Title

Evaluating Treatment Tolerability in Cancer Clinical Trials Using the Toxicity Index

Journal

Journal of the National Cancer Institute, 112(12)

ISSN

0027-8874

Authors

Gresham, Gillian
Diniz, Márcio A
Razaee, Zahra S
et al.

Publication Date

2020-12-14

DOI

10.1093/jnci/djaa028

Peer reviewed

Evaluating Treatment Tolerability in Cancer Clinical Trials using the Toxicity Index

Gillian Gresham, PhD[1,2]; Márcio A. Diniz, PhD[1]; Zahra S. Razaee, PhD[1]; Michael Luu, MPH[1]; Sungjin Kim, MS[1]; Ron D. Hays, PhD[3,4,5]; Steven Piantadosi, MD, PhD[2]; Mourad Tighiouart, PhD[1]; Greg Yothers, PhD[6]; Patricia A. Ganz, MD[4,7,8]; André Rogatko, PhD[1]

[1] Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles
[2] Brigham and Women's Hospital, Harvard University, Boston
[3] Department of Medicine, Division of General Internal Medicine & Health Services Research, David Geffen School of Medicine at University of California, Los Angeles
[4] Department of Health Policy and Management, UCLA Fielding School of Public Health
[5] RAND Corporation, Santa Monica
[6] University of Pittsburgh and NRG Oncology, Pittsburgh
[7] Center for Cancer Prevention and Control Research, Jonsson Comprehensive Cancer Center, University of California, Los Angeles
[8] Department of Medicine (Hematology/Oncology), David Geffen School of Medicine at University of California, Los Angeles

Corresponding Author:

André Rogatko, PhD
Director, Biostatistics and Bioinformatics Research Center
Samuel Oschin Comprehensive Cancer Institute, Cedars Sinai Medical Center
700 N. San Vincente Blvd., Suite G-588
West Hollywood, CA 90069
Andre.Rogatko@cshs.org
310-423-3316

Keywords: Toxicity; Adverse event; Toxicity analysis; Toxicity index; Clinical trial; Oncology

Abstract

Background: The National Cancer Institute Moonshot℠ research initiative calls for improvements in the analysis and reporting of treatment toxicity to advise key stakeholders on treatment tolerability and inform regulatory and clinical decision-making. This study illustrates alternative approaches to toxicity evaluation using the National Surgical Adjuvant Breast and Bowel Project (NSABP-R04) clinical trial as an example.

Methods: NSABP-R04 was a neoadjuvant chemo-radiation trial in stage II-III rectal cancer patients. A 2x2 factorial design was used to evaluate whether the addition of oxaliplatin (Oxa) to 5-fluorouracil (5FU) or capecitabine (Cape) with radiation therapy improved local-regional tumor control. The toxicity index (TI), which accounts for the frequency and severity of toxicities, was compared across treatments using multivariable probabilistic index models (PIMs), where Pr A < B indicates the probability that higher values of TI were observed for A when compared to B. Baseline age, gender, performance status (PS), body mass index (BMI), surgery type and stage were evaluated as independent risk factors.

Results: A total of 4,560 toxicities from 1,558 patients were analyzed. Results from adjusted PIMs indicate that oxaliplatin-containing regimens had statistically significant ($p<0.001$) probability for higher TI compared to regimens without oxaliplatin: Pr 5FU < 5FU + Oxa = 0.619 (95% CI 0.560-0.674); Pr 5FU< Cape + Oxa = 0.627 (95% CI 0.568-0.682); Pr Cape < 5FU + Oxa =0.587 (95% 0.527-0.644); and Pr Cape < Cape+ Oxa = 0.596 (95% 0.536-0.653).When compared to other existing toxicity analysis methods, TI provided greater power to detect differences between treatments.

Conclusions: This paper uses standard data collected in a cancer clinical trial to introduce descriptive and analytic methods that account for the additional burden of multiple toxicities. These methods may provide a more accurate description of a patient's treatment experience that could lead to individualized dosing for better toxicity control. Future research will evaluate the generalizability of these findings in trials with similar drugs.

For more than 60 years, cancer clinical trials have used an observer-rated toxicity grading system: Common Terminology Criteria for Adverse Events (CTCAE), which assesses the severity of various organ system toxicities associated with treatment.[1] CTCAE data collection in a trial provides detailed longitudinal information on the severity and types of toxicity. However, standard methods for summarizing CTCAE toxicities do not capture the complete toxicity experience over the course of treatment. For instance, maximum grade analysis involves the aggregation of toxicities by highest grade experienced over time and does not account for the cumulative burden that multiple toxicities may introduce, or the persistence and chronicity of some lower-grade toxicities.[2-6]

Toxicity reports are further limited by ignoring baseline risk factors that may contribute to treatment burden. Although information on demographic and clinical characteristics is collected for most clinical trials, it is often presented separately and seldom evaluated within the context of toxicity. Understanding which factors predict greater toxicity is critical to determining optimal treatment approaches and identifying those at higher risk for toxicity. For instance, host factors such as baseline performance status, older age, and gender, or disease-specific factors such clinical stage or surgery type received, are known predictors of survival and treatment outcomes, and should be considered when evaluating toxicity.[7,8,9,10] In recognition of deficiencies in toxicity reporting, the National Cancer Institute (NCI) launched a Cancer Moonshot℠ funding opportunity to accelerate research on improved approaches to evaluating the tolerability of cancer treatments.

This paper examines new strategies for understanding treatment toxicity applied to existing data from a large randomized clinical trial, the National Surgical Adjuvant Breast and Bowel Project (NSABP-R04).[11,12] Using new statistical approaches and graphical displays to summarize the toxicity data, we demonstrate how one can optimize the use of available information and provide a more complete and accurate account of which patients are at greatest risk for toxicity at the completion of a trial.

Methods

Methods for analyzing toxicity

We applied three different methods for analyzing toxicity: the toxicity index (TI)[13], the maximum grade and average toxicity. The TI was developed as a summary measure to better discriminate patients based on their overall toxicity experiences, accounting for all observed toxicity grades rather than just the most severe one.[13] A subject's TI score is defined as a function of the ordered toxicity grades, where the toxicity grades are represented in descending order by the sequence. The TI is computed according to the following algorithm:[14]

$$\mathrm{TI} = \sum_{i=1}^{n} \frac{X_i}{\prod_{j=1}^{i-1}(1 + X_j)}$$

The TI has the following properties: Any score 3 corresponds to the usual definition of dose limiting toxicity (DLT), and the maximum toxicity grade is the integer part of the final score. For example, a TI=3.0 indicates a single grade 3 toxicity, whereas a score of 3.5 indicates at least one grade 3 toxicity plus additional toxicity; all toxicity grades are represented in the score, though lower grades contribute less to the final score; the score is a number between 0 and 5.83 (*See Supplementary Material for explanation of upper limit)*; Multiple toxicities of the same grade yield a TI score slightly less than that generated by a single toxicity of the next higher grade; and when several patients are compared with relation to their toxicity profile, the TI preserves their ranking.

