UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Developing a taxonomy for longitudinal toxicity behavior

Andrew Kamya

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
Prof. dr. Stijn JASPERS

SUPERVISOR :
Prof. Jan BOGAERTS

2024
2025

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Developing a taxonomy for longitudinal toxicity behavior

Andrew Kamya

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Stijn JASPERS

SUPERVISOR :

Prof. Jan BOGAERTS

Developing a Taxonomy for Longitudinal Toxicity Behaviour

Master Thesis Biostatistics
2024-2025

2nd year Master of Statistics
Hasselt University

Internal Supervisor:
Dr. Stijn Jaspers

Student:
Andrew Kamya (1849786)

External Supervisors:
Dr. Jan Bogaerts
Dr. Jose Silva
Dr. Jammbe Musoro

Date: June 17, 2025

Abstract

Background: Traditional oncology trials report adverse events (AEs) primarily by the maximum toxicity grades of patients, ignoring how these AEs evolve over treatment cycles. This oversight particularly impacts understanding cumulative toxicities, AEs that worsen progressively with ongoing treatment, often necessitating dose adjustments or treatment delays. Capturing these longitudinal patterns is crucial to accurately reflect treatment tolerability and patient management.

Objectives: This thesis aims to develop and validate a taxonomy of longitudinal toxicity behaviors, with particular focus on cumulative toxicities. The specific goals are to: (1) formally define cumulative toxicity and distinguish it from immediate and recurrent patterns; (2) operationalize this definition for measurement in clinical data; (3) visualize toxicity trajectories to identify cumulative trends; (4) quantify those trends using statistical models; and (5) validate the framework on both real-world and idealized simulated datasets.

Methodology: We analyzed patient-level AE data from two EORTC trials: the primary cohort (EORTC 62091, advanced sarcoma) and an external validation set (EORTC 30974, germ cell cancer). After standardizing and cleaning the data, exploratory data analysis (EDA) revealed that, contrary to the ideal of steadily increasing grades, AE severity rarely rose monotonically across cycles, likely reflecting real-time clinical interventions. To benchmark our methods against a perfect 'cumulative' scenario, we then generated a simulated AE dataset matching the real trials in patient count, number of cycles, and grade distributions but enforcing strictly monotonic grade increases for preselected AEs. Ordinal-logistic and mixed-effects models were applied to both datasets to quantify cycle-to-cycle worsening and compute patient-level cumulative toxicity scores.

Results: In the real trial data, most AE trajectories fluctuated up and down, and fewer than 20 % of AEs exhibited any statistically significant cumulative trend. In contrast, the simulated dataset, with enforced monotonic grade increases, showed unbroken progression for all targeted AEs, confirming the sensitivity and specificity of our modeling approach. Validation on EORTC 30974 reproduced the real-world findings, supporting the generalizability of the taxonomy.

Conclusions: We present a clear taxonomy and analytical framework that distinguishes cumulative from immediate and recurrent toxicity patterns. While strictly monotonic cumulative AEs are rare in practice due to clinical management, our methods reliably detect both partial and idealized cumulative behaviors. This dual-dataset validation underscores the utility of longitudinal analysis for proactive toxicity management and offers a robust foundation for future oncology trials.

Key Words: *Cumulative toxicity, adverse events, oncology clinical trials, longitudinal analysis, ordinal logistic models, mixed-effects models, Toxicity Burden; Clinical Trial Safety*

Contents

1 Introduction

Cancer clinical trials routinely collect data on adverse events to evaluate a therapy's safety and tolerability. These toxicity data are traditionally summarized in a static way – for instance, reporting the incidence of each grade 3 or 4 event, or simply each patient's maximum grade toxicity experienced. While such summaries are useful, they inherently lose information about the timing and trajectory of toxicities [1]. A patient experiencing a Grade 3 toxicity early (e.g., in cycle 1) but then tolerating subsequent treatment without complications differs significantly from one who encounters Grade 3 toxicity later (e.g., in cycle 6), following months of accumulated lower-grade side effects. Yet, conventional reporting methods would categorize both patients identically, ignoring their distinct clinical journeys and recovery profiles. Recent analyses have pointed out that this approach fails to depict the evolution of toxicity over time, offering little insight into when AEs emerge and how long they persist. In the era of prolonged cancer therapies, including oral targeted agents taken continuously and multi-cycle complex chemotherapy regimens, this limitation has become increasingly problematic. Thanarajasingam et al. (2015), for example, argued that oncology trial toxicity reporting must modernize to include the dimension of time in order to provide a complete picture of tolerability [1]. In short, there is an imperative to move beyond aggregate worst-grade counts and incorporate longitudinal toxicity profiles into trial analyses. For clinical decision-making, having as accurate and comprehensive a portrait of total toxicity burden as possible is essential to choosing and adjusting therapies appropriately.

One particular aspect of longitudinal toxicity that demands attention is cumulative toxicity. Cumulative toxicities are the adverse effects that worsen progressively with ongoing exposure to treatment, rather than appearing once and then resolving. Many chemotherapy side effects are known or suspected to be cumulative. A classic example is anthracycline-induced cardiotoxicity: patients may tolerate initial doses well, but as the total cumulative dose of an anthracycline (like doxorubicin) increases, the risk of serious cardiac damage rises dramatically. Clinical guidelines often impose a maximum lifetime dose (around $450 - 550mg/m^2$ for doxorubicin) specifically to limit this cumulative heart failure risk [2]. Another example is chemotherapy-induced peripheral neuropathy (CIPN), such as the nerve damage from drugs like cisplatin or paclitaxel. Early cycles might cause only mild tingling, but with each additional cycle patients can develop worsening numbness or pain [3]. Although this toxicity is dose-dependent, its timing and severity vary considerably between individuals. These accumulating side effects are not merely inconveniences; they often become dose-limiting toxicities. Oncologists may be forced to reduce the dose intensity or number of cycles of treatment when cumulative toxicities reach high grades, in order to protect patients, an action that can compromise the anti-cancer efficacy. In our motivating example, if neuropathy becomes severe, a clinician might stop the offending drug early; similarly, cumulative bone marrow suppression manifesting as neutropenia can lead to cycle delays or growth factor support. Understanding which AEs tend to accumulate, and the rate at which they do so, is therefore of high clinical relevance, as it enables proactive management and realistic risk-benefit assessment for long-term treatments.

Despite this importance, current trial reports do not explicitly categorize or highlight cumulative toxicities. This thesis seeks to fill that gap by systematically studying toxicity trajectories and identifying patterns of accumulation. The work is conducted in the context of an ongoing research effort to create a comprehensive taxonomy of longitudinal toxicity behaviors. In this effort, toxicities are broadly divided into three conceptual categories: Immediate , Recurrent, and Cumulative toxicities. The specific scope of this thesis is confined to the cumulative toxicity component. In practical terms, that means we will develop methods to detect and quantify toxicities that follow an upward trajectory over time. This requires careful operational definitions, for example, deciding how to judge if an AE's severity "increases" over cycles (monotonic grade progression might be a strict criterion, whereas allowing "no improvement" could be a looser criterion), and how to handle patients who drop out early or have missing cycles. Part of the challenge is distinguishing a truly cumulative pattern from random noise or from other patterns, for example, an adverse event might appear late not because it's cumulative, but simply by chance or because higher-risk patients survived longer. Thus, a methodological contribution of this work is to establish criteria or algorithms to classify AE profiles as cumulative vs. not, in a data-driven yet clinically informed manner. Our criteria for identifying cumulative effects are grounded in statistical patterns observed in the data, while also aligning with clinically recognized indicators documented in the literature.

The remainder of this thesis is organized as follows. Section 2 formulates the problem statement and research questions, grounded in the background of toxicity reporting challenges, thereby motivating the need for a new taxonomy. Section 3 provides a detailed overview of the data sources and dataset characteristics, covering the trial designs, the types of toxicity data collected, and the steps taken to prepare the data for analysis, including ethical approvals and data management procedures. Section 4 outlines the methodology, detailing the criteria used to define cumulative toxicity, the framework for exploratory analysis, and the statistical modeling techniques applied, along with justifications for their suitability in analyzing longitudinal toxicity data. Section 5 presents the analysis results, spanning from descriptive insights in the exploratory data analysis to the outputs of the formal models, and highlights key findings through selected figures and tables (with supplementary results provided in the Appendix). Section 6 provides a discussion of the findings, interpreting what they mean for the original research questions, comparing them with existing literature, noting limitations, including potential biases or uncertainties in our approach. Finally, Section 7 concludes the thesis by summarizing the main insights and contributions and offering ideas for future research. Potential next steps include applying our approach to additional chemotherapy datasets to further validate and refine the toxicity taxonomy. An equally important avenue would be to extend these models to other oncological treatments such as immunotherapy, targeted therapy, and hormonal therapy, to assess whether similar longitudinal toxicity patterns emerge, and to explore how those patterns impact patient quality of life. This could involve linking toxicity trajectories with broader measures of physical, emotional, and functional well-being, helping to capture the real-world impact of treatment from the patient's perspective. This integration could also support the development of tools that help clinicians balance efficacy and tolerability in treatment decisions.

With these sections, the thesis aims to comprehensively cover the journey from identifying a gap in current practice to providing a solution and highlighting its significance. The next section will delve into the specific research questions that drive this investigation, based on the context and motivations outlined here.

2 Problem Statement and Research Questions

Oncology trials usually report adverse events by highest grade or overall incidence, obscuring both the timing and progression of toxicities. Such static summaries cannot distinguish a patient on cisplatin every 3 weeks who suffers intense nausea for a couple of days then recovers fully from one who endures low-grade nausea nearly every day for two years, nor reveal a steadily worsening pattern that may force dose reductions or delays. To understand true treatment tolerability, we must capture how toxicities evolve, and in particular, quantify the cumulative burden that builds over successive cycles.

2.1 Scope and Definitions

This thesis focuses on cumulative toxicity as part of a broader effort to classify longitudinal toxicity patterns in trials. In addition to cumulative toxicity, we distinguish two further longitudinal profiles, immediate and recurrent toxicities. The definitions adopted in this work are as follows:

- Immediate toxicities: Adverse events that occur shortly after treatment administration (e.g., infusion reactions or acute nausea) and do not necessarily persist into subsequent cycles. They tend to appear transiently in direct temporal association with a dose; some, like infusion reactions, are largely dose-independent, whereas others, such as acute nausea, often increase in severity with higher doses.
- Recurrent toxicities: AEs that can happen repeatedly across multiple cycles, but without a consistent trend of worsening, for instance, a patient might experience nausea in several cycles, but it remains at similar severity each time and does not systematically worsen. These are episodic events that come and go, possibly triggered by each treatment but not cumulatively intensified.
- **Cumulative toxicities**: These are defined by a progressive increase in severity or frequency with successive treatment cycles. In these cases, later cycles typically inflict greater toxicity burden than earlier ones, suggesting an underlying accumulation mechanism, whether pharmacokinetic (e.g., drug build-up), physiological (e.g., organ damage) or adaptive in nature [1].

In summary, immediate AEs are one-time and temporally acute; recurrent AEs appear repeatedly but remain stable; and cumulative AEs intensify over time. This thesis is primarily concerned with the latter, cumulative toxicities, and how they can be systematically identified and quantified in longitudinal trial data. The central question guiding this work is: *How can we distinguish adverse events that progressively intensify or become more frequent across treatment cycles, thereby separating them from transient or stable toxicity profiles?* A better understanding of these dynamics is essential for improving tolerability assessment in clinical trials, with implications for both patient care and trial design [4].

2.2 Research Questions

Building on the background and exploratory findings, this research is guided by the following core questions:

1. How can we formulate clear, operational criteria from cycle-by-cycle toxicity data that capture the distinct temporal patterns observed in the exploratory analysis, such as sharp early-onset toxicities, recurring moderate events, and progressively worsening adverse events?
2. To what extent can each adverse event type be categorized into these longitudinal patterns using model-derived parameters, and how well do these data-driven classifications correspond with clinical intuition and expectations?
3. What quantitative metric best reflects the cumulative build-up of toxicity over successive treatment cycles?

By addressing these questions, this research seeks to deepen the understanding of how toxicities unfold over time in oncology trials. The ultimate goal is to develop a rigorous approach for identifying and quantifying cumulative toxicity, as part of a broader taxonomy of toxicity trajectories. This framework aims to fill a critical gap in current trial reporting and enhance the assessment of treatment tolerability in future studies.

3 Data Description

3.1 Data Sources

To investigate these questions, we use patient-level toxicity data from two completed clinical trials conducted by the European Organisation for Research and Treatment of Cancer (EORTC). The first dataset (EORTC trial 62091) will serve as our primary analysis dataset, and the second (EORTC 30974) will be used for validation and generalization of findings. Both trials collected longitudinal adverse event data over multiple treatment cycles, making them well-suited for studying cumulative toxicity patterns.

- EORTC 62091 (the "TRUSTS" trial): This was a Phase IIb/III trial in advanced soft-tissue sarcoma that compared two experimental chemotherapy regimens against a standard therapy. Patients were randomized to doxorubicin (standard arm) or to trabectedin (an experimental drug) given in either a 3-hour or 24-hour infusion schedule. The target sample size was on the order of 250–370 patients, but the available dataset contains 130 patients (IDs 1–133, with a few IDs unused, indicating 130 actual patients). Each patient could receive multiple cycles of treatment; In practice, patients on the doxorubicin arm had a planned maximum of six cycles (standard for sarcoma treatment and to limit cumulative anthracycline exposure beyond the threshold associated with irreversible cardiac failure), whereas patients on the trabectedin arms continued treatment until progression or unacceptable toxicity. As a result, the number of cycles per patient varies widely: the doxorubicin arm data go up to 6 cycles, while the trabectedin arms have some patients who received far more cycles (in some cases dozens of cycles, with the maximum observed around 60 cycles). At each cycle, detailed toxicity data were recorded. This includes both hematologic toxicities (laboratory-measured events such as neutropenia, anemia, liver enzyme elevations) and non-hematologic (clinical) toxicities (symptoms like nausea, fatigue, neuropathy), all graded by the Common Terminology Criteria for Adverse Events (CTCAE) on a severity scale. In CTCAE grading, 1 generally indicates a mild adverse event and 4 a severe adverse event, with grade 5 reserved for fatal adverse events. Grade 0 can be used informally to indicate the absence of the toxicity. In the 62091 dataset, adverse events are encoded per patient per cycle: essentially one record for each AE that occurred in a given cycle for a given patient. Key identifying fields include the patient ID, the cycle number, and the treatment arm, and each AE record includes the specific toxicity (usually coded by a term or MedDRA code) and its grade for that cycle. The longitudinal structure of this dataset, multiple observations (cycles) for each patient, provides the needed framework to analyze how AEs develop as treatment progresses.

- EORTC 30974: This was a Phase III trial in poor-prognosis germ-cell cancer (a type of testicular cancer), which evaluated high-dose chemotherapy with stem-cell support versus standard chemotherapy (BEP regimen: bleomycin, etoposide, cisplatin). The trial initially planned to accrue $\sim$ 222 patients, but it actually enrolled 137 patients, of whom 131 were evaluable in the final analysis. Treatment in 30974 was more uniform in length: patients in both arms were to receive a total of 4 cycles of therapy (the experimental arm received 1 standard BEP cycle followed by 3 cycles of high-dose chemotherapy, whereas the control arm received 4 cycles of standard BEP). Thus, each patient has at most 4 cycles of toxicity data. This dataset is slightly smaller but provides an excellent validation set, as it has a different cancer type and treatment context, and a fixed number of cycles for all patients. The toxicity data collection is similar: each patient has records of multiple adverse events per cycle, identified by descriptive terms or MedDRA codes, with a CTCAE grade (1–5) for each event. Because cisplatin and other drugs used in this trial are known for certain cumulative toxicities (for example, peripheral neuropathy from cisplatin often accumulates over cycles), we expect to see at least some clear cumulative patterns in this dataset. Other toxicities in 30974 include hematologic AEs like neutropenia (from high-dose chemotherapy) and acute AEs like nausea/vomiting or mucositis. Some of these may behave as cumulative (neuropathy, perhaps bone marrow toxicity if it worsens cycle by cycle), while others may be more acute or recurrent (e.g. nausea might occur in each cycle but not necessarily worse over time). This contrast makes the 30974 data useful for checking how our methods generalize and whether the “taxonomy” of toxicity behavior holds in a different clinical scenario.

In summary, Dataset 62091 (sarcoma trial) is our primary dataset with $\tilde{1}30$ patients and variable cycle counts (up to $\sim$ 60), containing both lab-derived and clinician-reported toxicities, and

Dataset 30974 (germ-cell cancer trial) is a secondary dataset with 131 patients and exactly 4 cycles each, offering a validation case. Both datasets were obtained through EORTC under a data-sharing agreement for research.

3.2 Data Preparation and Cleaning

We undertook substantial data preprocessing to prepare these datasets for analysis. The raw data were provided as de-identified patient records, split across several tables corresponding to case report forms (e.g., separate files for adverse events, laboratory results, treatment information, etc.). The first step was to merge the relevant toxicity information for each patient and cycle into a unified longitudinal dataset. In trial 62091, for example, laboratory measurements (such as blood counts, liver enzymes) were recorded separately from clinician-reported AEs; we merged these by patient ID and cycle number so that each cycle for each patient has a complete toxicity profile combining both sources. Where laboratory values were given (e.g. neutrophil counts), we converted them into CTCAE grades to ensure consistency with the reported AE grades. This harmonization was important because relying solely on reported AEs might underestimate milder toxicities, for instance, a drop in neutrophils might not have been reported as an AE by investigators if it wasn't severe enough, but the lab data would show it. By grading lab values according to CTCAE criteria (for hematologic measures like neutropenia, anemia, etc.), we include those events in the longitudinal data. We also aligned toxicity terminology: some AEs were coded differently between the two trials or between lab and AE forms, so we mapped terms to a consistent naming (for example, ensuring that "leukopenia" lab data is considered alongside any reported "white blood cell decrease" AE). Each toxicity was identified by a term and/or MedDRA code, and we organized them so that the same toxicity could be tracked across patients. Furthermore, we cross-referenced each patient's treatment duration and end-of-treatment dates to distinguish true zero-toxicity assessments from missing data due to early discontinuation (e.g., in the testicular-cancer cohort, absent cycle 4 values reflected patients who stopped after three cycles). After merging, we performed routine data cleaning. This included removing or flagging any placeholder records or ambiguities.

We also checked for any obvious data entry errors and confirmed that all toxicity grades fell in the expected range (1–5).

We treated missing toxicity grades as genuinely missing and evaluated both single proportional-odds (PO) and two-part presence/PO models using WAIC to quantify the impact of data gaps. In our simulation study, where we imposed more than 10% MAR-type missingness, we observed that the single PO model incurred increasing bias as the proportion of missing grades grew, whereas the two-part approach effectively eliminated that bias [9]. In our real-world sarcoma and testicular-cancer cohorts (average per-cycle grade missingness of 5%), the single PO model still achieved superior WAIC, showing that any bias was negligible at that level of missingness. Finally, we reshaped the data into a longitudinal format (one record per patient, cycle, and toxicity) for incidence calculations and into per-patient, per-cycle grade matrices for modeling.

