UNIVERSITE PARIS-SUD 11
FACULTE DE MEDECINE

Année 2013

N° attribué par la bibliothèque

THESE DE DOCTORAT

Présentée par : **Monia EZZALFANI GAHLOUZI**

Soutenue le : **2 Octobre 2013**

Pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITE PARIS-SUD 11

Spécialité : Santé Publique
Option : Biostatistique

Ecole Doctorale de rattachement : **Ecole Doctorale 420**

Développement d'une méthode de recherche de dose modélisant un score de toxicité pour les essais cliniques de phase I en oncologie

Directeur de thèse : Marie-Cécile LE DELEY
Co-directeur : Sarah ZOHAR

JURY :

M. Jacques BENICHOU	Président
M. Raphaël PORCHER	Rapporteur
M. Gérard PONS	Rapporteur
M. Xavier PAOLETTI	Examinateur
Mme Marie-Cécile LE DELEY	Directeur de thèse
Mme Sarah ZOHAR	Co-directeur de thèse

Avant-propos

Cette thèse a été réalisée au sein du Service de Biostatistique et d'Epidémiologie (SBE), dirigé par le Dr Ellen Benhamou, à l'Institut Gustave-Roussy.
39 rue Camille Desmoulins, 94800 Villejuif, France.

Une collaboration avec le Dr Sumithra Mandrekar et le Dr Rui Qin du centre de recherche de Mayo Clinic (Rochester, USA) a été développée pendant ma deuxième année de thèse. Un séjour d'un mois a été effectué pendant la deuxième année de thèse.

Cette thèse a été financée par une allocation de recherche de l'Université Paris-Sud 11 pendant trois ans.

Une activité de doctorant-conseil (Prestataire en biostatistique) a été menée durant la première et la deuxième année de thèse auprès de l'équipe de Méta-analyse de SBE dirigée par le Dr Jean-Pierre Pignon. Ce projet a porté sur l'étude du rôle pronostique et prédictif de Kras et de la taille de la tumeur, dans les cancers bronchiques non à petites cellules réséqués (méta-analyse LACE-BIO) [1].

Une activité de "chargée de TD" a été également menée aux universités de :
Paris 6 : Cours de statistiques de première année PCEM1 (2 semestres/2ans : 2 fois 34 heures).
Paris 11 (Master 2 Méthodologie et statistique en recherche biomédicale) : Initiation au logiciel R et SAS (16h) et analyse de données de survie sous SAS (7 heures).

Directeur de thèse : Marie-Cécile Le Deley
Co-directeur de thèse : Sarah Zohar

1. Ce travail a mené à une publication dans la revue "Journal of Thoracic oncology"

Remerciements

Je remercie Dr Jacques Bénichou d'avoir accepté de présider mon jury de thèse. Je remercie également Dr Gérard Pons et Dr Raphaël Porcher d'avoir accepté d'évaluer ce travail et d'être rapporteurs de ma thèse. Ma gratitude s'adresse aussi au Dr Xavier Paoletti d'avoir accepté d'examiner ce travail.

Je leur en suis reconnaissante.

Je tiens à remercier Marie-Cécile Le Deley et Sarah Zohar d'avoir co-dirigé ce travail. Un "grand Merci" de tout le savoir que vous m'avez si généreusement transmis, d'encadrements si efficace, de vos conseils si pertinents, et de votre confiance et votre soutien.
Marie-Cécile : Merci d'avoir surpassé la fatigue momentanée de tes yeux pour prendre le temps de lire en détails mon manuscrit de thèse. Ta rigueur scientifique et ton savoir faire m'ont été très avantageux.
Sarah : Merci de m'avoir poussée à me surpasser. Merci d'avoir été compréhensive et de m'avoir encouragée, en particulier pendant cette dernière période de rédaction.

Ma reconnaissance va au Dr Jacques Grill, Dr Birgit Geoerger et tous les cliniciens qui ont contribué à ce travail, sans eux ce travail n'aurait pu être mené à bien.

Tous mes remerciements au Docteur Ellen Benhamou de m'avoir accueillie dans son service. Mes remerciements s'adressent aussi à Gwenaël Le Teuff pour son apport scientifique et sa bonne humeur ainsi qu'à tous les membres du service de Biostatistique et d'Epidémiologie de l'Institut Gustave Roussy.

J'exprime également ma gratitude à Jean Bouyer, directeur de l'école doctorale 420, ainsi qu'au ministère de l'enseignement supérieur et de la recherche et l'ENCCA (*European Network for Cancer Research in Children and Adolescents*) pour leur soutien financier.

Je tiens à remercier particulièrement Dr Sumithra Mandrekar et Rui Qin de Mayo Clinic pour leur collaboration précieuse.

Enfin, j'adresse mille mercis à mon mari et à nos enfants pour leur amour et leur soutien dans la réalisation de ce projet de longue haleine... je leur dédie cette thèse.

A ma Tunisie
Lorsqu'un jour le peuple veut vivre,
Force est pour le Destin de répondre,
Force est pour les ténèbres de se dissiper,
Force est pour les chaînes de se briser.
...
Qui n'aime pas gravir la montagne,
vivra éternellement au fond des vallées.

Extrait du poème "La volonté de vivre" à Abou El Kacem Chebbi[2].

2. Abou El Kacem Chebbi est l'un des plus grands poètes de la Tunisie (1909-1934), poète de la modernité, à la fois romantique et révolté. Ses poèmes chantaient en particulier l'amour et la liberté, et incitaient à la résistance. Les quatre premiers vers de ce poème font partie de l'hymne national tunisien, et ont été hissés et chantés par les protestataires pendant la révolution tunisienne.

A mon époux, ma famille, mes amours, mes amis
Pour qu'un enfant grandisse, il faut tout un village.
(Proverbe africain)

A mes chers enfants
... Shady & Lyne

Résumé de thèse

Le but principal d'un essai de phase I en oncologie est d'identifier, parmi un nombre fini de doses, la dose à recommander d'un nouveau traitement pour les évaluations ultérieures, sur un petit nombre de patients. Le critère de jugement principal est classiquement la toxicité. Bien que la toxicité soit mesurée pour différents organes sur une échelle gradée, elle est généralement réduite à un indicateur binaire appelé "toxicité dose-limitante" (DLT). Cette simplification très réductrice est problématiqu, en particulier pour les thérapies, dites "thérapies ciblées", associées à peu de DLTs.

Dans ce travail, nous proposons un score de toxicité qui résume l'ensemble des toxicités observées chez un patient. Ce score, appelé TTP pour *Total Toxicity Profile*, est défini par la norme euclidienne des poids associés aux différents types et grades de toxicités possibles. Les poids reflètent l'importance clinique des différentes toxicités. Ensuite, nous proposons la méthode de recherche de dose, QLCRM pour *Quasi − Likelihood Continual Reassessment Method*, modélisant la relation entre la dose et le score de toxicité TTP à l'aide d'une régression logistique dans un cadre fréquentiste.

A l'aide d'une étude de simulation, nous comparons la performance de cette méthode à celle de trois autres approches utilisant un score de toxicité : i) la méthode de Yuan et al. (QCRM) basée sur un modèle empirique pour estimer, dans un cadre bayésien, la relation entre la dose et le score, ii) la méthode d'Ivanova et Kim (UA) dérivée des méthodes algorithmiques et utilisant une régression isotonique pour estimer la dose à recommander en fin d'essai, iii) la méthode de Chen et al. (EID) basée sur une régression isotonique pour l'escalade de dose et l'identification de la dose à recommander. Nous comparons ensuite ces quatre méthodes utilisant le score de toxicité aux méthodes CRM basées sur le critère binaire DLT. Nous étudions également l'impact de l'erreur de classement des grades pour les différentes méthodes, guidées par le score de toxicité ou par la DLT. Enfin, nous illustrons le processus de construction du score de toxicité ainsi que l'application de la méthode QLCRM dans un essai réel de phase I. Dans cette application, nous avons utilisé une approche Delphi pour déterminer avec les cliniciens la matrice des poids et le score de toxicité jugé acceptable.

Les méthodes QLCRM, QCRM, UA et EID présentent une bonne performance en termes de capacité à identifier correctement la dose à recommander et de contrôle du surdosage. Dans un essai incluant 36 patients, le pourcentage de sélection correcte de la dose à recommander obtenu avec les méthodes QLCRM et QCRM varie de 80 à 90% en fonction des situations. Les méthodes basées sur le score TTP sont plus performantes et plus robustes aux erreurs de classement des grades que les méthodes CRM basées sur le critère binaire DLT. Dans l'application rétrospective, le processus de construction du score apparaît faisable facilement. Cette étude nous a conduits à proposer des recommandations pour guider les investigateurs et faciliter l'utilisation de cette approche dans la pratique.

En conclusion, la méthode QLCRM prenant en compte l'ensemble des toxicités s'avère séduisante pour les essais de phase I évaluant des médicaments associés à peu de DLTs *a priori*, mais avec des toxicités multiples modérées probables.

Mots clés : Méthode de Phase I, Oncologie, thérapie ciblée, CRM, QLCRM, score de toxicité, Quasi-Bernoulli.

Abstract

The aim of a phase I oncology trial is to identify a dose with an acceptable safety level. Most phase I designs use the Dose-Limiting Toxicity (DLT), a binary endpoint, to assess the level of toxicity. DLT might be an incomplete endpoint for investigating molecularly targeted therapies as a lot of useful toxicity information is discarded.

In this work, we propose a quasi-continuous toxicity score, the Total Toxicity Profile (TTP), to measure quantitatively and comprehensively the overall burden of multiple toxicities. The TTP is defined as the Euclidean norm of the weights of toxicities experienced by a patient, where the weights reflect the relative clinical importance of each type and grade of toxicity. We propose then a dose-finding design, the Quasi-Likelihood Continual Reassessment Method (QLCRM), incorporating the TTP-score into the CRM, with a logistic model for the dose-toxicity relationship in a frequentist framework. Using simulations, we compare our design to three existing designs for quasi-continuous toxicity scores : i) the QCRM design, proposed by Yuan et al., with an empiric model for the dose-toxicity relationship in a Bayesian framework, ii) the UA design of Ivanova and Kim derived from the "up-and-down" methods for the dose-escalation process and using an isotonic regression to estimate the recommended dose at the end of the trial, and iii) the EID design of Chen et al. using the isotonic regression for the dose-escalation process and for the identification of the recommended dose. We also perform a simulation study to evaluate the TTP-driven methods in comparison to the classical DLT-driven CRM. We then evaluate the robustness of these designs in a setting where grades can be misclassified. In the last part of this work, we illustrate the process of building the TTP-score and the application of the QLCRM method through the example of a paediatric trial. In this study, we have used the Delphi method to elicit the weights and the target toxicity-score considered as an acceptable toxicity measure.

All designs using the TTP-score to identify the recommended dose had good performance characteristics for most scenarios, with good overdosing control. For a sample size of 36, the percentage of correct selection for the QLCRM ranged from 80 to 90%, with similar results for the QCRM design. Simulation study demonstrates also that score-driven designs present an improved performance and robustness compared to conventional DLT-driven designs.

In the retrospective application of erlotinib trial, the consensus weights as well as the target-TTP were easily obtained, confirming the feasibility of the process. Some guidelines to facilitate the process in a real clinical trial for a better practice of this approach are suggested.

The QLCRM method based on the TTP-endpoint combining multiple graded toxicities is an appealing alternative to the conventional dose-finding designs, especially in the context of molecularly targeted agents.

Keywords : Phase I Designs, Oncology, targeted therapies, CRM, QLCRM, Toxicity score, Quasi-Bernoulli.

Productions scientifiques

Articles

- Article publié :

  Ezzalfani M, Zohar S, Qin R, Mandrekar SJ, Le Deley MC. Dose-finding designs using a novel quasi-continuous endpoint for multiple toxicities. Statistics in Medicine. 2013 Jan 21, doi : 10.1002/sim.5737

- Articles soumis :

  Ezzalfani M, Le Deley MC, Qin R, Mandrekar SJ, Zohar S. Comparison of the performance between score-driven and DLT-driven Phase I designs and sensitivity analysis to toxicity grade mis-classifications.

  Ezzalfani M, Zohar S, Geoerger B, Gilles V, Mandrekar SJ, Le Deley MC. Phase I trials of Molecularly Targeted agent : Which alternative to dose-limiting toxicity endpoint ?

Communications orales

- ISCB 2013, Munich
  Ezzalfani M, Le Deley MC, Qin R, Mandrekar S, Zohar S.
  Performance of toxicity score-driven versus DLT-driven oncology Phase I designs, and robustness to toxicity grading errors.

- EPICLIN, 2012, Lyon
  Ezzalfani M, Zohar S, Le Deley MC.
  Design des essais de phase I en oncologie : que gagne-t-on à considérer un score de toxicité plutôt que le critère binaire classique DLT ?

- ICSA (International Chinese Statistical Ass.), 2011, New York
  Ezzalfani M, Zohar S, Le Deley MC.
  A Novel Dose-Finding Design using a Quasi-Continuous Toxicity Endpoint, for Multiple Toxicity.

- SFDS, 2011, Tunis
  Ezzalfani M, Zohar S, Le Deley MC.
  Performance et comportements des différentes méthodes de recherche de dose en oncologie, avec prise en compte des grades de toxicité.

- EPICLIN, 2011, Marseille
  Ezzalfani M, Zohar S, Le Deley MC.
  Comparaison des différentes méthodes de recherche de dose en oncologie, avec

prise en compte de toxicités modérées et gradées.

- EPICLIN, 2010, Paris
  Ezzalfani M, Zohar S, Le Deley MC.
  Prise en compte des toxicités modérées et gradées dans les essais de recherche de dose.

Communication affichée

- ASCO, 2013, Chicago
  Ezzalfani M, Zohar S, Mandrekar S.J, Geoerger B, Vassal G, Le Deley MC.
  Novel toxicity endpoint for dose finding designs evaluating Molecularly Targeted Agents (MTA).

Abréviations utilisées

CRM : Méthode de réévaluation séquentielle en général, guidée par le critère DLT, *Continual Reassessment Method*. Le terme CRM est également utilisé plus spécifiquement pour la méthode de réévaluation séquentielle avec un modèle empirique dans un cadre bayésien, [1].

DR : Dose à recommander.

DR+1 : La dose au dessus de la dose à recommander.