The second approach, maximum grade analysis, yields an incidence rate that is summarized by the most severe grade observed across all events, independent of time of occurrence.[2-4] For example, a subject experiencing multiple high-grade toxicities across organ systems, is noted as having only experienced a single high-grade toxicity overall. We also compute the average toxicity, which is the summary statistic used in the Toxicity-Over-Time (Tox-T) approach,[13,15] which requires analysis across multiple treatment cycles.

Data Source

Data from the NSABP-R04 rectal cancer clinical trial were used as a case example for this research.[12] NSABP-R04 was a phase 3 trial conducted between July 2004 and August 2013 (NCT00058474). Eligible patients were diagnosed with surgically resectable stage II or III rectal

adenocarcinoma. The trial was approved by the local Institutional Review Boards and all patients provided written consent, as detailed in the main trial report; however, these secondary analyses were deemed exempt by our institutional IRB.[12] When the trial first opened, patients were randomized to two treatment groups: Infusional 5-fluorouracil (5FU) with pelvic radiation therapy (RT) compared to oral capecitabine (CAPE) with pelvic RT. In 2005, the protocol was amended to add an oxaliplatin (Oxa) option to 5FU and CAPE, resulting in a 2x2 factorial design with four treatment groups: 5FU + RT; 5FU + Oxa + RT; CAPE + RT; CAPE + Oxa+RT. The doses for 5FU and CAPE for the 4-arm amended trial were reduced from seven to five days a week post-amendment to allow for the addition of Oxa[12] (**Figure 1**). Surgery was performed within 6 to 8 weeks after RT completion. The primary outcome was local-regional tumor control, defined as time to local or regional recurrence or surgery if an R0 resection was not achieved. Oxa did not improve the primary outcome, and there was no statistically significant difference between the CAPE and 5FU alone arms.[12] As a result, CAPE with RT has now become the standard of care in subsequent trials.

Baseline assessments included demographics, medical history, height, weight, vitals, physical exam, quality-of-life, imaging, and bloodwork. Laboratory tests (e.g., CBC/Differential, Platelets, Bilirubin, ALP, AST, etc.) were evaluated weekly during treatment and two weeks prior to surgery. Toxicity assessment was conducted using CTCAE version 4.0 graded from 0 (least severe) to 5 (most severe) and grouped by 26 system organ classes. AEs were collected at a single time point after chemoradiation treatment within two weeks of surgery. Over 50 AEs of special interest were selected *a priori* based on clinical expertise concerning the study regimens and evaluated systematically during treatment **(Supplementary Table 1).** Quality of life questionnaire data was collected prior to treatment, at the end of chemoradiation prior to surgery, and then 12 months after surgery, and has been reported in part elsewhere, [11] and is not included here. Patient follow-up for survival and disease progression occurred at every 12 months from surgery for the first two years. The trial included 1,608 participants, with complete toxicity data available for 1,558 patients (our analysis sample). Additional information about the trial design and study population is reported elsewhere.[11,12]

Statistical Analysis

Graphical summaries of the toxicity data included box plots, histograms, and combinations of graphical and tabular results. All graphical summaries were produced in the R statistical package.

Probabilistic Index Models (PIMs), a rank-based method that generalizes the Kruskal-Wallis test, were fit to compute the probability of higher toxicity between groups.[16-20] For example, considering a score S for groups A and B, a probability $Pr(S_A < S_B)$ equal to 0.5 indicates that both groups have similar score S distributions; a probability statistically significantly greater than 0.5 ($Pr > 0.5$) gives evidence that group B has higher score S than group A; a probability statistically significantly less than 0.5 ($Pr < 0.5$) gives evidence that group B has a lower score S than group A. The probability that a score S for one group is greater than or equal to a score S for another group was estimated with a Wald-type 95% confidence interval. P-values were calculated using the Wald statistic, and p-values for multiple comparisons were corrected using Holm's adjustment.[21] In addition, we defined body system-specific TI as the TI calculated considering only toxicities in a given specific body system. Separate PIMs were then fit for each body system that had at least 10 non-zero TI values. All PIMs incorporated covariables of interest, including gender, age, Karnofsky Performance Status (KPS), clinical stage, and intended surgery (sphincter or non-sphincter preserving) at entry. Tests for interactions between gender and treatment were assessed, where an interaction effect was present if the interaction term in the PIMs was statistically significant ($p<0.05$). If the interaction term was not statistically significant, the term was removed from the model. To compare the performance of different analytic approaches, the power to detect treatment differences was estimated for sample sizes of 50, 75, 100, …, 300 patients for each method (TI, Maximum-Grade, Average Toxicity) based on 2000 resamples. Calculations were performed using the R-package pim,[22] and all hypotheses were two-tailed and tested at the 5% statistical significance level.

Results

Patient characteristics

The analytic sample consisted of 1,558 eligible patients. There were 141 subjects analyzed in each treatment group from the 2-arm trial (pre-amendment): (Group 1) 5FU + RT (2-Arm) and (Group 2) CAPE + RT (2-Arm). In the 4-arm 2x2 factorial trial, 316 subjects were randomized to 5FU + RT (Group 3), 321 to 5FU + Oxa + RT (Group 4), 318 to CAPE + RT (Group 5), and 321 to CAPE + Oxa + RT (Group 6) (**Figure 1**). Demographic characteristics including age, gender, clinical stage, and surgical treatment intent were well-balanced across groups as previously reported.[12]

Treatment toxicity

In this study, our only toxicity assessment timepoint was at the end of chemo-radiation therapy and before surgery. Among 1,558 eligible patients from all treatment groups (2-Arm and 4-Arm) there were a total of 4,560 toxicities, of which 3,720 toxicities occurred in the subgroup of 1,276 patients in the 4-Arm trial (post-amendment). **Figure 2** shows the relative proportion of toxicities for each toxicity severity (Y-axis) by the number of toxicities that occurred per patient (X-axis). From this figure, it can be observed that the number ranged from 0-24 toxicities per patient, with the most frequent and severe toxicities occurring in patients treated with Oxa combined with 5FU or CAPE (**Figure 2).**

TI was calculated to provide a quantitative measure of the cumulative burden of treatment toxicity. A summary of the mean, median and interquartile ranges for TI is provided in **Table 1**. TI was lowest in the 5FU 4-Arm Trial (Median = 2.33) and highest in the CAPE + Oxa and 5FU + Oxa arms (Median= 2.98) (**Table 1)**. The mode of the distribution of toxicities per patient was 0 for 5FU, 3 for CAPE, and 4 for both 5FU+Oxa and CAPE+Oxa (**Supplementary Figure 1**). Patients with 1 or 2 toxicities tended to have a median TI < 3, while patients with more than 4 toxicities displayed a median TI $\geq$ 3, which is typically classified as dose-limiting toxicity (DLT). **Figure 3** shows that TI increased with the increasing number of toxicities per patient in each treatment group, thus demonstrating that the severity of toxicities also increases with the number of toxicities occurring per patient.

Probabilistic Index Models

In univariable analysis (**Supplementary Table 2**)., older age, female gender, planned non-sphincter saving surgery, poor KPS (50-60), and BMI <18.5 were statistically significantly associated with increased probability (Pr > 0.5) for higher TI. Treatment with CAPE + Oxa and 5FU+ Oxa also had increased probability of higher TI than either 5FU or CAPE alone (4-arm) (Supplementary Table 2). Additionally, the higher dose of 5FU (2-arm) was associated with greater toxicity as compared to the 4-arm regimen. There were no statistical differences observed between 5FU and CAPE (4-arm) or 5FU + Oxa and CAPE + Oxa.