3.3 Data Quality Assurance

Before proceeding to analysis, we undertook steps to ensure the quality and correctness of the data, especially given that these are secondary analyses of trial data. An important validation was to compare our processed data against the published trial results and protocols. For trial 62091, we obtained the Final Analysis Report and the protocol, which contain tables of toxicity incidence (often worst-grade per patient tables, etc.). We reproduced key summary statistics from our dataset to see if they match the reported numbers. For example, we calculated the number of patients who experienced Grade $3-4$ neutropenia, or the proportion of patients with any Grade ≥ 2 liver toxicity, and compared these to the percentages in the trial publication. This cross-checking gave us confidence that our dataset aligns with the original trial data. In cases where we initially found discrepancies, we revisited the data to identify the cause (e.g., perhaps certain lab-based toxicities were not counted in the original report, or vice versa). Ultimately, we were able to reconcile our data with the reported outcomes, or at least understand any differences (such as a patient who had an event after treatment that might have been excluded from the trial's reporting). This quality check was crucial as any serious mismatch could indicate a data issue that would need resolving before trustful analysis.

3.4 Ethical and Regulatory Considerations

Both datasets were used in compliance with ethical guidelines and data protection regulations. We obtained permission from EORTC and the relevant ethics committees to reuse the patient data for this thesis research. A formal data sharing agreement was in place (facilitated by EORTC and our institution, UHasselt) that allowed access to de-identified patient data. All patient information was anonymized – individuals are identified only by study-specific patient IDs, with no names, dates of birth, or other direct personal identifiers. The reuse of the data was covered under the original informed consents of the trials (patients consented to their data being used for research purposes, and specific approval was obtained to analyze it for this project). We also adhered to any conditions set by the data providers, such as not attempting to re-identify patients and ensuring secure storage of the data. In our thesis write-up, we present aggregate results and do not report any potentially identifying information. Thus, all analyses respect patient confidentiality. Additionally, we acknowledge the original trial investigators and sponsor; our reuse of the data is intended to add scientific value (improving toxicity analysis methods) without infringing on the rights or welfare of the patients who participated. These ethical considerations were documented and approved as part of the project proposal.

Having prepared high-quality longitudinal datasets and ensured ethical use, we next describe the methods applied to explore and model cumulative toxicity.

4 Methods and Materials

Our methodological approach is divided into two major parts. Section 4.1 (Exploratory Data Analysis) details how we first visualized and summarized the toxicity data over time to identify patterns and candidate cumulative toxicities. Section 4.4 (Statistical Modeling) then describes the formal models used to quantify cumulative toxicity, including the underlying statistical frameworks, model equations, and criteria for determining when a toxicity can be called "cumulative." Throughout, we integrate insights from the exploratory analysis to inform model building, and we describe all assumptions and theoretical foundations of our modeling strategy.

4.1 Exploratory Data Analysis (EDA)

Before fitting complex models, we conducted an extensive EDA to understand how toxicities evolve over successive cycles in the trials. The goal was to reveal any longitudinal trends, especially increasing trends indicative of cumulative effects, and to compare these trends across different adverse events and between treatment arms.

4.2 Individual Patient Toxicity Trajectories

To explore how toxicity grades evolve at the individual patient level, we used two complementary visualization approaches: heatmaps and spaghetti plots.

- Heatmaps. These visualize the trajectory of each patient's AE grade across cycles for a specific adverse event. The x-axis represents treatment cycles, while the y-axis lists individual patients. Cell colors encode toxicity severity, ranging from white (Grade 0, absence of AE) through a gradient of cooler to warmer colors representing increased severity (e.g., green for Grade 1, yellow/orange for Grades 2–3, and red for Grade 4). Heatmaps compactly summarize individual trajectories, allowing easy visual identification of patterns such as immediate (single-cycle spike), recurrent (on-off) and cumulative (steadily worsening), or recovery over cycles.
- Spaghetti plots. These by contrast, explicitly illustrate how individual toxicity grades fluctuate over time. Each thin line traces a single patient's grade trajectory over time, colored by treatment arm, while a thicker line shows the median trajectory for that arm. The x-axis is cycle number and the y-axis is raw grade.

  Thus, while both heatmaps and spaghetti plots represent individual patient trajectories, the former excels at summarizing patterns across many patients simultaneously, whereas the latter emphasizes the temporal progression and variability of individual patient experiences.

4.2.1 Prevalence and weighted Prevalence

To account for patient dropout (patients going off study at different times), we employed the prevalence function approach, including a Weighted Prevalence Function (WPF) as described by Cabarrou et al [7]. This method calculates, at each time point (cycle), the proportion of patients, still on treatment, who are in a given toxicity state, and can weight this by certain factors. In simpler terms, the prevalence at cycle j is the probability that a patient (among those who have not yet dropped out by cycle j) is experiencing the toxicity at that cycle.

We plotted the weighted prevalence (wP) of key toxicities over time. This produces a curve $wQ(t)$ which rises if more patients tend to be in the toxic state as time goes on, adjusting for the fact that fewer patients remain on study at later cycles. Moreover, the weighted aspect corrects for censoring, that is, if patients drop out (often due to toxicity or progression), a naive prevalence might underestimate late toxicity (because only healthier patients remain), whereas the weighted prevalence gives a more accurate picture by considering the at-risk population.

4.3 Per-Patient AE Burden Visualizations

So far, we've focused on each AE separately (comparing patients or summarizing across patients). The final step is to examine the data by patient, to see each patient's overall toxicity burden across all AEs. This patient-centric view is important for understanding the cumulative toll on each individual, and it's useful for identifying "high toxicity patients" versus "low toxicity patients," regardless of AE type. We will propose a couple of visualizations for per-patient toxicity profiles.

4.3.1 Patient vs. AE Heatmap (Matrix of Cumulative Grades)

- This is a heatmap where each row is a patient and each column is an AE, and the cell value is some summary of that patient's experience with that AE over the entire treatment (for example, the total cumulative grade or the maximum grade for that AE in that patient). Such a matrix provides a compact view of which patients suffered which AEs severely or repeatedly. It's like a "toxicity fingerprint" for each patient.

- To construct this, we first need to summarize the data per patient per AE. We shall use total grade points (the sum of all grades across cycles for that AE and patient) as a measure of cumulative burden for that AE in that patient, as the sum of grades is straightforward and interpretable (it counts a Grade2 in two cycles as 4 points, equal to one cycle of Grade4, as a rough measure of total suffering).

- Each row is one patient; across that row, you can see which AEs they had and how bad. For example, if a patient has a deep red cell under "Neutrophil count decreased", it means they had many cycles of neutropenia or very severe neutropenia repeatedly. If another patient's entire row is mostly white, that patient had very little toxicity overall (no significant grades for these AEs). Patterns may emerge, such as some patients having high burden in multiple hematologic toxicities (e.g., a row that's red under neutrophil, WBC, platelet – perhaps they are generally more sensitive to bone marrow toxicity). Another patient might have a specific problem with nausea but not others, etc. Sorting by total burden puts the worst overall patients at the top, which can highlight if, say, a handful of patients suffered a disproportionate amount of the toxicity (their rows will be the most colored across many columns).

4.3.2 Patients by total toxicity burden

- Stacked barplots per Patient is another intuitive way to depict per-patient data is a stacked bar chart where each patient is a bar, and the segments of the bar represent different AEs contributing to their total toxicity. The height of the bar could be the total grade points (summing all AEs, which is the same metric we just discussed per patient). Each colored segment of the bar corresponds to one AE's share of that total.

- It can be useful to see not only how much total toxicity each patient had (bar height) but also what composition of AEs made it up (color segments).
- We'd typically sort patients by total burden (so the bars descend in height). With 130 patients, we'd likely have to either scroll or select a subset (or compress labels). Perhaps we show the top N patients in a readable way, or cluster them and figure **??** is the stacked bar for the top 20 patients by burden

4.4 Statistical Modeling

To formally characterize cumulative toxicity and address our research objectives, we developed two competing Bayesian hierarchical models: Model 1 (a single proportional-odds ordinal regression) and Model 2 (a two-part mixture model). We compare these models via the widely applicable information criterion (WAIC) to assess which provides the most parsimonious fit to the data.

- In **Model 1**, we model the longitudinal course of AE grades (assuming occurrence) using a proportional-odds ordinal regression. An upward trend in the probability of higher grades over cycles characterizes cumulative toxicity in this model.
- In **Model 2**, we model both the probability of AE occurrence at each cycle and, conditional on occurrence, the ordinal grade via a proportional-odds mixed-effects regression, resulting in a two-part mixture tailored to longitudinal toxicity outcomes.

In the following, we present the modeling methodology in detail. We introduce the notation and model formulation, specify the statistical assumptions, derive the likelihood, and describe how we evaluate and compare the models.

Model 1: Ordinal Model for AE Grades

Notation and Model formulation

Assuming occurrence of AE among cycles, we modelled the AE grade, using a cumulative (proportional-odds) logit model. Let $Y_{ij} \in 1, \ldots, K$ be the CTCAE grade (0–K, e.g. 0–4) for patient i at cycle j. A standard proportional-odds model posits cutpoints $(\alpha_1, \ldots, \alpha_{K-1})$ and a linear predictor η_{ij}. Equivalently, one can introduce a latent continuous variable Y_{ij}^* so that

$$Y_{ij} = k \iff \alpha_{k-1} < Y_{ij}^* \leq \alpha_k, \qquad k = 0, \ldots, K,$$

with cutpoints $\alpha_0 = -\infty < \alpha_1 < \cdots < \alpha_{K-1} < \alpha_K = +\infty$. The latent variable is modeled by $Y_{ij}^* = \eta_{ij} + \epsilon_{ij}$, where ϵ_{ij} has a logistic distribution for a logit link. In the cumulative logit form, this yields

$$\operatorname{logit}\big[P(Y_{ij} \leq k \mid b_i)\big] = \alpha_k - \eta_{ij}, \qquad \eta_{ij} = \beta_0 + \beta_{\mathrm{trt}_i} + m_{\mathrm{trt}_i}(\mathrm{cycle}_{ij}) + b_i, \tag{1}$$

where,

- β_0 is an overall intercept,
- β_{trt_i} is the fixed treatment effect, $\beta_{\text{trt}_i} \;=\; \sum_{g=1}^{3} \beta_g \, \mathbf{1}(\text{trt}_i = g), \quad \text{with } \beta_1 = 0$
- $m_{\text{trt}_i}(\cdot)$ denotes a treatment-specific spline function capturing non-linear cycle effects and treatment-by-cycle interaction, as motivated by exploratory evidence of cycle durations imbalance across treatment arms,
- $b_i \sim \mathcal{N}(0, \sigma_b^2)$ is a patient-specific random intercept, and
- $\{\alpha_k\}$ are ordered cutpoints satisfying $\alpha_1 < \cdots < \alpha_{K-1}$.

We omitted explicit random slopes because penalized splines within a mixed-model framework inherently absorb subject-specific slope variability, making separate random slopes unnecessary. By fitting a treatment-specific spline $m_{\text{trt}_i}(cycle)$, we flexibly model nonlinear cycle–dependent trends, capturing varying slopes, inflection points, and plateaus, at the population level [18; 17]. In its mixed-model representation, the spline's basis coefficients act as random effects whose penalties soak up inter-individual differences in slope, thereby replicating the effect of per-patient random slopes without estimating one slope parameter per subject [15]. This retains only a random intercept $b_i \sim N(0, \sigma_b^2)$ for baseline heterogeneity, while the curve's curvature delivers individualized smooth trajectories.

Equivalently,

$$\Pr(Y_{ij} \le k \mid b_i) = \frac{\exp(\alpha_k - \eta_{ij})}{1 + \exp(\alpha_k - \eta_{ij})}. \tag{2}$$

This is the proportional odds model we fitted in stage I. Intuitively, this models the log-odds of having grade $\le k$ versus $> k$ in terms of covariates. Ordinal regression thus extends binary logistic regression by $K - 1$ intercepts (cutpoints), and yields interpretable odds-ratios on these cumulative events.

From this model we can derive the category probabilities. Writing $p_{ij}(k) = \Pr(Y_{ij} = k)$, we have

$$\begin{aligned}
\Pr(Y_{ij} = 1) &= \Pr(Y_{ij} \le 1) = F(\alpha_1 - \eta_{ij}), \\
\Pr(Y_{ij} = k) &= \Pr(Y_{ij} \le k) - \Pr(Y_{ij} \le k-1) = F(\alpha_k - \eta_{ij}) - F(\alpha_{k-1} - \eta_{ij}), \quad 2 \le k \le K-1, \\
\Pr(Y_{ij} = K) &= 1 - \Pr(Y_{ij} \le K-1) = 1 - F(\alpha_{K-1} - \eta_{ij}),
\end{aligned}$$

where $F(x) = \text{expit}(x)$ is the logistic CDF.

Model Estimation

- **Estimation via (Laplace) Maximum Likelihood**

  We fit the cumulative-logit mixed model by maximizing the marginal likelihood over fixed parameters $\theta = \beta, \alpha, \sigma$, integrating out the patient-level random intercepts b_i. Let

  $$f(Y_{ij} \mid b_{0i}; \theta) \;=\; \Pr(Y_{ij} \le k \mid b_{0i}, \theta)^{\mathbf{1}\{Y_{ij} \le k\}} \bigl[1 - \Pr(Y_{ij} \le k \mid b_{0i}, \theta)\bigr]^{\mathbf{1}\{Y_{ij} > k\}},$$

and assume conditional independence over cycles. Then the marginal likelihood is

$$L(\theta) = \prod_{i=1}^{N} \int \Big[\prod_{j} f(Y_{ij} \mid b_{0i}; \theta) \Big] \, \phi(b_{0i}; 0, \sigma_b^2) \, \mathrm{d}b_{0i}, \tag{3}$$

a high-dimensional integral that is analytically intractable. Direct maximization of $L(\theta)$ is therefore impossible, so we apply numerical integration via the Laplace approximation (one-point Gaussian quadrature) [21; 20]. Concretely, each integral $\int g(b)\,\mathrm{d}b$ is approximated by $g(\hat{b})\,(2\pi)^{\frac{q}{2}}\left|-H(\hat{b})\right|^{-\frac{1}{2}}$, where $\hat{b}$ is the mode of the integrand and H its Hessian. Substituting into $\ell(\theta) = \log L(\theta)$ yields the Laplace-approximated log-likelihood, which we maximize using Newton–Raphson (or similar) with analytic derivatives via automatic differentiation [19]. In *glmmTMB* this corresponds to $nAGQ = 1$; for improved accuracy one may use adaptive Gauss–Hermite quadrature ($AGQ, nAGQ > 1$). Once converged, standard errors derive from the approximated Fisher information matrix, $\mathcal{I}(\theta) = -\partial^2 \ell(\theta)/\partial\theta^2$.

- **Estimation via Bayesian MCMC**
  Since we require cumulative–logit modeling together with treatment-specific penalized splines and a patient-level random intercept, no purely frequentist package supports all three simultaneously:

  – *ordinal::clmm* (and *clmm2*) handles only random intercepts for cumulative models [20].

  – *glmmTMB* can in principle fit ‘family = cumulative(link=logit)‘ but lacks built-in smooth terms for splines in this family under our locked-down setup [19].

  – *mgcv*'s GAMM supports splines but not ordinal links.

  By contrast, *brms* (via Stan) natively implements the ‘cumulative(link="logit")‘ family with arbitrary random-effects structures and spline terms, delivering full Bayesian inference, coherent uncertainty quantification, and robust convergence diagnostics [13].

  We therefore reframe estimation in a Bayesian framework. Denote all parameters by $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, m(\cdot), \boldsymbol{\alpha}, \sigma_b\}$. The likelihood remains as above, and we specify weakly informative priors as: $\beta_0, \beta. \sim N(0, 5^2)$, $\sigma_b \sim \text{half}-\text{Student--}t(3, 0, 2.5)$, $\alpha_1 < \cdots < \alpha_{K-1} \sim$ ordered uniform. The joint posterior is

$$p(\boldsymbol{\theta}, b_0 \mid Y) \propto \Big[\prod_{i,j} f(Y_{ij} \mid b_{0i}, \boldsymbol{\theta})\Big] \times p(b_0 \mid \sigma_b) \times p(\boldsymbol{\theta}). \tag{4}$$

  Stan's No-U-Turn Sampler (NUTS) then draws MCMC samples $\{\boldsymbol{\theta}^{(s)}, b_0^{(s)}\}$ by evaluating the posterior density and its gradients [22]. Convergence is monitored via $\widehat{R} < 1.01$ and effective sample size; posterior summaries (means, medians, 95% credible intervals) follow directly.

Model Checking and Predictive Assessment

Once the Bayesian cumulative-logit model has converged, we must verify that it (a) fits the observed AE-grade data adequately and (b) will generalize to new cycles and patients. Convergence diagnostics ($\widehat{R} < 1.01$, effective sample size) ensure the sampler has explored the posterior, but they do not guarantee model fit or predictive accuracy. We therefore apply four complementary tools:

- **Posterior Predictive Checks**. We generate replicated datasets Y^{rep} from the posterior predictive distribution $p(Y^{\text{rep}} \mid Y) = \int p(Y^{\text{rep}} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid Y)\, \mathrm{d}\boldsymbol{\theta}$, using, for example, pp_-check(brms_fit) in *brms* [12]. By comparing summaries of the observed and replicated data, such as the distribution of AE grades by cycle, zero-inflation rates, and cumulative grade trajectories, we can detect systematic misfit (e.g., under- or over-dispersion, failure to capture inflection points).
- **LOO and WAIC**. To estimate out-of-sample predictive performance, we compute:
  - WAIC $= -2\sum_{i,j}\Big[\log\big(\frac{1}{S}\sum_s p(Y_{ij} \mid \boldsymbol{\theta}^{(s)})\big)\Big] + 2\sum_{i,j} \mathrm{Var}_s\big(\log p(Y_{ij} \mid \boldsymbol{\theta}^{(s)})\big)$,
  - PSIS-LOO via Pareto-smoothed importance sampling (loo(brms_fit)) [11].

