Impact of follow-up on generalized pairwise comparisons for estimating the irinotecan benefit in advanced/metastatic gastric cancer

Ali N. Chamseddine[1, 2], Koji Oba[3], Marc Buyse[4], Narikazu Boku[5], Olivier Bouché[6], Tuvana Satar[7], Anne Auperin[1, 7], Xavier Paoletti[8, 9*]

1. OncoStat CESP, INSERM, Université Paris-Saclay, Équipe Labellisée Ligue Contre le Cancer, Villejuif, France

2. Department of Medical Oncology, Gustave Roussy Cancer Campus, Villejuif,

3. Department of Biostatistics, The University of Tokyo, Tokyo, Japan

4. International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium & Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), University of Hasselt, Hasselt, Belgium.

5. Division of Gastrointestinal Medical Oncology, National Cancer Center Hospital, Tokyo, Japan

6. Fédération Francophone de Cancérologie Digestive (FFCD) & Department of Digestive Oncology, CHU Reims, Reims, France

7. Service de Biostatistique et dEpidémiologie, Gustave Roussy, Université Paris-Saclay, Villejuif, France

8. Université de Versailles-St Quentin / Université Paris Saclay, France

9. Institut Curie & INSERM U900, Biostatistics for Precision Medicine (STAMPM), Saint-Cloud, France

* corresponding author: Xavier Paoletti, Institut Curie, 35 rue Dailly, 92210 St Cloud, France - xavier.paoletti@curie.fr

ABSTRACT (248 words)

Background and Objectives: The net treatment effect (Δ) is a new method to assess the treatment benefit that combines multiple time-to-event, binary and continuous endpoints according to a pre-specified sequence. It represents the net probability for a random patient treated in the experimental arm to have a better overall outcome than a random patient from the control arm does. We aimed at characterizing the impact of follow-up on Δ estimated from both time-to-event and binary toxicity endpoints, in randomized controlled trials (RCTs) of irinotecan-based regimen in advanced/metastatic gastric cancer (AGC).

Study Design: Three RCTs are reanalysed. The net treatment effect using from one to three outcomes (i.e. overall survival, time to progression and toxicity in this order) and the hazard ratio (HR) were estimated after various cut-off dates and compared to the values obtained after complete follow-up were reported.

Results: In all three RCTs (897 patients), the irinotecan-based regimen was superior to the non-irinotecan containing regimen in terms of HR and Δ. This superiority was lower when the net treatment effect also accounted for toxicity. The HR was slightly less influenced by an incomplete follow-up than Δ was, but correction proposed by Péron to account for censored observations showed quite robust results.

Conclusions: The net treatment effect using Péron's correction can be used in case of interim analyses or high censoring rates. In addition to relative measures such as the hazard ratio, it provides a simple mean to evaluate the net treatment effect with and without toxicity outcomes.

Keywords: Generalized pairwise comparisons; Irinotecan; Advanced metastatic gastric cancer; Net treatment effect; Follow-Up

Highlights

- In advanced gastric cancer, we studied the net effect of irinotecan-based regimen that accounts for overall survival, time to progression and toxicity by the means of the generalized pairwise comparisons (GPC) approach.
- The net benefit gives an insight of the relative influence of the various outcomes as hierarchized by the researcher or the patient.
- Incomplete follow-up has slightly stronger influence on the GPC analysis than it has on the hazard ratio.
- In the presence of incomplete follow-up, bias-corrected estimator proposed by Péron et al. provides the most stable results. Net benefit accounting for efficacy endpoints with and without toxicity endpoints should be provided.

1. INTRODUCTION (502 words)

The main aim of anticancer agents is to improve patients' survival while limiting the risk of adverse events. Overall survival (OS) represents the gold standard for the evaluation of new treatments; as OS may be long to observe, progression-free survival (PFS) is increasingly used as an intermediate endpoint. The hazard ratio (HR) quantifies the relative risk of death or progression after receiving a new treatment compared to the standard. Anticancer-related toxicities also critically impact both patients' quality of life and survival and influence the choice of the anticancer agent at the patient level [1], [2]. These various endpoints are reported separately according to a hierarchy established in terms of primary and secondary objectives in the protocol. For instance, clinical trials in advanced/recurrent gastric cancers generally rely on OS first, followed by PFS and toxicity [3]. The benefit-risk balance is then assessed in a qualitative way.

Recently, Buyse proposed a new statistical approach to combine several prioritised endpoints in randomised controlled trials (RCTs) using generalized pairwise comparisons (GPC) [4]. GPC provides an estimate of the net chance of a better overall outcome, also called the "net treatment effect" (denoted Δ) [4]. In brief, in all pairs of patients from the two treatment arms, the two patients are compared based on the first (highest priority) endpoint. A clinically significant better/worse outcome observed for the patient in the investigational arm yields to a favourable/unfavourable pair. If the two outcomes are equal or not different (difference below a pre-defined threshold), the pair is neutral. Finally, pairs that cannot be classified (due to missing or censored observations) are said uninformative; neutral and uninformative pairs are evaluated on the second prioritised endpoint and so forth. The net treatment effect is the probability that the investigational treatment is superior to the control minus the probability of the opposite situation. In contrast to HR, which is a relative measure of the treatment effect, Δ is an absolute difference between two probabilities.

The versatility of GPC to combine endpoints of various types in a simple measure, together with its good properties in case of non-proportional hazards has been demonstrated [4]. Nevertheless, the impact of follow-up on Δ remains a potential issue when the maturity of the data for the various endpoints is different, in particular, in the presence of censored observations. Indeed, shorter follow-up may increase the proportion of uninformative time-to-event pairs (i.e. for OS or time to progression (TTP)). When several prioritized endpoints are considered, lower priority endpoints such as toxicity are favoured over higher priority time-to-event endpoints. This makes it problematic, in a meta-analysis, to combine trials with different follow-up durations to estimate the overall net treatment effect. Péron proposed a corrected-estimator of Δ that accounts for censored observations in order to increase the proportion of informative time-to-event pairs [4], [5].

In this paper, we explore the impact of follow-up on Δ compared to HR. Three RCTs that investigated irinotecan-based regimen against control in advanced/metastatic gastric cancers (AGC) were reanalysed to estimate Δ and HR after multiple follow-up durations.