DR-1 : La dose en dessous de la dose à recommander.

DR+2 : La dose au-dessus de la dose DR+1.

DR-2 : La dose en dessous de la dose DR-1.

DLT : Toxicité limitant la dose ou Toxicité Dose-Limitante, *Dose Limiting-toxicity*.

θ_{DLT} : Toxicité jugée acceptable pour le critère DLT, c'est un pourcentage, jugé acceptable, de patients présentant une DLT.

EID : Méthode de recherche de dose proposée par Chen et al.[2] : c'est une méthode guidée par un score de toxicité utilisant une régression isotonique, appelée *Extended Isotonic Design*.

LCRM : Méthode de réévaluation séquentielle, guidée par le critère DLT, avec un modèle logistique dans un cadre fréquentiste, *Likelihood Continual Reassessment Method*.

nTTP : Score de toxicité normalisé, *Normalized Total Toxicity Profile*.

θ^* : Toxicité jugée acceptable pour le critère nTTP, c'est un score moyen de toxicité jugée acceptable.

PCS : Pourcentage de Sélection Correcte, *Percentage of Correct Selection*.

QCRM : Méthode de recherche de dose proposée par Yuan et al. [3] : c'est une extension de la CRM guidée par un score de toxicité utilisant un modèle empirique dans un cadre bayésien. Cette méthode est appelée *Quasi-CRM*, .

QLCRM : Méthode de recherche de dose proposée par Ezzalfani et al. (notre proposition, [4]) : c'est une extension de la CRM guidée par un score de toxicité utilisant

un modèle logistique dans un cadre fréquentiste. Cette méthode est appelée *Quasi-Likelihood CRM*.

TTP : Score de toxicité, *Total Toxicity Profile*.

θ_{TTP} : toxicité jugée acceptable pour le critère TTP.

UA : Méthode de recherche de dose proposée par Ivanova et Kim, [5] : c'est une méthode guidée par un score de toxicité utilisant un algorithme pour l'escalade de dose et une régression isotonique pour identifier la DR. Cette méthode est appelée *Unified Approach*.

A noter que nous avons préféré utiliser les abréviations en anglais pour la méthode CRM et le terme DLT car elles sont largement utilisées. Les abréviations des noms des méthodes publiées ont également été maintenues.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

Mon travail porte sur les essais de recherche de dose, appelés aussi essais cliniques de phase I, en cancérologie. Avant d'exposer le contexte actuel de ces études et les problématiques que j'ai étudiées, je présenterai brièvement les étapes du développement thérapeutique en oncologie.

Développement thérapeutique en oncologie

La méthodologie des essais cliniques s'est développée et structurée en grande partie autour des essais thérapeutiques d'évaluation de nouveaux médicaments, mais le champ des essais cliniques est plus large. Après la phase pré-clinique réalisée in-vitro ou sur l'animal, évaluant la toxicologie et le mécanisme d'action d'un nouveau médicament, l'autorisation de mise sur le marché de ce dernier est le plus souvent subordonnée à quatre phases d'évaluation distinctes chez l'homme :

1. Les essais de phase I permettent d'évaluer la toxicité d'un nouvel agent thérapeutique ou d'une combinaison de médicaments chez l'homme. Le but principal de ces essais est de définir, parmi un nombre fini de doses, une dose à recommander pour les phases ultérieures d'évaluation du médicament. En oncologie, ces essais sont menés chez des malades en impasse thérapeutique avec les traitements validés. Les essais de phase I nécessitent en général l'inclusion de 15 à 40 malades.
2. Les essais de phase II sont centrés sur l'évaluation de l'efficacité d'un nouveau médicament, tout en continuant d'informer sur la toxicité de la molécule évaluée. De taille encore limitée par rapport aux études ultérieures, ces essais jouent essentiellement un rôle de filtre pour décider ou non de la poursuite du développement thérapeutique.
3. Les essais de phase III sont des essais comparatifs de confirmation. Ils permettent de démontrer l'efficacité du nouveau traitement par rapport à un traitement de référence ou à un placebo. Si les résultats de ces essais sont en faveur du nouveau traitement, une demande est soumise aux autorités de santé afin que ce traitement soit commercialisé. Les essais de phase III nécessitent l'inclusion d'un

grand nombre de malades afin d'assurer une puissance statistique suffisante et un bon niveau de preuve pour mettre en évidence l'effet du traitement.

4. Les essais de phase IV, appelés aussi étude de pharmacovigilance ou de post-marketing, évaluent le médicament en prescription de routine dans le but, en particulier, de dépister des effets secondaires rares ou tardifs.

Principes généraux des essais de phase I en oncologie

Plusieurs hypothèses sous-tendent les essais de phase I en oncologie, en particulier quand il s'agit de médicaments cytotoxiques conventionnels. Il est admis qu'il existe une relation monotone et croissante entre la dose et la toxicité d'une part, et la dose et l'efficacité d'autre part. Ceci définit le paradigme "Plus est mieux" (*More is better*) ; c'est-à-dire plus la dose est élevée, plus elle est supposée efficace, mais également plus elle est supposée toxique [6]. L'idée est alors d'identifier la dose la plus élevée possible entraînant une toxicité jugée acceptable, cette dose étant appelée dose maximale tolérée, supposée associée à une efficacité maximale. Compte tenu des différences de définition de la dose maximale tolérée en fonction des pays et de l'époque, nous avons choisi de travailler avec la terminologie de la dose à recommander pour les phases II (en anglais, *Recommended Phase* 2 *Dose*, RP2D), ou plus simplement la dose à recommander (DR).

Notre travail est centré sur les essais de phase I de recherche de dose pour lesquels la toxicité est le critère de jugement principal.

Les essais de phase I sont des essais séquentiels d'escalade de dose, c'est-à-dire que des cohortes successives de patients sont traités séquentiellement à différents paliers de dose du médicament. Les patients de la première cohorte reçoivent une dose initiale faible. La dose à allouer à chaque nouvelle cohorte est définie en fonction des toxicités rapportées chez les patients précédents. La dose peut être escaladée à une dose supérieure, répétée à la même dose ou dés-escaladée (suivant le schéma d'allocation de dose pré-défini avant le début de l'essai).

Les événements toxiques sont évalués de manière standardisée pour chaque organe de chaque type possible de toxicité. La classification NCI-CTC ("National Cancer Institute-Common Toxicity Criteria" V4.0 [7]) est la classification la plus utilisée actuellement en oncologie. Elle définit les grades pour chaque type de toxicité sur une échelle ordinale, par ordre croissant de sévérité, de 0 (pas de toxicité) à 5 (toxicité létale). Un grade 1 correspond à une toxicité minime (en anglais *mild*), un grade 2 à une toxicité modérée (*moderate*), un grade 3 à une toxicité sévère (*serious*), et un grade 4 à une toxicité mettant en jeu le pronostic vital du patient (*life-threatening*).

Contrastant avec l'énorme quantité d'information collectée habituellement dans un essai de phase I, le critère de toxicité considéré pour guider l'escalade de dose et définir la dose à recommander est classiquement défini comme une variable binaire : survenue ou non d'une toxicité jugée sévère sur une période d'observation donnée. Ce critère appelé "toxicité limitant la dose" ou "toxicité dose-limitante" (en anglais *Dose-Limiting Toxicity*, DLT) correspond donc à un critère composite agrégeant différents types pos-

sibles de toxicité sévère. Sa définition varie en fonction du contexte clinique, mais il est assez habituel de classer comme DLT l'apparition d'une toxicité hématologique de grade 4 et l'apparition d'une toxicité extra-hématologique de grade supérieur ou égal à 3 [8]. La catégorie "pas de DLT", inclut l'absence complète de toxicité ainsi que les toxicités de grade minime ou modéré. Ces dernières sont généralement acceptées comme "un mal nécessaire" en regard du bénéfice espéré pour les traitements cytotoxiques.

Principales méthodes de recherche de dose en oncologie

Parmi les méthodes d'escalade de dose basées sur le critère DLT, nous distinguons deux principales approches :

- La première approche est une approche non paramétrique basée sur un algorithme d'escalade de dose. Dérivée de l'approche proposée par Dixon et Mood en 1948 [9], le schéma 3+3, appelé schéma standard ou "*Design A*", publié par Storer en 1989 en est l'archétype [10]. Cette méthode peut être décrite comme suit : après avoir traité trois patients à un palier de dose, les patients suivants sont traités au palier de dose supérieur si aucun patient ne présente de DLT, tandis que le même palier de dose est répété si on observe 1 DLT parmi les 3 patients. Si on observe au moins 2 DLTs sur 3 ou 6 alors la recherche de dose est arrêtée et la dose à recommander est définie comme étant la dose juste inférieure. Il s'agit d'une méthode sans mémoire puisque la dose à allouer aux patients suivants est définie uniquement en utilisant les observations des derniers patients [11]. Plusieurs versions ont été dérivées du schéma standard, parmi lesquelles les schémas de type "up-and-down" où la dose peut être dés-escaladée selon le nombre de DLT observé. Ce type de schémas est largement utilisé dans les essais de phase I compte tenu de sa simplicité d'implémentation.
- La seconde approche est une approche paramétrique guidée par des modèles mathématiques. Elle repose sur une modélisation de la relation entre la dose et la probabilité d'observer une DLT en utilisant toute l'information acquise à l'inclusion de chaque nouvelle cohorte dans l'essai(O'Quigley 90, Rogatko EWOC 98, whitehead 98). Ceci permet une meilleure estimation de la dose à recommander [12, 13, 14, 15, 16, 17]. O'Quigley, Pepe et Fisher ont publié en 1990 la première "version" de la méthode de réévaluation séquentielle (connue par *Continual Reassessment Method*, CRM) [1]. Plusieurs extensions de cette méthode ont été ensuite proposées [18, 19, 20, 21, 22, 23]. Le chapitre 2 est consacré à présenter et détailler le principe de la CRM ainsi que ses avantages et ses limites par rapport aux schémas algorithmiques.

Limites du critère binaire DLT

Que ce soient les schémas algorithmiques de type "3+3" ou les méthodes paramétriques de type "CRM", l'escalade de dose ainsi que l'identification de la dose à recommander reposent sur le critère binaire DLT. Même si la performance des méthodes basées sur un modèle est améliorée par rapport aux schémas algorithmiques,

ces méthodes sont intrinsèquement limitées par la simple variablité du critère de jugement binaire, particulièrement problématique dans le contexte des essais de phase I où peu de patients sont inclus. La performance de ces méthodes en termes d'identification correcte de la dose à recommander reste insuffisante puisqu'en moyenne la bonne dose est identifiée avec une probabilité d'à peine 50% [24].

Bien que tous les grades de toxicité observés soient généralement collectés dans le cahier d'observation d'un essai de phase I, toute cette information n'est pas exploitée. Ceci peut engendrer une perte importante d'information, particulièrement préjudiciable dans ces essais de petite taille : i) En effet, la DLT considère seulement le maximum des toxicités sévères observées. ii) La multiplicité possible des événements toxiques n'est pas prise en compte dans la définition de la DLT. iii) La DLT ne fait pas la différence entre les organes atteints et les types de toxicité, par exemple une toxicité rénale de grade 4 et une fatigue de grade 3 peuvent être classées, sans distinction, comme DLT en les considérant comme échangeables et de même importance clinique. iv) Enfin, les toxicités minimes et modérées ne sont pas prises en compte dans l'information résumée DLT puisque elles sont inférieures au seuil définissant la DLT.

Développements thérapeutiques actuels en oncologie

La prise en charge des patients en oncologie a été profondément modifiée ces dernières années par l'émergence des thérapies ciblées qui bouleversent le paradigme "Plus est mieux" établi pour les médicaments cytotoxiques. Ces nouvelles thérapies possèdent en général un spectre de toxicité très différent de celui des agents anticancéreux cytotoxiques [25, 26, 27, 28, 29, 30]. En effet, contrairement aux chimiothérapies classiques qui agissent de façon non spécifique sur les cellules cancéreuses et les cellules non cancéreuses, les thérapies ciblées sont des molécules plus "intelligentes" qui agissent en inhibant une cible biologique spécifique impliquée dans la prolifération des cellules cancéreuses. Elles sont ainsi supposées ne pas interagir avec les cellules saines. A la différence de la chimiothérapie conventionnelle qui doit être administrée de façon discontinue afin de permettre la régénération des tissus sains, les thérapies ciblées sont administrées, le plus souvent, de façon continue et prolongée. Ces nouvelles thérapies sont généralement bien tolérées présentant souvent très peu, voire même une absence de toxicité sévère quand la période d'observation est limitée aux premières semaines du traitement [29, 30]. La dose maximale tolérée, au sens strict du terme, peut ne pas être atteinte puisque l'escalade de dose n'est stoppée par aucune DLT [29]. Cependant, ces thérapies ne sont pas dépourvues de toxicité et il est possible d'observer des toxicités modérées multiples qui posent d'autant plus de souci que les traitements peuvent être administrés de façon prolongée [31]. Ces toxicités non prises en compte dans les essais de phase I basés sur le critère DLT peuvent conduire à des arrêts temporaires ou à des diminutions de doses répétées au cours des essais de phase II ou de phase III. Une mauvaise identification de la dose à recommander au terme de l'essai de phase I peut contribuer à l'échec des essais ultérieurs [32].

La méthodologie des essais de phase I développée pour les thérapies cytotoxiques est ainsi remise en question dans le contexte de l'évaluation des thérapies ciblées. Bien

que de nombreux travaux soient en cours pour intégrer des critères d'activité biologique ou d'efficacité afin de définir la dose biologique optimale [25, 26, 27, 30, 33, 34, 35, 36], la toxicité reste le critère de jugement principal. Ceci a été confirmé par une revue des essais de phase I en oncologie publiés entre 1997 et 2008, dont 99 essais évaluant les thérapies moléculaires ciblées [29] : la toxicité était le critère principal utilisé pour identifier la dose à recommander dans toutes ces études.