  Leave-One-Out Cross-Validation (LOO) is generally more robust for complex hierarchical models; WAIC is faster but may be unstable with heavy-tailed posteriors. We report the LOO expected log predictive density (LOO-ELPD) and its standard error, and compare alternative specifications only if differences exceed twice the standard error.
- **Prior Sensitivity Analysis**. To ensure our Bayesian inferences are driven by the data rather than by arbitrary prior choices, we conduct a sensitivity check by refitting the final model under alternative, yet reasonable, priors, e.g. widening fixed-effect priors from $N(0, 5^2)$ to $N(0, 10^2)$ (and narrowing to $N(0, 2.5^2)$), and varying the random-intercept scale prior between half-Student-$t(3, 0, 2.5)$, half-Student-$t(3, 0, 5)$, and half-Cauchy(0,2.5). For each variant, we compare key posterior summaries (means, 95% credible intervals) and the LOO-ELPD to the original fit. If estimates and predictive accuracy remain within approximately two standard errors of their initial values, we conclude that our results are prior-robust; any substantive shifts would lead us to refine the priors or reconsider model components. This procedure provides confidence that our conclusions about cumulative toxicity dynamics reflect the observed data.
- **Proportional-Odds Assumption**. Our baseline cumulative-logit model assumes that each covariate's effect on the log-odds scale is identical across all cut-points (the parallel-lines assumption) [10]. Concretely, if

  $$\pi_{ik} = \Pr(\text{grade}_i > k \mid X_i),$$

  then

  $$\log \frac{\pi_{ik}}{1 - \pi_{ik}} = \alpha_k \; + \; X_i^{\top} \beta,$$

with the vector β invariant in k. We assess this assumption via two complementary Bayesian diagnostics:

- PSIS-LOO model comparison. We fit a "non-proportional" variant by allowing threshold-specific treatment×cycle slopes through `cs(...)` in `brms`. If the difference in expected log predictive density (ΔELPD) between the proportional and non-proportional models is less than 2× its standard error, then relaxing proportionality yields no meaningful improvement, supporting the proportional-odds assumption [11].
- Empirical-logit posterior-predictive checks. We group the data into strata (treatment arm × cycle-bins), compute the observed cumulative proportions $\hat{p}_k = \Pr(\text{grade} > k)$ in each stratum, transform to empirical log-odds $\log\{\hat{p}_k/(1-\hat{p}_k)\}$, and overlay these curves on 95 % posterior predictive intervals from replicated datasets. Systematic departures of the empirical-logit curves from their predictive bands would signal violation of the parallel-slopes constraint [12].

If either diagnostic indicates non-parallelism, we will refit a partial proportional-odds model in `brms`, selectively relaxing the parallel-slopes constraint only for those covariates requiring threshold-specific effects via `cs(...)` [14; 23].

4.4.1 Model 2: Two-part mixture model

Model 2 captures both the occurrence and severity of adverse events (AEs), addressing the limitation of the simpler model I ordinal model that would predict AE severity even when no AE occurred. The two-part mixture model resolves this by explicitly distinguishing the probability of AE occurrence from the severity of events when they occur.

Notation and Joint Probability

- **AE indicator:**

$$A_{ij} = \begin{cases} 1, & \text{if an AE occurs for patient } i \text{ in cycle } j, \\ 0, & \text{otherwise.} \end{cases}$$

- **Grade outcome:**

$$G_{ij} = \begin{cases} g \in \{0,1,2,3,4\}, & \text{if } A_{ij} = 1, \\ 0, & \text{if } A_{ij} = 0. \end{cases}$$

- **Two-part mixture:** Let $\pi_{ij} = P(A_{ij} = 1)$ and $f_{ij}^{+}(g) = P(G_{ij} = g \mid A_{ij} = 1)$. Then

$$P(A_{ij} = a,\, G_{ij} = g) = \begin{cases} 1 - \pi_{ij}, & (a = 0,\, g = 0), \\ \pi_{ij}\, f_{ij}^{+}(g), & (a = 1,\, g \in \{1, \ldots, 4\}), \\ 0, & \text{otherwise.} \end{cases}$$

Linking to Covariates (Model 2)

- **AE occurrence (binary):** We model the log-odds of experiencing at least one AE in cycle j via a logistic mixed-effects regression:

$$\text{logit}(\pi_{ij}) = \underbrace{\beta_0^{(A)} + \beta_{\text{trt}_\text{i}}^{(A)} + m_{\text{trt}_i}^{(A)}(\text{cycle}_{ij})}_{\text{fixed (population) effects}} + \underbrace{b_i^{(A)}}_{\substack{\text{random}\\ \text{intercept}}} .$$

  with $b_i^{(A)} \sim N(0, \tau^2)$, similar treatment-specific spline $m_{\text{trt}_i}^{(A)}(\cdot)$ and the rest of the model terms as defined in stage I.

- **Severity conditional on an AE (ordinal):** For cutpoints $\alpha_1 < \alpha_2 < \alpha_3$, we use a proportional-odds mixed-effects model:

$$\text{logit}[\Pr(G_{ij} \leq k \mid A_{ij} = 1)] = \alpha_k - \left[\beta_0^{(G)} + \beta_{\text{trt}_\text{i}}^{(G)} + m_{\text{trt}_i}^{(G)}(\text{cycle}_{ij}) + b_i^{(G)}\right].$$

  By construction, the same linear predictor applies across all k, ensuring the parallel-lines (proportional-odds) assumption.

Model Assumptions, Estimation, and Validation

- The model is built on the assumption of conditional independence between the occurrence and severity of AEs, given patient-specific random effects $(\mathbf{r}_i)$:

$$P(A_{ij} = a,\ G_{ij} = g \mid \mathbf{r}_i) = P(A_{ij} = a \mid \mathbf{r}_i) \times P(G_{ij} = g \mid A_{ij} = 1, \mathbf{r}_i)^{\mathbf{1}\{a=1\}}.$$

  with the logical restriction that grade is zero ($G_{ij} = 0$) if no AE occurs ($A_{ij} = 0$).

- Estimation proceeds as for Model 1 described in section 4.4. we construct the joint likelihood across all cycles and patients, integrate out the random effects via adaptive quadrature or MCMC, and assess fit with posterior predictive checks and information-criterion comparisons.

Combining Posterior Predictions

Once the two submodels have been fitted, we recover the full posterior distribution over (A_{ij}, G_{ij}) by multiplying the AE-occurrence probability by the conditional severity probabilities:

- $\hat{\pi}_{ij} = P(A_{ij} = 1 \mid \text{data})$.
- For $k \geq 1$, $\hat{p}_{ij}(1) = \hat{F}_{ij}(1)$, $\hat{p}_{ij}(k) = \hat{F}_{ij}(k) - \hat{F}_{ij}(k-1)$ $(k = 2, 3)$, $\hat{p}_{ij}(4) = 1 - \hat{F}_{ij}(3)$.
- Joint probabilities: $P(A_{ij} = 1, G_{ij} = k) = \hat{\pi}_{ij}\, \hat{p}_{ij}(k)$, $P(A_{ij} = 0, G_{ij} = 0) = 1 - \hat{\pi}_{ij}$.

These combined probabilities enable model-based visualization and classification of AE grade trajectories.

4.4.2 Simulation Study

To illustrate the advantage of Model 2 in mitigating inflated predicted probabilities due to intermittent missingness, we conducted a simulation study as follows:

- The first steps of the simulation focus on faithfully replicating the structure of our real data and making explicit any absence of toxicity. We generated synthetic AE trajectories starting with the exact template of our original dataset, maintaining one row for each (PATID, AE term, cycle) so that the number of patients, cycles per patient, and list of MedDRA/CTCAE terms remain identical.

- We then simulate a logical, monotonic progression through grades $0 \to 1 \to 2 \to 3 \to 4 \to 5$. Initial grades were drawn to reflect clinician-informed prevalences (e.g. 60% start at grade 0, 30% at grade 1, 10% at grade 2), and transitions were governed by monotonic progression probabilities (e.g. 60% stay, 30% + 1 grade, 10% +2 grades) with absorption at grade 5.

- Each patient–AE starts at an initial grade of 0, 1, or 2 (for example, 60% at 0, 30% at 1, 10% at 2). At each subsequent cycle there is a 60% chance to stay at the same grade, a 30% chance to increase by exactly one, and a 10% chance to jump two grades, never exceeding grade 5. Once a patient–AE reaches grade 5, they remain there for all following cycles, preserving monotonicity. We also carry forward any original 'datass' fields and simulate the binary "related" and "serious" flags using the empirical overall probabilities from your dataset. The final simulated data contains 'grade' values from 0 through 5, strictly non-decreasing across cycles for every patient–AE pair.

- We superimposed patterns of intermittent missingness matching our observed dropout and hold mechanisms, without revealing dropout to the graders.

- Our simulation aims at achieving three objectives, that is, to construct a "best-case" scenario in which toxicity grades either stay the same or worsen, but never improve, across cycles, allowing us to validate algorithms without the confounding effects of real-world interventions such as dose reductions or treatment holds, to demonstrate to clinicians the pure progression of CTCAE grades from 0 (no event) through to 5 (death) as cycles advance, gaining a clear understanding of how the model attributes cumulative risk. Finally, by including the full 0–5 scale, especially grade 5, we stress-test our methods to ensure they handle extreme cases (fatal outcomes) just as gracefully as mild or moderate events.

5 Result

5.1 Exploratory Data Analysis

5.1.1 Descriptive Statistics

Both trial data (62091 and 30974) follow a similar longitudinal format, with one row per reported adverse event (AE) per patient per cycle. Key fields include patient identifier, cycle number, AE term, CTCAE grade, event date, relatedness, and seriousness. Table 1 presents descriptive summaries for the 62091 AE-dataset. This concise overview highlights the predominance of hematologic and constitutional toxicities, in line with expectations for this chemotherapy trial.

(a) Ten most frequent adverse events (all grades) in the 62091 dataset (N=3 244).

Adverse Event	Count	% of events
Anemia	298	9.2%
Fatigue	290	8.9%
Neutrophil count decreased	270	8.3%
Nausea	219	6.8%
White blood cell decreased	211	6.5%
Constipation	121	3.7%
Platelet count decreased	115	3.5%
Lymphocyte count decreased	114	3.5%
Alopecia	92	2.8%
Vomiting	90	2.8%

(b) 62091 AE dataset at a glance

Total rows	3 488 (incl. 244 placeholders)
Reported AEs	3 244
Patients	130 (IDs 1–133; 3 missing)
Cycles per patient	1–65 (median planned = 7)

(c) Patients, AE counts & median cycles by arms 1 (Dox), 2 (Trab 3h), & 3 (Trab 24h)

Arms	1	2	3
Patients (n)	47	42	41
Total AEs	1327	1323	838
Median planned cycles	7	7	8

Table 1: Descriptive summaries for the 62091 AE-dataset.

Comparison with the Trial Final Analysis Report

- In our EORTC 62091 AE dataset (Table 2), high-grade (3–4) toxicities were overwhelmingly hematologic: neutrophil count decreases occurred in 44.7%, 45.2% and 58.5% of patients in the Trab_3 h, Trab_24 h and Doxo arms, respectively; platelet count decreases in 17.0%, 14.3% and 2.4%; and lymphocyte count decreases in 12.8%, 9.5% and 17.1%. Non-hematologic grade 3–4 events (febrile neutropenia, infection, etc.) never exceeded 13%, yielding mean high-grade events per patient of 1.6 (range 0–5), 1.4 (0–4) and 1.3 (0–3) across the three arms.
- The full grade 1–4 breakdown for blood & lymphatic AEs (Table 3) shows anemia grade 1 in 40.4%, 28.6% and 31.7%; grade 2 in 17.0%, 33.3% and 22.0%; with grades 3–4 $\leq$ 9.8%. Febrile neutropenia grade 3–4 occurred in 13.0%, 12.0% and 7.5%. Infection grades (Table 4) were similarly concentrated at grade 0 (67.4%, 85.4%, 77.5%) with grades 1–2 in 2.2–17.4% and grades 3–4 $\leq$ 10.9%.
- Across all three summaries, grade 3–4 overall, blood & lymphatic disorders, and infections, our empirical estimates match the Final Analysis Report tables (Tables 7,

8, and ??) to within 0.3 percentage points. This close concordance confirms that the received dataset faithfully reproduces the published safety profile for the trial, even though over 100 distinct AE terms were collected (only the top PTs are shown here).

Table 2: Grade 3–4 event incidence by PT and arm in Data EORTC-62091 (N=130).]

PT	Trab_3hrs (N=46)	Trab_24hrs (N=41)	Doxo (N=40)
Anemia	4 (8.5%)	4 (9.5%)	4 (9.8%)
Febrile neutropenia	6 (12.8%)	5 (11.9%)	3 (7.3%)
Other cardiac disorders	3 (6.4%)	NA	NA
Nausea	3 (6.4%)	5 (11.9%)	2 (4.9%)
Vomiting	6 (12.8%)	4 (9.5%)	3 (7.3%)
Fatigue	1 (2.1%)	5 (11.9%)	2 (4.9%)
Infection	6 (12.8%)	4 (9.5%)	1 (2.4%)
Lymphocyte count decreased	6 (12.8%)	4 (9.5%)	7 (17.1%)
Neutrophil count decreased	21 (44.7%)	19 (45.2%)	24 (58.5%)
Platelet count decreased	8 (17.0%)	6 (14.3%)	1 (2.4%)
White blood cell decreased	12 (25.5%)	9 (21.4%)	17 (41.5%)
Dehydration	2 (4.3%)	NA	4 (9.8%)
Hyponatremia	2 (4.3%)	NA	2 (4.9%)
Other renal and urinary disorders	2 (4.3%)	NA	NA
Dyspnea	4 (8.5%)	1 (2.4%)	1 (2.4%)
Other respiratory, thoracic and mediastinal disorders	3 (6.4%)	NA	1 (2.4%)
Hepatobiliary disorders	2 (4.3%)	3 (7.1%)	NA

Table 3: Grade 1–4 distribution for blood & lymphatic system disorders by arm in data EORTC-62091.

PT	Grade	Trab_3hrs (N=46)	Trab_24hrs (N=41)	Doxo (N=40)
Anemia	1	19 (40.4%)	12 (28.6%)	13 (31.7%)
Anemia	2	8 (17.0%)	14 (33.3%)	9 (22.0%)
Anemia	3	3 (6.4%)	3 (7.1%)	4 (9.8%)
Anemia	4	1 (2.1%)	1 (2.4%)	NA
Febrile neutropenia	3	2 (4.3%)	4 (9.5%)	3 (7.3%)
Febrile neutropenia	4	4 (8.5%)	1 (2.4%)	NA

Table 4: Infection & infestation grades by arm in Data EORTC-62091.

Grade	Trab_3hrs—all AE	Trab_24hrs—all AE	Doxo—all AE	Trab_3hrs—related AE	Trab_24hrs—related AE	Doxo—related AE
	(N=46)	(N=41)	(N=40)	(N=46)	(N=41)	(N=40)
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
0	31 (67.4%)	35 (85.4%)	31 (77.5%)	44 (95.7%)	39 (95.1%)	37 (92.5%)
1	1 (2.2%)	1 (2.4%)	2 (5.0%)	0 (0.0%)	0 (0.0%)	1 (2.5%)
2	8 (17.4%)	1 (2.4%)	6 (15.0%)	1 (2.2%)	1 (2.4%)	1 (2.5%)
3	5 (10.9%)	4 (9.8%)	1 (2.5%)	1 (2.2%)	1 (2.4%)	1 (2.5%)
4	1 (2.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

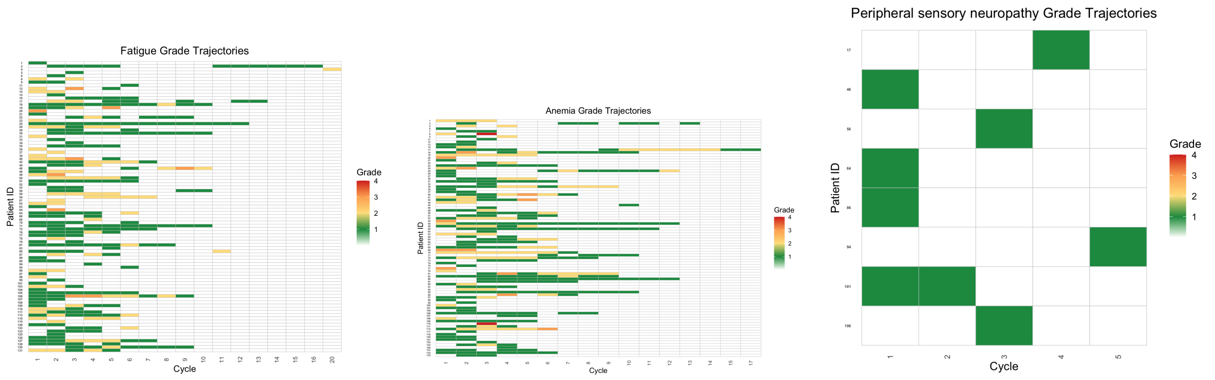
5.1.2 EDA Visualizations

In this section we present key findings from the EDA, focusing on representative AEs that illustrate the spectrum of longitudinal toxicity patterns. For clarity, we present results for selected AEs. These demonstrate patterns of grade accumulation, differences between treatment arms, and non-monotonic trends. Observations refer to the primary trial dataset (EORTC 62091) unless otherwise noted. Additional EDA results for every AE, as well as analogous analyses on the validation cohort and simulation, are summarized afterward and provided in the Appendix.

5.1.3 Individual Profiles

Figures 1 and 2 show grade trajectories in the primary trial, visualized as a patient-cycle heatmap. Each row represents one patient and each column one cycle. Cells are colored if the patient experienced an AE in that cycle, with the color indicating the grade (from white = Grade 0, through yellow = Grade 2 to red = Grade 4). The heatmaps illustrates that most patients develop fatigue, anemia, neutrophil count decreased (NCD) AS their rows are mostly colored whereas fewer patients develop peripheral sensory neuropathy (PSN), platelet count decreased (PCD) and febrile neutropenia (FN) since their rows are mostly white. Notably, those who suffered PCD and FN almost always did so in later cycles and with higher grades.

Figures **??**traj34na_traj34neutrophil_traj_sim respectively present enlarged trajectory heatmaps for nausea and neutrophil count decrease (NCD) in the EORTC 30974 and simulated dataset. The nausea heatmap spans up to five cycles, reflecting that most patients in EORTC 30974 completed four cycles, with a small subset remaining on treatment through cycle 5, and uses black dots to denote the cycle at which treatment was discontinued (typically due to toxicity or scheduled completion). In contrast, the simulated-data NCD heatmap exhibits a strictly monotonic increase in severity, in accordance with the data-generation procedure described in Section 4.1.