2. METHODS (923 words)

2.1 Patients

Patients with AGC have poor prognosis. The median OS is below 12 months [6]–[9]. Although irinotecan-based regimens in AGC were not statistically superior to their comparators in terms of OS, they were associated with some efficacy and an acceptable severe toxicity profile [7], [10]–[12]. The rate of severe chemo-induced diarrhea (CID) nevertheless ranges between 10% to 20% [10], [13]–[17]. CID is mainly generated by irinotecan complex activation and subsequent metabolism pathways [13]. Furthermore, irinotecan displays a dose-response relationship for its anti-neoplastic activity and most tumour responses are observed at the highest doses administered [15]. CIDs may lead to early treatment interruption in approximately 40% of patients, which reduces its efficacy and raises the question of the quantification of the net benefit [1], [18]–[20].

Individual patient data (IPD) from three RCTs that evaluated the irinotecan-based regimen in AGC have been provided to the GASTRIC group (JGOG 9205 [21], FFCD 9803 [22] and a Sanofi-sponsored trial [14]. All studies were approved by ethics committee and all patients provided informed consent.

2.2 Endpoints

The efficacy outcomes were OS and TTP defined as the time from the date of randomisation to respectively the date of death, whatever the cause or the date of progression. Patients without event were censored at the cut-off; event indicator took value 1. For the analysis at the final cut-off date (i.e. complete follow-up), time to event endpoints were censored at the last assessment date; for analyses with truncated follow-up, we used the cut-off date as the censoring date. The toxicity outcome was the presence or absence of severe CID, i.e. grade ≥3 according to the National Cancer Institute Common Terminology Criteria, possibly related to the regimen. CID were adverse event related to the treatment administration as assessed by the investigator.

2.3 Measures of treatment effect

In addition to the hazard ratios for OS (HR_{OS}), we estimated the net effect of irinotecan vs. non-irinotecan-based regimens. Up to three outcomes were ranked according to their perceived clinical importance: OS, TTP and toxicity. A pair of two randomly selected patients from each treatment arm

was regarded as:

- Favourable (or unfavourable) in case of:
  - a difference in OS or TTP larger than two months in the irinotecan arm compared to the non-irinotecan arm (or vice-versa). Smaller differences were not considered clinically meaningful [7];
  - the absence of severe CID in the irinotecan arm compared to the presence of severe CID in the non-irinotecan arm (or vice-versa) taken as a binary variable.
- Neutral: if the difference between endpoints was null or not clinically significant.
- Uninformative: if the pair was not assessable because of a missing outcome for at least one of the two patients. With time-to-event data, the uninformative status depends on how the censoring is tackled. We considered two approaches, Gehan and Péron, which are introduced below.

The endpoint with the highest rank was analysed first; uninformative/neutral pairs on the first priority outcome were then evaluated on the second (or even the third) ranked outcome. Three scenarios were defined that account for an increasing number of outcomes: 1 outcome where the single time-to-event outcome OS was used to rank pairs, 2 outcomes, where TTP was also used for pairs that could not be ranked for OS and 3 outcomes where OS, TTP and severe CID were analysed. A value of $\Delta = 0$ indicates the absence of net effect. Positive or negative values reflect the beneficial or detrimental effect of the investigational arm, respectively.

Let us denote T_i and T_j the times to event for patients i (i=1,..., n) from the irinotecan arm and j (j=1,…, m) from the control arm and (δ_i , δ_j) the two indicators of event. A pairwise scoring indicator $S_{ij}(\ l\ )$ is defined for the l^{th} ranked outcome as follows:

$$S_{ij}\ (\ l\ ) = \begin{cases} 1 \text{ if the pair is favorable} \\ -1 \text{ if the pair is unfavorable} \\ 0 \text{ if the pair is neutral} \end{cases}$$

With the Gehan's estimator, $S_{ij}(\ l\ )$ is uninformative if ($T_i \leq T_j$) and (δ_i=0, δ_j=1) or if (δ_i=0, δ_j=0) for the l^{th} outcome and the next outcome in the hierarchy is considered [23]. The pairs that remain uninformative or neutral after the full sequence of outcomes are excluded; this is equivalent to treating the outcomes as completely missing at random [24]. The "net chance of a better outcome" is estimated directly as $\hat{\Delta}_{Gehan} = \frac{\sum_{ij} s_{ij}}{n.m}$.

In the Péron's procedure, the probability for uninformative pairs to be classified as favourable, unfavourable, or neutral is estimated from the Kaplan–Meier distribution at the censored time [5]. Pairs that remain neutral or uninformative after the analysis of the last outcome are excluded. The formulas for the various patterns of censoring are detailed in the original publication.

2.3 Statistical Analyses

To investigate the impact of incomplete follow-up time on both Δ and HR, we reanalysed the three RCTs after increasing the cut-off dates for follow-up, denoted τ. τ ranged from the date of the last inclusion to the follow-up at the final analysis, denoted CFU for complete follow-up time, as performed in each RCT, by three-month increments. Variations of both Δ and HR after follow-up τ relative to values after the complete follow-up were computed as follows:

$$\%\ \mathrm{Var} = \frac{[\text{Treatment effect measure}]\ \text{at CFU} - [\text{Treatment effect measure}]\ \text{at } \tau}{[\text{Treatment effect measure}]\ \text{at CFU}}$$

$$\text{with } 0 \leq \tau < \text{CFU}.$$

Median follow-up was calculated by inverse Kaplan–Meier; HRs were calculated under the Cox proportional hazard model. The proportional hazard assumption was tested by the Therneau test [25]. All analyses were performed in R; the net benefit was calculated with the Buysetest package.

3. RESULTS (790 words)

Patients' characteristics and survival data

Individual data of 897 patients were obtained from three RCTs (labelled in tables and figures as Boku, Bouché and Dank trials) that assessed investigational (containing irinotecan, N=453) vs. control (no irinotecan, N=444) regimens [14], [21], [26]. Patients' baseline characteristics were well balanced between the two groups (Table 1).

Table 1 here

Severe CID were more frequent in patients in the investigational groups (from 8.9% to 22.2%) than in the control groups (from 0.4% to 7.3%). The numbers of progressions were 453 out of 470 patients (Boku trial), 80 out of 90 patients (Bouché trial) and 265 out of 337 patients (Dank trial) after a median follow-up of respectively 46.6, 39.8 and 22.7 months. The numbers of deaths were 444, 80 and 296, respectively. Median OS and median TTP were moderately increased for the irinotecan arms vs. the non-irinotecan arms in all three trials. The test for non-proportional hazard assumption was not significant in both Dank (p=0.43) and Bouché (p=0.24) trials but was significant (p<0.001) for both OS and TTP in the Boku trial.