Méthodes de recherche de dose basées sur un score de toxicité

Définition des scores

Récemment, de nouvelles méthodes de recherche de dose ont été proposées pour prendre en compte la toxicité comme variable ordinale ou continue [2, 3, 4, 5, 37, 38, 39, 40]. Bekele et Thall sont les premiers à considérer un score de toxicité appelé TTB pour *Total Toxicity Burden*. Ce score prend en compte l'ensemble des toxicités observées chez un patient [38]. Leur approche consiste à définir un poids associé à chaque grade de chaque type de toxicité lié au traitement. Ces poids, fixés par les cliniciens, reflètent l'importance clinique relative des différentes toxicités possibles. Le score est défini par la somme arithmétique de ces poids. En 2012, Lee et al. ont proposé un score de toxicité dérivé de l'approche de Bekele et Thall, appelé TBS pour *Toxicity Burden Score* [40]. Dans leur processus de construction, Lee et al. ont estimé les poids à l'aide d'une régression des scores sur les différents grades des différents types de toxicité considérés, en utilisant des données historiques. Ces deux scores diffèrent dans le processus de construction, cependant ils sont tous les deux définis par la somme arithmétique des poids de toxicité. A noter que selon ces deux mesures, le score d'un patient ne présentant que des toxicités minimes et modérées peut être plus élevé que le score d'un patient présentant une DLT isolée. Par ailleurs, le score d'un patient présentant deux toxicités est égal à la somme des scores de deux patients présentant chacune de ces deux toxicités séparément.

Ultérieurement, Chen et al. ont défini un nouveau score de toxicité appelé ETS pour *Equivalent Toxicity Score* [2]. Pour ceci, les auteurs ont proposé dans un premier temps une nouvelle échelle définissant la sévérité des grades. Cette échelle est dérivée de celle de NCI mais elle différencie les grades 3 non-DLT des grades 3 DLT, d'une part, et les grades 4 non-DLT des grades 4 DLT, d'autre part. Ces nouveaux grades, appelés "grades ajustés" (*adjusted grades*) sont classés entre 0 et 6 :

- Le grade ajusté 0 est le grade NCI 0,
- Le grade ajusté 1 est le grade NCI 1,
- Le grade ajusté 2 est le grade NCI 2,
- Le grade ajusté 3 est le grade NCI 3 non classé DLT,
- Le grade ajusté 4 est le grade NCI 4 non classé DLT,
- Le grade ajusté 5 est le grade NCI 3 classé DLT,
- Le grade ajusté 6 est le grade NCI 4 classé DLT,

Le score ETS est défini par le grade ajusté maximum observé moins 1 plus une valeur positive comprise entre 0 et 1 calculée à l'aide d'une fonction logistique. Cette

valeur représente les toxicités supplémentaires observées chez le même patient. Cela conduit aux observations suivantes : le score d'un patient présentant une DLT est supérieur à 4 et le score d'un patient ne présentant pas de DLT est inférieur à 4. La multiplicité des toxicités observées chez un même patient ne contribue qu'à la partie décimale du score. A noter que le score ETS ne prend pas en compte l'importance clinique relative entre différents types de toxicité de même grade.

Définition des scores-cibles

Par analogie à la définition du pourcentage jugé acceptable de DLT pour définir la dose à recommander, le score-cible correspondant au score de toxicité jugé acceptable, pour chaque proposition des scores définis ci-dessus, doit être défini préalablement au commencement de l'essai.
Bekele et Thall ont proposé de définir ce score-cible en utilisant un ensemble de cohortes hypothétiques. Ainsi, ils invitent les cliniciens à se déterminer quant à la dose à allouer aux patients suivants après observation de différentes cohortes hypothétiques. Ces cohortes fictives doivent couvrir un large spectre de profils de toxicité, certaines cohortes ne présentant aucune toxicité, d'autres à l'inverse, présentant plusieurs toxicités sévères. Un même patient peut cumuler plusieurs toxicités. Le nombre de cohortes doit être suffisamment élevé pour offrir une bonne représentation de l'ensemble des profils de toxicité possibles tout en restant raisonnable pour les cliniciens. Pour chaque cohorte, la question suivante est posée aux cliniciens : "Au vu des toxicités observées chez ces patients, quelle décision prenez-vous pour la cohorte suivante : augmenter, diminuer ou répéter la dose ?". Les cohortes sont ordonnées ensuite par ordre croissant de score moyen, en indiquant les décisions prises par les cliniciens. Les décisions seront jugées cohérentes si elles se répartissent en trois blocs distincts, les premières décisions étant d'escalader, les dernières de désescalader, le bloc central correspondant aux décisions de répéter. Si les décisions apparaissent incohérentes avec l'ordre des scores moyens, le processus doit être revu et corrigé (définition des poids, décisions prises pour certaines cohortes hypothétiques). Une fois l'ensemble de décisions cohérent avec l'ordonnancement des scores moyens des cohortes, le score de toxicité jugée acceptable (score-cible) est défini par la moyenne des scores associés aux cohortes pour lesquelles la décision des cliniciens est de répéter la dose.
En 2010, Chen et al. ont proposé une nouvelle approche pour définir le score-cible correspondant à leur proposition de score ETS. La détermination du score-cible est également une décision clinique basée sur les profils cibles de toxicité. Afin de définir ces profils-cibles de toxicité, les cliniciens sont invités à répondre à un questionnaire de quatre questions :

1. Si le traitement devenait standard, quelle serait la proportion acceptable de patients présentant une DLT ? (A) 20%, (B) 33%, (C) 50%, (D) Autres, précisez
2. Pour une probabilité-cible de DLT à la dose à recommander, quel ratio de grades 3 DLT et de grades 4 DLT, trouveriez-vous acceptable ? (A) 1 :1, (B) 2 :1, (C) 1 :2, (D) Autres, précisez
3. Parmi les patients traités à la dose à recommander, beaucoup n'auront pas de

DLT, mais auront tout de même des effets secondaires. Quel est le plus petit pourcentage acceptable de patients qui n'auront pas de toxicité ? (A) 0%, (B) 5%, (C) 10%, (D) Autres, précisez

4. Pour les patients présentant une toxicité non-DLT à la dose à recommander, quel ratio des toxicités de grade 1, grade 2, grade 3 non-DLT et grade 4 non-DLT serait acceptable ? (A) 1 :1 :1 :1, (B) 4 :3 :2 :1, (C) 1 :2 :3 :4, (D) Autres, précisez

Le score-cible de toxicité est déduit des réponses à ces quatre questions.

A noter que les auteurs de ces deux différentes propositions n'ont pas détaillé les étapes d'élaboration du score-cible dans un essai réel de phase I.

Pour information, Lee et al. n'ont pas défini de score-cible correspondant au score TBS, car ils n'ont pas utilisé ce score comme variable continue. En effet, ils ont proposé de dichotomiser le score TBS et d'utiliser la méthode classique CRM basé sur un critère binaire.

Méthodes basées sur un score de toxicité

Les méthodes de recherche de dose habituelles qui utilisent la variable binaire DLT pour l'escalade de dose et l'identification de la dose à recommander, telles que les méthodes de réévaluation séquentielles (CRM) et le schéma d'escalade de dose 3+3, ne sont pas applicables pour ces différentes propositions de score. Elles doivent être adaptées ou modifiées pour tenir compte d'un score de toxicité. Plusieurs méthodes ont été récemment proposées pour considérer ce type de critère de toxicité [2, 3, 5, 38, 40]. Bekele et Thall sont les premiers à proposer une méthode de recherche de dose basée sur un score de toxicité [38]. Cette méthode est une méthode paramétrique utilisant des variables latentes pour décrire la distribution de probabilité conjointe des différentes toxicités. Le modèle, de dimension égale au nombre de toxicités considérées, est très complexe conduisant à une méthode de calcul très lourde et difficile à appliquer dans la pratique. Il nous a semblé préférable d'explorer dans le cadre de ce travail des méthodes utilisant directement le score de toxicité. Ainsi en 2007, Yuan et al. ont développé une extension de la CRM qui permet d'intégrer un score de toxicité compris entre 0 et 1 [3]. Cette méthode a été développée dans un cadre bayésien et avec un modèle empirique pour estimer la relation entre la dose et le score. En 2009, Ivanova et Kim ont proposé une méthode de recherche de dose, simple et applicable à tout type de critère de toxicité. Après une phase d'escalade de dose dérivée de la méthode algorithmique "*up and-down*" où la règle de décision est basée sur la statistique de test t, la dose à recommander est estimée en fin d'essai par une régression isotonique [5]. En revanche, ces derniers auteurs ne font pas de proposition pour la construction du critère de toxicité. Dans l'article où ils définissent le score ETS, Chen et al. ont également proposé une méthode de recherche de dose pour laquelle l'escalade de dose et l'identification de la dose à recommander sont basées sur une régression isotonique [2].

Objectif de la thèse et cheminement scientifique

Mon travail de recherche a été motivé par la demande des cliniciens insatisfaits devant la réduction d'information en tout ou rien dans le contexte d'un essai clinique de phase I pédiatrique évaluant l'erlotinib, inhibiteur des récepteurs d'un facteur de croissance épithélial (NCT00418327) [8]. Le premier objectif de ce travail était d'élaborer un nouveau score de toxicité qui prend en compte l'ensemble des grades des différentes toxicités observées chez un patient. Etant donné que la toxicité est mesurée et gradée pour plusieurs organes, nous avons considéré qu'elle peut définir un espace multidimentionel d'information. Séduits par la proposition de Bekele et Thall qui considère l'importance clinique relative entre les différents type de toxicité à travers la définition des poids associés à chaque grade de chaque type de toxicité, nous avons proposé un nouveau score de toxicité, appelé TTP (*Total Toxicity Profile*), défini par la norme euclidienne des poids. Nous avons choisi de travailler avec la norme euclidienne puisqu'il s'agit de la mesure la plus adéquate pour refléter des informations multidimentionelles sur une même échelle. L'étape suivante était de développer une méthode de recherche de dose modélisant ce score. Parmi les méthodes existantes, nous avons été attirés par l'approche de Yuan et al. qui propose une extension de la méthode de référence CRM pour un score de toxicité. Comme mentionné précédemment, leur méthode utilise un modèle empirique dans un cadre bayésien. Notre choix initial était de proposer une extension de la CRM avec un modèle logistique dans un cadre fréquentiste. Le choix initial d'un modèle logistique était motivé par l'utilisation large de ce type de modèle dans les essais de phase I paramétriques ; l'inférence fréquentiste présentait l'avantage d'éviter de spécifier une distribution *a priori* du paramètre du modèle. Nous avons appelé cette méthode QLCRM pour *Quasi-Likelihood CRM*. D'autres extensions ont été étudiées par la suite. Cette première partie du travail a abouti à un article publié dans la revue *Statistics in Medicine*, joint en annexe de ce manuscrit [4]. Nous avons ensuite évalué cette méthode en comparaison des méthodes existantes susceptibles d'utiliser un score de toxicité d'une part, et en comparaison des méthodes basées sur le critère binaire DLT d'autre part. En intégrant tous les grades de toxicité observés dans le score, il est légitime de s'inquiéter de la sensibilité de la méthode aux éventuelles erreurs d'observation, c'est-à-dire aux erreurs de classement des grades. L'étude de robustesse des méthodes aux erreurs d'observation des grades fait l'objet du deuxième volet de ce travail.
Nous avons illustré la construction du score de toxicité ainsi que l'application de la méthode QLCRM dans un essai de phase I évaluant l'erlotinib chez les enfants traités pour une tumeur cérébrale. Dans cette application, nous avons proposé une approche Delphi pour déterminer avec les cliniciens la matrice des poids et le score de toxicité jugé acceptable. Ceci nous a conduits à proposer des recommandations pour guider les investigateurs et faciliter l'utilisation de cette approche dans la pratique.

Organisation du manuscrit

Etant donné que notre méthode de recherche de dose est dérivée de la méthode

CRM, nous présentons dans la première partie du manuscrit le principe de la méthode CRM, ainsi que ses avantages et ses limites en comparaison des méthodes algorithmiques. Dans la deuxième partie, nous présentons la construction du score et la méthode d'escalade de dose que nous avons proposées, ainsi que les méthodes auxquelles notre approche sera comparée. Nous détaillons ensuite le plan de simulations et les différents scénarios considérés pour comparer les différentes méthodes. Après avoir exposé les résultats de l'étude de simulation et illustré la méthode dans le contexte d'un essai clinique réel, nous terminons ce manuscrit par une discussion au regard de la littérature et des perspectives envisageables.

Chapitre 2

Méthodes de Réévaluation Séquentielle (CRM) pour un critère binaire

La méthode de réévaluation séquentielle (en anglais *Continual Reassessment Method*, CRM) a été initialement proposée par O'Quigley, Pepe et Fisher en 1990 [1]. C'est une méthode séquentielle et adaptative guidée statistiquement par une modélisation paramétrique de la relation entre la dose et la probabilité d'observer une DLT. Cette relation sera notée ultérieurement par la relation dose-toxicité. Le principe de cette méthode consiste à réévaluer, après l'inclusion de chaque nouvelle cohorte de patients, la relation dose-toxicité, à partir des observations de toxicité de tous les patients inclus dans l'essai jusqu'au moment donné. La dose à administrer aux prochains patients est la dose à laquelle la probabilité prédite par le modèle pour avoir une DLT est la plus proche de la probabilité de toxicité jugée acceptable, appelée toxicité-cible.

Dans sa version originale, la CRM a été développée avec un modèle tangente hyperbolique dans un cadre bayésien pour estimer la relation dose-toxicité. Différentes extensions ont été ensuite proposées, se distinguant principalement par le modèle utilisé d'une part et par la méthode d'estimation des paramètres du modèle, d'autre part. Parmi les méthodes développées, les modèles les plus fréquemment utilisés sont : le modèle logistique et le modèle empirique [13, 41, 42]. Deux approches d'estimation existent : l'approche bayésienne proposée dans la publication d'origine et l'approche fréquentiste basée sur l'estimation par le maximum de vraisemblance [18]. Par des études de simulation, différents auteurs ont montré que les modèles à un seul paramètre présentent une meilleure performance comparée aux modèles à deux paramètres [13, 24]. Compte tenu de la multiplicité des variantes possibles, nous avons choisi, pour la suite du travail, de considérer uniquement deux variantes : le modèle empirique avec une approche bayésienne et le modèle logistique à un seul paramètre avec une approche fréquentiste. Ces deux méthodes sont détaillées dans les paragraphes suivants.

2.1 Notations

Nous considérons les notations suivantes, valables également pour la suite du travail :

$D = \{d_1, \ldots, d_k, \ldots, d_K\}$: l'ensemble de doses à explorer dans l'essai de phase I.

n : le nombre total de patients à inclure dans l'essai, fixé préalablement au commencement de l'essai.