Multivariable PIMs for the 2-Arm and 4-Arm trials are shown in **Tables 2 and 3**, respectively. The adjusted probability that the 5FU 2-arm had a higher TI than 4-arm (Pr= 0.57, 95% CI 0.51, 0.63, p=0.02) and the probability the CAPE 2-arm had a higher TI than the 4-arm (Pr=0.56, 95% CI 0.50, 0.62, p=0.05) were greater than 0.5, showing that 2-arm single treatments were more toxic than 4-arm single treatments, but only the comparison between the 5FU 2-arm and 4-arm trials was statistically significant (**Table 2**).

Oxaliplatin-containing regimens also had statistically significant probability (Pr>0.5) for higher TI compared to regimens without oxaliplatin in the 4-arm trials: Pr 5FU < 5FU + Oxa= 0.619 (95% CI 0.560-0.674); Pr 5FU< Cape + Oxa = 0.627 (95% CI 0.568-0.682); Pr Cape < 5FU + Oxa =0.587 (95% 0.527-0.644); and Pr Cape < Cape+ Oxa = 0.596 (95% 0.536-0.653) (**Table 3)**. Baseline characteristics independently associated with increased probability of higher toxicity included female gender, poor KPS, low BMI (<18.5) and planned non-sphincter preserving surgery **(Table 3)**. No statistically significant interaction between gender and treatment was observed (p=0.97). We did observe that females had statistically significant higher toxicity than males using body system-specific TI for the following body systems: Blood, Gastrointestinal, General, Investigations, Metabolism, and Reproductive (**Table 4)**.

Comparison with existing toxicity methods

Results from adjusted PIMs for each analysis method (TI, Max-grade, Average Toxicity) are graphically represented in **Figure 4**. The corresponding numerical estimates and 95% confidence intervals are available in **Supplementary Table 3**. Overall, point estimates for the probability of higher score were of greater or equal value when TI was used as compared to maximum grade for all comparisons, except CAPE (2 Arm) < CAPE (4 Arm).

While point estimates and measures of precision were comparable for TI and maximum grade, TI had greater power to detect differences between treatments (**Figure 5**). Thus, the use of TI results in a smaller number of patients needed to detect differences in treatments. For example, a sample size of 95 would be required to detect a difference between 5FU and 5FU + Oxa using TI. The same comparison would require a sample of 117 patients for the maximum grade method or 137 for the Average Toxicity method, resulting in a 19% and 31% difference in required sample sizes, respectively (**Figure 5, Supplementary Table 3**).

Discussion

Current approaches for analyzing and reporting clinical trial toxicity data are limited and do not capture the complete picture of a patient's treatment experience. Most analyses have defaulted to the maximum grade approach, which collapses toxicities across all grades, organ systems, and ignores the extensive toxicity data and baseline risk factors that are available. In this paper, we demonstrate the feasibility of a more comprehensive approach to the presentation and analysis of toxicity data using the NSABP R04 clinical trial as a case example.

Findings from this analysis revealed important differences in toxicity across treatment arms. By supplementing visual displays with the computation of TI scores, we were able to demonstrate the positive relationship between the frequency and severity of toxicities.

The TI also allowed for treatment comparisons, where the probability of a treatment having higher toxicity can be adjusted for baseline factors using PIMs. Applying this method to NSABP R04 toxicity data resulted in statistically significant differences in toxicity between treatment arms that combined 5FU or CAPE with Oxa and RT. The TI was also sensitive to differences in doses of 5FU where toxicity in the 2-Arm trial, at a higher dose (7 days), was statistically greater than the lower dose (5 days) in the 4-arm trials. While the primary NSABP R04 trial publication described differences in the percentage of grade 3-4 toxicities in the 2-Arm and 4-Arm trials, it did not reach statistical significance.[12] Further, there was no information about the frequency or occurrence of less-severe toxicities, using the standard maximum grade approach to present safety results.

Existing trial reports also failed to describe the additional risk that baseline factors may contribute to our understanding of the overall toxicity burden and tolerability of treatment regimens on subgroups of patients within the setting of a randomized trial. Using adjusted PIMs,

we compared TI across treatments and patient characteristics. We found that older age, female gender, worse KPS, and clinician intent for non-sphincter-preserving surgery were statistically associated with higher probability of subsequent toxicity. While the prognostic values of some of these host factors for survival (e.g., age, KPS) are established in the literature, we know of few reports that describe the impact of baseline host factors on treatment toxicity.[23,24,25,26] When reported, these analyses usually occur in secondary analyses long after the primary trial result, and thus may not be promptly reported to the clinicians adopting a treatment regimen. Reporting on baseline characteristics that are risk factors for greater toxicity can better prepare clinicians who apply trial results to the treatment of patients in their clinical practice.

Our analysis also uncovered interesting differences in toxicity that are independently associated with gender. Overall, females had statistically significantly higher toxicity across treatments and body systems than males. It is unclear whether these differences are a result of differences in clinician reports by gender, or if females are at greater risk for toxicity. Earlier studies, more than two decades ago, reported gender-related differences in 5FU toxicity related to hematological toxicity and mucositis, but the sample size and quality of these studies were limited.[23,27-29] Thus, this analysis greatly expands on these past observations, showing multi-system toxicities. We plan to evaluate our gender-related findings in another adjuvant colon cancer trial comparing 5FU with or without Oxa. We have also begun to explore whether there are gender differences in the PRO data that were collected in the R04 trial.[11] The National Institutes of Health now requires all research applications to discuss gender as a biological variable and there are increasing reports of gender differences in the newer immunotherapy treatment trials, where there are known differences between men and women with regard to the immune system as well as other factors.[30] In retrospect, we may have missed an opportunity to identify an important variable that is closely related to treatment toxicity and tolerability for some regimens, and future evaluations of treatment toxicity should consider gender-specific evaluations of toxicity.

There are several strengths associated with the use of TI for toxicity analysis. We show that it contains more information than other toxicity analysis methods by accounting for both the multiplicity and severity of toxicities, without losing the natural interpretability of the maximum grade approach. This added information provides greater power to examine comparisons across treatment types when compared to the maximum grade and average toxicity approaches,

resulting in the least number of patients required to detect differences between treatments, and consequently saving trial resources and time.

As a limitation, the use of TI requires rank-based methods because it does not follow any well-known probability distribution such as the normal distribution. These methods are less powerful than parametric approaches, and rank-based regressions such as PIM are less disseminated. Although one could argue that the decreased power is mitigated due the large sample sizes used in phase III clinical trials, the lack of a distribution assumption makes our conclusions more robust. Furthermore, the TI can be applied to other ordinal scales such as the PRO-CTCAE, which is increasingly being introduced into clinical trial data collection and analysis.[31,32] The use of patient-reported toxicities also addresses deficiencies of clinician-rated CTCAE toxicities that lack standardization, are not systematically rated, are difficult to assess due to their subjective nature (e.g. pain, fatigue, anxiety), leading to underreporting of the frequency and severity of symptoms. Detailed analysis of PRO data from the quality of life questionnaire from the NSABP R04 trial will be presented in an independent report. Future applications of TI may also incorporate weights for different toxicities as determined *a priori* by investigators or patients and included in the analysis of toxicities in clinical trials.