Impact of the follow-up time

As shown in Table 2, in all three trials, whatever the measure of the treatment effect, the irinotecan-based regimen was superior to the non-irinotecan based regimen. Of note, the net treatment effect, Δ, was expressed as percentages and the confidence intervals are provided as supplementary Table. For instance, in the Bouché trial, $\Delta_{Péron}$ (computed on OS alone) and HR_{OS} estimated after complete follow-up (CFU), were 25% (CI 95%, 13.2–46.1) and 0.71 (CI 95%, 0.46–1.10), respectively. A random patient in the irinotecan treatment group had a 25% higher probability of a longer survival than a patient in the control group and a 29% hazard reduction, only if a single OS endpoint is used to compute the net treatment benefit. The net treatment effect computed on three outcomes (OS, TTP and toxicity) compared to the net effect computed on OS and TTP only were slightly lower in all three trials as irinotecan-based regimen is associated with some severe CID. Conversely, adding TTP in the GPC (two outcomes versus one outcome) led to strong improvement of the net effect. This illustrates the added value of using richer outcomes to assess the benefit and the sensitivity of the net treatment effect to all three components.

Table 2 here

To some extent, Δ and HR depended on the follow-up. Relative variations were more limited for HR than for Δ values. Between the two estimators of the net benefit, the bias corrected estimator by Péron was less influenced by the truncation of follow-up than Gehan estimator was. This better stability is illustrated in Table 3 and Figure 1. In the Dank trial, the net benefit after complete follow-up was close to 0, which magnifies any relative variations. Table 3 gives the relative variation of the treatment effect after increasing follow-up relative to the final analyses. With Gehan estimator, the value of Δ analysed at the time of the last inclusion (Δ_0) was larger than Δ_{CFU}; conversely, fluctuations were more limited with Péron's procedure and Δ_τ remained close to Δ_{CFU} for all τ. As an illustration, in the Boku trial, GPC applied to OS only (scenario 1 outcome) gave Δ_0 = 21% and Δ_{CFU} =15% with Gehan, while it was Δ_0 = 17% and Δ_{CFU} = 15% with Péron estimator. The impact of follow-up on Gehan estimator, compared to the impact on Péron estimator, was more important as the rate of severe CID was larger. As an example, 22.2% of severe CID in the irinotecan arm was reported in the Bouché trial while only 8.9% was reported in the Boku trial, which translated in Δ_τ larger when estimated with the Gehan procedure compared to the estimate with Péron's procedure, for all τ (Table 3).

Table 3 and Figure 1 here

The better robustness of the Péron estimator in the presence of incomplete follow-up is related to the much higher percentages of informative pairs. Table 4 shows the distribution of both informative pairs (favourable or unfavourable) and uninformative/neutral pairs, according to the follow-up time τ. For instance, in the Boku trial, if the GPC analysis of OS only was carried out just after the last inclusion at $\tau = 0$, 50.7% of pairs were uninformative with the Gehan estimator and 13.8% with the Péron estimator. After complete follow-up, 13.9%% and 13.5% remained uninformative with both Gehan and Péron estimator, respectively, which led to very similar estimated net benefits whatever the estimator (Table 4). More variations in the distribution of patterns of pairs according to the follow-up were observed with Gehan estimator than with Péron estimator. This is in line with the documented robustness of the bias-corrected procedure when the rate of censored observations increases [5].

Table 4 here

4. DISCUSSION (858 words)

We investigated the impact of follow-up on the effect of irinotecan-based regimens as measured by Δ compared to HR. Irinotecan-based regimens was superior to the control in the setting of AGC, in terms of OS, with both measures of treatment effects, as suggested in the literature [7], [11]. Whatever the considered sequence of endpoints (using OS alone, OS and TTP, or OS, TTP and CID), a randomly selected patient from the irinotecan group would have a chance of better clinical benefit than a patient from the control group. Interestingly, the net treatment effect reflected the analysed outcomes: it was lower when severe diarrhea was incorporated compared to OS and TTP, which is in line with the known toxicity profile of irinotecan [13]. Additionally, there was a between-trial difference in severe CID rates, which might be related to the patient ethnicity (European vs. Asian), differences in schedule or dose and the disparity in the intestinal bacterial β-glucuronidase activity [13]. The comparison of the net effect with and without including toxicity gives insight on the risk-benefit trade off. The second result is that HR was slightly less influenced by incomplete follow-up than Δ was. Finally, when using GPC with incomplete follow-up, Péron's estimate of Δ provided the most stable results and is thus recommended.

The impact of follow-up is related to the modification of the distribution of the endpoints that serves to declare a pair favourable. Whereas toxicity is supposed to be measured immediately after the treatment initiation and hence takes the same value whatever the follow-up, this is not the case with time-to-event endpoints. If time-to-event outcomes have the highest priorities, shorter follow-up entails more uninformative pairs for the first outcome that get ranked by the toxicity outcome. To

a large extent, Péron's estimate limits this effect as it strongly decreases the proportion of uninformative pairs.

More generally, our study raises the question of the true quantity that we estimate with GPC. The estimands depends on the censoring distribution and on the duration of follow-up. In the case of several trials with different follow-up durations (or different event rates), we may end-up with different estimates even though the relative effect of treatment is similar. Conversely, Δ captures the benefit that is related to the populations' characteristics, which provides a more informative estimate of the treatment effect for a given population. As underlined in the PRISMA recommendation, the relative risk should be complemented with absolute measures [27]. This is especially important with diseases such as AGC, in which a 30% reduction in HR_{PFS} (or HR_{TTP}) yields an absolute gain of no more than one month.

One trial departed from the proportional hazard assumption. In the case of non-proportional hazards, Péron et al. and Parmar et al. have repeatedly demonstrated the superiority of absolute measures over HR [5], [28]. However, as with any measure, insufficient follow-up is a critical issue in the case of non-proportional hazards. With GPC, the score of a pair assessed on a time-to-event may then be favourable at a given time point and unfavourable at another one.

The best choice for priorities is disease specific, or even patient-specific, as proposed by Buyse [4]. In this study, the "net treatment effect" is mainly driven by the death or the progression components, as CID was ranked third. This is in line with the setting of AGC, where treatment goals are to prolong the survival and to optimise the quality of life [29]. A strength of the GPC approach is to integrate the notion of clinically meaningful differences in the assessment of the treatment effect. An additional advantage of combining endpoints in a single measure of treatment effect is to limit the number of tests. This can be seen as desirable in an inference framework.