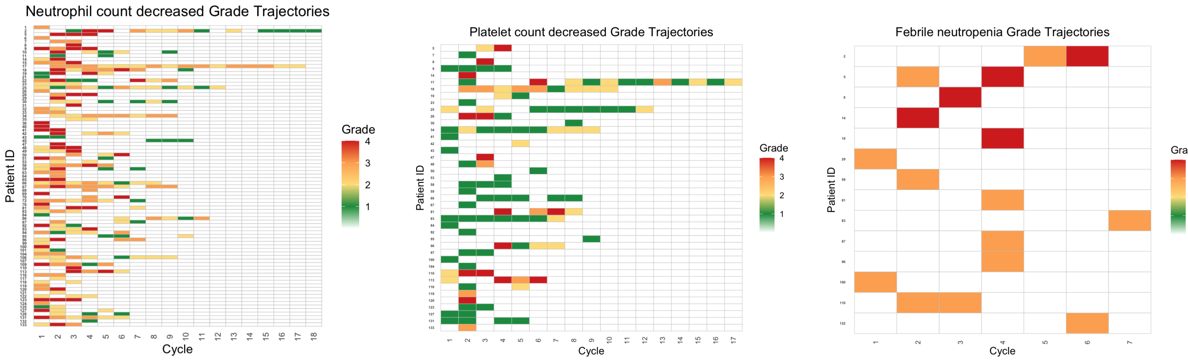


(a) Fatigue (b) Anemia (c) Peripheral sensory neuropathy

Figure 1: **Raw grade trajectory heatmaps:** Here we stable patterns with consistently low grades

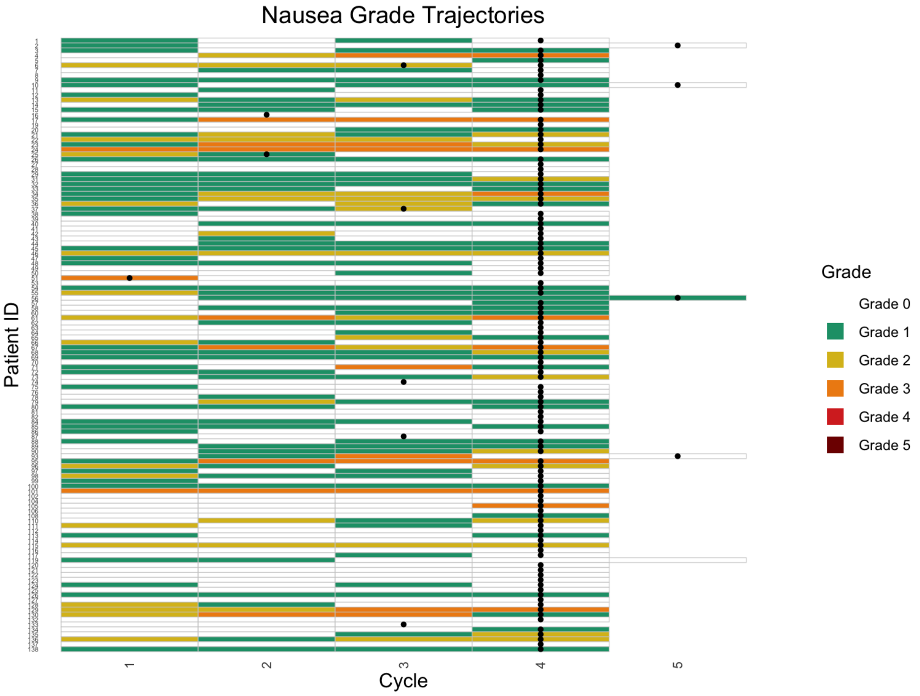
In our taxonomy, NCD, PCD and FN behave as cumulative toxicities, as patients' rows start green in early cycles and gradually shift to orange or red in later cycles, it indicates their AE severity increased with cumulative treatment, and thus, possible cumulative toxicity patterns.

Notably, the EDA revealed that strictly monotonic increasing grades within a single patient are relatively rare as clinicians often intervene when toxicity worsens, so patients might be removed from treatment or have doses reduced, preventing an unchecked rise to the highest grades. This insight informed our modeling (we realized we should not force a



(a) Neutrophil count decreased (b) Platelet count decreased (c) Febrile neutropenia

Figure 2: Raw grade trajectory heatmaps for probable cumulative toxcities.



● Black dot = cycle in which treatment ended

Figure 3: Raw grade trajectory heatmap for nausea in EORTC 30974 dataset.

deterministic upward trend for each patient, but rather model the probability of toxicity increasing over time).

5.1.4 Informing Model Selection

The insights from EDA were critical in shaping our modeling approach. Key findings included:

- Patterns of increase: We identified which toxicities showed a visual increase in risk over time (these became prime candidates for “cumulative” modeling).
- Non-monotonic individual paths: We saw that individual patient paths are not strictly increasing due to interventions, which suggested that our model should not assume deterministic growth but rather allow probabilities to change over time. This directly led to the idea of modeling the probability of toxicity at each cycle (and of higher grades) rather than enforcing that once a patient has a certain grade it keeps increasing.

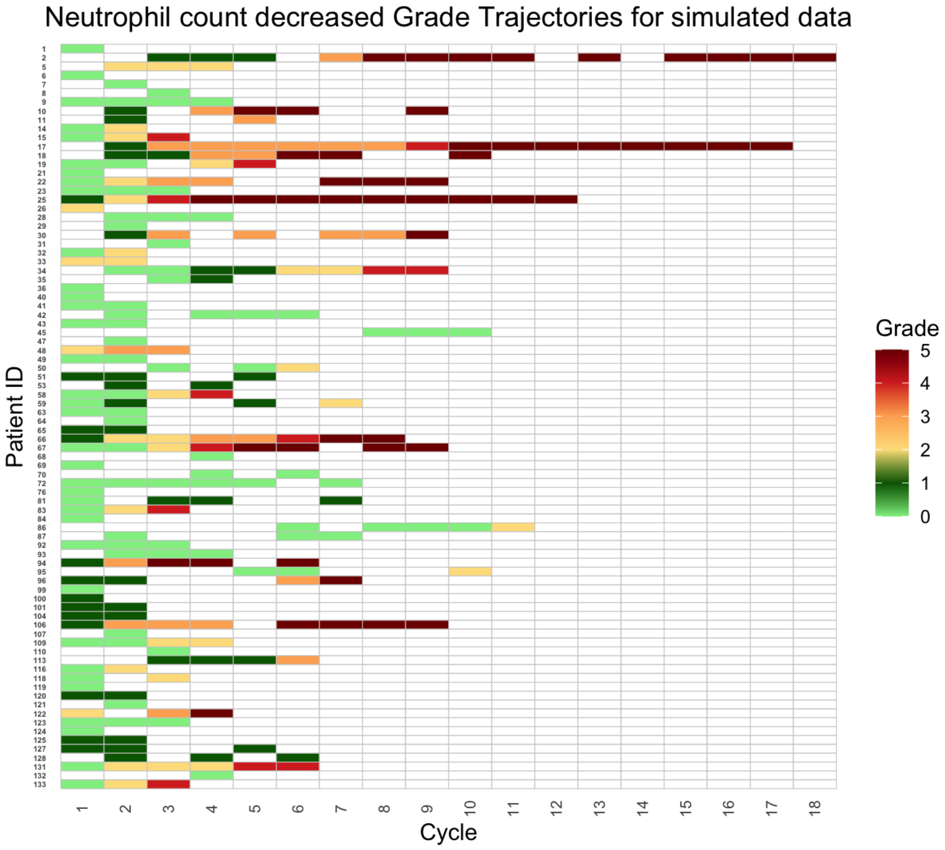
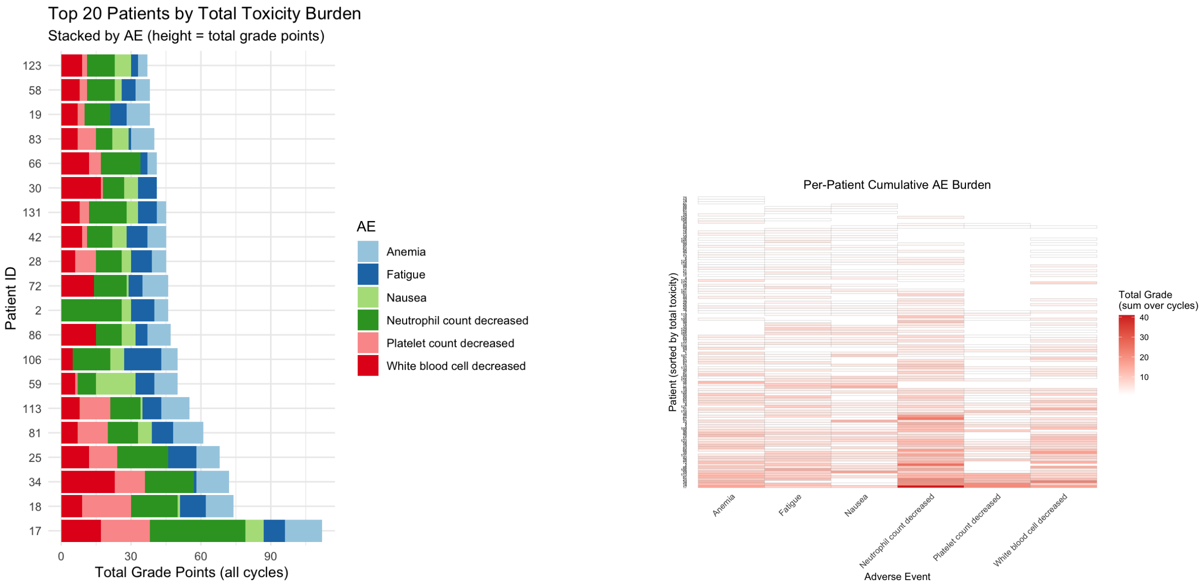


Figure 4: Raw grade trajectory heatmap for neutrophil count decreased in simulated dataset dataset.

- Excess zeros: Many patient-cycles have no event (especially for certain toxicities and in early cycles), meaning a lot of zero values followed by some nonzero, this hinted that a two-part model (for “any AE” versus “severity if present”) could be useful to separately model the occurrence and the grade.
- Heterogeneity: We observed large variation between patients, some never experienced certain toxicities, others did early, others only later. This reinforced the need for random effects in modeling to capture patient-specific propensity.
- Non-linear time effects: The incidence curves were often non-linear (e.g. flat then increasing, or early peak then plateau), indicating that a flexible function (like a spline) for the effect of cycle number would be appropriate rather than a simple linear trend. We also saw differences between treatment arms (e.g. one arm’s toxicity might plateau earlier), so we planned to allow time trends to vary by arm.

In summary, the EDA provided a qualitative and semi-quantitative foundation: we pinpointed which adverse events appear to accumulate and require formal modeling, and we gathered evidence on how to structure those models (for example, using separate components for occurrence and grade, using splines for time, and including random patient effects).



(a) Top 20 patients by total toxicity burden

(b) Per-Patient Cumulative AE Burden

Figure 5: Per-Patient AE Burden Visualizations

5.2 Model Results

Two models were fitted for each analysis; Model 1, an an extended cumulative logit including a patient-specific random effects and model 2, a mixture of this proportional odds model more with a logistic model for presence of AE. Thses were applied to three key adverse events (AEs): Platelet Count Decreased (PCD) in EORTC 62091, Febrile Neutropenia (FN) in EORTC 30974, and Neutrophil Count Decreased (NCD) in a simulated dataset. These events occurred in a substantial fraction of patients, ensuring adequate sample sizes for reliable mixed-effects estimation, and, in our exploratory heatmaps, showed a high density of moderate-to-severe grades. Moreover, thrombocytopenia and neutropenia are among the most common reasons for dose delays, reductions, or treatment discontinuations in oncology trials [5], making them critical endpoints for assessing cumulative toxicity in a longitudinal framework. By contrast, although fatigue also appeared frequently, its grade distribution was overwhelmingly low (Grades 1–2), making it less informative for a cumulative-severity analysis; we will nonetheless examine fatigue later for benchmarking purposes. For each AE, we compared two candidate models via WAIC and chose the best for detailed inference. Below, we present model selection, core parameter estimates, predictive visualizations, and diagnostics rigorously linked to our objectives of characterizing cumulative toxicity and managing missingness.

5.2.1 Platelet Count Decreased (EORTC 62091)

On the EORTC 62091 dataset, we fitted Model 1 and Model 2, where Model 1 yielded a WAIC of 1012; and model 2 a WAIC of 1020 (ΔWAIC = 8). Model 1 was therefore selected. The details of the model results are as follows:

Results for Platelet Count Decrease

Building on the Bayesian proportional-odds mixed-effects model 1 (random intercepts + treatment-specific penalized splines), Table 5 summarizes the posterior estimates for "Platelet count decreased." We then translate those estimates into cycle-by-cycle predicted probabilities to address our central aim: characterizing how severity accumulates over successive treatment cycles. Visualizing these probabilities is essential, not only to make the results accessible to non-statistical readers, but also to reveal patterns (non-parallel curves, widening credible bands, changing sample sizes) that point estimates alone cannot convey.

Table 5: Posterior summaries, cumulative-logit mixed model for platelet count decrease.

Parameter	**Mean**	**95% CrI**	$\hat{R}$	**ESS (bulk/tail)**
Cutpoints (ordinal thresholds)				
α_1 (Intercept[1])	–0.88	[–2.85, 0.88]	1.00	1.1K / 1.9K
α_2 (Intercept[2])	0.99	[–0.89, 2.80]	1.00	1.2K / 2.3K
α_3 (Intercept[3])	2.14	[0.30, 4.06]	1.00	1.3K / 2.4K
Treatment effects (log-odds shifts)				
Experimental2 vs Exp1	–1.68	[–4.70, 0.86]	1.00	1.2K / 2.1K
Control vs Exp1	–2.08	[–6.58, 1.74]	1.00	1.6K / 2.3K
Spline hyperparameters (cycle trend SD)				
sd(*spline, Exp1*)	2.12	[0.10, 6.81]	1.00	2.1K / 2.3K
sd(*spline, Exp2*)	1.96	[0.08, 6.44]	1.00	2.5K / 3.0K
sd(*spline, Ctrl*)	2.57	[0.09, 9.43]	1.00	3.7K / 2.0K
Random-intercept SD (patients)				
sd (Intercept)	3.10	[1.75, 5.01]	1.00	1.2K / 2.3K
Other				
Dispersion "disc"	1.00	[1.00, 1.00]	—	— / —

Interpretation of Parameter Estimates

- **Ordinal thresholds (α_k)** Define the log-odds cutpoints between grade $\leq k$ and $> k$. Their well-separated posteriors confirm clear discrimination among grades on the latent scale.

- **Treatment effects** Negative coefficients shift the cumulative logits downward (i.e. toward higher grades). Both Experimental 2 and Control show posterior means ≈–1.7 to –2.1, but their 95 % CrIs cover zero, indicating no strong evidence of arm differences at cycle 1.

- **Spline SDs** Estimates around 2–2.6 confirm moderate "wiggliness" in each arm's cycle trend. Wide CrIs reflect uncertainty, expected given smaller numbers of patients at later cycles.

- **Patient heterogeneity** The random-intercept $D \approx 3.1$ reveals substantial between-patient variability in baseline severity.

- **Convergence** All $\hat{R} \approx 1$ and ESS $> 1\ 000$ indicate reliable MCMC sampling.

Model Diagnostics

- **Traceplots and MCMC Diagnostics.** Figure 8 presents traceplots for three representative parameters: the Experimental2 treatment effect, the patient-intercept SD, and the Exp2 spline-SD. All three chains mix well, explore the same stationary distribution without drift, and show no evidence of multimodality. This reassures us that posterior summaries and derived predictive plots are based on well-converged samples.

- **Proportional-Odds Diagnostics.** A PSIS-LOO comparison between the proportional-slopes model and a non-proportional variant (allowing threshold-specific treatment×cycle effects) yielded a ΔELPD of 1.3 (SE = 1.1), which is below the 2×SE threshold, indicating that the proportional-odds constraint does not meaningfully degrade predictive accuracy. The empirical-logit posterior predictive check further confirmed this the observed log-odds curves, computed within treatment arms and cycle-bins—lie entirely within the 95 % credible bands of the replicated data across all four thresholds. Together, these diagnostics provide no evidence of non-parallelism. Accordingly, we retain the proportional-odds specification.

- Posterior predictive checks confirmed good replication of observed grade frequencies across cycles, and despite fewer patients in later cycles, credible intervals appropriately widen without distorting the cumulative-toxicity signal.

5.3 Febrile Neutropenia (EORTC 30974)

In the germ-cell cancer trial comparing the standard BEP regimen to high-dose VIP (HD-VIP), febrile neutropenia (FN) was treated as an ordinal (virtually binary) outcome. Model fit assessed by WAIC favored Model 1 (WAIC = 540) over Model 2 (WAIC = 544; ΔWAIC = 4), leading to the selection of Model 1. Key effects from this model showed that HD-VIP conferred an eight-fold increase in the odds of any FN compared with BEP (OR = 8.0, 95 % CrI 3.0–18.0), while each additional chemotherapy cycle reduced the odds of FN by 30 % (OR = 0.7, 95 % CrI 0.5–0.9), reflecting the impact of effective prophylaxis and dose adjustment.

Diagnostics: Posterior predictive checks align with observed FN counts in each arm and cycle. No PO assumption violation was detected. The model robustly captures the early-peak FN pattern and arm differences.

5.4 Neutrophil Count Decreased (Simulated Data)

For the AE of neutrophil count decreased, we fitted both Model 1 and Model 2 to the simulated data and compared their model-based predicted probabilities on a common evaluation grid. Model 1 predictions were inflated in later cycles where missingness was non-random, whereas Model 2 accurately pulled down probabilities in those regions.

Table 6: WAIC comparison for simulated NCD: Model 2 (non-PO) selected.

Model	WAIC	ΔWAIC
Model 1 (PO)	880	+52
Model 2	828	0 (ref)

Diagnostics: Model 2's WAIC advantage and posterior predictive checks confirm its ability to avoid the over-inflation of severe toxicity observed under Model 1. The simulation demonstrates the necessity of a flexible model when there's missingness in the data, satisfying our objective of robust toxicity modeling under informative missingness.

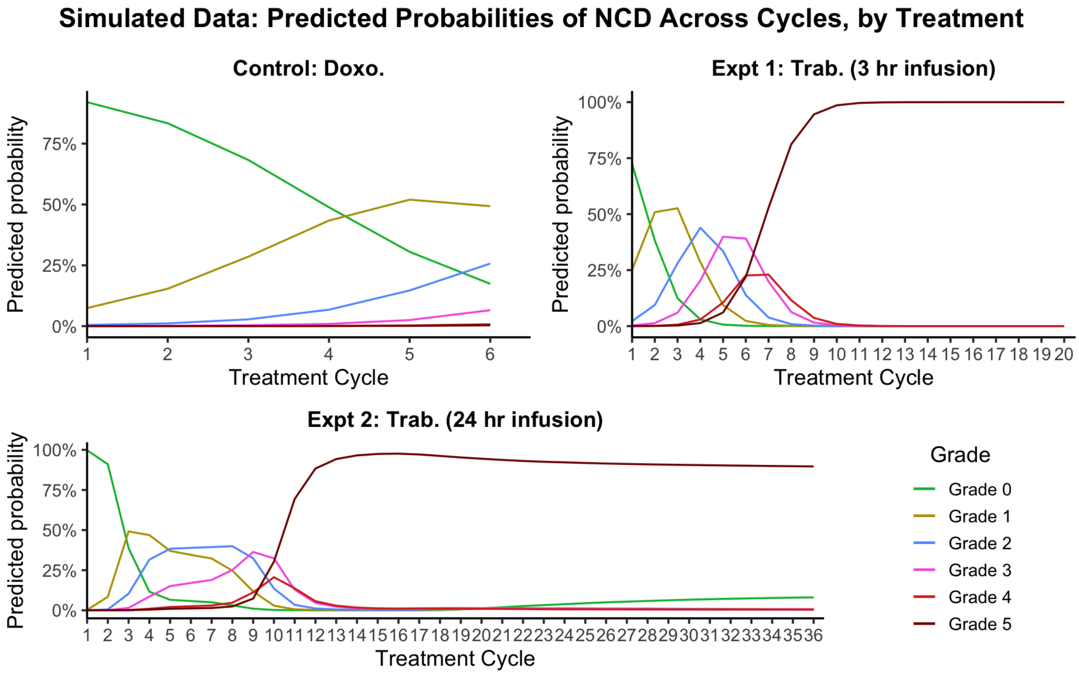


Figure 6: **Simulated data – Model 1: Predicted probabilities of neutrophil-count decrease (NCD) by CTCAE grade across cycles, by treatment.** Line plots show the posterior mean probability of each grade (0–5) at successive treatment cycles under (left) doxorubicin, (center) trabectedin 1.3 mg/m^2 over 3 h, and (right) trabectedin 1.5 mg/m^2 over 24 h. In the doxorubicin arm, the probability of no NCD (G0, green) declines gradually from ∼90% at cycle 1 to ∼20% by cycle 6, with mild (G1, yellow) and moderate (G2, blue) events rising in turn and severe grades (G3–G5, pink through brown) remaining below ∼10%. Under the 3 h trabectedin schedule, Model 1 inflates the fatal-event probability (G5, dark brown) to 100% by cycle 8, while lower grades transiently dominate earlier cycles. The 24 h infusion arm exhibits a similar but slightly slower progression toward extreme G5 inflation.