A limitation of this study was the availability of only a single AE assessment time point in the NSABP R04 clinical trial. We plan to explore the use of TI for longitudinal evaluations and compare to other methods that require repeated measures such as Tox-T[15] and TAME[33] to assess whether the benefits of the TI approach hold. While this was previously challenging using rank-based approaches, there are recently-developed methods that can allow for this analysis and be applied to the comparison of treatments using TI.[34,35]

In conclusion, this research used standard data collected in a cancer clinical trial to introduce descriptive and analytic methods that account for the additional burden of multiple toxicities. Our findings demonstrate initial feasibility of TI and its added value in the analysis of toxicity data to improve our understanding of the comparative tolerability across different treatments. These methods may provide a more accurate account of treatment tolerability that could lead to individualized dosing for better toxicity control. Future research will validate the clinical findings observed in the R04 trials with additional trials that used similar drugs.

Funding

This work was supported in part by the National Cancer Institute of the National Institutes of Health (NIH) (1U01CA232859-01) (G.G., M.A.D., Z.R, M.L., S.K., R.D.H., S.P., M.T., G.Y., P.A.G., A.R.); National Cancer Institute (R01 CA188480-01A1) (M.T., A.R.) and NIH National Center for Advancing Translational Science (NCATS) UCLA CTSI (UL1 TR001881-01) (M.A.D., M.T., A.R.); Additional funding included support from the NIH for the original trial (U10-CA180868, U10-CA180822, UG1-CA189867, U10-CA180888, U10-CA180820, and U10-CA180821)

Notes

Role of the funder: The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication had no role in the

Competing interests: Authors report no conflicts of interest (G.G., M.A.D., Z.R, M.L., S.K., M.T., A.R.)

Acknowledgements: Authors would like to acknowledge members of the scientific advisory committee (Lari Wenzel, PhD; Claire Snyder, PhD, Michael Brundage, MD; N. Lynn Henry, MD PhD, Elisa Long, PhD), and NRG Oncology statisticians (Hanna Bandos, PhD and Reena Cecchini, PhD).

References

1. Health UDo, Services H. Common terminology criteria for adverse events (CTCAE) version 4.0. *National Institutes of Health, National Cancer Institute.* 2009;4(03).
2. Forastiere AA, Goepfert H, Maor M, et al. Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer. *N Engl J Med.* 2003;349(22):2091-2098.
3. Adelstein DJ, Li Y, Adams GL, et al. An intergroup phase III comparison of standard radiation therapy and two schedules of concurrent chemoradiotherapy in patients with unresectable squamous cell head and neck cancer. *J Clin Oncol.* 2003;21(1):92-98.
4. Baselga J, Trigo JM, Bourhis J, et al. Phase II multicenter study of the antiepidermal growth factor receptor monoclonal antibody cetuximab in combination with platinum-based chemotherapy in patients with platinum-refractory metastatic and/or recurrent squamous cell carcinoma of the head and neck. *J Clin Oncol.* 2005;23(24):5568-5577.
5. Goldberg RM, Sargent DJ, Morton RF, Mahoney MR, Krook JE, O'Connell MJ. Early detection of toxicity and adjustment of ongoing clinical trials: the history and performance of the North Central Cancer Treatment Group's real-time toxicity monitoring program. *J Clin Oncol.* 2002;20(23):4591-4596.
6. Mahoney MR, Sargent DJ, O'Connell MJ, Goldberg RM, Schaefer P, Buckner JC. Dealing with a deluge of data: an assessment of adverse event data on North Central Cancer Treatment Group trials. *J Clin Oncol.* 2005;23(36):9275-9281.
7. Hurria A, Togawa K, Mohile SG, et al. Predicting Chemotherapy Toxicity in Older Adults With Cancer: A Prospective Multicenter Study. *Journal of Clinical Oncology.* 2011;29(25):3457-3465.
8. Laghousi D, Jafari E, Nikbakht H, Nasiri B, Shamshirgaran M, Aminisani N. Gender differences in health-related quality of life among patients with colorectal cancer. *Journal of gastrointestinal oncology.* 2019;10(3):453-461.
9. Burkeen J, Pan T, Dalia Y, et al. Gender Differences in Health-Related Quality of Life (hrQOL) in Patients Undergoing Intracranial RT. *International Journal of Radiation Oncology• Biology• Physics.* 2018;102(3):e730.
10. Unger JM, Vaidya R, Albain KS, et al. Sex differences in adverse event reporting in SWOG chemotherapy, biologic/immunotherapy, and targeted agent cancer clinical trials. American Society of Clinical Oncology; 2019.
11. Russell MM, Ganz PA, Lopa S, et al. Comparative effectiveness of sphincter-sparing surgery versus abdominoperineal resection in rectal cancer: patient-reported outcomes in National Surgical Adjuvant Breast and Bowel Project randomized trial R-04. *Annals of surgery.* 2015;261(1):144-148.
12. Allegra CJ, Yothers G, O'connell MJ, et al. Neoadjuvant 5-FU or capecitabine plus radiation with or without oxaliplatin in rectal cancer patients: a phase III randomized clinical trial. *Journal of the National Cancer Institute.* 2015;107(11):djv248.
13. Rogatko A, Babb JS, Wang H, Slifker MJ, Hudes GR. Patient characteristics compete with dose as predictors of acute treatment toxicity in early phase clinical trials. *Clinical cancer research : an official journal of the American Association for Cancer Research.* 2004;10(14):4645-4651.
14. Knuth D. The Art of Computer Programming, Vol. 3, Sorting and Searching, Addison-Wessley. *Reading Mass.* 1973.