In our three databases, only 39 patients (4.3%) died without documented progression; we could have then considered the PFS outcome rather than the time to progression, as death from another cause as progression was unlikely; this was the approach used by Péron and Buyse in their analysis of the benefit-risk balance of Nab-paclitaxel in metastatic pancreatic cancer [30]. More generally, correlation between outcomes may affect the estimate of the net treatment effect, but without adding any biases [31]. This is rather a concern for the computation of the sample size in trials that would use the GPC for the analysis of the primary objective. So far, the net treatment effect has only been applied in retrospective studies to better quantify the benefit-ratio of a new treatment [32]. Likewise, to our knowledge, the win-ratio a closely related measure has been used only for retrospective analyses [33].

Irinotecan-based regimen has a positive benefit-risk ratio using GPC in the AGC setting, despite their salient toxicity profile. Even though the HR was the most stable of the compared measure, the benefit was robust with incomplete follow-ups, which opens the door to its implementation in meta-

analyses of trials with various data maturity. Δ is a promising measure of treatment effect at the time of the final analysis. If Δ is selected, the estimator proposed by Péron and colleagues is recommended; net effect with and without toxicity outcomes should be provided.

ACKNOWLEDGEMENTS: We are strongly indebted to the Fédération Francophone de Cancérologie Digestive (FFCD) Group, the Gastrointestinal Oncology Study Group of the Japan Clinical Oncology Group and Sanofi for providing individual patients data to the GASTRIC collaboration. The trials' sponsors were not involved in the design, the analysis and interpretation of this contribution. We thank the GASTRIC collaboration for giving access to their data collection.

ROLE OF THE FUNDING SOURCE: This work was partially funded by a French governmental grant from the *Programme Hospitalier pour la Recherche Clinique* (PHRC). The funder had no involvement neither in the data collection, design, analysis nor interpretation of the data.

COMPETING INTERESTS STATEMENT: All authors declared they had no competing interest in the conduct of this methodological research.

Table 1. List of included trials and patients' characteristics.

	Boku trial		***Bouché trial***		***Dank trial***	
	Non-irinotecan	**Irinotecan**	**Non-irinotecan**	**Irinotecan**	**Non-irinotecan**	**Irinotecan**
Number of patients	234	236	45	45	165	172
Male (%)	356 (76%)		75 (83%)		236 (70%)	
Median age (range), years	63 (24–75)		64 (37–75)		58 (28–77)	
Histology						
Adenocarcinoma diffuse type	255 (54%)		0		106 (31%)	
Adenocarcinoma intestinal type	213 (45%)		84 (93%)		93 (28%)	
Adenocarcinoma mixed type	0		1 (1%)		7 (2%)	
Unknown	2 (0.4%)		5 (6%)		131 (39%)	
Type of chemotherapy regimen	**5FU**: 5-FU 800 mg/m2 i.v., daily for 5 days. Cycles were repeated every four weeks.	**IC:** irinotecan 70 mg/m2 i.v., on days 1 and 15; and cisplatin 80 mg/m2 i.v., day 1. Cycles were repeated every four weeks. After six cycles, the same dose of irinotecan alone was continued every two weeks.	**FL:** leucovorin 200 mg/m2 i.v.; and bolus 5-FU 400 mg/m2 i.v.; and 5-FU 600 mg/m2 i.v. on days 1 and 2. Cycles (15 days) were repeated every two weeks.	**IFL:** irinotecan 180 mg/m2 i.v.; leucovorin 200 mg/m2 i.v.; and bolus 5-FU 400 mg/m2 i.v.; and 5-FU 600 mg/m2 i.v. on days 1 and 2. Cycles (15 days) were repeated every two weeks.	**CF**: cisplatin 100 mg/m2 i.v. day 1; and 5-FU 1000 mg/m2 i.v., days 1 through 5. The cycles were repeated every four weeks.	**IFL:** irinotecan 80 mg/m2 i.v.; and leucovorin 500 mg/m2 i.v. and 5-FU 2000 mg/m2 i.v., on days 1, 8, 15, 22, 29 and 36. Cycles were repeated every seven weeks.
Median Follow-up, months (95%CI)	46.6 (36.4–71.1)	39.8 (35.2–51.4)	22.7 (15.4–33.6)	NR (17.3–NR)	25.7 (24.8–32.3)	25.3 (22.3–29.5)
Median OS, months (95%CI)	10.70 (8.70–12.00)	12.20 (11.20–14.20)	6.85 (4.45–10.4)	11.21 (8.80–14.2)	8.50 (7.50–9.50)	9.00 (8.10–9.70)
Median TTP, months (95%CI)	2.90 (2.10–3.70)	4.85 (4.20– 5.60)	3.85 (2.50–5.25)	7.30 (5.50–8.00)	4.00 (3.70–5.10)	410 (3.70–5.15)
Rate of severe CID (%) (Grade ≥3)	0.4	8.9	2.2	22.2	7.3	20.9

i.v., intra-venous; *5FU*, 5 Fluorouracil; *IC*, irinotecan and cisplatin; *CF*, Cisplatin and fluorouracil; *FL*, fluorouracil and leucovorin; *IFL*: irinotecan, leucovorin and fluorouracil; *NR*, not reached; *OS*, Overall survival; *TTP*, Time to Progression; *CID*, chemo-induced diarrhea

Table 2. Net treatment effect estimated with Gehan (Δ_{Gehan}) and Péron ($\Delta_{Péron}$) generalized pairwise procedures, informative pairs (%) with each generalized pairwise procedures and hazard ratio for overall survival (HR_{OS}), after increasing cut-off times τ (at 0, 3, 6, 9 months after the last included patient and complete follow-up time noted CFU) for three sets of outcomes.