Y_i : variable binaire indiquant la survenue ou non de DLT chez le patient i ; $Y_i \in \{1, 0\}$ où $i \in \{1, \ldots, n\}$.

$\Omega_i = \{(d_1, Y_1), (d_2, Y_2), \ldots, (d_i, Y_i)\}$: l'ensemble des observations recueillies après l'inclusion de i sujets.

θ_{DLT} : le pourcentage-cible de toxicité dose-limitante. C'est un pourcentage, jugé acceptable, de patients présentant une DLT. Cette valeur est définie préalablement en collaboration étroite avec les cliniciens de l'essai.

Nous supposons que la variable aléatoire Y_i suit une loi de Bernoulli, prenant la valeur 1 avec la probabilité $P(Y_i = 1)$.

2.2 Modèle empirique avec une inférence bayésienne

Dans le modèle empirique, défini ci-dessous, nous cherchons à estimer le paramètre b de la fonction de puissance :

$$P(Y = 1|d_k) = \Psi(d_k, b) = \alpha_k^b \tag{2.1}$$

Où $b \in \Re_+$ et $0 \leq \alpha_1 \leq \ldots \leq \alpha_K \leq 1$.

Les α_k sont les probabilités de toxicité à chaque niveau de dose k, reflétant l'opinion initiale des cliniciens. Cet ensemble de probabilités est appelé modèle de travail ou squelette du modèle (connu en anglais par *working model* ou *initial guesses*). La définition de ce modèle sera détaillée dans l'annexe (8.1).

2.2.1 Inférence bayésienne

Le paramètre du modèle que l'on cherche à estimer est considéré comme une variable aléatoire de densité de probabilité *a priori*, $g_0(b)$. Cette dernière est préalablement définie à partir de l'avis des cliniciens de l'essai et à l'aide des informations disponibles sur le traitement à évaluer.

Après l'observation du $i^{ème}$ sujet, la distribution *a posteriori* du paramètre à estimer est définie comme suit :

$$g_i(b|\Omega_i) = \frac{L_i(\Omega_i|b)g_0(b)}{\int_{\Re_+} L_i(\Omega_i|v)g_0(v)dv} \tag{2.2}$$

Où L_i est la fonction de vraisemblance définie, après l'observation du $i^{ème}$ patient, par :

$$L_i(\Omega_i|b) = \prod_{r=1}^{i} [\Psi(d_r, b)]^{y_r} [1 - \Psi(d_r, b)]^{1-y_r} \tag{2.3}$$

Les probabilités de toxicité prédites $\hat{P}(Y = 1)$ (appelées ultérieurement probabilités estimées) sont calculées, à chaque niveau de dose, comme suit :

$$\hat{P}(Y = 1) = \int_{\Re_+} \Psi(d_k, b) g_i(b|\Omega_i) db \approx \Psi(d_k, \hat{b}) \tag{2.4}$$

où $\hat{b}$ est l'estimateur bayésien du paramètre b. Il est calculé à partir de l'espérance *a posteriori* de b comme suit :

$$\hat{b} = E(b|\Omega_i) = \int_{\Re_+} b \times g_i(b|\Omega_i) db \tag{2.5}$$

2.2.2 Règle d'administration de la prochaine dose et identification de la dose à recommander

Les informations recueillies sur la toxicité des patients permettent de réestimer les probabilités de toxicité à chaque niveau de dose en utilisant l'équation 2.4. La dose à administrer aux patients suivants est la dose pour laquelle l'estimation *a posteriori* de la probabilité de toxicité est la plus proche du pourcentage cible θ_{DLT}.

Différentes règles d'arrêt ont été proposées pour les études conduites par cette méthode [43, 44, 45]. Dans ce travail, nous nous limitons aux essais de taille fixée préalablement au commencement de l'essai.

A la fin de l'essai, la dose à recommander est définie comme suit :

$$\text{DR} = \arg\min_{d_k} \left| \Psi(d_k, \hat{b}) - \theta \right| ;\ k \in \{1, \ldots, K\} \tag{2.6}$$

La dose à recommander est ainsi la dose qui pourrait être attribuée au prochain patient et qui est associée à un pourcentage estimé de DLT le plus proche du pourcentage cible θ_{DLT}.

2.3 Modèle logistique avec une inférence fréquentiste

Dans la suite de ce travail, nous considérons le modèle logistique à un seul paramètre ci-dessous :

$$P(Y = 1|d_k) = \Psi(x_k, b) = \frac{\exp(a + bx_k)}{1 + \exp(a + bx_k)} \tag{2.7}$$

Où a est une valeur fixe définissant l'ordonnée à l'origine du modèle ($a \in \Re$) ; b est la pente du modèle que l'on cherche à estimer ($b \in \Re_+$). A noter qu'un modèle logistique avec une pente fixe et une ordonnée à l'origine à estimer peut également être utilisé. x_k sont les pseudo-doses ($x_k \in \Re$). Ils sont obtenus à partir du modèle de travail α_k.

En supposant que la moyenne *a priori* du paramètre b est égal à 1, les pseudo-doses sont calculées en résolvant l'équation ci-dessous :

$$x_k = \ln\left(\frac{\alpha_k}{1-\alpha_k}\right) - a, \forall k \in (1, \ldots, K) \tag{2.8}$$

Où ln est le logarithme népérien.

2.3.1 Inférence fréquentiste

Afin de pallier les difficultés à définir les distributions *a priori* du paramètre à estimer et pour limiter leur subjectivité, O'Quigley et Shen ont proposé une méthode d'estimation par inférence fréquentiste qui ne nécessite aucune distribution *a priori* sur le paramètre du modèle que l'on cherche à estimer [18]. Le paramètre est ainsi estimé, par le maximum de vraisemblance, comme suit :

$$\hat{b} = \underset{\Re_+}{\arg\max}\ L_i\left(\Omega_i|b\right) \tag{2.9}$$

Où $L_i(\Omega_i|b)$ est la fonction de vraisemblance (cf. la formule 2.3 de la page 12).

Les probabilités de toxicité estimées à chaque niveau de dose sont calculées comme suit :

$$\hat{P}(Y=1|d_k) = \Psi(x_k, \hat{b}) = \frac{\exp(a+\hat{b}x_k)}{1+\exp(a+\hat{b}x_k)} \tag{2.10}$$

Une certaine hétérogénéité de la réponse est nécessaire pour utiliser le modèle de la CRM dans un cadre fréquentiste. En effet, la fonction de la vraisemblance est strictement croissante si toutes les observations sont égales à 0 (Y=0) et elle est strictement décroissante si toutes les observations sont égales à 1 (Y=1), dans ces deux cas de figure, le maximum de la fonction de vraisemblance se trouve aux frontières du domaine de définition du paramètre à estimer (soit 0, soit $+\infty$). La CRM est ainsi proposée en deux étapes :

- Dans la première étape, la dose est escaladée à l'aide d'un schéma algorithmique jusqu'à l'observation de la première DLT.
- La deuxième étape est guidée par la CRM avec une inférence fréquentiste.

Ce processus est problématique pour l'évaluation des thérapies ciblées qui sont moins toxiques que les traitements cytotoxiques et pour lesquels peu de DLTs sont attendues.

2.3.2 Règle d'administration de la prochaine dose et identification de la dose à recommander

Le but de la CRM est le même quelle que soit la méthode d'estimation. La dose à administrer aux prochains patients est la dose pour laquelle la probabilité de toxicité estimée est la plus proche de la cible θ_{DLT}.

A la fin de l'essai, la dose à recommander est définie selon la formule 2.6.

2.4 Avantages et limites de la méthode CRM en comparaison des méthodes algorithmiques

Dans le contexte classique où la DLT est le critère de jugement, les méthodes CRM ont été largement étudiées par différents auteurs [13, 46, 47, 48, 49, 50]. Les conclusions étaient divergentes.

Le principal inconvénient de ces méthodes est d'ordre pratique : cette technique nécessite la collaboration étroite avec un biostatisticien et l'utilisation prospective d'un logiciel spécifique après le traitement de chaque cohorte de patients.

Certaines études de simulation ont montré que la performance de ces méthodes est comparable à celle des méthodes algorithmiques, en termes d'identification correcte de la dose à recommander, avec un risque élevé d'exposer les patients à des doses trop toxiques [51, 52].

Cependant, de nombreuses études de simulation, en réponse à ces critiques, ont montré l'inverse. Lorsque la méthode CRM est bien spécifiée (choix du modèle, inférence d'estimation, définition du modèle de travail, nombre de sujets par cohorte), elle présente une performance supérieure à celle des méthodes algorithmiques, en termes de capacité à identifier correctement la dose à recommander [13, 46, 47, 50, 53]. Ces méthodes permettent également de traiter plus de patients à la vraie dose à recommander en réduisant le nombre de patients traités à des doses inférieures à la vraie dose à recommander [15, 54]. Le comportement asymptotique de la CRM a été également évalué, montrant que l'estimation de la dose à recommander converge vers la dose cible contrairement aux méthodes algorithmiques [12, 48].

Notons aussi que contrairement aux méthodes algorithmiques, la CRM permet de définir explicitement la toxicité cible.

Cependant ces méthodes utilisent le critère binaire DLT qui ne semble pas adéquat pour l'évaluation des thérapies moléculaires ciblées pour les raisons détaillées dans l'introduction. De plus, le faible pourcentage de sélection correcte de la dose à recommander avec la CRM (à peine 50% en moyenne) est dû à l'usage d'un critère binaire très simpliste [24]. Ceci peut expliquer en partie le taux élevé d'échec des essais ultérieurs [32].

Chapitre 3

Méthodes utilisant un score de toxicité

3.1 Critère de jugement de toxicité

3.1.1 Construction du score de toxicité TTP

Pour résumer les différentes toxicités observées, Bekele et Thall ont proposé de construire un score de toxicité, appelé TTB (*Total Toxicity Burden*), qui prend en compte l'ensemble des toxicités [38]. Ce score de toxicité est défini comme la somme des poids associés à chaque grade et chaque type de toxicité. Les poids quantifient l'importance clinique relative des différentes toxicités possibles.

Un même score peut, de ce fait, correspondre à des situations cliniquement très différentes : un patient A ayant une seule toxicité de poids égal à 6 aura le même score qu'un patient B ayant 3 toxicités de poids égal à 2. En discutant avec les cliniciens, il est apparu que ces cas n'étaient pas interchangeables dans leur processus de décision, le premier étant considéré comme plus grave que le second.

Ceci nous a conduit à proposer un score de toxicité, appelé TTP (*Total Toxicity Profile*), défini par la norme euclidienne des poids.

Pour les deux profils de toxicité définis ci-dessus aboutissant à la même somme des poids, la norme est de 6 dans le premier cas *versus* 3.5 ($=\sqrt{2^2+2^2+2^2}$) dans le second, ce qui reflète mieux la gravité relative de ces profils de toxicité aux yeux des cliniciens.

Nous considérons $W = \{w_{t,j}\}$ la matrice de poids où $w_{t,j}$, est le poids défini pour chaque grade j, $j \in \{0, \ldots, 4\}$, de chaque type de toxicité t ($t \in \{0, \ldots, T\}$, où T est fixe). Afin que ce score capture l'information pertinente, il est demandé aux investigateurs de l'essai de sélectionner préalablement les toxicités significatives attendues dans le contexte de l'essai. Nous avons travaillé uniquement avec les grades de toxicité variant de 0 à 4, puisque la survenue d'un grade 5 (décès) nécessite une collaboration étroite entre les investigateurs de l'essai pour bien interpréter cet événement et pour prendre une décision concernant la poursuite ou non de l'essai. La matrice W est complétée par des zéros pour les grades qui n'existent pas dans la classification de NCI : par exemple, grade 4 fatigue ou grade 1 ou 2 d'insuffisance hépatique.

$$W = \begin{pmatrix} w_{1,0} & \dots & w_{1,4} \\ \vdots & \ddots & \vdots \\ w_{T,0} & \dots & w_{T,4} \end{pmatrix}$$

Ces poids peuvent être différents des grades NCI-CTC ([7]) puisque l'importance clinique des différentes toxicités dépend de l'organe touché, d'une part, et du contexte clinique, d'autre part. Ainsi une fatigue de grade 3 peut être considérée cliniquement moins inquiétante qu'une insuffisance rénale de grade 3.

La définition de la matrice W est réalisée en collaboration étroite avec les cliniciens avant de commencer l'essai. Pour ce faire, il est demandé aux investigateurs de l'essai de donner pour chaque grade de chaque type de toxicité possible un poids compris entre 0 et une valeur maximale, correspondant à la toxicité la plus sévère possible dans le contexte de l'essai.

Le score de toxicité TTP prend en compte l'ensemble des toxicités observées chez un patient en utilisant la norme euclidienne des poids définis dans la matrice W. Pour le patient i traité au palier de dose d_k, le score de toxicité, $TTP_{i,k}$, est défini par :

$$TTP_{i,k} = \sqrt{\sum_{t=1}^{T} \sum_{j=0}^{4} w_{t,j}^2 \ \mathbb{1}(G_{i,k,t} = j)} \tag{3.1}$$

où $\mathbb{1}(G_{i,k,t} = j)$ est l'indicatrice définie pour le patient i traité à la dose d_k comme suit,

$$\mathbb{1}(G_{i,k,t} = j) = \begin{cases} 1 & \text{si le grade de la toxicité t est égal à j,} \\ 0 & \text{sinon} \end{cases}$$

Notons que la norme euclidienne fait référence à la distance des différentes toxicités observées par rapport à leur absence sur chacun des axes de toxicité. La réponse de chaque patient est définie par un ensemble des grades correspondant aux différentes toxicités observées, définissant ainsi un vecteur dans un espace multidimensionnel.

Parmi les propriétés importantes de la norme euclidienne, l'inégalité triangulaire est une propriété séduisante dans le cadre de notre travail. En effet, elle conduit à supposer que deux événements toxiques comptent plus s'ils sont observés chez deux patients différents que s'ils sont associés chez un même patient. Par exemple, un patient cumulant deux toxicités a un score moins élevé que la somme des scores de deux patients présentant chacun l'une des deux toxicités. Ceci différencie le score TTP du score TTB proposé par Bekele et Thall.