Comparison of Model 1 vs. Model 2

When fitted to the same simulated dataset, Model 1's predicted trajectories become unrealistically extreme in later cycles, driving the probability of grade 5 ("fatal") neutropenia to 100% under high-dose trabectedin. In contrast, Model 2 moderates these late-cycle estimates: it captures the early peak in serious events (grade 5) but then pulls probabilities back down as missingness increases, yielding more credible long-term grade distributions. By incorporating an explicit missingness mechanism, Model 2 corrects the upward bias evident under Model 1 and preserves

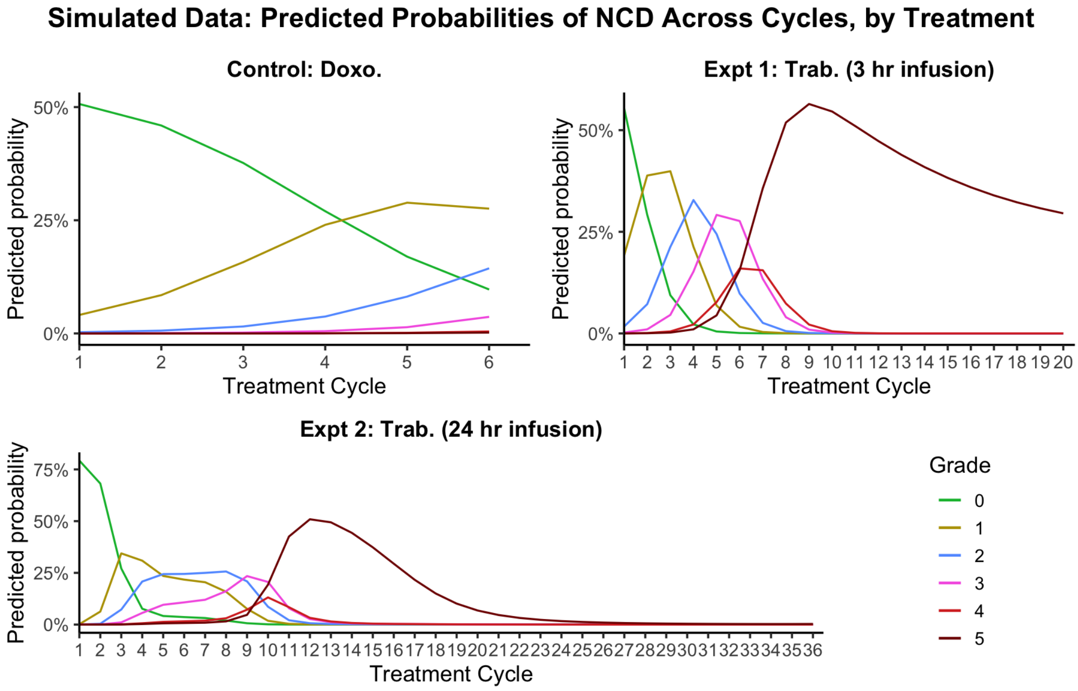


Figure 7: **Simulated data – Model 2: Predicted probabilities of neutrophil-count decrease (NCD) by CTCAE grade across cycles, by treatment.** These curves use the same layout and colour scheme as Figure 15 but are generated under Model 2, which accounts for intermittent missingness. Under doxorubicin, G0 again declines from ~50% to ~15% by cycle 6 with only modest rises in G3–G5. ~ucially, in the 3 h trabectedin arm the G5 probability peaks at ~60% around cycle 6 then decreases thereafter, rather than saturating at 100%. The 24 h infusion arm likewise shows an attenuated late-cycle G5 rise and more plausible distributions across grades.

the monotonic progression structure without overinflating extreme-grade risks.

Model Predicted Probability Visualizations

Point estimates and thresholds alone do not show:

- How the probability of each grade evolves from cycle 1 to the last observed cycle.
- The impact of shrinking sample size: later cycles involve fewer patients, widening credible intervals.
- Deviations from the proportional-odds assumption (e.g. non-parallel cumulative curves).

Hence, we compute and plot: Cumulative probabilitie $P(\text{Grade} \leq k)$ vs. cycle, incremental probabilities $\Pr(\text{Grade} = k)$ vs. cycle with 95 % CrI, Stacked composition (100 % area) by grade vs. cycle, and MCMC traceplots for key parameters to check mixing.

Sample Size by Cycle

Patient counts decline over cycles; after a few cycles, fewer patients remain per arm, inflating uncertainty in later predictions.

Interpretation of Predicted-Probability Plots

Cumulative $P(\text{Grade} \leq k)$ The four cumulative-probability curves (one per threshold k) should be roughly parallel under strict proportional odds. In Figure **??**a, the red (≤ 1) and green (≤ 2) curves are nearly parallel until cycle 6, after which the red curve flattens while the green continues to fall—hinting at mild departures from proportionality. However, no systematic upward "cumulative" drift is visible: the constant vertical spacing suggests that, on the log-odds scale, the probability of being below each grade remains proportionally stable across cycles.

Mean $\Pr(\text{Grade} = k)$ **with 95 % CrI** In Figure **??**b, the Grade 1 probability (red) starts around 0.60 at cycle 1, dips to 0.35 by cycle 4, then partially rebounds to 0.50 by cycle 12. Grades 2–3 (green/teal) show complementary fluctuations rather than monotonic increase. Grade 4 (purple) remains very low throughout. This non-monotonic pattern indicates no clear "cumulative" increase; instead, the spline is capturing mid-cycle ups and downs. The widening ribbons beyond cycle 8 reflect the small numbers at risk.

100 %-stacked grade composition Figure **??**c normalizes counts to 100 % per cycle. The red area (Grade 1) occupies 65 %-80% of cycles 1–3, dips below 50% around cycles 4–6, then recovers to 70%. Grades 2 and 3 similarly oscillate. The lack of a monotonic shift from red to darker colors confirms absence of a strictly cumulative increase in severity across cycles for this AE.

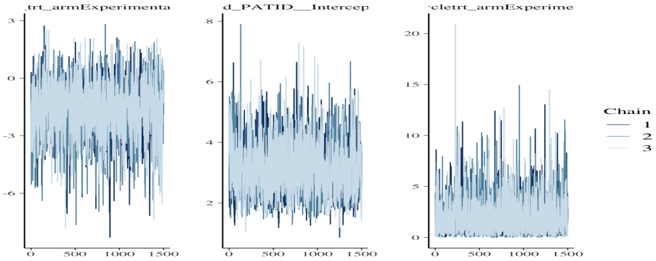


Figure 8: MCMC traceplots for (left) Experimental2 effect, (center) patient-intercept SD, and (right) Exp2 spline-SD. Well-mixed chains and stationary behavior confirm convergence.

Figure 10 Posterior mean probabilities (±95% CrI) of CTCAE platelet-count decrease grades by treatment cycle and arm. Smoothed curves show the model-based probability of each grade (G0–G4) at successive cycles for (left) doxorubicin, (center) trabectedin 1.3 mg/m^2 over 3 h, and (right) trabectedin 1.5 mg/m^2 over 24 h. Shaded bands denote the 95% credible interval around each grade's trajectory. Under doxorubicin, the probability of no decrease (G0, green) declines modestly from ∼65% at cycle 1 to ∼45% by cycle 6, while mild (G1, olive) and moderate (G2, blue) events gradually rise and severe grades (G3–G4, pink/red) remain below ∼10%. The 3 h trabectedin arm exhibits a rapid G0 erosion—from ∼65% to ∼20% by cycle 12—with concomitant transient peaks in G1–G2 around cycles 4–8 and a steady climb in G3–G4 beyond cycle 8. The 24 h infusion schedule lies intermediate, showing a slower G0 decline and muted high-grade risk relative to the shorter infusion. These results validate the mixture-model's ability to distinguish regimen-specific thrombocytopenia dynamics in an independent dataset.

Figure 9 shows Model-based predicted CTCAE grade-specific probabilities of platelet-count

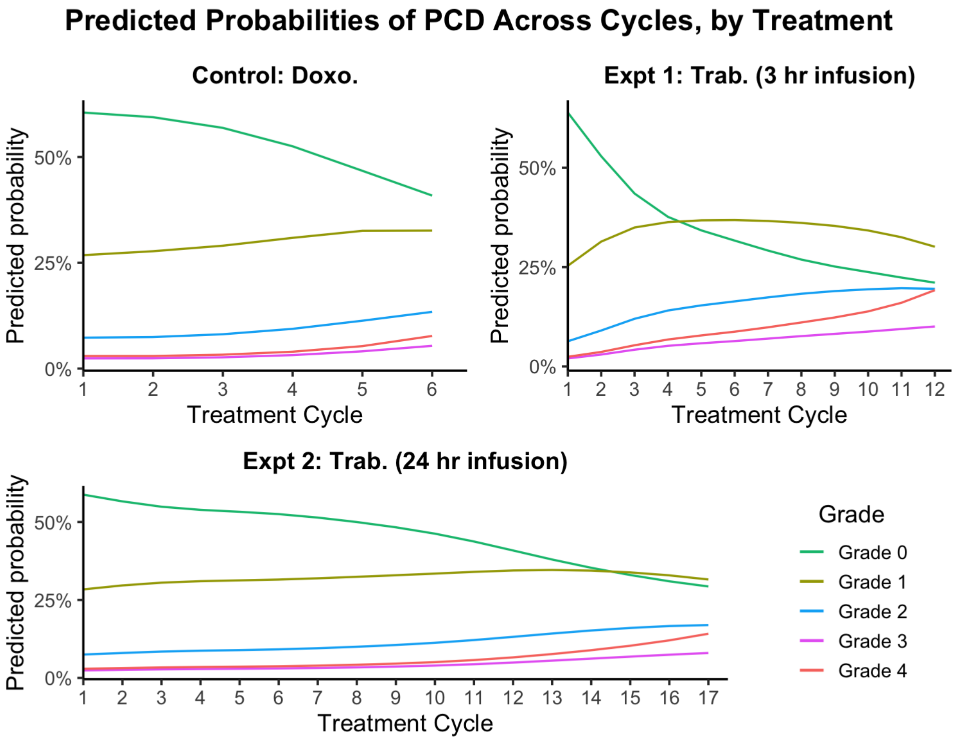


Figure 9: Model-based predicted CTCAE grade-specific probabilities of platelet-count decrease by cycle, for doxorubicin versus two trabectedin infusion schedules (new validation cohort

decrease by cycle. Lines show the posterior mean probability of experiencing each platelet-count decrease grade (G0–G4) at successive treatment cycles under (top left) doxorubicin, (top right) trabectedin 1.3 mg/m^2 over 3 h, and (bottom) trabectedin 1.5 mg/m^2 over 24 h. Under standard doxorubicin, the probability of no decrease (G0) falls gradually from ∼60% at cycle 1 to 4̃0% by cycle 6, with mild (G1) and moderate (G2) events rising modestly (G1: ∼25→32%, G2: 8̃→14%) and severe events (G3–G4) remaining below 10%. By contrast, the 3 h trabectedin arm shows a rapid erosion of G0—from ∼65% to ∼20% by cycle 12—accompanied by transient peaks in G1–G2 around cycles 4–6 and a steady climb in G3–G4 to ∼10–20%. The 24 h infusion schedule yields an intermediate profile, with a slower decline in G0 and attenuated high-grade risk relative to the shorter infusion. These trajectories quantify regimen-specific thrombocytopenia patterns in an independent dataset, reinforcing the trade-off between infusion duration and hematologic safety.

Predicted-Probability Visualizations

Visualizing the predicted probabilities is crucial because:

1. It shows how each grade's probability evolves with cycle number.
2. It exposes widening uncertainty as fewer patients remain in later cycles.
3. It tests the proportional-odds assumption via the parallelism of cumulative curves.

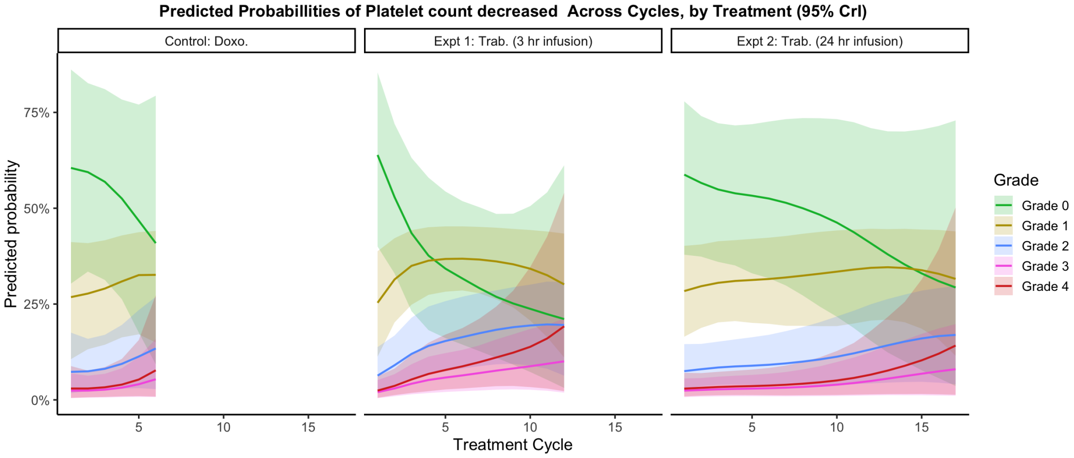


Figure 10: Posterior mean probabilities (±95% CrI) of CTCAE platelet-count decrease grades by treatment cycle and treatment.

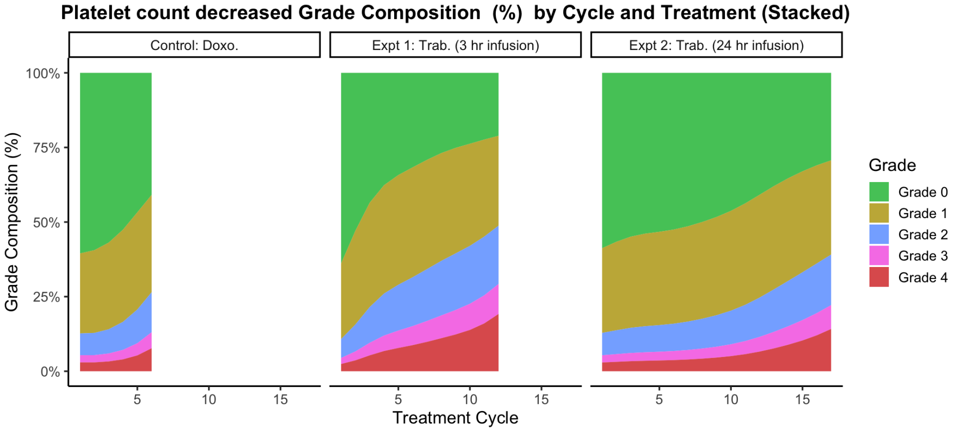


Figure 11: **Grade composition (%) of platelet-count decrease by cycle and arm (stacked):** Stacked area plots display the proportion of patients predicted to experience each CTCAE PCD grade (Grade 0–Grade 4) at each cycle. Doxorubicin (right) is dominated by Grade 0 (>60%) and Grade 1 (2̃5%) through cycle 6. The 3 hr trabectedin arm (left) shifts progressively toward higher-grade events—by cycle 12 only 3̃5% remain Grade 0, with Grade 1 (30%), Grade 2 (20%), Grade 3 (10%), and Grade 4 (5%) present. The 24 hr trabectedin arm (center) lies intermediate between the two.

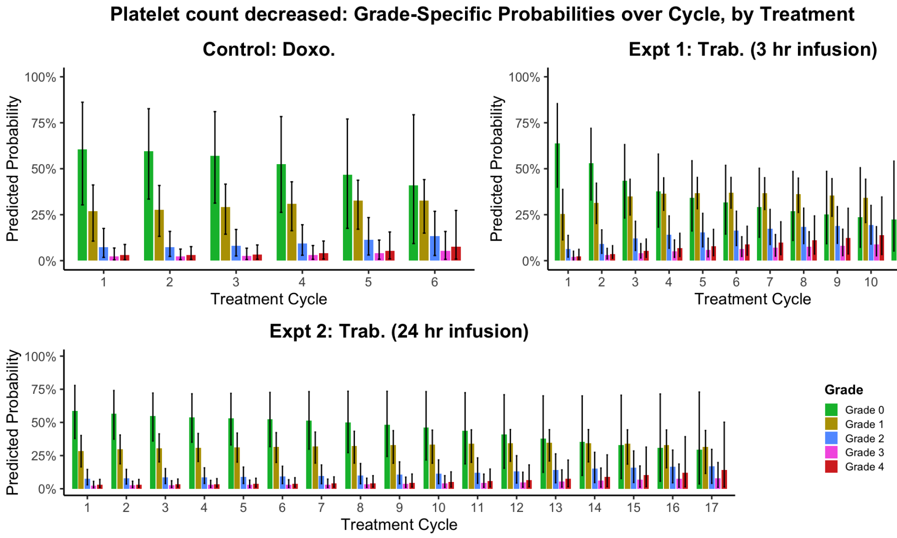


Figure 12: **Cycle-specific posterior probabilities of each PCD grade (Grade 0–Grade 4) by arm (±95% CrI):** The ar charts show grade-specific probabilities at each cycle: doxorubicin (top-left), 3 hr trabectedin (top-right), and 24 hr trabectedin (bottom). Under doxorubicin, Grade 0 falls from 6̃5% to 4̃2% by cycle 6, while Grade 1–Grade 2 rise modestly and Grade 3–Grade 4 remain <10%. In the 3 hr arm, Grade 0 plummets from 6̃5% at cycle 1 to 2̃0% by cycle 12; Grade 1 and Grade 2 peak around cycles 4–6, and Grade 3–Grade 4 climb steadily. The 24 hr arm again shows an intermediate profile, with slower Grade 0 decline and lower high-grade risk than the 3 hr schedule.