		Δ_{Gehan} (%)			$\Delta_{Péron}$ (%)			Informative pairs (%) with Δ_{Gehan}			Informative pairs (%) with $\Delta_{Péron}$			HR		
Scenarios	**τ (months)**	**Boku**	**Bouché**	**Dank**	**Boku**	**Bouché**	**Dank**	**Boku**	**Bouché**	**Dank**	**Boku**	**Bouché**	**Dank**	**Boku**	**Bouché**	**Dank**
1 outcome (OS)	**0**	21	32	5	17	28	10	49.27	45.18	33.53	86.22	85.76	78.49	0.80	0.63	0.84
	3	20	29	4	17	25	5	58.97	60.15	48.86	86.66	84.23	78.49	0.78	0.66	0.93
	6	18	28	6	14	25	6	68.26	69.14	62.45	86.84	84.02	78.49	0.83	0.67	0.89
	9	16	27	4	13	27	3	75.22	76.84	70.93	86.9	84.73	78.49	0.85	0.67	0.96
	CFU	15	26	4	15	25	4	86.12	80.25	78.08	86.51	69.30	78.49	0.83	0.71	0.93
2 outcomes (OS - TTP)	**0**	29	41	7	19	32	11	69.62	59.26	49.95	93.87	92.56	91.15			
	3	27	36	9	19	29	5	81.01	75.75	67.71	93.77	91.36	91.00			
	6	24	36	9	15	29	7	87.49	84.84	80.92	93.63	90.63	90.1			
	9	21	33	6	15	30	3	90.45	88.1	85.55	93.48	90.45	90.26			
	CFU	16	30	5	17	28	4	93.23	89.04	88.30	93.30	90.61	90.24			
3 outcomes (OS - TTP - Toxicity)	**0**	27	31	1	19	31	9	72.15	69.58	62.06	94.52	93.82	93.65			
	3	26	32	4	18	27	4	82.47	81.29	75.95	94.51	92.78	93.53			
	6	23	33	6	15	27	5	88.58	87.76	85.77	94.4	92.16	92.89			
	9	20	31	4	14	28	2	91.38	90.18	88.59	94.25	92.09	92.97			
	CFU	16	28	3	16	27	2	94.02	90.82	91.42	94.09	92.16	92.98			

OS, Overall survival; *TTP*, Time to Progression; *Toxicity corresponds to grade ≥3* chemo-induced diarrhea

Supplementary Table: Estimates and 95% confidence intervals (95%CI) of the net treatment effect after complete follow-up, noted Δ. *OS*, Overall survival; *TTP*, Time to Progression; *Toxicity corresponds to grade ≥3* chemo-induced diarrhea

		Δ_{Gehan} (%); (95%CI)			$\Delta_{Péron}$ (%); (95%CI		
Scenarios	**τ (months)**	**Boku**	**Bouché**	**Dank**	**Boku**	**Bouché**	**Dank**
1 outcome (OS)	**CFU**	15; (7, 23)	25; (13, 44)	4; (-5, 11)	15; (7, 23)	25; (13, 45)	4; (-5, 11)
2 outcomes (OS - TTP)	**CFU**	16; (9, 25]	30; (12, 50)	5; (-3, 13)	17; (9, 25)	28; (12, 50)	4; (-4, 12)
3 outcomes (OS - TTP - Toxicity)	**CFU**	16; (8, 24)	28; (14, 50)	3; (-5, 12)	16; (8, 25)	27; (13, 48)	2; (-6, 11)

Table 3. Variation of the net treatment effect Δ (%var) and of hazard ration (HR) after increasing follow-up compared to the value estimated at the end of the trial for three RCTs. The net treatment effect, Δ, is calculated with two different estimators Δ_{Gehan} and $\Delta_{Péron}$ for the three sets of outcomes. Negative value indicates that Δ or HR were lower at complete follow-up than at cut-off τ follow-up after the last included patient**.** *OS*, Overall survival; *TTP*, Time to Progression; *Toxicity corresponds to grade ≥3* chemo-induced diarrhea

		Boku		**Bouché**		**Dank**		**Boku**	**Bouché**	**Dank**
		Δ_{Gehan}	$\Delta_{Péron}$	Δ_{Gehan}	$\Delta_{Péron}$	Δ_{Gehan}	$\Delta_{Péron}$	**HR**	**HR**	**HR**
Scenarios	**τ (months)**	**%Var**		**%Var**		**%Var**		**%Var**	**%Var**	**%Var**
1 outcome (OS)	**0**	-40.00	-13.33	-23.08	-12.00	-25.00	-150.00	3.61	11.27	9.68
	3	-33.33	-13.33	-11.54	0.00	0.00	-25.00	6.02	7.04	0.00
	6	-20.00	6.67	-7.69	0.00	-50.00	-50.00	0.00	5.63	4.30
	9	-6.67	13.33	-3.85	-8.00	0.00	25.00	-2.41	5.63	-3.23
2 outcomes (OS - TTP)	**0**	-81.25	-11.76	-37.00	-14.28	-40.00	-175.00			
	3	-68.75	-11.76	-20.00	-3.50	-40.00	-25.00			
	6	-50.00	11.76	-20.00	-3.50	-80.00	-75.00			
	9	-31.25	11.76	-10.00	-7.14	-20.00	25.00			
3 outcomes (OS - TTP - Toxicity)	**0**	-68.75	-18.75	-10.71	-14.80	-67.00	-350.00			
	3	-62.50	-12.50	-14.28	0.00	-33.00	-100.00			
	6	-43.75	6.25	-17.86	0.00	-100.00	-150.00			
	9	-25.00	12.50	-10.71	-3.70	-33.00	0.00			

Table 4. Distribution of the type of pairs in the calculation of the net treatment effect according to Gehan or Péron procedures for the three sets of outcomes (OS, OS-TTP and OS-TTP - Toxicity). τ denotes the cut-off follow-up after the last included patient; *OS*, Overall Survival; *TTP*, Time to Progression;