3.1.2 Détermination du score de toxicité acceptable (score-cible)

Par analogie avec le pourcentage-cible de DLT, le score-cible correspond au score de toxicité jugé acceptable. Il doit être défini préalablement au commencement de l'essai.

De façon comparable à la démarche de Bekele et Thall, nous avons défini le score-cible comme étant la moyenne des scores moyens associés aux cohortes pour lesquelles la décision des cliniciens serait de répéter la dose.

Pour cela, nous présentons aux cliniciens un nombre fini, M, de cohortes de patients hypothétiques avec différents profils de toxicité. Ces cohortes sont présentées dans un ordre aléatoire, sans dévoiler le score individuel ni le score moyen. Nous invitons alors les cliniciens à définir au vu des toxicités observées chez ces patients la dose qu'ils administreraient aux patients suivants (dose supérieure, dose inférieure ou la même dose). Les cohortes peuvent être ordonnées par ordre croissant des scores moyens, $\overline{TTP}$. Les décisions des cliniciens sont jugées cohérentes si les valeurs les plus basses de $\overline{TTP}$ sont associées à la décision d'escalader, les valeurs les plus élevées à la décision de désescalader, les valeurs intermédiaires correspondant à la décision de répéter la dose.

Si les décisions apparaissent incohérentes avec l'ordre des scores moyens, le processus doit être revu et réajusté (définition des poids, décisions prises pour certaines cohortes hypothétiques), en utilisant la méthode Delphi, jusqu'à obtenir des décisions cohérentes.

Une fois les décisions cohérentes obtenues, le score-cible est défini par la moyenne des scores associés aux cohortes pour lesquelles la décision des cliniciens est de répéter la dose :

$$\theta_{TTP} = \frac{\sum_{m=1}^{M} \overline{TTP}\ \mathbb{1}_{(\text{Décision=répéter})}}{\sum_{m=1}^{M}\ \mathbb{1}_{(\text{Décision=répéter})}}$$

3.1.3 Normalisation du score de toxicité et du score-cible

Nous avons normalisé le score de toxicité, TTP, dans le but de travailler avec un score de toxicité compris entre 0 et 1 plus facile à modéliser (cf. paragraphe 3.2). Ce score normalisé, noté nTTP (*normalized Total Toxicity Profile*), est défini pour chaque patient i traité au palier de dose d_k par :

$$nTTP_{i,k} = \frac{TTP_{i,k}}{\nu} \tag{3.2}$$

avec $\nu = TTP_{max} + \xi$ et TTP_{max} est le score de toxicité correspondant au profil de toxicité le plus sévère, calculé à partir de la matrice W, et ξ est une valeur positive qui permet d'introduire, si nécessaire en cours d'essai, une toxicité inattendue plus sévère que le profil de toxicité le plus grave calculé à partir de la matrice W.
De la même manière, nous avons défini le score-cible normalisé par :

$$\theta^* = \frac{\theta_{TTP}}{\nu}$$

3.2 Méthode de recherche de dose pour un score de toxicité : QLCRM

Pour les essais de recherche de dose basés sur le critère DLT, il a été montré que les méthodes utilisant une modélisation de l'ensemble des observations avaient une performance supérieure aux méthodes algorithmiques. Par analogie, nous avons proposé de modéliser le score de toxicité nTTP en fonction de la dose pour l'escalade de dose et l'identification de la dose à recommander.

Il est important dans un premier temps d'identifier le type de la variable à expliquer avant d'aller plus loin dans la démarche de modélisation. Dans les situations classiques, la variable DLT est supposée distribuée selon une loi de Bernoulli. Dans ce travail, cette hypothèse n'est pas adéquate puisqu'on travaille sur un score de toxicité. Comme décrit précédemment, ce score est dérivé des différentes variables définissant les poids associés à chaque grade de chaque type de toxicité. Nous avons décidé de travailler directement sur la résultante du calcul, à savoir le score de toxicité préalablement normalisé, et de ne pas tenir compte des informations élémentaires de toxicité collectées.
Les valeurs prises par ces scores normalisés sont, par construction, des valeurs comprises entre 0 et 1. Il est possible de les énumérer puisqu'il y a un nombre limité de combinaisons de poids. Nous parlons ainsi de variable quantitative discrète. Étant donné que le nombre de combinaisons est grand, cette variable peut être traitée comme une variable quasi-continue. La loi de la distribution théorique de ce type de variable n'est pas connue. Un exemple de la distribution des scores est présenté dans la partie de résultats. Wedderburn est le pionnier à s'intéresser à modéliser ce type de variable [55]. En 1974, il a proposé de la considérer comme une fraction d'événement (puisque sa valeur est comprise entre 0 et 1) et de la traiter comme une variable pseudo-Bernoulli ou quasi-Bernoulli. Le premier et le deuxième moment (espérance et variance) du score normalisé sont considérés, par défaut, comme ceux d'une variable aléatoire distribuée selon une loi de Bernoulli.

Ayant défini le type de la variable à expliquer, la deuxième étape est de développer un modèle statistique qui permet de modéliser ce score normalisé nTTP. Etant donné que ce type de variable n'est pas gaussien, nous avons proposé de modéliser l'espérance de ce score (qui est équivalente à la moyenne attendue) en utilisant les modèles linéaires généralisés. Ces derniers permettent d'étendre les modèles linéaires simples. Initialement développés en 1972 par Nelder et Wedderburn, ces modèles sont détaillés dans le livre de McCullagh et Nelder publié en 1989 [55].

Les modèles catalogués dans la classe des modèles linéaires généralisés se caractérisent par les trois composantes suivantes :

1. La loi de distribution de la variable à expliquer :
   Dans notre cas, notons la variable aléatoire à expliquer, le score normalisé nTTP, par Z. La réalisation $\{Z = z\}$ définit alors une valeur donnée du score de toxicité normalisé à partir de l'ensemble des résultats possibles d'une expérience aléatoire.
   La distribution de la variable Z est caractérisée par :

1.a. La fonction d'espérance, $E(Z|x_k)$ (premier moment),

x_k, représentant les variables explicatives, désignent les doses à explorer dans l'essai.

1.b. La fonction de variance, $V(Z|x_k)$ (deuxième moment).
Rappelons que contrairement aux modèles linéaires simples définissant un bruit additif dans le modèle, ce bruit n'apparaît pas directement dans le modèle linéaire généralisé puisqu'on ne peut pas explicitement déduire la variance de la variable à expliquer à partir du modèle. La variance de la variable à expliquer est ainsi spécifiée conditionnellement à la valeur de x_k.
Comme énoncé précédemment, nous avons supposé dans notre proposition initiale que l'espérance et la variance de la variable Z suivent une distribution de Bernoulli.

2. Le prédicteur linéaire défini la composante déterministe du modèle, noté $a + bx_k$

3. La fonction de lien, R, reliant la composante aléatoire au prédicteur linéaire : $R(E(Z|x_k)) = a + bx_k$
Cette fonction doit être une fonction inversible.
Le choix de la fonction de lien dépend de la nature de la variable à expliquer. Par exemple, il est classique de modéliser une probabilité par une fonction de répartition de forme sigmoïdale. La forme de cette sigmoïde change en fonction des paramètres caractérisant cette fonction de répartition. Lorsque cette fonction de répartition est celle de la loi logistique, on obtient le modèle de régression logistique, appelé aussi le modèle *Logit*. Pour modéliser des variables comprises entre 0 et 1, il existe d'autres choix pour la fonction de lien, telles que la fonction de lien cloglog, la fonction de lien probit et la fonction de lien de puissance définissant le modèle empirique.

3.2.1 Modèle statistique de la QLCRM

Nous avons appelé notre méthode QLCRM pour *Quasi-Likelihood CRM* [4].

Cette méthode est une extension de la CRM. Elle permet de modéliser l'espérance du score normalisé, supposé analogue à une probabilité, en utilisant le modèle linéaire généralisé avec une fonction de lien *Logit*. Nous avons choisi cette fonction de lien car le modèle logistique est largement utilisé en épidémiologie et biostatistique [56], les courbes sigmoïdes présentant un bon ajustement des relations dose-réponse.
Le modèle est spécifié au premier et au deuxième moment comme suit :

1.
$$E(Z|x_k) = \frac{\exp(a + bx_k)}{1 + \exp(a + bx_k)}$$

Avec $a \in \Re$ et $b \in \Re^+$. x_k sont les pseudo-doses, définies de la même manière que dans la méthode classique CRM (cf. paragraphe 2.3).

2. $Var(Z|x_k) = \phi \times E(Z|x_k) \times (1 - E(Z|x_k))$

Où ϕ est le paramètre permettant d'estimer la dispersion, appelé paramètre de

dispersion. En fait, en considérant le score normalisé comme une fraction d'évènement, la loi de Bernoulli est utilisée par défaut pour décrire la variabilité aléatoire de cette variable. Or, il se peut que ce ne soit pas le cas et que la dispersion soit en réalité supérieure ou inférieure à celle prédite par le modèle.
L'estimation du paramètre ϕ de la variance de quasi-Bernoulli est détaillée en annexe 8.5.

Suivant les recommandations des différents auteurs pour la CRM classique [13], nous avons choisi de fixer l'ordonnée à l'origine et d'estimer la valeur de la pente b.

3.2.2 Méthode d'estimation : inférence fréquentiste

Nous avons choisi d'estimer le paramètre du modèle en utilisant l'inférence fréquentiste.
Pour écrire de façon littérale la vraisemblance du modèle, il faudrait spécifier la distribution de la variable à expliquer. Cependant, celle-ci est inconnue. Une solution consiste à bien spécifier le modèle au premier et au deuxième ordre. La fonction dite de quasi-vraisemblance permet d'estimer le paramètre du modèle. Il a été montré que les propriétés asymptotiques de cet estimateur sont satisfaisantes [57, 58]. Cette fonction, particulièrement intéressante dans notre travail, a été initialement introduite par Wedderburn en 1974 [59].
La fonction de quasi-vraisemblance du patient i est donnée par (cf. annexe 8.3 pour le détail) :

$$L_i = z_i \times \ln(E(Z|x_i)) + (1 - z_i) \times \ln(1 - E(Z|x_i)) \tag{3.3}$$

La fonction de quasi-vraisemblance pour n patients s'écrit comme suit :

$$L = \sum_{i=1}^{n} L_i \tag{3.4}$$

Le paramètre du modèle est estimé par le maximum de la quasi-vraisemblance :

$$\hat{b} = \underset{\Re_+}{\arg\max}\ L \tag{3.5}$$

Les scores normalisés moyens estimés, $\hat{E}(Z|x_k)$, à chaque niveau de dose sont obtenus comme suit :

$$\hat{E}(Z|x_k) = \frac{\exp(a + \hat{b}x_k)}{1 + \exp(a + \hat{b}x_k)} \tag{3.6}$$

En utilisant une inférence fréquentiste, une certaine hétérogénéité de la réponse est nécessaire pour estimer le paramètre b du modèle logistique (cf. chapitre 2). La QLCRM est ainsi proposée en deux étapes :

- Dans la première étape, la dose est escaladée à l'aide d'un schéma algorithmique jusqu'à l'observation d'un score différent de 0.
- La deuxième étape est ensuite guidée par la méthode QLCRM.

Un des avantages à utiliser un score de toxicité dans un cadre fréquentiste, par comparaison au critère binaire DLT, est que l'estimation du modèle est possible dès la première observation de toxicité, même si cette toxicité est d'intensité minime.

Règle de décision d'escalade de dose et identification de la dose à recommander

Pour l'escalade de dose, la dose à allouer aux patients suivants est la dose associée à un score estimé le plus proche du score-cible, θ^*.

A la fin de l'essai, la dose à recommander est définie comme la dose qui pourrait être attribuée à la prochaine cohorte, c'est-à-dire la dose associée à un score de toxicité normalisé le plus proche de θ^* :

$$\mathrm{DR} = \arg\min_{d_k} \left| \hat{E}(Z|x_k) - \theta^* \right|$$

Avec $k \in \{1, \ldots, K\}$

3.3 Calibration de la méthode QLCRM

3.3.1 Choix de l'ordonnée à l'origine pour la méthode QLCRM

Comme énoncé précédemment, la méthode QLCRM est une méthode basée sur un modèle à un seul paramètre où l'ordonnée à l'origine est fixée et la pente est à estimer. Suivant les recommandations des différents auteurs pour la CRM classique, nous avons choisi de travailler initialement avec une ordonnée à l'origine égale à 3 [13]. Deux autres valeurs ont été étudiées pour évaluer l'impact de l'ordonnée à l'origine sur la performance de la méthode. Ceci est détaillé ultérieurement dans la partie simulation.

3.3.2 Choix de la fonction de la variance pour la méthode QLCRM

Par défaut, la méthode QLCRM est basée sur la variance de Bernoulli. Une autre fonction de la variance est utilisée dans le cadre des réponses aléatoires comprises entre 0 et 1. Il s'agit de la variance de Wedderburn [55]. Elle est définie par :

$$Var(Z|x_k) = \phi \times E(Z|x_k)^2 \times (1 - E(Z|x_k))^2$$

Où ϕ est le paramètre de dispersion.
L'estimation du paramètre ϕ de la variance de quasi-Bernoulli est détaillée en annexe 8.5.

La variance de Wedderburn n'est autre que le carré de la variance de Bernoulli à une constante près.

Si l'on utilise la variance de Wedderburn, la fonction de quasi-vraisemblance du patient i est donné par (cf. annexe 8.4 pour le détail) :

$$L_i = \left[(2z_i - 1) \times \ln\left(\frac{E(Z|x_i)}{1 - E(Z|x_i)}\right) - \frac{z_i}{E(Z|x_i)} - \frac{1 - z_i}{1 - E(Z|x_i)}\right] \tag{3.7}$$

La fonction de quasi-vraisemblance pour n patients s'écrit comme suit :

$$L = \sum_{i=1}^{n} \left[(2z_i - 1) \times \ln\left(\frac{E(Z|x_i)}{1 - E(Z|x_i)}\right) - \frac{z_i}{E(Z|x_i)} - \frac{1 - z_i}{1 - E(Z|x_i)}\right] \tag{3.8}$$

La méthode QLCRM basée sur la variance de Wedderburn, notée QLCRMW, a été évaluée en comparaison avec celle basée sur la variance de Bernoulli.