15. Thanarajasingam G, Atherton PJ, Novotny PJ, Loprinzi CL, Sloan JA, Grothey A. Longitudinal adverse event assessment in oncology clinical trials: the Toxicity over Time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254. *Lancet Oncol.* 2016;17(5):663-670.
16. De Neve J, Thas O. A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association.* 2015;110(511):1276-1283.
17. De Neve J, Thas O. A Mann–Whitney type effect measure of interaction for factorial designs. *Communications in Statistics - Theory and Methods.* 2017;46(22):11243-11260.
18. De Neve J, Thas O, Ottoy J-P. Goodness-of-Fit Methods for Probabilistic Index Models. *Communications in Statistics - Theory and Methods.* 2013;42(7):1193-1207.
19. Fay MP, Malinovsky Y. Confidence intervals of the Mann-Whitney parameter that are compatible with the Wilcoxon-Mann-Whitney test. *Statistics in Medicine.* 2018;37(27):3991-4006.
20. Thas O, Neve JD, Clement L, Ottoy JP. Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2012;74(4):623-671.
21. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics.* 1979:65-70.
22. Meys J, De Neve J, Sabbe N, Amorim G. Pim: Fit Probabilistic Index Models. R Package Version 2.0. 0.2. 2016.
23. Ilich AI, Danilak M, Kim CA, et al. Effects of gender on capecitabine toxicity in colorectal cancer. *Journal of oncology pharmacy practice : official publication of the International Society of Oncology Pharmacy Practitioners.* 2016;22(3):454-460.
24. Greenlee H, Hershman DL, Shi Z, et al. BMI, Lifestyle Factors and Taxane-Induced Neuropathy in Breast Cancer Patients: The Pathways Study. *J Natl Cancer Inst.* 2017;109(2).
25. Hershman DL, Till C, Wright JD, et al. Comorbidities and Risk of Chemotherapy-Induced Peripheral Neuropathy Among Participants 65 Years or Older in Southwest Oncology Group Clinical Trials. *J Clin Oncol.* 2016;34(25):3014-3022.
26. Bandos H, Melnikow J, Rivera DR, et al. Long-term Peripheral Neuropathy in Breast Cancer Patients Treated With Adjuvant Chemotherapy: NRG Oncology/NSABP B-30. *J Natl Cancer Inst.* 2018;110(2).
27. Sloan JA, Loprinzi CL, Novotny PJ, Okuno S, Nair S, Barton DL. Sex differences in fluorouracil-induced stomatitis. *J Clin Oncol.* 2000;18(2):412-420.
28. Chansky K, Benedetti J, Macdonald JS. Differences in toxicity between men and women treated with 5-fluorouracil therapy for colorectal carcinoma. *Cancer.* 2005;103(6):1165-1171.
29. Zalcberg J, Kerr D, Seymour L, Palmer M. Haematological and non-haematological toxicity after 5-fluorouracil and leucovorin in patients with advanced colorectal cancer is significantly associated with gender, increasing age and cycle number. Tomudex International Study Group. *European journal of cancer (Oxford, England : 1990).* 1998;34(12):1871-1875.
30. Colli LM, Morton LM, Chanock SJ. Sex-related effect on immunotherapy response: implications and opportunities. *J Natl Cancer Inst.* 2019.
31. Basch E, Abernethy AP, Mullins CD, et al. Recommendations for Incorporating Patient-Reported Outcomes Into Clinical Comparative Effectiveness Research in Adult Oncology. *Journal of Clinical Oncology.* 2012;30(34):4249-4255.

32. Basch E, Dueck AC, Rogak LJ, et al. Feasibility assessment of patient reporting of symptomatic adverse events in multicenter cancer clinical trials. *JAMA Oncology*. 2017;3(8):1043-1050.
33. Trotti A, Pajak TF, Gwede CK, et al. TAME: development of a new method for summarising adverse events of cancer treatment by the Radiation Therapy Oncology Group. *The Lancet Oncology*. 2007;8(7):613-624.
34. Chen T, Kowalski J, Chen R, et al. Rank-preserving regression: a more robust rank regression model against outliers. *Statistics in Medicine*. 2016;35(19):3333-3346.
35. Jung SH, Ying Z. Rank- based regression with repeated measurements data. *Biometrika*. 2003;90(3):732-740.

Figure Titles and Legends

Figure 1. CONSORT Diagram.

5FU: 5-fluorouracil; CAPE: Capecitabine; Oxa: Oxaliplatin

*Main study description available in Allegra et al. (2015). All treatment arms included radiation therapy.
†Pre-amendment: In 2004, patients were randomly assigned to either RT+5FU or RT+CAPE for 7 days a week beginning the day of RT start and ending on the last dose of RT.
‡Post-amendment: In 2005, the protocol was amended to add Oxa and resulted in a 2x2 factorial design. Doses were reduced from 7 days to 5 days.

Figure 2. Relative percentages of toxicities per patient by treatment arm

5FU: 5-fluorouracil; CAPE: Capecitabine; Oxa: Oxaliplatin
The x-axis represents the number of adverse events observed per patient. The y-axis represents the percentage of the total associated with each number of adverse events within each treatment arm. Each column is further broken down and color coded by grade, from no adverse events (Grade 0) represented as green to most severe (Grade 5) represented as red. The labeling within each column represents the relative percentage of a given grade among patients with the specified number of adverse events (grades with less than 1 percent are omitted). The table inset presents the grade (G), count (C), and percent (%) of each grade observed for a given treatment arm.

Figure 3. Relationship between Toxicity Index and Number of Toxicities per Patient by Treatment Arm

5FU: 5-fluorouracil; CAPE: Capecitabine; Oxa: Oxaliplatin

Figure 4. Multivariable Probabilistic Index Model Results by Treatment Comparison and Analytic Method.

5FU: 5-fluorouracil; CAPE: Capecitabine; Oxa: Oxaliplatin
Each bar represents the probability index and 95% confidence intervals.

Figure 5. Power to Detect Treatment Differences for Toxicity Index, Maximum Grade, and Mean Toxicity Methods

5FU: 5-fluorouracil; CAPE: Capecitabine; Oxa: Oxaliplatin

Tables

Table 1. Measures of Central Tendency of the Toxicity Index by Treatment

Treatment*	Number of Patients	Number of toxicities	Mean (SD)	Median [IQR]
5FU (2 Arm)	141	385	3.56 (2.81)	3 [1, 5]
5FU (4 Arm)	316	706	3.28 (2.85)	2 [1, 4]
5FU + Oxa (4 Arm)	321	1121	4.36 (3.65)	3 [2, 6]
CAPE (2 Arm)	141	455	4.21 (3.83)	3 [2, 5]
CAPE (4 Arm)	318	761	3.43 (2.59)	3 [1, 5]
CAPE + Oxa (4 Arm)	321	1132	4.51 (3.73)	3 [2, 6]

*5FU: 5-fluorouracil; CAPE: Capecitabine; Oxa: Oxaliplatin

SD = standard deviation, IQR = Interquartile range.

Table 2. Multivariable Probabilistic Index[*] for Toxicity Index Comparing 4-Arm and 2-Arm trials

Variable	Comparison*	5-FLUOROURACIL		CAPECITABINE	
	A < B	Probability (95% CI)	P-Value[§]	Probability (95% CI)	P-Value[§]
Treatment	4 Arm[†] < 2 Arm	0.570 (0.513 – 0.625)	0.02	0.558 (0.499 – 0.615)	0.054
Gender	Male < Female	0.571 (0.513 – 0.628)	0.02	0.609 (0.551 – 0.664)	<0.001
Age (Years)	(Every five years)	0.511 (0.499 – 0.522)	0.07	0.508 (0.496 – 0.520)	0.18
Karnofsky PS	90-100 < 70-80	0.605 (0.532 – 0.673)	0.005	0.579 (0.503 – 0.653)	0.043
	90-100 < 50-60	N/A	N/A	0.933 (0.916 – 0.947)	<0.001
Clinical stage N	Negative < Positive	0.492 (0.438 – 0.547)	0.78	0.522 (0.468 – 0.576)	0.41
Sphincter-saving surgery	Yes < No	0.503 (0.445 – 0.562)	0.92	0.508 (0.448 – 0.568)	0.79
Clinical stage T	T1/T2/T3 < T4	0.537 (0.383 – 0.684)	0.64	0.505 (0.384 – 0.626)	0.93
BMI (kg/m^2)	LT[‡] 18.5 < 18.5 - 25	0.467 (0.241 – 0.708)	0.80	0.369 (0.226 – 0.541)	0.13
	LT 18.5 < 25-30	0.445 (0.224 – 0.690)	0.67	0.362 (0.221 – 0.532)	0.11
	LT 18.5 < GE 30	0.417 (0.206 – 0.664)	0.52	0.309 (0.183 – 0.471)	0.02

[*]Probabilistic Model Interpretation: Comparison A<B denotes the probability that toxicity index for B is higher than toxicity index for A. Probability of 0.5 indicates no difference between comparisons (A=B). If Probability > 0.5, then probability of toxicity index for B is greater than A is high indicating that B has higher toxicity. If the probability < 0.5, then probability of toxicity index for B greater than A is small indicating that A has higher toxicity. Multivariable models were adjusted for gender, 4-arm treatments, age, BMI, clinical T stage, clinical N stage, sphincter-saving surgery, and Karnofsky performance status (PS).