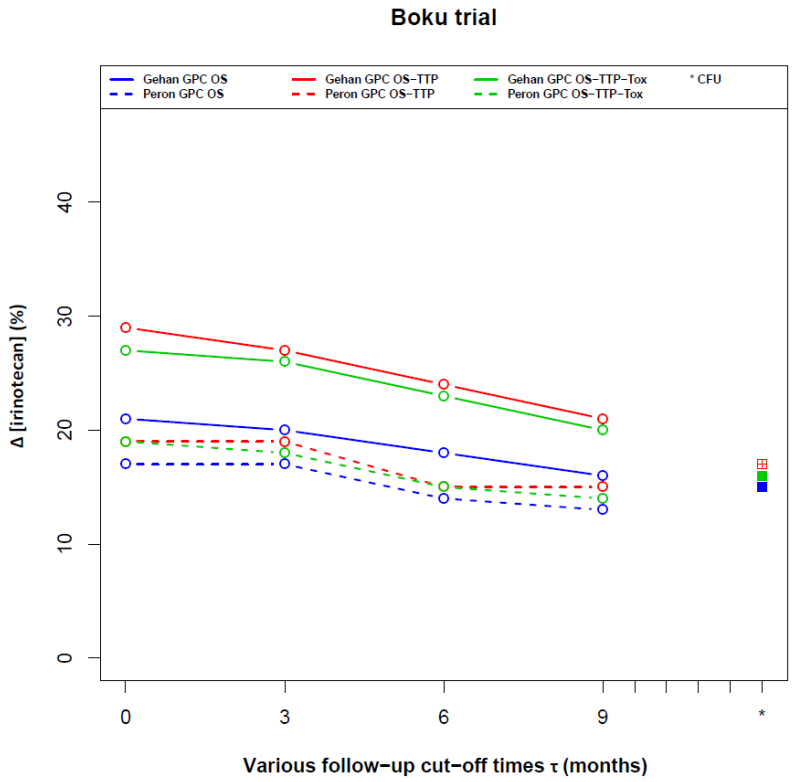
		1 outcome (OS)														
		Favourable pairs (%)					Unfavourable pairs (%)					Neutral and Uninformative pairs (%)				
	τ (months)	0	3	6	9	CFU	0	3	6	9	CFU	0	3	6	9	CFU
Boku (N = 470)	Gehan	30.59	36.39	41.12	44.60	50.29	18.68	22.58	27.14	30.62	35.83	50.73	41.03	31.74	24.78	13.88
	Péron	50.99	51.26	49.99	49.89	50.55	35.23	35.40	36.85	37.01	35.96	13.78	13.34	13.16	13.10	13.49
Bouché (N = 136)	Gehan	30.81	40.05	45.63	50.72	52.35	14.37	20.10	23.51	26.12	27.90	54.82	39.85	30.86	23.16	19.75
	Péron	55.73	53.61	53.78	55.56	55.06	30.03	30.62	30.24	29.17	14.24	14.24	15.77	15.98	15.27	14.76
Dank (N = 337)	Gehan	17.84	25.46	33.46	37.10	40.77	15.69	23.40	28.99	33.83	37.31	66.47	51.14	37.55	29.07	21.92
	Péron	43.49	42.58	42.98	42.17	42.71	35.00	38.32	37.12	39.28	39.25	21.51	19.10	19.90	18.55	18.04

		2 outcomes (OS - TTP)														
		Favourable pairs (%)					Unfavourable pairs (%)					Neutral and Uninformative pairs (%)				
	τ (months)	0	3	6	9	CFU	0	3	6	9	CFU	0	3	6	9	CFU
Boku (N = 470)	Gehan	43.97	50.56	53.52	54.50	54.81	25.65	30.45	33.97	35.95	38.42	30.38	18.99	12.51	9.55	6.77
	Péron	55.99	55.78	54.30	54.06	54.90	37.88	37.99	39.33	39.42	38.40	6.13	6.23	6.37	6.52	6.70
Bouché (N = 136)	Gehan	40.79	51.11	57.43	59.07	58.52	18.47	24.64	27.41	29.03	30.52	40.74	24.25	15.16	11.90	10.96
	Péron	60.44	58.72	58.64	59.42	58.65	32.12	32.64	31.99	31.03	31.96	7.44	8.64	9.37	9.55	9.39
Dank (N = 337)	Gehan	26.55	36.89	44.04	45.63	46.63	23.40	30.82	36.88	39.92	41.67	50.05	32.29	19.08	14.45	10.70
	Péron	50.51	48.33	48.50	46.98	47.15	40.64	42.67	41.60	43.28	43.09	8.85	9.00	9.90	9.74	9.76

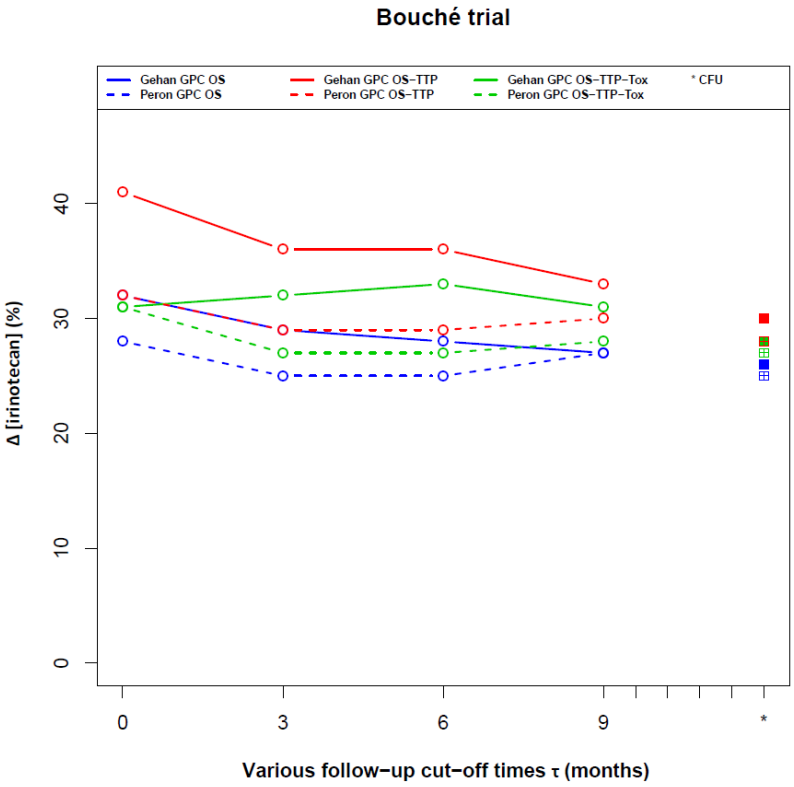
		3 outcomes (OS - TTP - Toxicity)														
		Favourable pairs (%)					Unfavourable pairs (%)					Neutral and Uninformative pairs (%)				
	τ (months)	0	3	6	9	CFU	0	3	6	9	CFU	0	3	6	9	CFU
Boku (N = 470)	Gehan	44.07	50.63	53.57	54.54	54.85	28.08	31.84	35.01	36.84	39.17	27.85	17.53	11.42	8.62	5.98
	Péron	56.02	55.81	54.34	54.10	54.94	38.50	38.70	40.06	40.15	39.15	5.48	5.49	5.60	5.75	5.91
Bouché (N = 136)	Gehan	41.28	51.51	57.68	59.32	58.77	28.30	29.78	30.08	30.86	32.05	30.42	18.71	12.24	9.82	9.18
	Péron	60.60	58.92	58.84	59.62	58.85	33.22	33.86	33.32	32.47	33.31	6.18	7.22	7.84	7.91	7.84
Dank (N = 337)	Gehan	29.40	38.61	45.03	45.63	47.25	32.66	37.34	40.74	42.96	44.17	37.94	24.05	14.23	10.61	8.58
	Péron	51.05	48.86	49.07	47.56	47.71	42.60	44.67	43.82	45.41	45.27	6.35	6.47	7.11	7.03	7.02

Figure 1. Net treatment effect under irinotecan-based regimen (Δ [irinotecan]) estimated with Gehan and Péron approaches after increasing follow-up for three sets of outcomes for the Boku, Bouché and Dank trials (panels A, B and C respectively).