3.3.3 Choix de la fonction de lien et de l'inférence d'estimation

Nous avons étendu notre travail en choisissant des fonctions de lien autres que le modèle *Logit*. Etant donné que le modèle probit et le modèle *Logit* donnent souvent des

résultats très similaires, nous avons préféré explorer, comme alternatives, la fonction de lien cloglog et la fonction de puissance.

Le modèle statistique s'écrit ainsi :

1. Pour la fonction de lien cloglog :

$$\ln(-Ln(1 - E(Z|x_k))) = a + bx_k \Rightarrow E(Z|x_k) = 1 - \exp(-\exp(a + bx_k))$$

   Où l'ordonnée à l'origine a est fixée et la pente b est à estimer.

2. Pour la fonction de lien de puissance, définissant un modèle empirique :

$$E(Z|x_k) = \alpha_k^b$$

Ces différents modèles sont basés sur la variance de Bernoulli.

Pour les différentes fonctions de lien, nous avons utilisé les deux approches d'inférence, fréquentiste et bayésienne, pour estimer le paramètre du modèle. Au total, cinq variantes ont été définies comme suit :

- La modélisation du score nTTP avec la fonction de lien *Logit* dans un cadre fréquentiste, cette méthode, notée QLCRM, a été précédemment détaillée dans le paragraphe 3.2.
- La modélisation du score nTTP avec une fonction de lien *Logit* dans un cadre bayésien, cette méthode est notée QCRM-LB pour *Quasi CRM-Logistic Bayesian*.
- La modélisation du score nTTP avec une fonction de lien de puissance dans un cadre fréquentiste, cette méthode est notée QCRM-EF pour *QCRM-Empiric Frequentist*.
- La modélisation du score nTTP avec une fonction de lien de puissance dans un cadre bayésien, cette méthode a été développée par Yuan et al., et est notée QCRM (elle sera détaillée dans le paragraphe 3.4.1).
- La modélisation du score nTTP avec une fonction de lien cloglog dans un cadre fréquentiste, cette méthode est notée QCRM-cl pour QCRM-cloglog.

Ces différents modèles à un seul paramètre peuvent être résumés dans le tableau 3.1 :

TABLE 3.1 – Méthodes paramétriques évaluées selon la fonction de lien et l'inférence d'estimation

Fonction de lien	Inférence fréquentiste	Inférence bayésienne
Logit	QLCRM	QCRM-LB
Puissance	QCRM-EF	QCRM*
cloglog	QCRM-cl	

∗ Méthode publiée par Yuan et al. (2007), voir le paragraphe suivant 3.4.1.

NB : Nous avons conscience que les acronymes utilisés pour différencier les méthodes ne sont pas cohérents. La QCRM pré-existait à la QLCRM. Les autres variantes ont été proposées ultérieurement.

La même règle de décision (escalade de dose et identification de la dose à recommander) que la méthode QLCRM a été considérée pour les différentes variantes.

3.4 Méthodes publiées utilisant un score de toxicité

Nous avons comparé la performance de la méthode QLCRM à la méthode paramétrique proposée par Yuan et al. ainsi qu'à deux méthodes non paramétriques basées sur un score de toxicité.

3.4.1 Méthode de recherche de dose proposée par Yuan et al. 2007 (QCRM)

Cette méthode est notée QCRM pour "*Quasi-CRM*".

Il s'agit d'une méthode paramétrique proposée en 2007 par Yuan et al. combinant la CRM et la fonction de quasi-vraisemblance, dans un cadre bayésien, avec un modèle empirique pour estimer la relation entre la dose et le score de toxicité [3] :

$$E(Z|d_k) = \alpha_k^b$$

Où $b \in \Re_+$ et $0 \leq \alpha_1 \leq \ldots \leq \alpha_K \leq 1$.

Par analogie à la CRM classique utilisant le critère binaire, les α_k sont les valeurs initiales des scores reflétant l'opinion *a priori* des cliniciens (cf. parargraphe 2.2). Elles sont définies préalablement au commencement de l'essai.

Le paramètre du modèle à estimer b est considéré comme variable aléatoire de densité de probabilité *a priori*, $g_0(b)$. Cette dernière est préalablement définie à partir de l'avis des cliniciens de l'essai et à l'aide des informations disponibles sur le traitement à évaluer.

La distribution *a posteriori* du paramètre à estimer pour un patient i est définie comme suit :

$$g_i(b) = \frac{L_i g_0(b)}{\int_{\Re_+} L_i g_0(v) dv} \tag{3.9}$$

Où L_i est la fonction de quasi-vraisemblance.

L'estimateur bayésien $\hat{b}$ du paramètre b est calculé à partir de l'espérance *a posteriori* de b comme suit :

$$\hat{b} = E(b) = \int_{\Re_+} b g_i(b) db \tag{3.10}$$

Les scores normalisés de toxicité sont estimés, à chaque niveau de dose, comme suit :

$$\hat{E}(Z|x_k) = \alpha_k^{\hat{b}} \tag{3.11}$$

3.4.2 Méthode de recherche de dose proposée par Ivanova et Kim 2009 (UA)

Cette méthode est notée UA pour "*Unified Approach*".

C'est une méthode non paramétrique proposée par Ivanova et Kim en 2009 pour tout type de variable de toxicité (binaire, ordinal ou continu) [5]. Elle est utilisable en particulier pour le score normalisé nTTP. Dans cette méthode, la dose à allouer aux patients suivants n'est pas définie sur la base d'un modèle utilisant l'ensemble des observations acquises, mais en fonction de l'écart entre le score-cible et le score de toxicité moyen observé chez tous les patients traités à la dernière dose attribuée.

Aucune hypothèse n'est faite sur la loi de distribution de la variable de toxicité. On considère seulement que les observations de toxicité des différents sujets sont indépendantes.

L'écart entre la moyenne des scores normalisés observés à la dose d_k et le score-cible θ^* peut être testé par la statistique de test T_k définie par :

$$T_k = \frac{\overline{nTTP}_k - \theta^*}{s_k/\sqrt{n_k}}$$

Où $\overline{nTTP}_k$ et s_k^2 sont, respectivement, la moyenne et la variance de nTTP, estimées sur l'ensemble des patients n_k traités à la dose d_k.

Supposons que le dernier sujet inclus a reçu la dose d_k, la dose à administrer à la cohorte suivante suit l'algorithme ci-dessous :

- Si $T_k \leq -\Delta$, la cohorte suivante reçoit la dose d_{k+1}.
- Si $T_k \geq \Delta$, la cohorte suivante reçoit la dose d_{k-1}.
- Si $-\Delta < T_k < \Delta$, la cohorte suivante reçoit la dose d_k.

Le paramètre Δ est appelé "paramètre de *design*" par les auteurs. Des études de simulation ont conduit les auteurs à fixer Δ à 1 [5].

A la fin de l'essai, la dose à recommander est la dose associée à un score moyen estimé le plus proche du score-cible. Ce score moyen est estimé en utilisant une régression isotonique.

La régression isotonique consiste à identifier, dans l'ensemble des fonctions monotones et croissantes possibles f(d), la fonction f^* minimisant la somme des écarts pondérés entre les observations et les estimations :

$$\min \sum_{k=1}^{K} n_k \left(\overline{nTTP}_k - f(d_k)\right)^2$$

Aucune hypothèse n'est faite quant à la forme de la relation.

Si le score de toxicité est croissant avec la dose, le score estimé par la régression isotonique, $\widehat{nTTP}_k$, à la dose d_k correspond à la moyenne des scores observés à cette dose. Dans le cas contraire, f^* est définie en utilisant l'algorithme PAVA (*Pool Adjacent Violators Algorithm*) qui permet de résoudre l'équation ci-dessus.

3.4.3 Méthode de recherche de dose proposée par Chen et al. 2010 (EID)

La méthode de Chen et al. proposée en 2010 est notée EID pour " *Extended Isotonic Design*" [2].

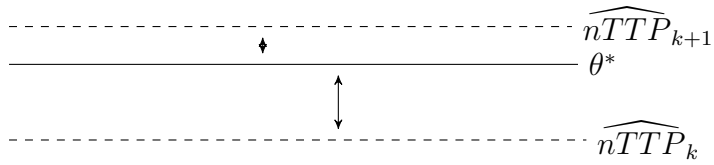
Il s'agit également d'une méthode non paramétrique. L'algorithme d'escalade de dose ainsi que l'identification de la dose à recommander en fin d'essai sont également basés sur la méthode de régression isotonique. Si la dose d_{k+1} n'a pas encore été explorée, le score estimé à cette dose est égal au score correspondant à la dose d_k.

L'algorithme d'escalade de dose est détaillé selon les règles suivantes :

– Si $\widehat{nTTP_k} < \theta^*$, alors
si $\theta^* - \widehat{nTTP_k} \geq \widehat{nTTP}_{k+1} - \theta^*$ et $k < K$, alors
la cohorte suivante reçoit la dose d_{k+1}.
Sinon, la cohorte suivante reçoit la dose d_k.

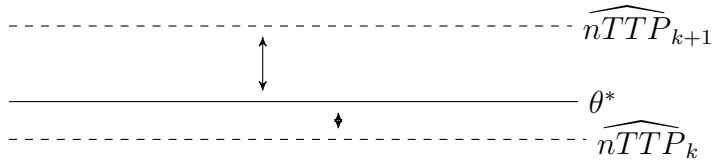
Ceci peut être schématisé comme suit :

Si $\theta^* - \widehat{nTTP_k} \geq \widehat{nTTP}_{k+1} - \theta^*$



→ Dose suivante = d_{k+1}

Sinon ($\theta^* - \widehat{nTTP_k} < \widehat{nTTP}_{k+1} - \theta^*$)

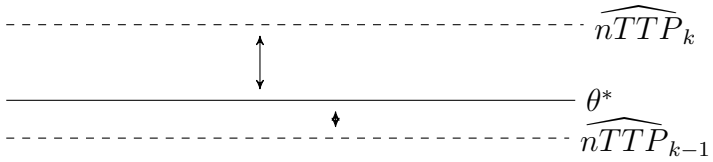


⟶ Dose suivante = d_k

– Si $\widehat{nTTP_k} \geq \theta^*$, alors
si $\theta^* - \widehat{nTTP}_{k-1} < \widehat{nTTP_k} - \theta^*$ et $k > 1$, alors
la cohorte suivante reçoit la dose d_{k-1}.
Sinon, la cohorte suivante reçoit la dose d_k.

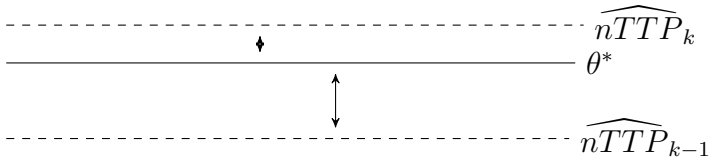
Ceci peut être schématisé comme suit :

Si $\theta^* - \widehat{nTTP}_{k-1} < \widehat{nTTP}_k - \theta^*$



→ Dose suivante = d_{k-1}

Sinon ($\theta^* - \widehat{nTTP}_{k-1} \geq \widehat{nTTP}_k - \theta^*$)



⟶ Dose suivante = d_k

A la fin de l'essai, la dose à recommander est définie comme la dose qui pourrait être attribuée à la cohorte prochaine. Suivant les recommandations des auteurs, cette dose devrait être parmi les doses allouées pendant l'essai.

Chapitre 4

Etude de simulation

Une des particularités des essais cliniques de phase I est d'être réalisés, pour des raisons éthiques, sur un nombre faible d'observations. Pour comparer les différentes méthodes, nous avons alors entrepris une étude de simulation puisque l'évaluation des propriétés asymptotiques des méthodes de recherche de dose est insuffisante. En supposant connue la vraie relation entre la dose et la toxicité, la dose à recommander est associée à un "niveau" (probabilité ou score) de toxicité le plus proche de la valeur cible de toxicité préalablement fixée. Cependant, la vraie relation dose-toxicité n'est pas connue. Nous avons donc évalué le comportement des méthodes sous différentes relations dose-toxicité plausibles, notées ultérieurement scénarios. Un exemple de relation entre la dose et le score de toxicité normalisé, nTTP, est illustré dans la figure 4.1.

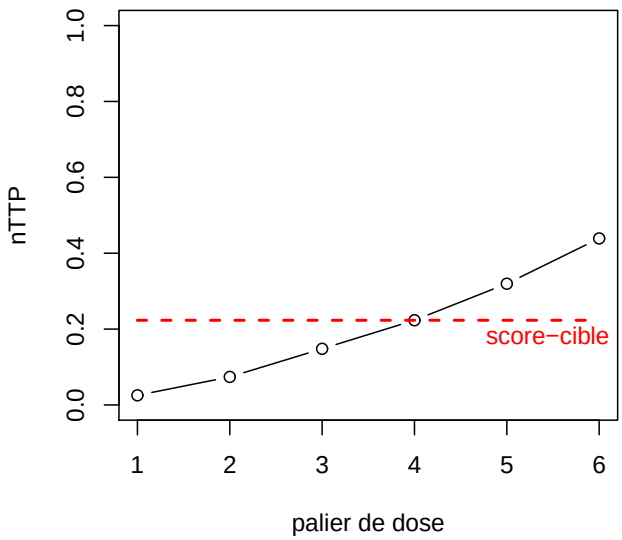


FIGURE 4.1 – Exemple de scénario (relation entre la dose et le score normalisé).

Dans cet exemple, la vraie dose à recommander (appelée aussi dose cible) est la quatrième dose.

Le principe des études de simulation est de répéter un grand nombre de fois, de façon indépendante, une expérience aléatoire à partir d'un scénario donné et de mesurer pour chaque répétition l'issue de l'expérience. Différents scénarios ont été étudiés dans le cadre de cette étude. Pour chaque scénario, nous avons simulé 5000 essais de phase I indépendants. Nous avons considéré des essais évaluant 6 doses. Le nombre de sujets par essai a été fixé à 36. Les patients ont été traités séquentiellement, à chaque palier de dose, par cohorte de trois [15, 16], sans saut de dose (*no skipping*).

4.1 Méthodes à évaluer

Dans cette étude de simulation, la méthode QLCRM (variance de Bernoulli, fonction de lien *Logit* et inférence fréquentiste) est évaluée en comparaison avec différentes méthodes.