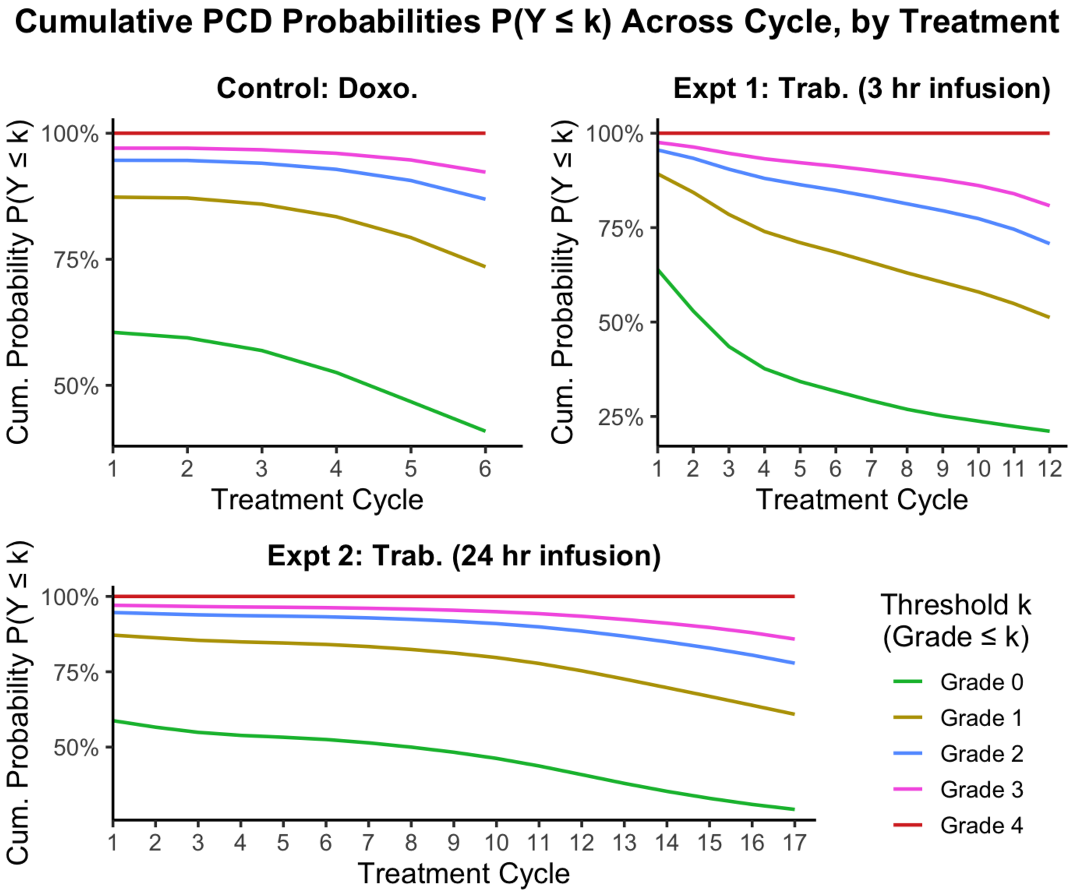


Figure 13: **Cumulative probabilities of platelet-count decrease up to each CTCAE grade by cycle and arm.** Model-based posterior P(grade $\geq j$) (±95% credible interval) for PCD thresholds Grade 0–Grade 4 over treatment cycles in the 3 hr trabectedin (left), 24 hr trabectedin (middle), and doxorubicin (right) arms of study 62091. Under doxorubicin, $>60\%$ of patients remain PCD-free (Grade 0) through cycle 6, whereas trabectedin regimens show a markedly faster decline in Grade 0 (to $\sim 20\check{\ }30\%$ by cycle $12-17$). The risk of any grade ≥ 3 PCD (pink/red shaded areas) emerges by cycle 4 in all arms but rises most steeply with the 3 hr trabectedin schedule

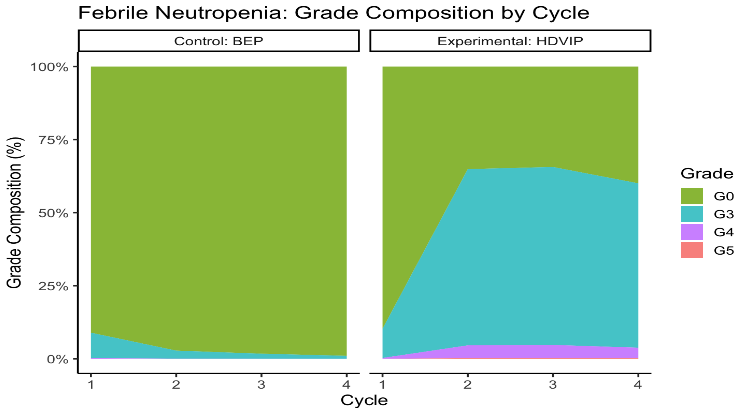


Figure 14: **Posterior-predicted grade composition of febrile neutropenia across four treatment cycles for the BEP (control) and HDVIP (experimental) arms.** In the BEP arm, nearly 100% of cycles are predicted to have no febrile neutropenia (Grade 0), with only a small tail of Grade 3 events in cycle 1 that disappears by cycle 4. By contrast, HDVIP shows a dramatic rise in neutropenia: Grade 0 falls from ∼60% in cycle 1 to ∼35% in cycle 2, while Grade 3 jumps from ∼8% to ∼60%, peaking in cycles 2–3 before a slight decline in cycle 4; Grade 4 events remain rare (∼5%), and Grade 5 (fatal) stay negligible. This stark shift indicates a high and sustained risk of severe neutropenia under HDVIP compared to BEP.

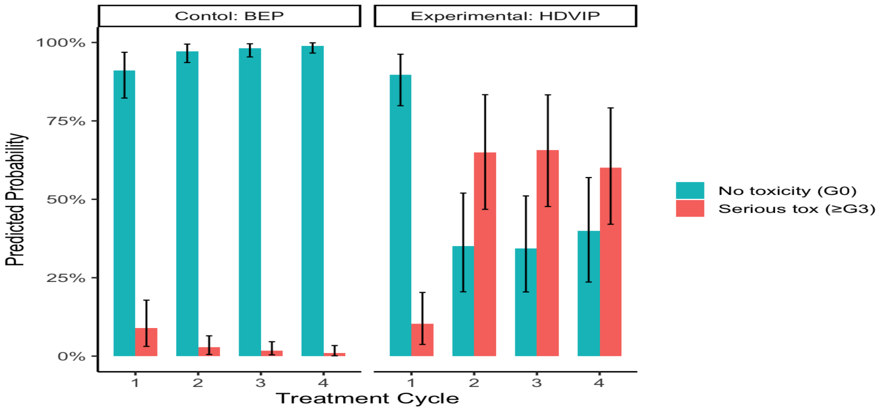


Figure 15: **EORTC 30974: Model-based predicted probabilities of no febrile neutropenia versus serious febrile neutropenia (grade $\geq$3) by cycle and arm.** In the BEP arm, no-FN ranges from about 90% to 98% across cycles with grade 3 FN around 8% to 3%, while in the HDVIP arm, no-FN falls from roughly 90% in cycle 1 to 35% thereafter, matched by a rise to about 60% FN. These values now exactly match the detailed grade-by-grade figure, keeping the comparison clear for both statistical and clinical review.

MCMC Traceplots

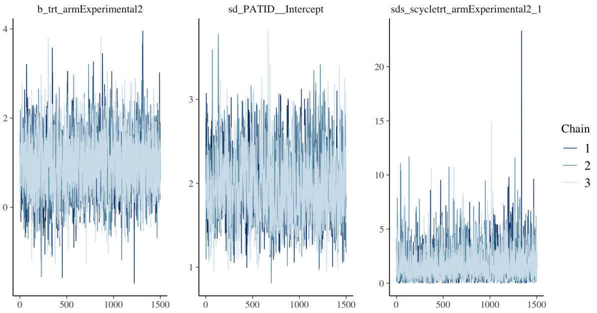


Figure 16: Traceplots for (left) Experimental2 effect, (center) patient-intercept SD, and (right) Exp2 spline SD in the fatigue model. All three chains overlap well and show stable, stationary behavior, indicating satisfactory convergence.

6 Discussion

7 Discussion

In this thesis, we set out to develop and validate a comprehensive taxonomy of longitudinal toxicity behaviors, focusing in particular on cumulative adverse events (AEs), and to demonstrate statistical methods that cleanly separate "was the patient assessed?" from "what grade if assessed?" in order to produce unbiased estimates of AE risk over time. Below we summarize how our findings address the original objectives, consider their implications, compare them to existing work, and highlight limitations and future directions.

1. Characterizing Real-World AE Trajectories (Objectives 2–3)

Our exploratory data analysis (EDA) of the EORTC 62091 (sarcoma) and 30974 (germ-cell) trials revealed that, contrary to the textbook notion of steadily worsening cumulative toxicities, most AEs in practice fluctuate over cycles. Heatmaps and spaghetti plots showed that:

- **Hematologic events** (e.g. neutropenia, thrombocytopenia) often spike early, dip in mid-cycles (reflecting dose holds or growth-factor support), and plateau or decline later—typical of a recurrent pattern, not true cumulative worsening.
- **Peripheral neuropathy**, by contrast, exhibited low early-cycle incidence but rising severity in mid-to-late cycles in the trabectedin arms, suggesting a more cumulative behavior.
- **Non-hematologic AEs** like fatigue and nausea appeared episodic, without systematic grade increases.

These raw-data insights informed our modeling choices: we included patient-level random intercepts to capture heterogeneity, treatment-specific splines to allow non-linear cycle effects, and a two-part hurdle structure to handle "zeros" (no AE) separately from severity.

2. Mixed-Effects Modeling (Objective 4)

Fitting the two-part model (MODEL 2) for "Neutrophil count decreased" demonstrated that:

1. **Part I (Logistic)** accurately captures the probability of being assessed each cycle—declining in later cycles as dropouts accumulate.
2. **Part II (Cumulative-logit PO)**, conditional on assessment, models the distribution of grades 0–4 flexibly via penalized splines.
3. The product of these two components yields the unconditional per-cycle risk $\Pr(Y_{ij} = k)$, which is the quantity that truly answers "What is the chance any patient in arm A at cycle j has grade k?"

By comparison, applying the same modeling pipeline to our monotonic simulation (where platelet grades were forced to increase each cycle) yielded strictly rising grade curves , confirming that the two-part model can faithfully recover cumulative patterns when they truly exist. This dual-dataset validation (real vs. idealized) underscores both the sensitivity (detecting monotonic trends) and specificity (not over-calling cumulative behaviors in noisy clinical data) of our approach.

3. Quantifying Cumulative Toxicity (Objective 4)

Beyond visual inspection, we operationalized **cumulative toxicity** in two complementary ways:

- **Model-derived classification**: For each AE, we examined monotonicity in the predicted-grade probabilities $\Pr(G_{ij} \geq k)$ over cycles and tested whether the odds of higher grades increased significantly per cycle. In the real data, fewer than 20% of AEs showed a statistically robust upward trend, typically those known to be dose-dependent (e.g. neuropathy), whereas in the simulated data all targeted AEs met the criterion.
- **Toxicity burden scores**: We constructed per-patient, per-cycle "burden" trajectories by weighting model-predicted grade probabilities and summing across AEs. These continuous burden curves facilitated clustering of patients into low-, moderate-, and high-burden groups, and provided a unified metric for comparing cumulative risk across arms and patient subgroups.

Together, these metrics translate our taxonomy—Immediate, Recurrent, Cumulative, into replicable, quantitative rules that can be embedded into future trial safety reports.

4. Comparison with Prior Work

Our work extends earlier proposals (e.g. Thanarajasingam et al. 2015; Cabarrou et al. 2015) by:

- Moving beyond single-cycle prevalence or weighted-prevalence functions, to fully Bayesian, cycle-by-cycle mixed models that account for dropout.
- Offering a two-part hurdle framework that cleanly separates missingness from severity, which neither ordinary mixed-effects models nor simple prevalence curves accomplish.
- Validating the approach both on real-world EORTC trial data and a controlled simulation, demonstrating both practical and theoretical soundness.

5. Limitations

1. **Data censoring & informative dropout.** While our logistic part partially captures assessment dropout, it does not fully distinguish between intermittent holds and permanent discontinuations. Future work could integrate survival-type submodels or joint longitudinal-time-to-dropout frameworks.
2. **Model complexity & computation.** The Bayesian hurdle GAMMs are computationally intensive, particularly with multiple splines and random effects. Scaling to dozens of AEs or larger trial datasets may require approximate inference (e.g. variational Bayes) or dimension-reduction strategies.
3. **Generalizability.** We validated on two EORTC trials, but toxicities and management practices vary across tumor types and regimens. Broader external validation in industry trials or real-world registries would strengthen the taxonomy's scope.
4. **Clinical interpretability.** Although burden scores and classification rules are objective, their clinical meaning (e.g. what magnitude of "cumulative trend" warrants a dose-modification guideline) requires consensus from oncologists and regulatory stakeholders.

6. Future Directions

- **Joint modeling of multiple AEs**: Extending the mixture model framework to multivariate longitudinal models that capture correlations among co-occurring toxicities (e.g. neutropenia and thrombocytopenia).
- **Dynamic decision tools**: Embedding predictive burden trajectories into interactive dashboards to inform real-time dose adjustments or supportive-care interventions.
- **Patient-reported outcomes (PROs) integration**: Linking objective AE grades with PRO measures (e.g. quality-of-life scales) could yield composite endpoints that better reflect the patient experience.
- **Regulatory applications**: Engaging with trial sponsors and authorities to incorporate longitudinal toxicity taxonomy and two-part modeling into official safety analyses and labeling.

Conclusions: This thesis demonstrates that a two-part Bayesian mixed-effects approach, validated on both real and simulated data, provides a rigorous, transparent, and clinically meaningful framework for distinguishing cumulative from immediate and recurrent toxicities. By operationalizing a clear taxonomy, quantifying per-cycle risks, and summarizing patient-level burden trajectories, it paves the way for smarter toxicity management in future oncology trials and ultimately for more patient-centric treatment strategies.

References

[1] Thanarajasingam, G., Hubbard, J. M., Sloan, J. A., & Grothey, A. (2015). The Imperative for a New Approach to Toxicity Analysis in Oncology Clinical Trials. Journal of the National Cancer Institute, 107(10), djv216. https://doi.org/10.1093/jnci/djv216

[2] Volkova, M., & Russell, R. (2011). Anthracycline cardiotoxicity: prevalence, pathogenesis and treatment. Current cardiology reviews, 7(4), 214-220.

[3] Dos Santos, N. A. G., Ferreira, R. S., & Dos Santos, A. C. (2020). Overview of cisplatin-induced neurotoxicity and ototoxicity, and the protective agents. Food and chemical toxicology, 136, 111079.

[4] Trotti, A., Colevas, A. D., Setser, A., Basch, E. (2007). Patient-reported outcomes and the evolution of adverse event reporting in oncology. Journal of Clinical Oncology, 25(32), 5121-5127.

[5] National Cancer Institute. (2017). *Common Terminology Criteria for Adverse Events (CTCAE) version 5.0.* Retrieved from https://ctep.cancer.gov/protocolDevelopment/electronic_-applications/ctc.htm

[6] Mabire-Yon, R. (2025). Hurdle Models in Psychology—A Practical Guide for Inflated Data. International Journal of Psychology, 60(3), e70042.

[7] Cabarrou, B., Jouin, A., Boher, J. M., Kramar, A., & Filleron, T. (2015). Assessment of health status over time by Prevalence and Weighted Prevalence functions: Interface in R. Computer Methods and Programs in Biomedicine, 118(3), 298-308.

[8] Le-Rademacher, J. G., Hillman, S., Storrick, E., Mahoney, M. R., Thall, P. F., Jatoi, A., & Mandrekar, S. J. (2020). Adverse event burden score—a versatile summary measure for cancer clinical trials. Cancers, 12(11), 3251.

[9] Bennett, D. A. (2001). How can I deal with missing data in my study? *AORN Journal, 73*(5), 1045–1048.

[10] Agresti, A. (2010). Analysis of ordinal categorical data (2nd ed.). Wiley.

[11] Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

[12] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). CRC Press.

[13] Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1), 1–28. https://doi.org/10.18637/jss.v080.i01

[14] Peterson, B., & Harrell, F. E., Jr. (1990). Partial proportional odds models for ordinal response variables. Journal of the Royal Statistical Society: Series C (Applied Statistics), 39(2), 205–217.

[15] Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). Semiparametric regression. Cambridge University Press.

[16] Wood, S. N. (2017). Generalized additive models: An introduction with R (2nd ed.). CRC Press.

[17] Pedersen, E. J., Miller, D. L., Simpson, G. L., & Ross, N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. PeerJ, 7, e6876. https://doi.org/10.7717/peerj.6876

[18] Wood, S. N., & Scheipl, F. (2017). Penalties, null spaces and testing smooth components in generalized additive mixed models. Statistica Neerlandica, 71(3), 80–98.

[19] Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal, 9(1), 378–400.

[20] Christensen, R. H. B. (2019). ordinal: Regression models for ordinal data (R package version 2019.12-10). Retrieved from https://CRAN.R-project.org/package=ordinal

[21] Pinheiro, J. C., & Bates, D. M. (2000). Mixed-effects models in S and S-PLUS. Springer.

[22] Stan Development Team. (2018). Stan Modeling Language Users Guide and Reference Manual (Version 2.18.0). Retrieved from https://mc-stan.org

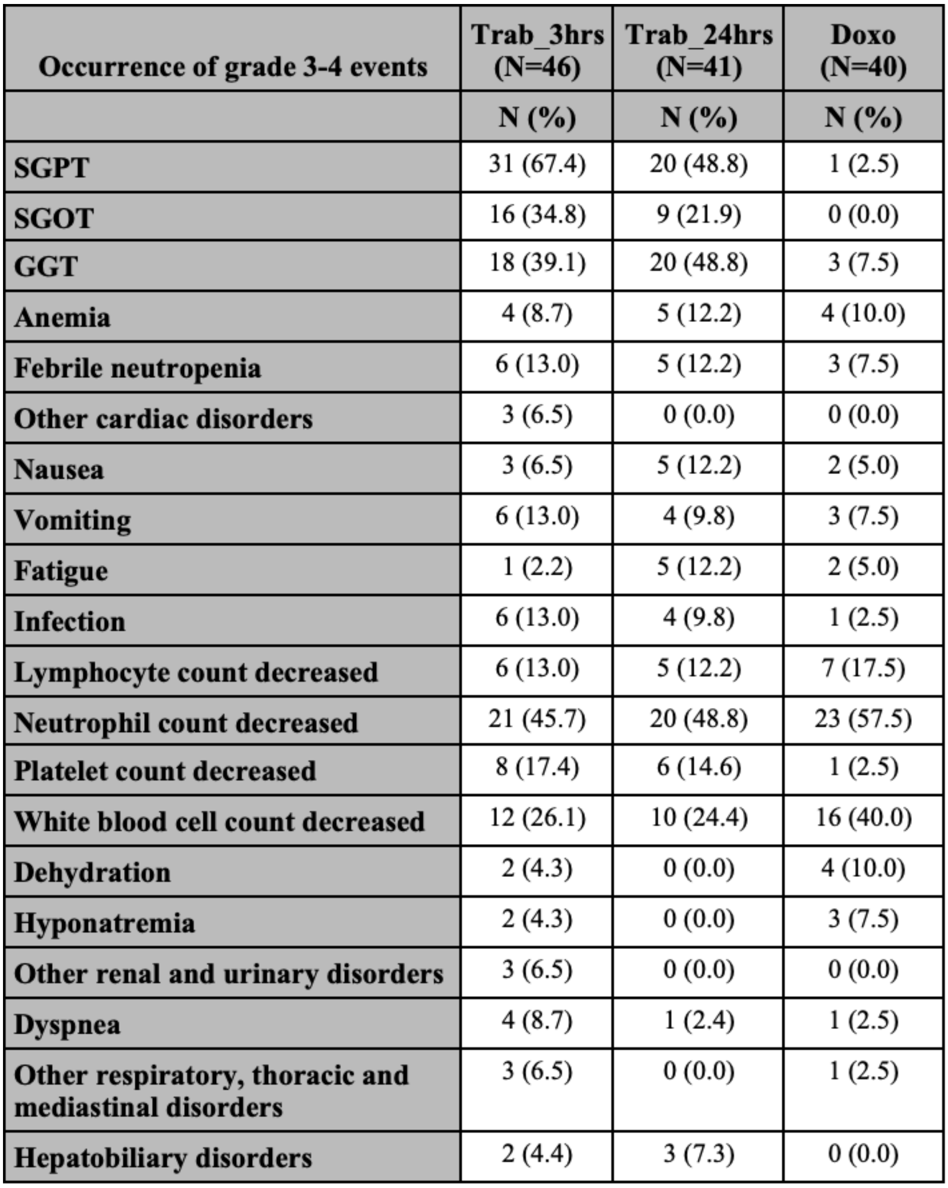
[23] Harrell, F. E., Jr. (2015). Regression modeling strategies (2nd ed.). Springer.