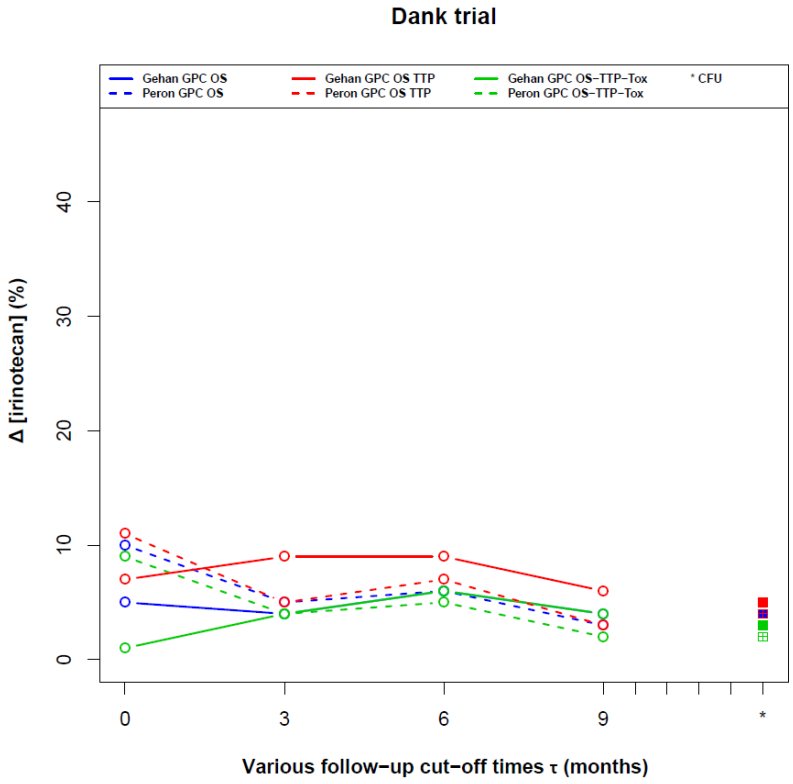
Panel A



Panel B



Panel C



CFU: Complete follow-up time

References

[1] T. Delaunoit *et al.*, "Mortality associated with daily bolus 5-fluorouracil/leucovorin administered in combination with either irinotecan or oxaliplatin: results from Intergroup Trial N9741," *Cancer*, vol. 101, no. 10, pp. 2170–2176, Nov. 2004, doi: 10.1002/cncr.20594.

[2] T. Tachi *et al.*, "The Impact of Outpatient Chemotherapy-Related Adverse Events on the Quality of Life of Breast Cancer Patients," *PLOS ONE*, vol. 10, no. 4, p. e0124169, Apr. 2015, doi: 10.1371/journal.pone.0124169.

[3] GASTRIC (Global Advanced/Adjuvant Stomach Tumor Research International Collaboration) Group *et al.*, "Role of chemotherapy for advanced/recurrent gastric cancer: an individual-patient-data meta-analysis," *Eur. J. Cancer*, vol. 49, no. 7, pp. 1565–1577, May 2013, doi: 10.1016/j.ejca.2012.12.016.

[4] M. Buyse, "Generalized pairwise comparisons of prioritized outcomes in the two-sample problem," *Stat Med*, vol. 29, no. 30, pp. 3245–3257, Dec. 2010, doi: 10.1002/sim.3923.

[5] J. Péron, M. Buyse, B. Ozenne, L. Roche, and P. Roy, "An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring," *Stat Methods Med Res*, vol. 27, no. 4, pp. 1230–1239, Apr. 2018, doi: 10.1177/0962280216658320.

[6] K. D. Miller *et al.*, "Cancer treatment and survivorship statistics, 2016," *CA Cancer J Clin*, vol. 66, no. 4, pp. 271–289, 2016, doi: 10.3322/caac.21349.

[7] A. D. Wagner *et al.*, "Chemotherapy for advanced gastric cancer," *Cochrane Database Syst Rev*, vol. 8, p. CD004064, 29 2017, doi: 10.1002/14651858.CD004064.pub4.

[8] H. Yamamoto, "An updated review of gastric cancer in the next-generation sequencing era: Insights from bench to bedside and *vice versa*," *World Journal of Gastroenterology*, vol. 20, no. 14, p. 3927, 2014, doi: 10.3748/wjg.v20.i14.3927.

[9] L. de Mestier, S. Lardière-Deguelte, J. Volet, R. Kianmanesh, and O. Bouché, "Recent insights in the therapeutic management of patients with gastric cancer," *Dig Liver Dis*, vol. 48, no. 9, pp. 984–994, Sep. 2016, doi: 10.1016/j.dld.2016.04.010.

[10] R. Guimbaud *et al.*, "Prospective, randomized, multicenter, phase III study of fluorouracil, leucovorin, and irinotecan versus epirubicin, cisplatin, and capecitabine in advanced gastric adenocarcinoma: a French intergroup (Fédération Francophone de Cancérologie Digestive, Fédération Nationale des Centres de Lutte Contre le Cancer, and Groupe Coopérateur Multidisciplinaire en Oncologie) study," *J. Clin. Oncol.*, vol. 32, no. 31, pp. 3520–3526, Nov. 2014, doi: 10.1200/JCO.2013.54.1011.

[11] E. C. Smyth, M. Verheij, W. Allum, D. Cunningham, A. Cervantes, and D. Arnold, "Gastric cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†," *Annals of Oncology*, vol. 27, no. suppl_5, pp. v38–v49, Sep. 2016, doi: 10.1093/annonc/mdw350.

[12] A. Zaanan *et al.*, "Gastric cancer: French intergroup clinical practice guidelines for diagnosis, treatments and follow-up (SNFGE, FFCD, GERCOR, UNICANCER, SFCD, SFED, SFRO)," *Dig Liver Dis*, vol. 50, no. 8, pp. 768–779, Aug. 2018, doi: 10.1016/j.dld.2018.04.025.