1. Dans un premier temps, nous avons évalué la méthode QLCRM en étudiant différentes valeurs de l'ordonnée à l'origine du modèle logistique. Une seule valeur est retenue pour la suite de l'évaluation.
   Nous avons ensuite évalué la méthode QLCRM (variance de Bernoulli) en comparaison avec la méthode QLCRMW (variance de Wedderburn).
   L'impact de l'inférence d'estimation et la fonction de lien a été également évalué à travers les méthodes ci-dessous :
   - La méthode QLCRM (fonction de lien *Logit* et inférence fréquentiste).
   - La méthode QCRM-LB avec une fonction de lien *Logit* et une inférence bayésienne.
   - La méthode QCRM (fonction de lien de puissance et inférence bayésienne), cette méthode a été développée par de Yuan et al.
   - La méthode QCRM-EF avec une fonction de lien de puissance et une inférence fréquentiste.
   - La méthode QCRM-cl avec une fonction de lien cloglog et une inférence fréquentiste.

   Seules les méthodes QLCRM et QCRM sont retenues pour la suite de comparaison.

2. L'étape suivante consiste à comparer les méthodes QLCRM et QCRM aux deux autres méthodes non paramétriques :
   - La méthode UA d'Ivanova et Kim (cf. paragraphe 3.4.2).
   - La méthode EID de Chen et al. (cf. paragraphe 3.4.3).

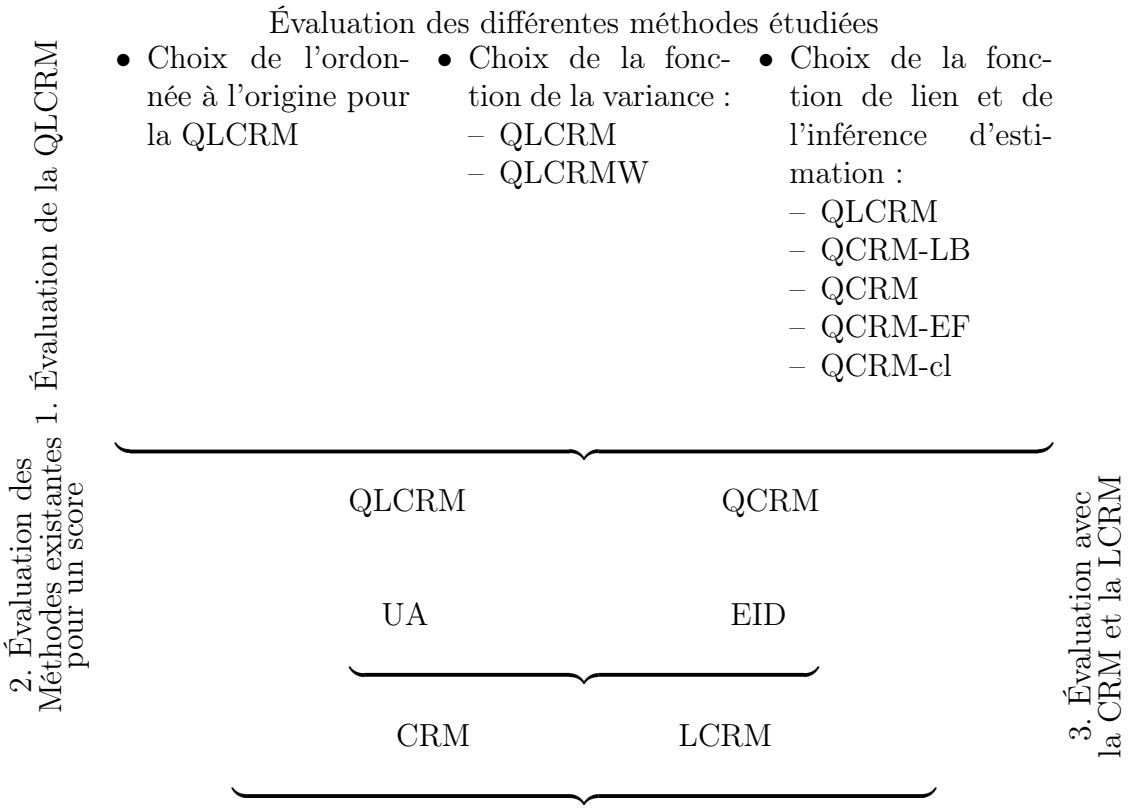
3. Dans le but de comparer les méthodes basées sur le score de toxicité nTTP aux méthodes utilisant le critère binaire DLT, nous avons considéré deux variantes de la CRM :

– La CRM proposée avec un modèle empirique pour modéliser la relation dose-toxicité et une inférence bayésienne pour estimer le paramètre du modèle. Cette méthode sera notée ultérieurement par CRM.
– La CRM développée avec un modèle logistique et une inférence fréquentiste. Cette méthode sera notée ultérieurement par LCRM.

Ces deux méthodes ont été détaillées dans le chapitre 2.

4. Dans une dernière étape, nous avons étudié la sensibilité des méthodes basées sur le score et les méthodes basées sur la DLT aux erreurs d'observation des grades.

Ces différents axes de comparaison peuvent être schématisés comme suit :



Dans ce travail, nous avons fixé les paramètres suivants :

1. Pour la méthode QLCRM :
   - Nous avons défini le modèle de travail (*working model*) en utilisant la fonction "getprior" du package dfcrm (cf. annexe 8.1), en considérant la troisième dose comme dose à recommander *a priori* et en fixant le paramètre δ définissant l'intervalle d'indifférence à 0.04 selon les recommandations publiées dans l'article de Lee et Cheung [41]. Le modèle de travail de la méthode QLCRM avec une ordonnée à l'origine de 3 est égal à (0.14, 0.20, 0.28, 0.36, 0.44, 0.52).

Nous avons également étudié la performance de la méthode QLCRM en utilisant le modèle de travail proposé par Yuan et al. (0.11, 0.25, 0.40, 0.55, 0.75, 0.85) [3]. Les résultats sont détaillés en annexe (cf. annexe 8.3).

Le modèle de travail est défini de la même manière grâce à la fonction *getprior* pour une ordonnée à l'origine égale à 2 et à 5, les autres paramètres étant inchangés.

2. Pour la méthode QCRM de Yuan et al. :
   - Nous avons considéré que la distribution *a priori* du paramètre b du modèle empirique est une distribution exponentielle du paramètre égal à 1.
   - Nous avons également défini le modèle de travail en considérant la troisième dose comme dose à recommander *a priori* et en fixant le paramètre δ à 0.04. Le modèle de travail est égal à (0.136, 0.203, 0.280, 0.362, 0.444, 0.523).
     La performance de la méthode QCRM a été également étudiée en utilisant le modèle de travail proposé par Yuan et al. [3].
3. Pour la méthode QLCRMW et QCRM-cl, nous avons fixé le même modèle de travail que celui proposé pour la méthode QLCRM. Pour la méthode QCRM-LB, nous avons fixé le même modèle de travail que celui défini pour la méthode QCRM.
4. Pour la méthode CRM (modèle empirique et cadre bayésien) :
   - Nous avons défini également le modèle de travail en considérant la fonction "getprior" avec la troisième dose comme dose à recommander *a priori* et une valeur de δ égale à 0.05 (0.147, 0.233, 0.330, 0.431,0.527, 0.615).
5. Pour la méthode LCRM (modèle logistique et cadre fréquentiste) :
   - Le modèle de travail de cette méthode est défini de la même façon que celui de la méthode CRM. Il est égal à (0.150, 0.233, 0.330, 0.430, 0.524, 0.606).

Pour être cohérent dans la comparaison des différentes méthodes, nous avons utilisé la même définition de la dose à recommander, à savoir la dose qui pourrait être administrée à la prochaine cohorte hypothétique, une fois le nombre maximum de sujets atteint. Selon cette définition, cette dose peut être une dose qui n'a jamais été allouée. Nous avons effectué une analyse de sensibilité pour étudier l'impact de cette règle de décision sur la performance des différentes méthodes. Pour cette analyse, nous avons considéré la règle de décision proposée par Chen et al., définissant la dose à recommander dans l'ensemble des doses explorées. La méthode d'Ivanova et Kim ne peut pas recommander une dose qui n'a jamais été explorée puisque la régression isotonique, utilisée en fin d'essai, est appliquée seulement sur les doses déjà explorées. Par conséquent, la méthode d'Ivanova et Kim n'est pas considérée dans cette analyse de sensibilité.

4.2 Elaboration des scénarios

Une étape importante dans ce travail était de générer des données élémentaires de toxicité selon des scénarios plausibles. Nous avons supposé une relation monotone et croissante entre la dose et la toxicité (cf. figure 4.1). Un scénario est défini par les probabilités d'observer chaque grade de chaque type de toxicité. Nous avons résumé

chaque scénario par le score moyen associé à chaque palier de dose. Avant de générer les scénarios et les données de toxicité, il est nécessaire de définir la matrice de poids et le score jugé acceptable.

4.2.1 Définition de la matrice de poids et du score jugé acceptable

La matrice de poids ainsi que le score-cible ont été proposés uniquement pour le travail méthodologique (partie simulation) et non pas pour un essai clinique réel.
Nous avons contacté un clinicien expert travaillant à l'Institut Gustave Roussy (Dr Jacques Grill) pour établir la matrice de poids et le score de toxicité jugé acceptable.

Nous avons supposé que la toxicité attendue porte essentiellement sur trois types principaux de toxicité indépendants :
- toxicité rénale,
- toxicité neurologique et
- toxicité hématologique.

Nous avons limité notre travail aux grades variant de 0 à 4, puisque la survenue du grade 5, correspondant à une toxicité létale, nécessite une discussion étroite avec les investigateurs de l'essai et le comité de sécurité afin de prendre une décision concernant la poursuite de l'essai.
La DLT a été définie par la survenue d'une toxicité rénale ou neurologique de grade 3 ou d'une toxicité hématologique de grade 4.

Les poids associés à chaque grade de chaque type de toxicité sont définis dans la matrice W. Les lignes de cette matrice indiquent les trois types de toxicité (R pour toxicité rénale, N pour toxicité neurologique et H pour toxicité hématologique). Les colonnes de la matrice correspondent aux cinq grades possibles (de G0 pour grade 0 à G4 pour grade 4).

$$W = \begin{array}{c} \\ \text{R} \\ \text{N} \\ \text{H} \end{array} \begin{array}{c} \begin{array}{ccccc} G_0 & G_1 & G_2 & G_3 & G_4 \end{array} \\ \begin{pmatrix} 0 & 0.5 & 0.75 & 1 & 1.5 \\ 0 & 0.5 & 0.75 & 1 & 1.5 \\ 0 & 0 & 0 & 0.5 & 1 \end{pmatrix} \end{array}$$

Notons qu'avec ces poids, le score d'un patient avec une DLT isolée de grade 3 rénal ou neurologique est égal au score d'un patient avec une DLT hématologique (grade 4) et ceci est très proche du score d'un patient ne présentant pas de DLT mais deux toxicités de grade 2 rénale et neurologique.
Le score maximum calculé à partir de cette matrice correspond aux profils de toxicité les plus sévères, c'est-à-dire à une toxicité rénale de grade 4 associée une toxicité neurologique de grade 4 et une toxicité hématologique de grade 4. Ce score est alors égal à :

$$TTP_{max} = \sqrt{1.5^2 + 1.5^2 + 1^2} = 2.35 \tag{4.1}$$

Nous avons choisi de normaliser les scores par 2.5 ($\nu = 2.5$).
Les scores de toxicité sont ainsi calculés selon la formule 3.1 en utilisant les poids définis

dans la matrice W.
Ayant défini un score de toxicité, la deuxième étape consiste à définir le score de toxicité jugé acceptable (score-cible). Pour ceci, nous avons invité le clinicien à répondre à un questionnaire quant à sa décision (escalader, répéter la dose ou désescalader) pour différentes cohortes hypothétiques ayant des scores moyens compris entre 0.20 et 0.41. Chaque cohorte est définie par trois patients présentant des profils de toxicités différents. Ces cohortes ont été présentées dans un ordre aléatoire sans montrer au clinicien les scores de toxicité associés.

Le score-cible est défini comme étant la moyenne des scores associés aux cohortes pour lesquelles la décision du clinicien était de répéter la dose :

$$\theta^* = 0.28 \tag{4.2}$$

A titre d'exemple, la décision était d'escalader pour une cohorte associant :
- Un patient avec un grade 2 de toxicité rénale, un grade 0 de toxicité neurologique et un grade 2 de toxicité hématologique.
- Un patient avec un grade 0 de toxicité rénale, un grade 2 de toxicité neurologique et un grade 1 de toxicité hématologique.
- Un patient avec un grade 0 de toxicité rénale, un grade 0 de toxicité neurologique et un grade 0 de toxicité hématologique.

Le score moyen associé à cette cohorte est égal à 0.20.

La décision était de répéter pour une cohorte associant :
- Un patient avec un grade 2 de toxicité rénale, un grade 0 de toxicité neurologique et un grade 1 de toxicité hématologique.
- Un patient avec un grade 0 de toxicité rénale, un grade 1 de toxicité neurologique et un grade 3 de toxicité hématologique.
- Un patient avec un grade 0 de toxicité rénale, un grade 2 de toxicité neurologique et un grade 0 de toxicité hématologique.

Le score moyen associé à cette cohorte est égal à 0.29.

La décision était de désescalader pour une cohorte associant :
- Un patient avec un grade 3 de toxicité rénale, un grade 1 de toxicité neurologique et un grade 0 de toxicité hématologique.
- Un patient avec un grade 2 de toxicité rénale, un grade 2 de toxicité neurologique et un grade 0 de toxicité hématologique.
- Un patient avec un grade 1 de toxicité rénale, un grade 1 de toxicité neurologique et un grade 3 de toxicité hématologique.

Le score moyen associé à cette cohorte est égal à 0.41.

Parallèlement, l'expert a fixé le pourcentage-cible de DLT à 33%.

4.2.2 Probabilité des différents profils de toxicité

Comme énoncé précédemment, un scénario est défini par la probabilité d'observer les différents profils de toxicité. Un profil de toxicité d'un patient est défini par la combinaison d'une :

- toxicité rénale de grade j_R, notée G_{R,j_R}, plus
- une toxicité neurologique de grade j_N, notée G_{N,j_N}, et plus
- une toxicité hématologique de grade j_H, notée G_{H,j_H}.