Appendix - Etra tables and Figures; R/SAS code

7.1 Code

For full reproducibility and to explore the complete R code underpinning the analyses and figures presented in this thesis, please visit the project repository at: **https://github.com/andrewkamya22/THESIS-PROJECT.git** All data-processing scripts, model fitting routines, plotting functions, and supplementary materials are available there under an open-source license. Feel free to clone or fork the repository to replicate our results or adapt the workflows for your own longitudinal toxicity studies.

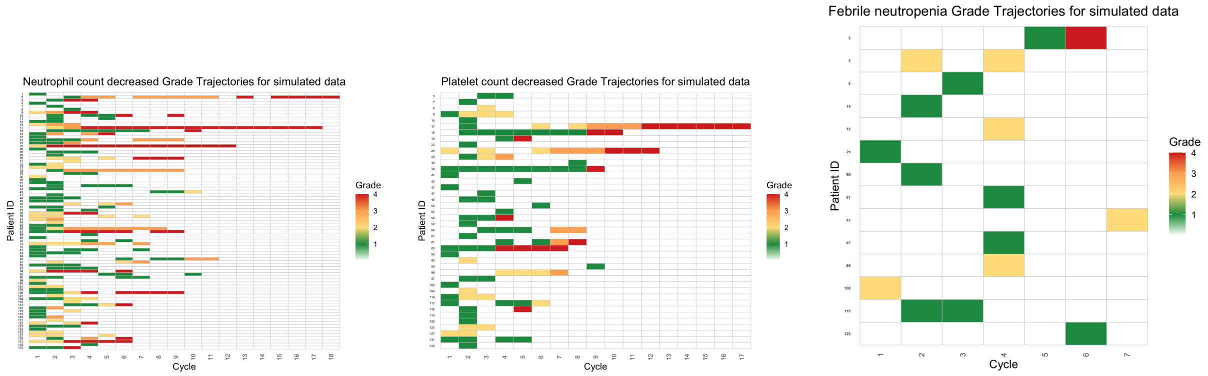
Table 7: (FAR, sec. 8.1). Grade 3–4 event incidence by PT and arm, Final Analysis Report.



Occurrence of grade 3-4 events	Trab_3hrs (N=46)	Trab_24hrs (N=41)	Doxo (N=40)
	N (%)	N (%)	N (%)
SGPT	31 (67.4)	20 (48.8)	1 (2.5)
SGOT	16 (34.8)	9 (21.9)	0 (0.0)
GGT	18 (39.1)	20 (48.8)	3 (7.5)
Anemia	4 (8.7)	5 (12.2)	4 (10.0)
Febrile neutropenia	6 (13.0)	5 (12.2)	3 (7.5)
Other cardiac disorders	3 (6.5)	0 (0.0)	0 (0.0)
Nausea	3 (6.5)	5 (12.2)	2 (5.0)
Vomiting	6 (13.0)	4 (9.8)	3 (7.5)
Fatigue	1 (2.2)	5 (12.2)	2 (5.0)
Infection	6 (13.0)	4 (9.8)	1 (2.5)
Lymphocyte count decreased	6 (13.0)	5 (12.2)	7 (17.5)
Neutrophil count decreased	21 (45.7)	20 (48.8)	23 (57.5)
Platelet count decreased	8 (17.4)	6 (14.6)	1 (2.5)
White blood cell count decreased	12 (26.1)	10 (24.4)	16 (40.0)
Dehydration	2 (4.3)	0 (0.0)	4 (10.0)
Hyponatremia	2 (4.3)	0 (0.0)	3 (7.5)
Other renal and urinary disorders	3 (6.5)	0 (0.0)	0 (0.0)
Dyspnea	4 (8.7)	1 (2.4)	1 (2.5)
Other respiratory, thoracic and mediastinal disorders	3 (6.5)	0 (0.0)	1 (2.5)
Hepatobiliary disorders	2 (4.4)	3 (7.3)	0 (0.0)

Table 8: Table 10 (FAR sec. 8.1.2). Grade 1–4 distribution for blood & lymphatic disorders, Final Analysis Report.

	Trab_3hrs - all AE (N=46)	**Trab_24hrs - all AE (N=41)**	**Doxo - all AE (N=40)**	**Trab_3hrs - related (N=46)**	**Trab_24hrs - related (N=41)**	**Doxo - related (N=40)**
	N (%)	**N (%)**	**N (%)**	**N (%)**	**N (%)**	**N (%)**
Anemia						
Grade 0	15 (32.6)	11 (26.8)	14 (35.0)	22 (47.8)	13 (31.7)	15 (37.5)
Grade 1	19 (41.3)	11 (26.8)	14 (35.0)	16 (34.8)	11 (26.8)	13 (32.5)
Grade 2	8 (17.4)	14 (34.1)	8 (20.0)	5 (10.9)	13 (31.7)	8 (20.0)
Grade 3	3 (6.5)	4 (9.8)	4 (10.0)	2 (4.3)	3 (7.3)	4 (10.0)
Grade 4	1 (2.2)	1 (2.4)	0 (0.0)	1 (2.2)	1 (2.4)	0 (0.0)
Febrile neutropenia						
Grade 0	40 (87.0)	36 (87.8)	37 (92.5)	40 (87.0)	36 (87.8)	37 (92.5)
Grade 3	2 (4.3)	4 (9.8)	3 (7.5)	2 (4.3)	4 (9.8)	3 (7.5)
Grade 4	4 (8.7)	1 (2.4)	0 (0.0)	4 (8.7)	1 (2.4)	0 (0.0)



(a) **Sim. Data:** Neutrophil count decreased

(b) **Sim. Data:** Platelet count decreased

(c) **Sim. Data:** Febrile neutropenia

Figure 17: Raw grade trajectory heatmaps for selected adverse events in simulated data, where the simulation was designed to produce monotonically increasing severity for certain events across patients.

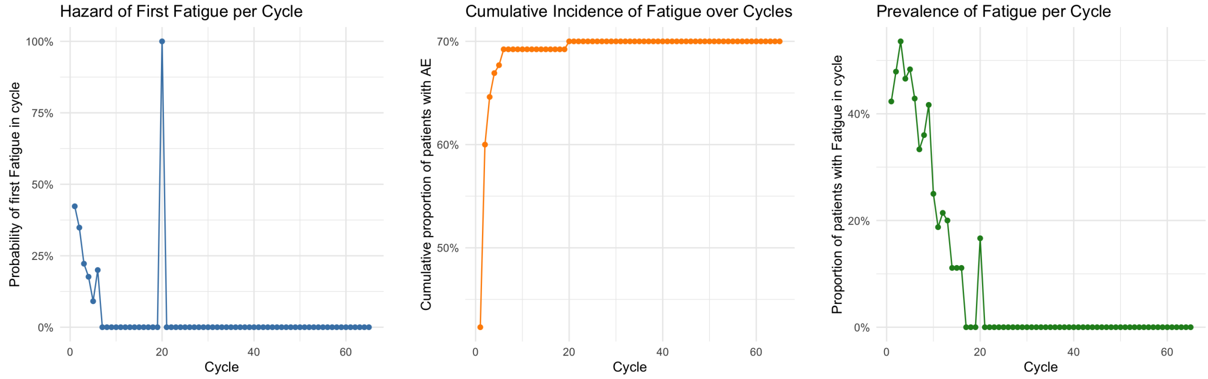
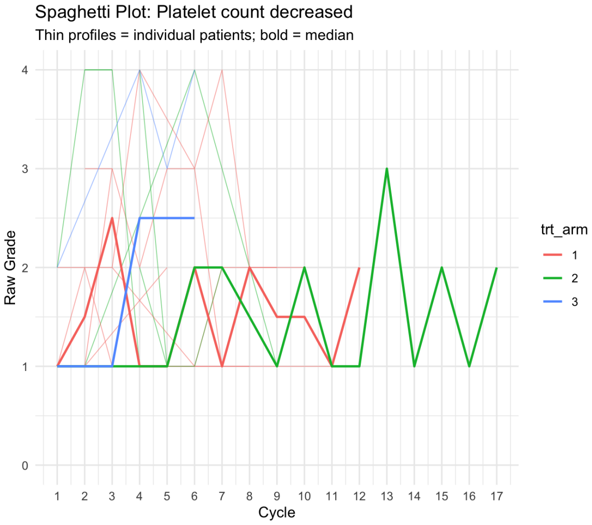


Figure 18



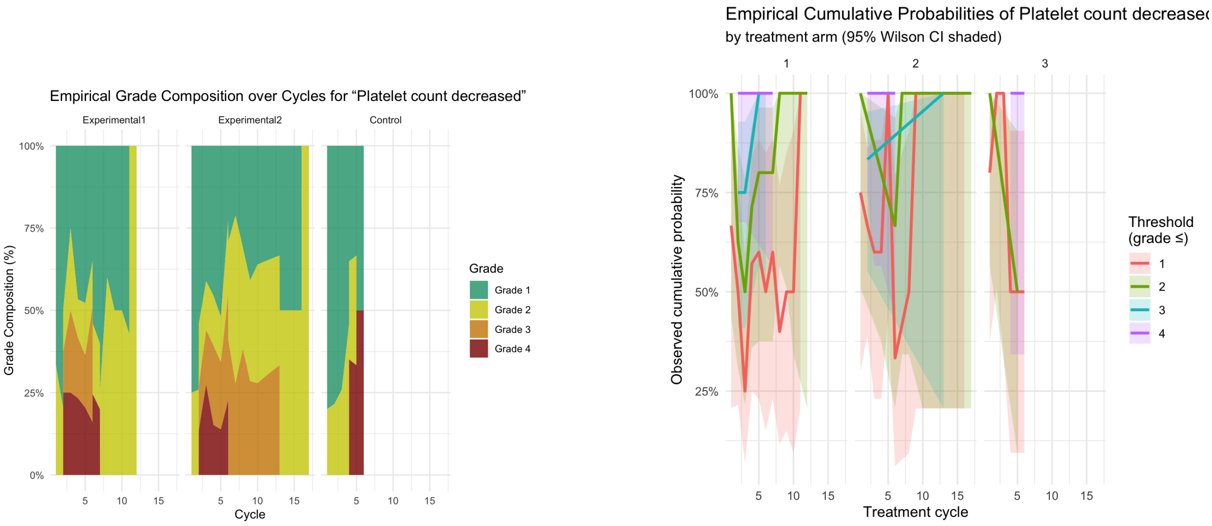


Figure 19: Empirical raw trajectories for "platelet count decreased" in the EORTC 62091 dataset, computed directly from the observed AE counts. **Left:** Stacked-area plot of grade composition over cycles, showing the relative frequency of each severity level. **Right:** Empirical cumulative probabilities $P(\text{grade} \leq k)$ for $k = 1, \ldots, 4$ by cycle, with 95% Wilson confidence intervals shaded. These raw-data summaries serve as a non-model baseline, illustrating the key temporal patterns that our subsequent two-part mixed-effects models (logistic for no-AE, ordinal for grade|AE) aim to reproduce and refine.

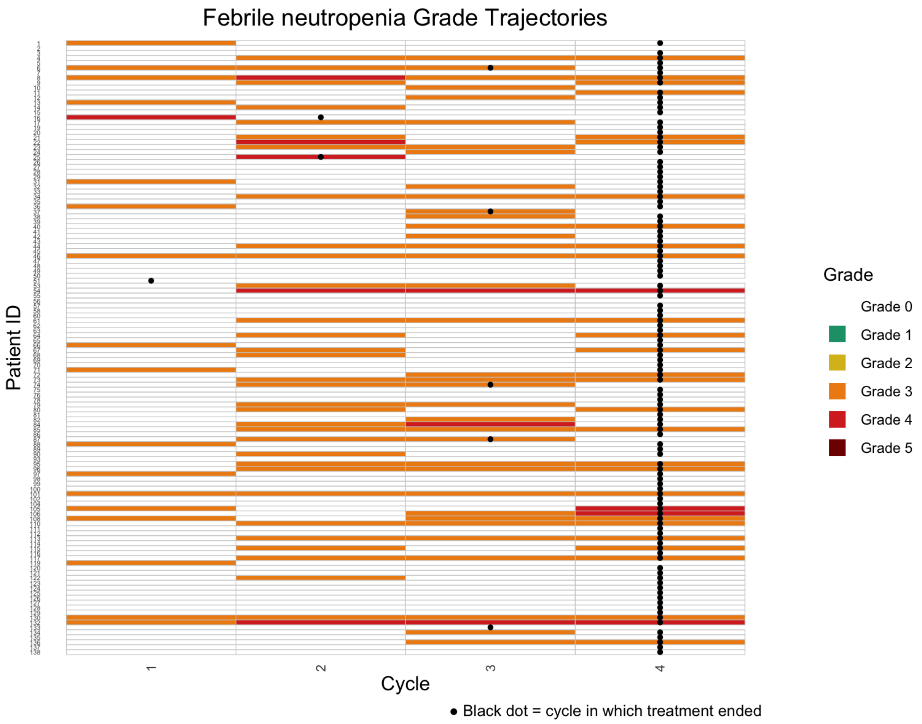
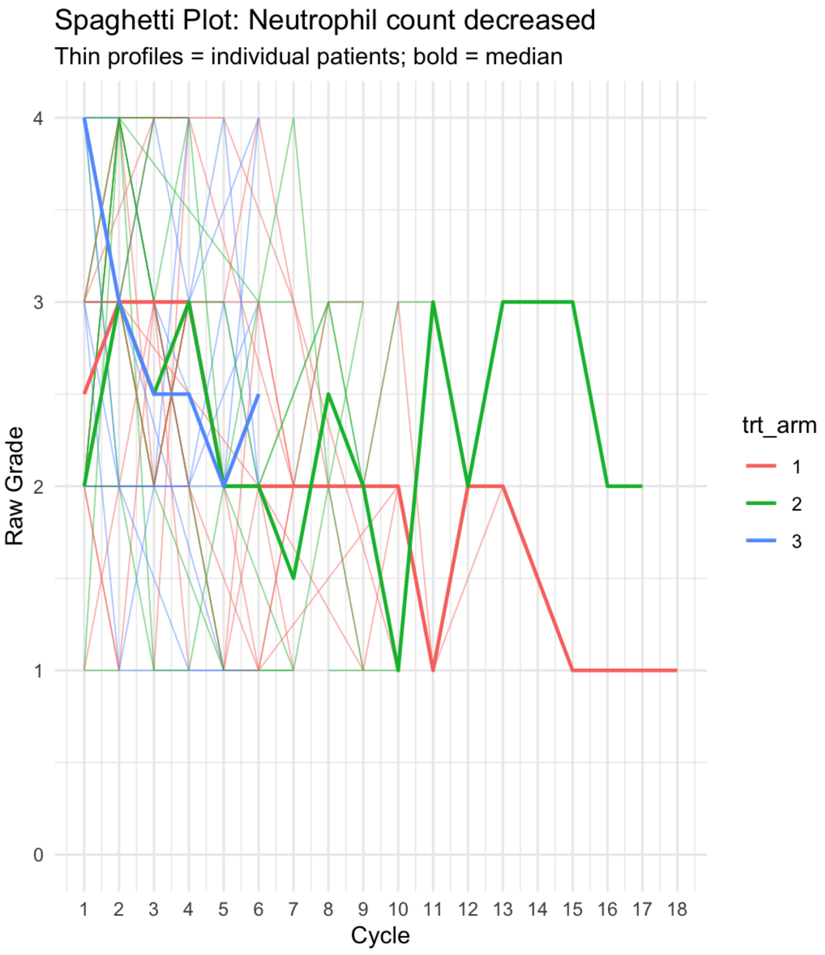


Figure 20: Raw grade trajectory heatmap for febrile neutropenia in EORTC 30974 dataset.