[†]The trial was amended in 2005 to add Oxaliplatin to each of the arms. The doses for 5FU and CAPE for the 4-arm clinical trial were reduced from seven days (2-Arm trial) to five days (4-Arm trials) a week at the same daily dose.

[‡]LT=Less Than; GE= Greater or equal to

[§]All P-values are two-sided and were calculated using the Wald statistic. P-values for multiple comparisons were corrected using Holm's adjustment.

Table 3. Multivariable Probabilistic Index[*] for Toxicity Index Comparing 4-Arm and 2-Arm trials[†]

Variable	Comparison A < B	Probability (95% CI)	P-value[§]
Treatment	5FU < 5FU + Oxa	0.619 (0.560 – 0.674)	<0.001
	5FU < CAPE	0.533 (0.472 – 0.593)	0.30
	5FU < CAPE + Oxa	0.627 (0.568 – 0.682)	<0.001
	CAPE < 5FU + Oxa	0.587 (0.527 – 0.644)	<0.001
	CAPE < CAPE + Oxa	0.596 (0.536 – 0.653)	<0.001
	5FU + Oxa < CAPE + Oxa	0.509 (0.449 – 0.569	0.70
Gender	Male < Female	0.623 (0.589 – 0.655)	<0.001
Age, Years	(Every five years)	0.507 (0.500 – 0.515)	0.04
Karnofsky PS	90-100 < 70-80	0.575 (0.529 – 0.619)	0.001
Clinical stage N	Negative < Positive	0.480 (0.447 – 0.513)	0.24
Sphincter- Saving Surgery	Yes < No	0.540 (0.504 – 0.577)	0.03
Clinical stage T	T1-3 < T4	0.551 (0.468 – 0.632)	0.23
BMI (kg/m^2)	LT[‡] 18.5 < 18.5 - 25	0.441 (0.311 – 0.580)	0.49
	LT 18.5 < 25-30	0.403 (0.278 – 0.542)	0.17
	LT 18.5 < GE 30	0.360 (0.243 – 0.495)	0.04

[*]Probabilistic Model Interpretation: Comparison A<B denotes the probability that toxicity index for B is higher than toxicity index for A. Probability of 0.5 indicates no difference between comparisons (A=B). If Probability > 0.5, then probability of toxicity index for B is greater than A is high indicating that B has higher toxicity. If the probability < 0.5, then probability of toxicity index for B greater than A is small indicating that A has higher toxicity. Multivariable models were adjusted for gender, 4-arm treatments, age, BMI, clinical T stage, clinical N stage, sphincter-saving surgery, and Karnofsky performance status (PS).

[†]The trial was amended in 2005 to add Oxaliplatin (Oxa) to each of the arms. The doses for 5-fluorouracil (5FU) and Capecitabine (CAPE) for the 4-arm clinical trial were reduced from seven days (2-Arm trial) to five days (4-Arm trials) a week at the same daily dose.

[‡]LT=Less Than; GE= Greater or equal to

[§]All P-values are two-sided and were calculated using the Wald statistic. P-values for multiple comparisons were corrected using Holm's adjustment.

Table 4. Multivariable Probabilistic Index Models[*] for System Organ Class-specific Toxicity Index Comparing Gender

System Organ Class (SOC)	Number of observations with non-zero SOC-specific Toxicity Index[†]	Probability[‡] (95% CI[§])	P-Value[¶]
Blood	136	0.553 (0.522 – 0.584)	<.001
Cardiac	14	Not examined due to a small number of nonzero values	
Ear	2	Not examined due to a small number of nonzero values	
Endocrine	1	Not examined due to a small number of nonzero values	
Eye	8	Not examined due to a small number of nonzero values	
Gastrointestinal	672	0.616 (0.566 – 0.662)	<.001
General	362	0.563 (0.521 – 0.604)	<.001
Hepatobiliary	3	Not examined due to a small number of nonzero values	
Immune	26	0.501 (0.489 – 0.514)	1.00
Infections	91	0.518 (0.494 – 0.543)	0.26
Injury	198	0.511 (0.479 – 0.544)	1.00
Investigations	299	0.572 (0.532 – 0.612)	<.001
Metabolism	251	0.554 (0.517 – 0.591)	<.001
Musculoskeletal	71	0.507 (0.486 – 0.528)	1.00
Nervous	116	0.506 (0.480 – 0.533)	1.00
Psychiatric	62	0.500 (0.480 – 0.519)	1.00
Renal	130	0.500 (0.472 – 0.527)	1.00
Reproductive	22	0.522 (0.506 – 0.538)	0.001
Respiratory	29	0.501 (0.488 – 0.515)	1.00
Skin	114	0.512 (0.486 – 0.538)	1.00
Vascular	70	0.519 (0.497 – 0.540)	0.10

[*]Multivariable models were adjusted for gender, 4-arm treatments, age, BMI, clinical T stage, clinical N stage, sphincter-saving surgery, and Karnofsky performance status.

[†]From a total of 1,276 observations.

[‡]Probability that SOC-specific toxicity index for females is higher than SOC-specific toxicity index for males.

[§]Adjusted for multiple tests using the Bonferroni procedure.

[¶] All P-values are two-sided and were calculated using the Wald statistic. P-values for multiple comparisons were corrected using Holm's adjustment.

