this approach depends on the adequacy of fitted models that can predict the probability of response in each treatment group. In other words, the key is how well we can collect the covariates that affect the response.

CONCLUSION

DOR is a robust metric that integrates both response status and response duration information [4]. For DOR estimation, it is important to determine how to handle intermediate events based on the purpose of the study and provide appropriate interpretations accordingly. When comparing DORs, we recommend methods that reduce bias originating from the difference in response rates between treatment groups, such as PBIR and Matsuyama's approach within the principal stratum. By appropriately reporting DOR, we can gain insights into the efficacy of a particular treatment and support decision-making in patient care.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Seonok Kim. Writing—original draft: Seonok Kim. Writing—review & editing: Min-Ju Kim, Jooae Choe.

ORCID IDs

Seonok Kim
https://orcid.org/0000-0001-9010-5460
Min-Ju Kim
https://orcid.org/0000-0003-4600-5352
Jooae Choe
https://orcid.org/0000-0003-0486-4626

Funding Statement

None

REFERENCES

1. Lee BC, Kim GM, Park J, Chung JW, Choi JW, Chun HJ, et al. Comparison of chemoembolization outcomes using 70–150 μm and 100–300 μm drug-eluting beads in treating small hepatocellular carcinoma: a Korean multicenter study. *Korean J Radiol* 2024;25:715-725
2. Shah BD, Ghobadi A, Oluwole OO, Logan AC, Boissel N, Cassaday RD, et al. KTE-X19 for relapsed or refractory adult B-cell acute lymphoblastic leukaemia: phase 2 results of the single-arm, open-label, multicentre ZUMA-3 study. *Lancet* 2021;398:491-502
3. Levy S, Verbeek WHM, Eskens FALM, van den Berg JG, de Groot DJA, van Leerdam ME, et al. First-line everolimus and cisplatin in patients with advanced extrapulmonary neuroendocrine carcinoma: a nationwide phase 2 single-arm clinical trial. *Ther Adv Med Oncol* 2022;14:17588359221077088
4. Hu C, Wang M, Wu C, Zhou H, Chen C, Diede S. Comparison of duration of response vs conventional response rates and progression-free survival as efficacy end points in simulated immuno-oncology clinical trials. *JAMA Netw Open* 2021;4:e218175
5. U.S. Food and Drug Administration. Clinical trial endpoints for the approval of cancer drugs and biologics: guidance for industry [accessed on July 31, 2024]. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics
6. Pilié PG, Jonasch E. Durable complete response in renal cell carcinoma clinical trials. *Lancet* 2019;393:2362-2364
7. Weber HJ, Corson S, Li J, Mercier F, Roychoudhury S, Sailer MO, et al. Duration of and time to response in oncology clinical trials from the perspective of the estimand framework. *Pharm Stat* 2024;23:91-106
8. Dimopoulos MA, Beksac M, Pour L, Delimpasi S, Vorobyev V, Quach H, et al. Belantamab mafodotin, pomalidomide, and dexamethasone in multiple myeloma. *N Engl J Med* 2024;391:408-421
9. Huang B, Tian L. Utilizing restricted mean duration of response for efficacy evaluation of cancer treatments. *Pharm Stat* 2022;21:865-878
10. Huang B, Tian L, Talukder E, Rothenberg M, Kim DH, Wei LJ. Evaluating treatment effect based on duration of response for a comparative oncology study. *JAMA Oncol* 2018;4:874-876
11. Huang B, Tian L, McCaw ZR, Luo X, Talukder E, Rothenberg M, et al. Analysis of response data for assessing treatment effects in comparative clinical studies. *Ann Intern Med* 2020;173:368-374
12. Tian L, Jin H, Uno H, Lu Y, Huang B, Anderson KM, et al. On the empirical choice of the time window for restricted mean survival time. *Biometrics* 2020;76:1157-1166
13. Luo X, Huang B, Tian L. PBIR: estimating the probability of being in response and related outcomes [accessed on July 31, 2024]. Available at: https://cran.r-project.org/web/packages/PBIR
14. Korn EL, Othus M, Chen T, Freidlin B. Assessing treatment efficacy in the subset of responders in a randomized clinical trial. *Ann Oncol* 2017;28:1640-1647
15. Matsuyama Y, Morita S. Estimation of the average causal effect among subgroups defined by post-treatment variables. *Clin*

https://doi.org/10.3348/kjr.2024.0589

Trials 2006;3:1-9
16. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002;58:21-29
17. Robins JM. *Marginal structural models versus structural nested models as tools for causal inference*. In: Halloran ME, Berry D, eds. *Statistical models in epidemiology, the environment, and clinical trials*. New York: Springer, 2000:95-133
18. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550-560

 https://doi.org/10.3348/kjr.2024.0589