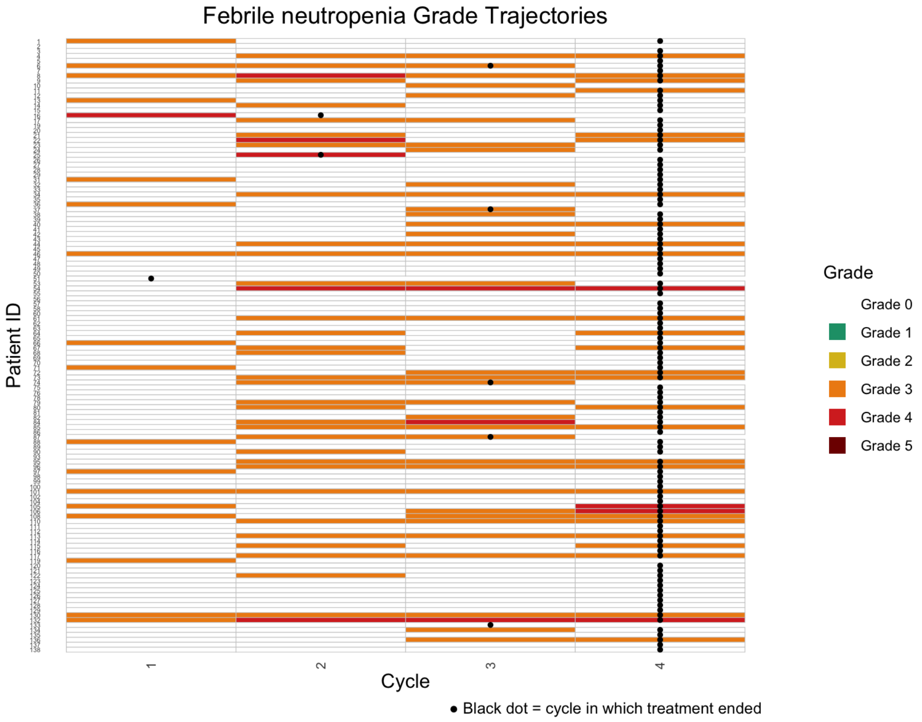
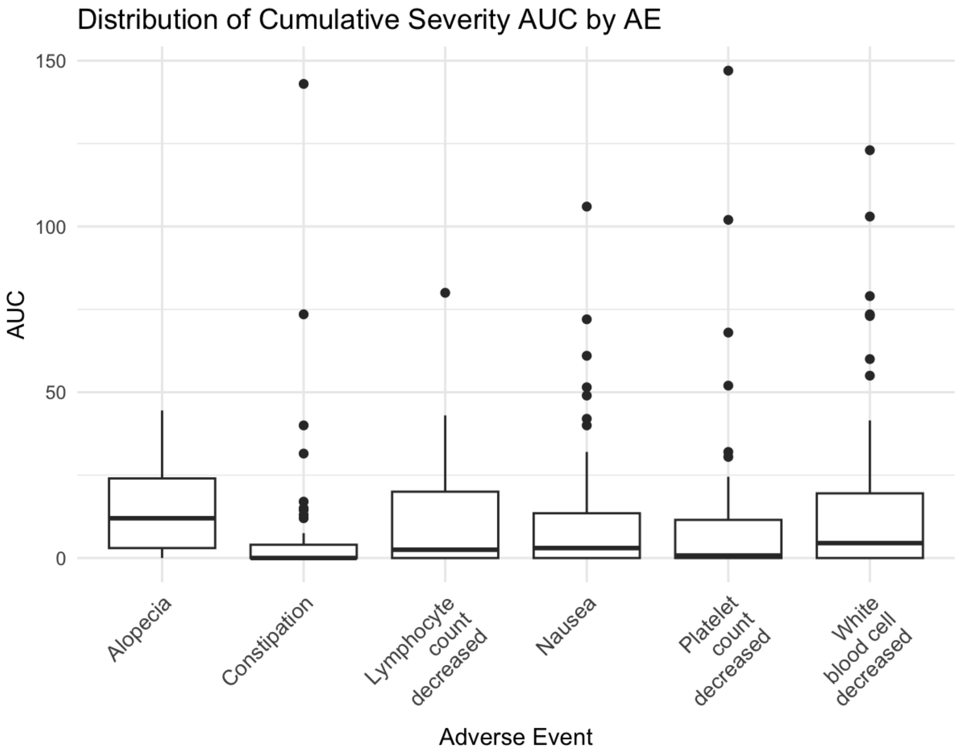
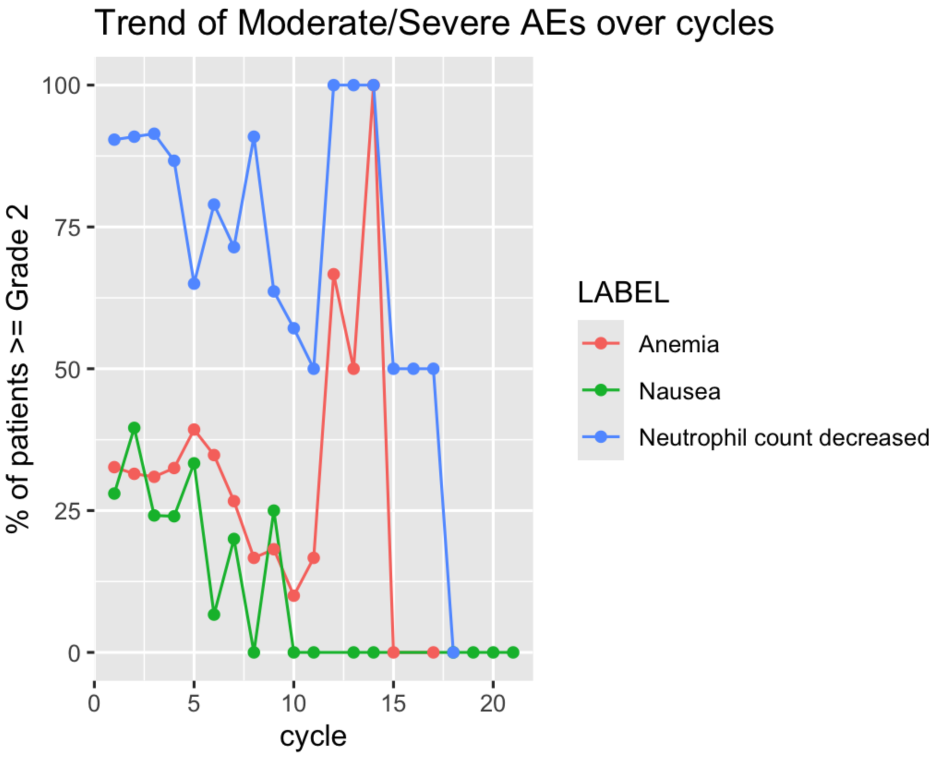
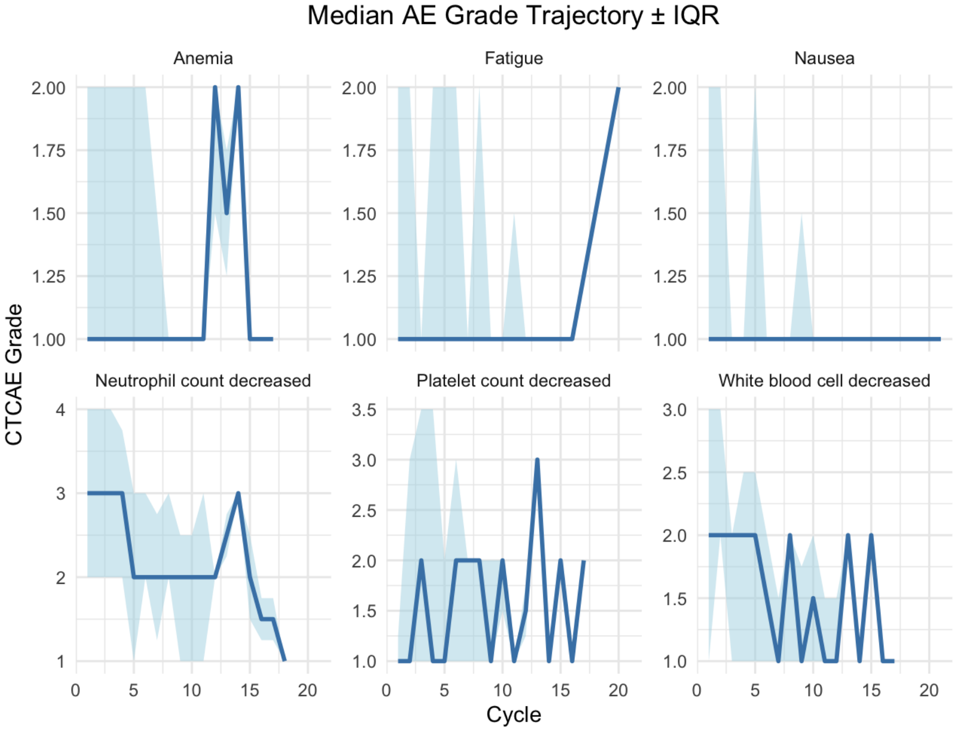


Figure 21: Raw grade trajectory heatmap for febrile neutropenia in EORTC 30974 dataset.




Trend of Moderate/Severe AEs over cycles
% of patients >= Grade 2
100
75
50
25
0
0
5
10
15
20
cycle
LABEL
Anemia
Nausea
Neutrophil count decreased


Median AE Grade Trajectory ± IQR
Anemia
Fatigue
Nausea
Neutrophil count decreased
Platelet count decreased
White blood cell decreased
CTCAE Grade
Cycle
2.00
1.75
1.50
1.25
1.00
4
3
2
1
3.5
3.0
2.5
2.0
1.5
1.0
0
5
10
15
20

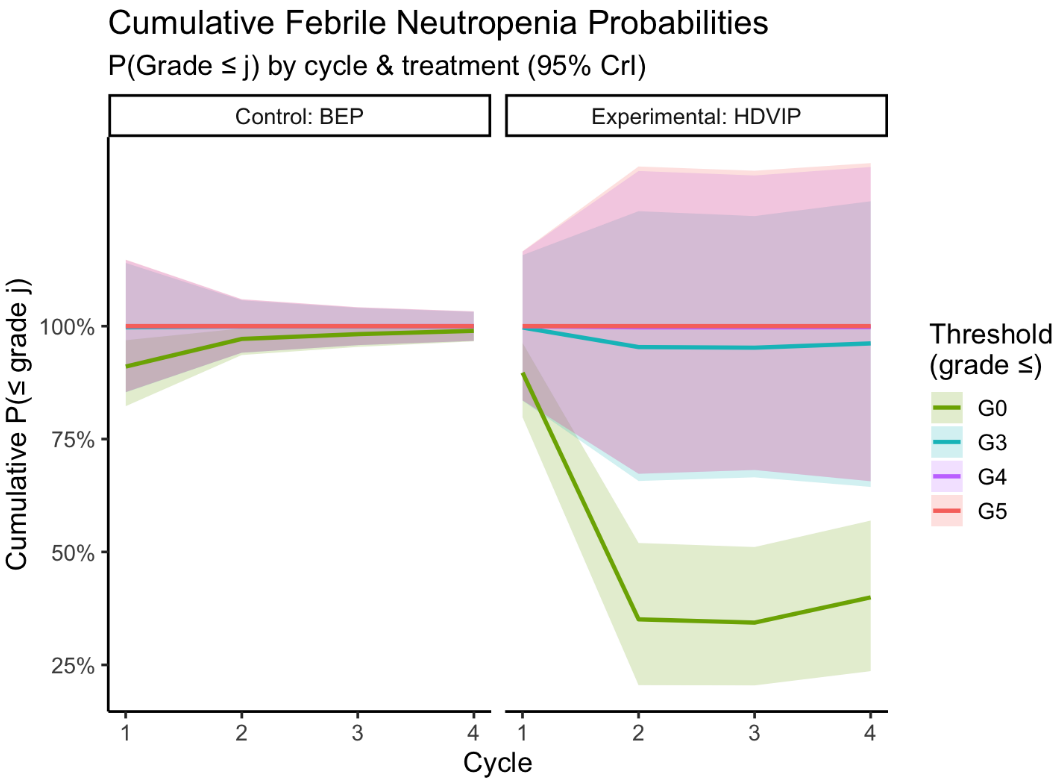


Figure 22: MCMC traceplots for (left) Experimental2 effect, (center) patient-intercept SD, and (right) Exp2 spline-SD. Well-mixed chains and stationary behavior confirm convergence.

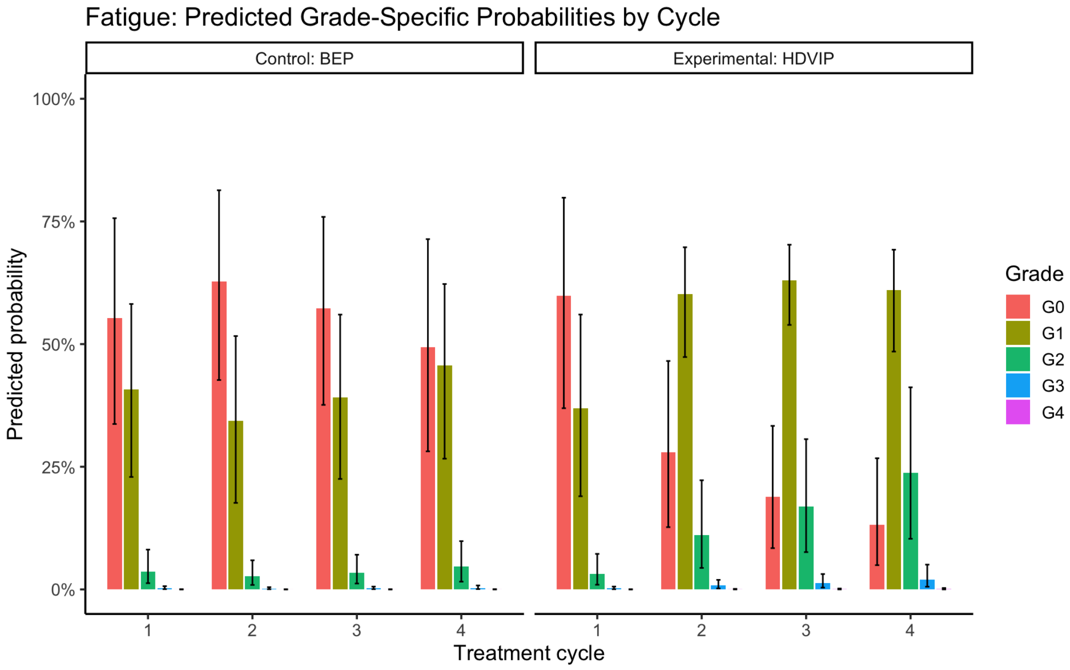


Figure 23: **Posterior-predicted probabilities of fatigue grades over four treatment cycles for the BEP (control) and HDVIP (experimental) arms.** In BEP, the probability of no fatigue (G0, red) remains around 50–60% through cycle 4 while mild fatigue (G1, olive) stays near 35–40%; higher grades (G2–G4) remain rare. In contrast, HDVIP shows a striking shift: G0 falls from ∼60% in cycle 1 to $\tilde{1}$5% by cycle 4, with G1 rising from ∼35% to over 60%; moderate (G2) and severe (G3–G4) fatigue also steadily increase. Whiskers are 95% credible intervals. This pattern indicates accumulating patient fatigue under HDVIP relative to BEP.

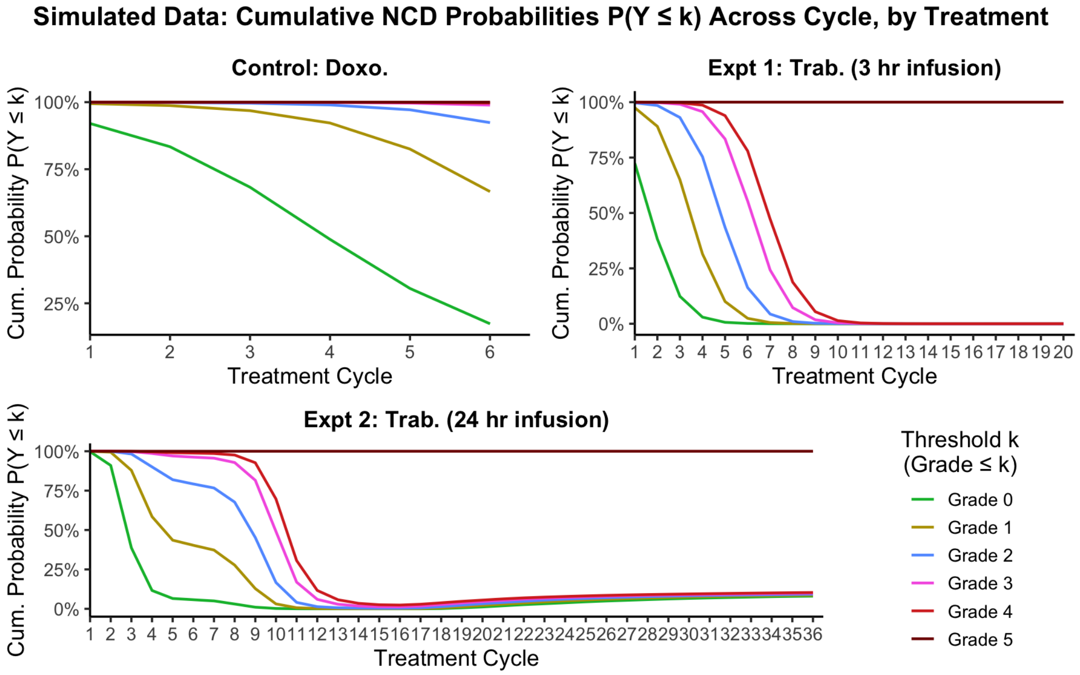


Figure 24

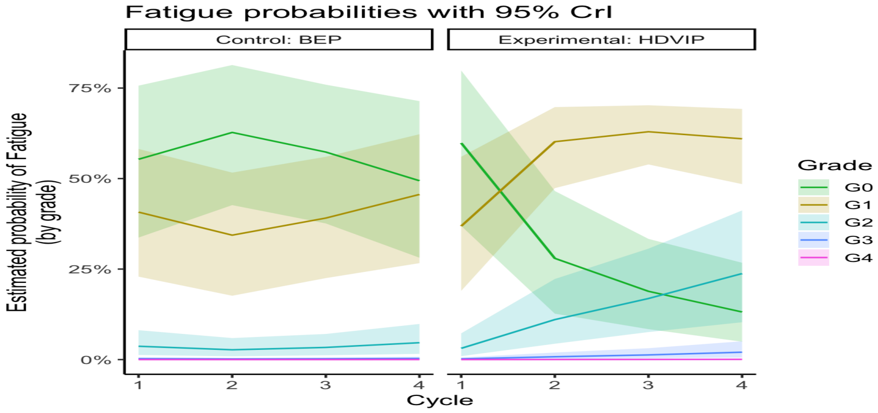


Figure 25: **Estimated probabilities of each fatigue grade over treatment cycles, separately for the Control (BEP) and Experimental (HDVIP) arms.** Solid lines show posterior median probabilities for grades 0–4 at cycles 1–4; shaded bands are 95% credible intervals. The plot illustrates how lower grades dominate early cycles in both arms, while higher-grade fatigue (especially grade 2) increases over time, most notably in the HDVIP arm.

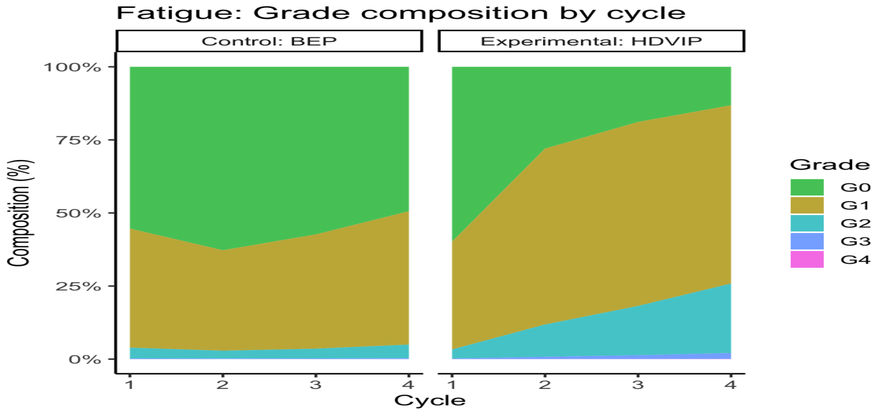


Figure 26: EORTC 30974: Fatigue grade composition (%) by cycle and treatment arm (stacked)