Figure 1

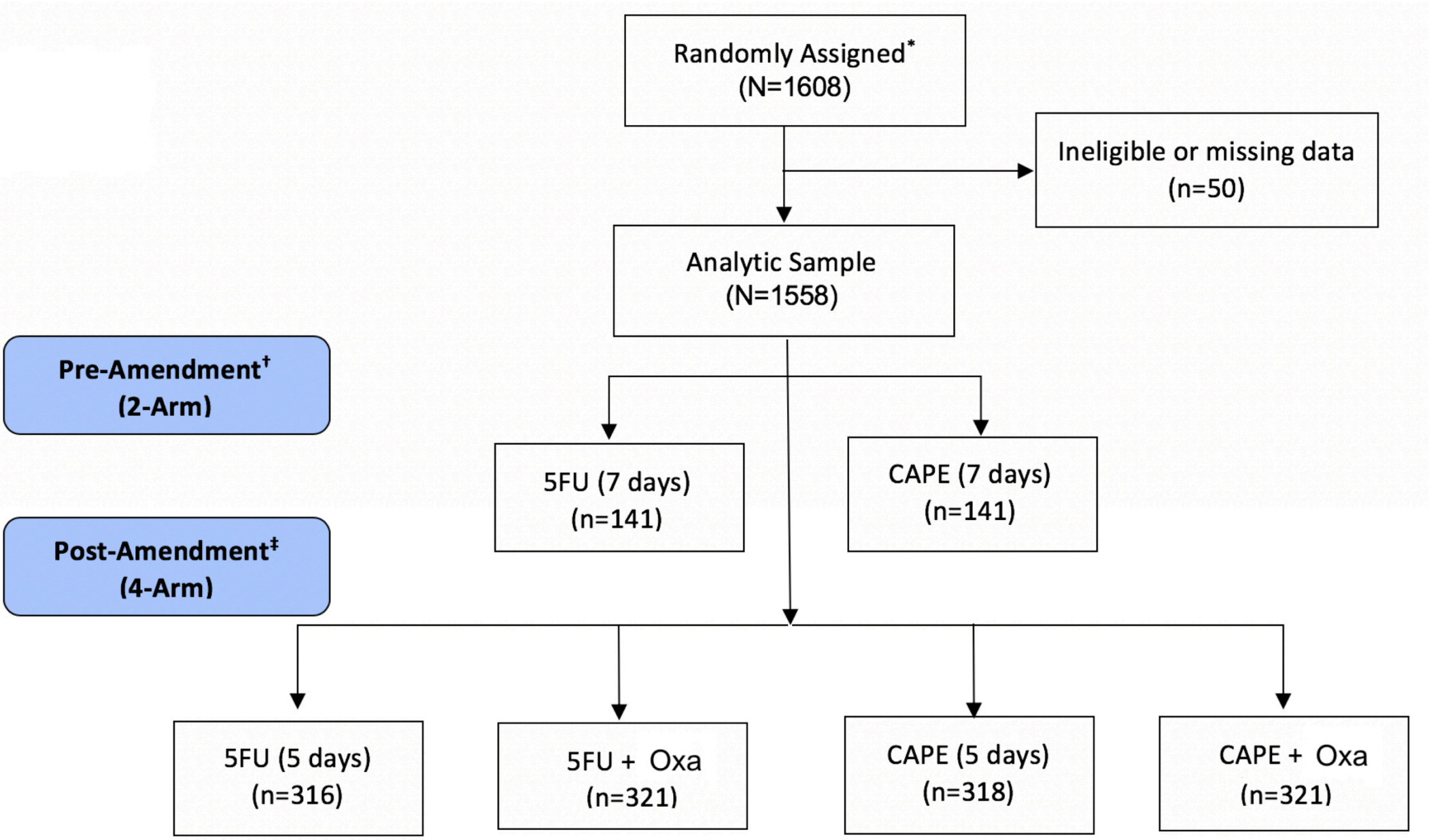

Figure 2

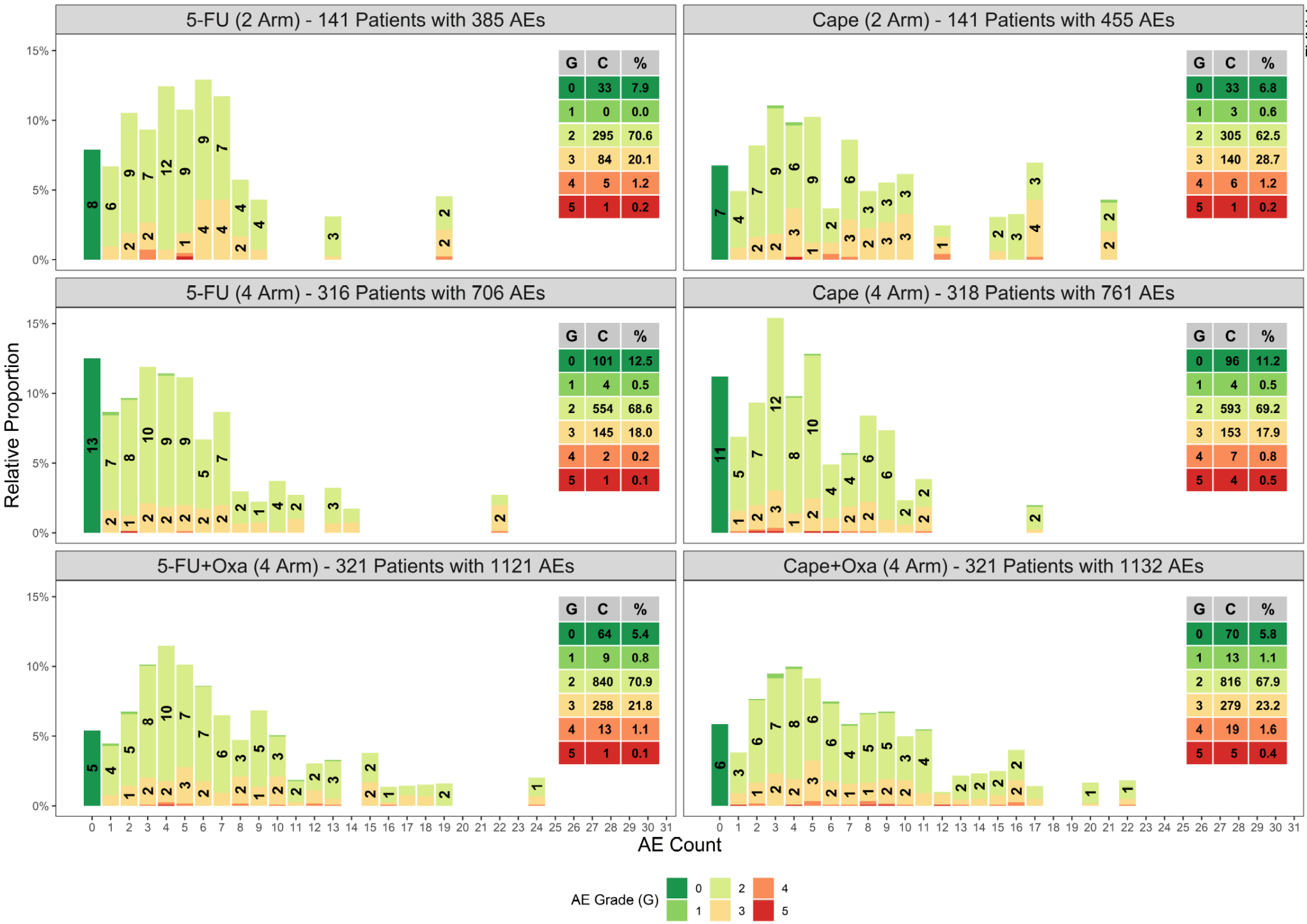

Figure 3