[13] A. N. Chamseddine *et al.*, "Intestinal bacterial β-glucuronidase as a possible predictive biomarker of irinotecan-induced diarrhea severity," *Pharmacol. Ther.*, Mar. 2019, doi: 10.1016/j.pharmthera.2019.03.002.

[14] M. Dank *et al.*, "Randomized phase III study comparing irinotecan combined with 5-fluorouracil and folinic acid to cisplatin combined with 5-fluorouracil in chemotherapy naive patients with advanced adenocarcinoma of the stomach or esophagogastric junction," *Ann. Oncol.*, vol. 19, no. 8, pp. 1450–1457, Aug. 2008, doi: 10.1093/annonc/mdn166.

[15] Y. Merrouche *et al.*, "High dose-intensity of irinotecan administered every 3 weeks in advanced cancer patients: a feasibility study," *J. Clin. Oncol.*, vol. 15, no. 3, pp. 1080–1086, Mar. 1997, doi: 10.1200/JCO.1997.15.3.1080.

[16] M. Moehler *et al.*, "Randomised phase II evaluation of irinotecan plus high-dose 5-fluorouracil and leucovorin (ILF) vs 5-fluorouracil, leucovorin, and etoposide (ELF) in untreated metastatic gastric cancer," *Br. J. Cancer*, vol. 92, no. 12, pp. 2122–2128, Jun. 2005, doi: 10.1038/sj.bjc.6602649.

[17] T. Shiozawa *et al.*, "Risk Factors for Severe Adverse Effects and Treatment-related Deaths in Japanese Patients Treated with Irinotecan-based Chemotherapy: A Postmarketing Survey†," *Japanese Journal of Clinical Oncology*, vol. 43, no. 5, pp. 483–491, May 2013, doi: 10.1093/jjco/hyt040.

[18] H. C. Pitot *et al.*, "Phase II trial of irinotecan in patients with metastatic colorectal carcinoma," *J. Clin. Oncol.*, vol. 15, no. 8, pp. 2910–2919, Aug. 1997, doi: 10.1200/JCO.1997.15.8.2910.

[19] D. Abigerges *et al.*, “Irinotecan (CPT-11) high-dose escalation using intensive high-dose loperamide to control diarrhea,” *J. Natl. Cancer Inst.*, vol. 86, no. 6, pp. 446–449, Mar. 1994.
[20] M. Ducreux, C.-H. Köhne, G. K. Schwartz, and U. Vanhoefer, “Irinotecan in metastatic colorectal cancer: dose intensification and combination with new agents, including biological response modifiers,” *Ann. Oncol.*, vol. 14 Suppl 2, pp. ii17-23, 2003.
[21] N. Boku *et al.*, “Fluorouracil versus combination of irinotecan plus cisplatin versus S-1 in metastatic gastric cancer: a randomised phase 3 study,” *Lancet Oncol.*, vol. 10, no. 11, pp. 1063–1069, Nov. 2009, doi: 10.1016/S1470-2045(09)70259-1.
[22] O. Bouché *et al.*, “Randomized multicenter phase II trial of a biweekly regimen of fluorouracil and leucovorin (LV5FU2), LV5FU2 plus cisplatin, or LV5FU2 plus irinotecan in patients with previously untreated metastatic gastric cancer: a Federation Francophone de Cancerologie Digestive Group Study--FFCD 9803,” *J. Clin. Oncol.*, vol. 22, no. 21, pp. 4319–4328, Nov. 2004, doi: 10.1200/JCO.2004.01.140.
[23] E. Gehan, “A generalized two-sample Wilcoxon test for doubly censored data.,” *Biometrika*, vol. 52, no. 3, pp. 650–653, 1965.
[24] R. Latta, “Generalized Wilcoxon statistics for the two sample problem with censored data.,” *Biometrika*, vol. 63, pp. 663–635, 1977.
[25] P. M. Grambsch and T. M. Therneau, “Proportional hazards tests and diagnostics based on weighted residuals,” *Biometrika*, vol. 81, no. 3, pp. 515–526, 1994, doi: 10.1093/biomet/81.3.515.
[26] O. Bouché *et al.*, “Randomized multicenter phase II trial of a biweekly regimen of fluorouracil and leucovorin (LV5FU2), LV5FU2 plus cisplatin, or LV5FU2 plus irinotecan in patients with previously untreated metastatic gastric cancer: a Federation Francophone de Cancerologie Digestive Group Study--FFCD 9803,” *J. Clin. Oncol.*, vol. 22, no. 21, pp. 4319–4328, Nov. 2004, doi: 10.1200/JCO.2004.01.140.
[27] L. A. Stewart *et al.*, “Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement,” *JAMA*, vol. 313, no. 16, pp. 1657–1665, Apr. 2015, doi: 10.1001/jama.2015.3656.
[28] P. Royston and M. K. B. Parmar, “The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt,” *Stat Med*, vol. 30, no. 19, pp. 2409–2421, Aug. 2011, doi: 10.1002/sim.4274.
[29] E. D. Saad, J. R. Zalcberg, J. Péron, E. Coart, T. Burzykowski, and M. Buyse, “Understanding and Communicating Measures of Treatment Effect on Survival: Can We Do Better?,” *J. Natl. Cancer Inst.*, vol. 110, no. 3, pp. 232–240, 01 2018, doi: 10.1093/jnci/djx179.
[30] J. Péron, J. Giai, D. Maucort-Boulch, and M. Buyse, “The Benefit-Risk Balance of Nab-Paclitaxel in Metastatic Pancreatic Adenocarcinoma,” *Pancreas*, vol. 48, no. 2, pp. 275–280, Feb. 2019, doi: 10.1097/MPA.0000000000001234.
[31] J. Giai, D. Maucort-Boulch, B. Ozenne, J.-C. Chiêm, M. Buyse, and J. Péron, “Net benefit in the presence of correlated prioritized outcomes using generalized pairwise comparisons: A simulation study,” *Stat Med*, vol. 40, no. 3, pp. 553–565, Feb. 2021, doi: 10.1002/sim.8788.
[32] S. Piantadosi and C. L. Meinert, Eds., “Generalized pairwise comparisons for prioritized outcomes,” in *Principles and Practice of Clinical Trials*, Cham: Springer International Publishing, 2021.
[33] J. P. Ferreira *et al.*, “Use of the Win Ratio in Cardiovascular Trials,” *JACC Heart Fail*, vol. 8, no. 6, pp. 441–450, Jun. 2020, doi: 10.1016/j.jchf.2020.02.010.