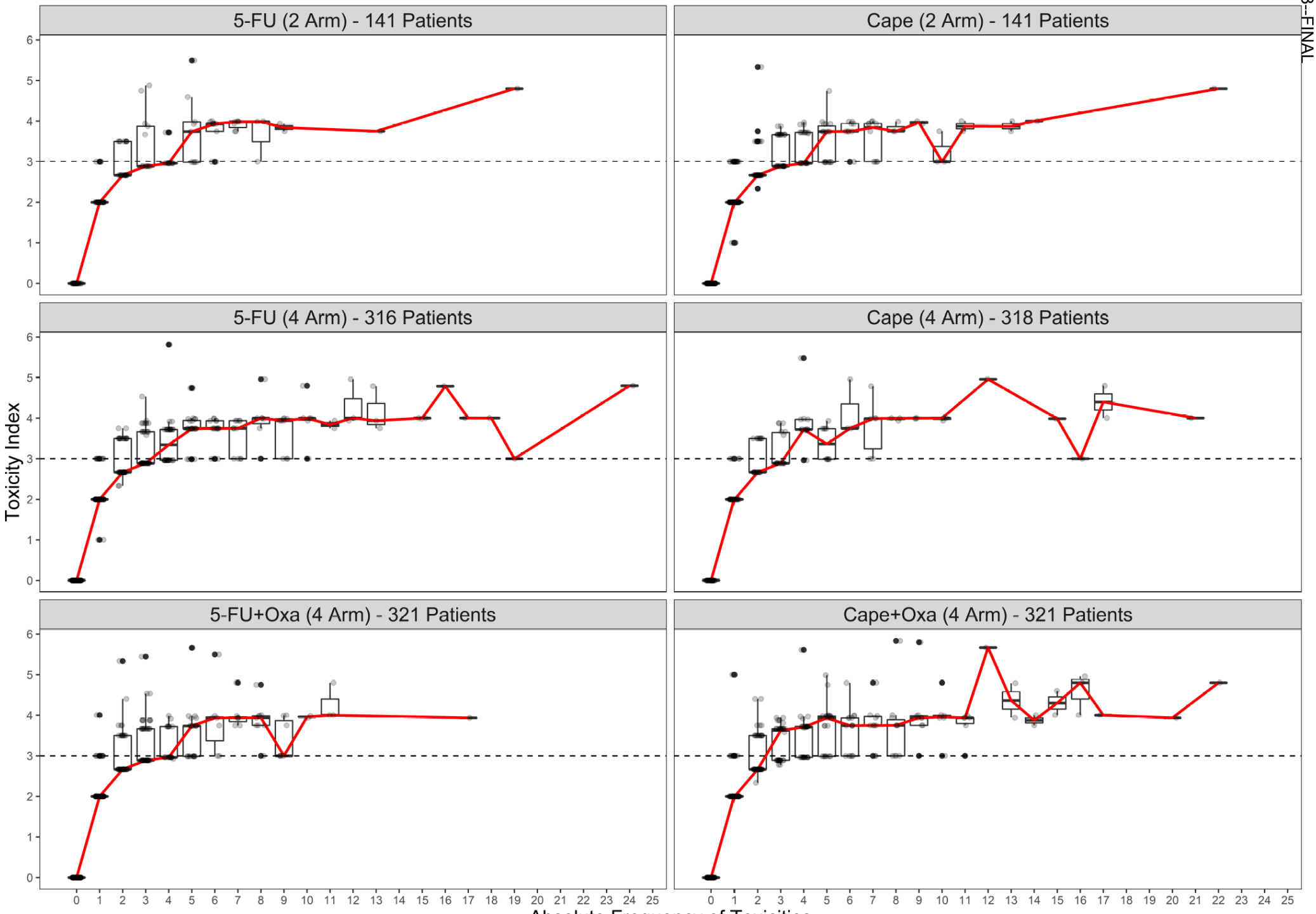

Figure 4

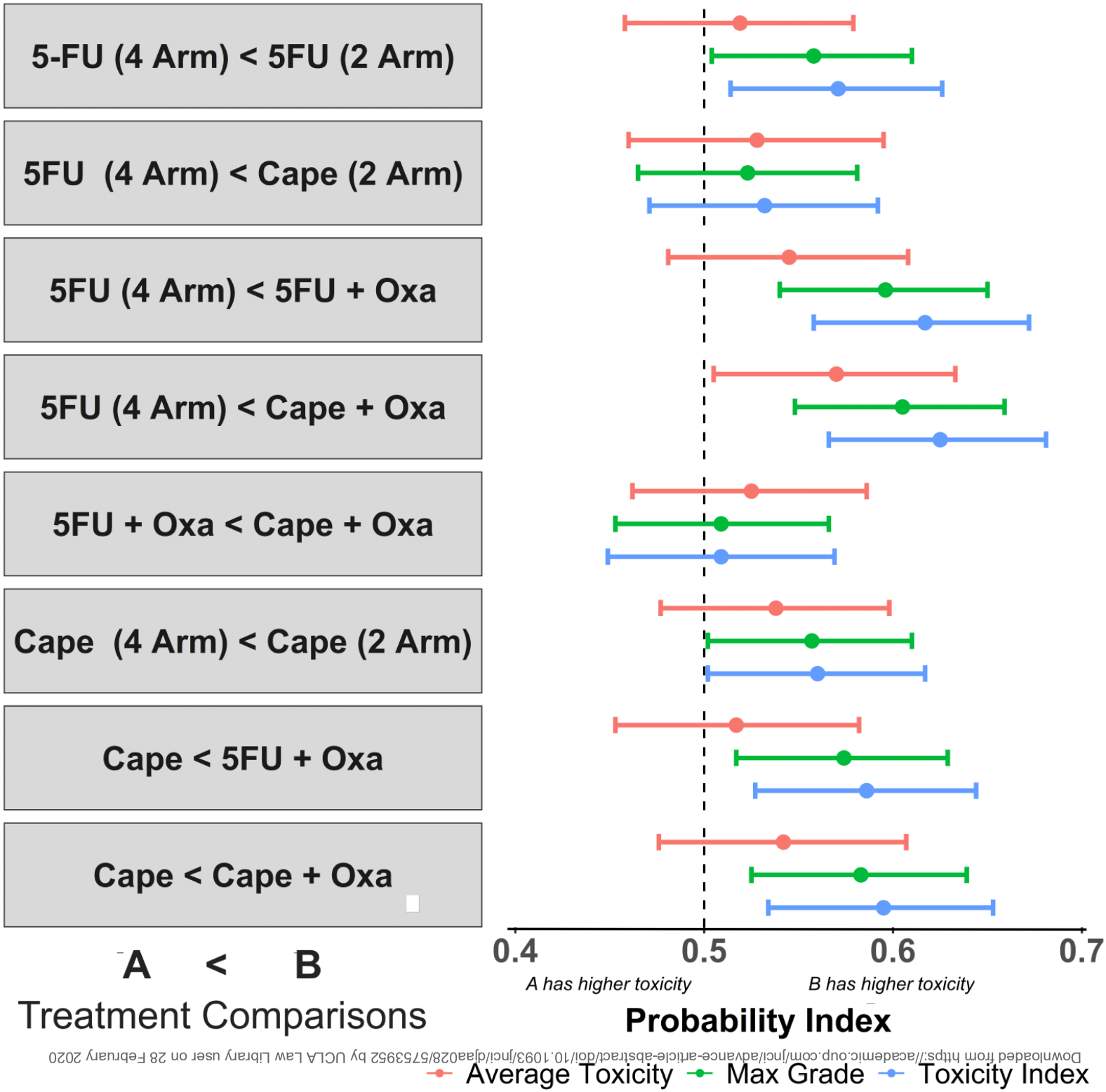

Figure 5--FINAL