Où j_R, j_N et $j_H \in \{0, 1, 2, 3, 4\}$. Considérant les trois types de toxicité, il existe au total $5^3 = 125$ profils de toxicité possibles.

En supposant l'indépendance des trois types de toxicité, la probabilité d'observer un profil de toxicité donné est calculée, à chaque palier de dose d_k, comme suit :

$$P_k(G_{R,j_R}, G_{N,j_N}, G_{H,j_H}) = P_k(G_{R,j_R})P_k(G_{N,j_N})P_k(G_{H,j_H})$$

Où $P_k(G_{R,j_R})$ est la probabilité d'observer une toxicité rénale de grade j_R au palier de dose d_k. Les probabilités $P_k(G_{N,j_N})$ et $P_k(G_{H,j_H})$ sont définies de la même manière pour les toxicités neurologique et hématologique.

4.2.3 Probabilité de toxicité pour chaque type de toxicité

Pour chaque type de toxicité, nous avons considéré les hypothèses "intuitives" suivantes :

- La distribution de la probabilité d'observer un grade 0 est strictement décroissante en d_k. Cette probabilité est maximale pour le palier de dose d_1.
- La distribution de la probabilité d'observer un grade 4 est strictement croissante en d_k. Cette probabilité est maximale pour le dernier palier de dose (d_6).
- La distribution de la probabilité d'observer un grade 1, 2 ou 3 est une distribution unimodale en d_k.

Pour générer les probabilités d'observer les grades 0 à 4 de chaque type de toxicité, nous avons supposé que les grades correspondent à une transformation d'une variable aléatoire L, distribuée selon une loi normale, en variable catégorielle selon des seuils prédéfinis, l_0, l_1, l_2 et l_3. Nous avons considéré que la moyenne de la loi normale est croissante avec la dose. Différentes variances ont été considérées (fixe, croissante ou décroissante avec la dose).

Les grades sont définis comme suit :

– Le grade 0 est défini si la variable $l < l_0 \Rightarrow$ la probabilité d'observer le grade 0 est ainsi égale à la probabilité d'observer la réalisation $l < l_0$.

– Le grade 1 est défini si la variable $l_0 < l < l_1 \Rightarrow$ la probabilité d'observer le grade 1 est ainsi égale à la probabilité d'observer la réalisation $l_0 < l < l_1$.

– Le grade 2 est défini si la variable $l_1 < l < l_2 \Rightarrow$ la probabilité d'observer le grade 2 est ainsi égale à la probabilité d'observer la réalisation $l_1 < l < l_2$.

- Le grade 3 est défini si la variable $l_2 < l < l_3 \Rightarrow$ la probabilité d'observer le grade 3 est ainsi égale à la probabilité d'observer la réalisation $l_2 < l < l_3$.

- Le grade 4 est défini si la variable $l > l_3 \Rightarrow$ la probabilité d'observer le grade 4 est ainsi égale à la probabilité d'observer la réalisation $l > l_3$.

La figure 4.2 illustre les différentes distributions de la probabilité d'observer chaque grade pour un type donné de toxicité dont le processus sous-jacent serait la variable L de moyenne comprise entre -2 et 6, avec $l_0 = 0$, $l_1 = 1$, $l_2 = 2$ et $l_3 = 3$.

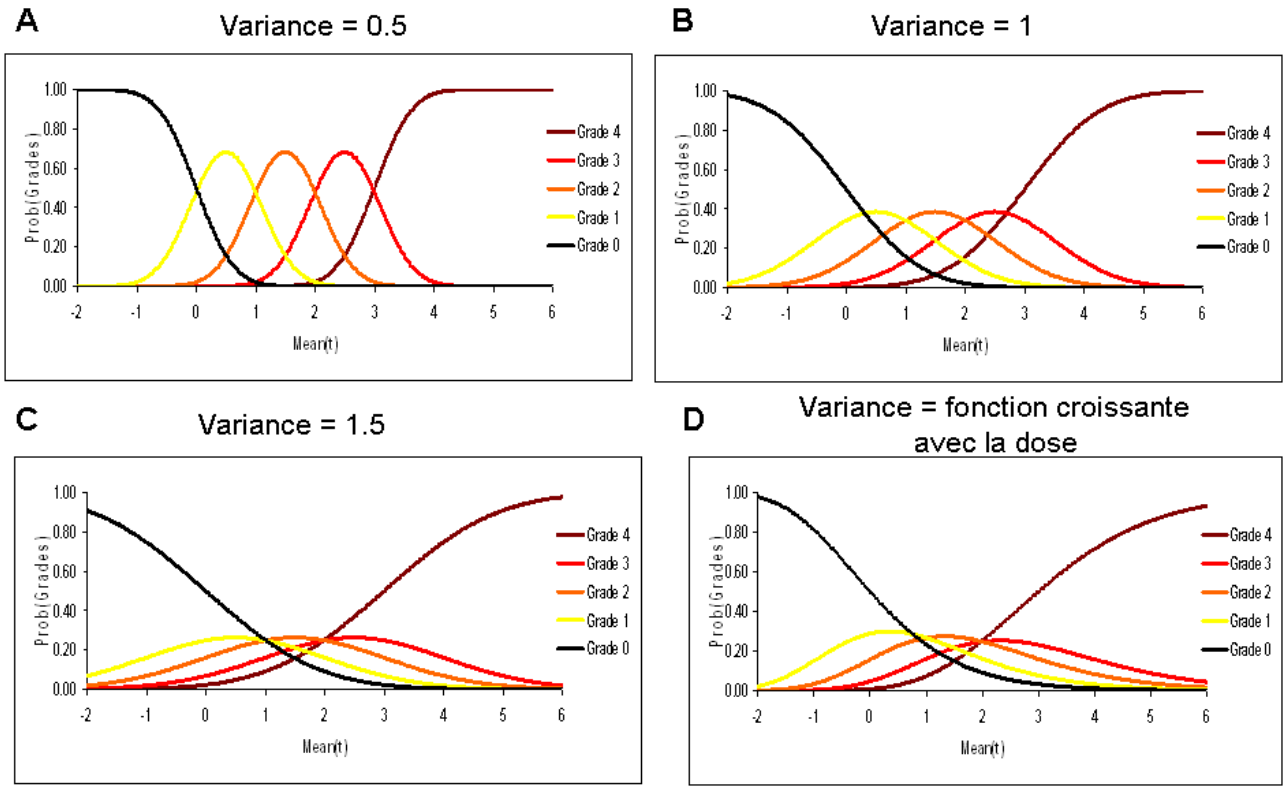


FIGURE 4.2 – Différentes distributions de la probabilité d'observer les grades 0, 1, 2, 3 et 4

.

A titre d'exemple (cf. panel B de la figure 4.2), en supposant que L suit une loi normale de variance 1 de moyenne égale respectivement à -1.5, -0.5, 1, 2, 4, et 6 pour le palier de dose d_1 à d_6. On obtient ainsi les probabilités suivantes pour chaque grade :

$$\begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} \begin{array}{c} \begin{array}{ccccc} G_0 & G_1 & G_2 & G_3 & G_4 \end{array} \\ \begin{pmatrix} 0.933 & 0.061 & 0.006 & 0.000 & 0.000 \\ 0.691 & 0.242 & 0.061 & 0.006 & 0.000 \\ 0.159 & 0.341 & 0.341 & 0.140 & 0.023 \\ 0.023 & 0.136 & 0.341 & 0.341 & 0.159 \\ 0.000 & 0.001 & 0.021 & 0.1361 & 0.841 \\ 0.000 & 0.000 & 0.000 & 0.001 & 0.999 \end{pmatrix} \end{array}$$

Pour un scénario donné, nous avons donc défini une matrice de probabilités d'observer les grades de 0 à 4 à chaque palier de dose (de d_1 à d_6) pour chaque type de toxicité (rénale, neurologique, et hématologique). Ces matrices sont notées respectivement P_R, P_N et P_H. Chaque matrice présente 6 lignes indiquant les paliers de doses à explorer et 5 colonnes indiquant la probabilité d'observer chaque grade.

4.2.4 Définition des scénarios

Comme énoncé précédemment, nous avons résumé un scénario par une valeur moyenne de score à chaque palier de dose. Le score moyen est défini à chaque palier de dose d_k comme suit :

$$\overline{nTTP}_k = \sum_{j_H=0}^{4}\sum_{j_N=0}^{4}\sum_{j_R=0}^{4} P_k(G_{R,j_R}, G_{N,j_N}, G_{H,j_H}) \times nTTP(G_{R,j_R}, G_{N,j_N}, G_{H,j_H})$$

A partir des mêmes profils de toxicité, nous avons généré les probabilités d'observer une toxicité dose-limitante (DLT), à chaque palier de dose d_k, comme suit :

$$P_k(DLT) = \sum_{j_H=0}^{4}\sum_{j_N=0}^{4}\sum_{j_R=0}^{4} P_k(G_{R,j_R}, G_{N,j_N}, G_{H,j_H}) \times \mathbf{DLT}(G_{R,j_R}, G_{N,j_N}, G_{H,j_H})$$

Où

$$\mathbf{DLT}(G_{R,j_R},G_{N,j_N},G_{H,j_H}) = \begin{cases} 1 & \text{si } \max(G_{R,j_R}, G_{N,j_N}) \geq 3 \text{ ou } \max(G_{H,j_H}) \geq 4 \\ 0 & \text{sinon} \end{cases} \tag{4.3}$$

A chaque scénario sont donc associées à la fois une relation dose-score et une relation dose-probabilité de DLT.

4.2.5 Exemple d'un scénario

Soient P_R, P_N et P_H les matrices des probabilités définies respectivement pour la toxicité rénale, neurologique et hématologique, comme suit :

$$P_R = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} \begin{array}{c} \begin{array}{ccccc} G_0 & G_1 & G_2 & G_3 & G_4 \end{array} \\ \begin{pmatrix} 0.791 & 0.172 & 0.032 & 0.004 & 0.001 \\ 0.758 & 0.180 & 0.043 & 0.010 & 0.009 \\ 0.685 & 0.190 & 0.068 & 0.044 & 0.013 \\ 0.662 & 0.200 & 0.078 & 0.046 & 0.014 \\ 0.605 & 0.223 & 0.082 & 0.071 & 0.020 \\ 0.390 & 0.307 & 0.201 & 0.073 & 0.028 \end{pmatrix} \end{array}$$

$$P_N = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} \begin{array}{c} \\ \begin{pmatrix} 0.968 & 0.029 & 0.002 & 0.001 & 0.000 \\ 0.813 & 0.172 & 0.006 & 0.009 & 0.000 \\ 0.762 & 0.183 & 0.041 & 0.010 & 0.004 \\ 0.671 & 0.205 & 0.108 & 0.010 & 0.005 \\ 0.397 & 0.258 & 0.277 & 0.061 & 0.008 \\ 0.260 & 0.377 & 0.281 & 0.073 & 0.008 \end{pmatrix} \end{array}$$

$$P_H = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} \begin{array}{c} \\ \begin{pmatrix} 0.917 & 0.070 & 0.007 & 0.000 & 0.005 \\ 0.652 & 0.280 & 0.010 & 0.020 & 0.037 \\ 0.536 & 0.209 & 0.031 & 0.091 & 0.134 \\ 0.015 & 0.134 & 0.240 & 0.335 & 0.276 \\ 0.005 & 0.052 & 0.224 & 0.372 & 0.347 \\ 0.004 & 0.022 & 0.220 & 0.344 & 0.409 \end{pmatrix} \end{array}$$

A titre d'exemple, la probabilité d'observer une toxicité rénale de grade 2, $G_{R,2}$, une toxicité neurologique de grade 0, $G_{N,0}$, et une toxicité hématologique de grade 1, $G_{H,1}$, à la dose d_1 est calculée comme suit :

$$P_1(G_{R,2}, G_{N,0}, G_{H,1}) = P_1(G_{R,2})P_1(G_{N,0})P_1(G_{H,1}) = 0.032 \times 0.968 \times 0.070 = 0.002$$

Le score moyen défini à la dose d_1 est égal :

$$\overline{nTTP}_1 = \sum_{j_H=0}^{4} \sum_{j_N=0}^{4} \sum_{j_R=0}^{4} P_1(G_{R,j_R}, G_{N,j_N}, G_{H,j_H}) \times nTTP(G_{R,j_R}, G_{N,j_N}, G_{H,j_H}) = 0.054$$

La probabilité d'observer une DLT à la dose d_1 est égale à :

$$P_1(DLT) = \sum_{j_H=0}^{4} \sum_{j_N=0}^{4} \sum_{j_R=0}^{4} P_1(G_{R,j_R}, G_{N,j_N}, G_{H,j_H}) \times DLT(G_{R,j_R}, G_{N,j_N}, G_{H,j_H}) = 0.011$$

4.3 Scénarios pour l'évaluation des différentes méthodes

4.3.1 Scénarios pour comparer les méthodes basées sur un score de toxicité

Nous avons considéré 8 scénarios principaux notés de A à H. Le scénario détaillé dans le paragraphe précédent correspondant au scénario F. Ces scénarios sont donnés dans le tableau 4.1. Ils sont triés par ordre décroissant de toxicité, commençant par les scénarios les plus toxiques pour lesquels la dose à recommander est parmi les premières doses à explorer. La dose à recommander est la deuxième dose pour le scénario A, et la cinquième dose pour le scénario H. Les scénarios C, F et G représentent une translation du scénario A vers la droite. Ces quatre scénarios sont utilisés pour étudier l'impact de la position de la dose à recommander dans l'échelle des doses. Ils présentent la même pente de la courbe dose-score autour de la dose à recommander. Pour chacun d'eux, le score nTTP à la vraie dose à recommander est égal à la valeur exacte de la cible (0.28). Les scénarios B et E représentent une variation de scénarios C et F avec un score nTTP à la vraie dose à recommander légèrement au-dessus de la valeur cible. De même, les scénarios D et H représentent une variation de scénarios C et G avec un score nTTP à la vraie dose à recommander légèrement en dessous de la valeur cible.

Table 4.1 – Description des scénarios (score nTTP et probabilité de DLT à chaque palier de dose)

	d_1	d_2	d_3	d_4	d_5	d_6
Scenario A						
nTTP	0.183	**0.280**	*0.359*	0.409	0.432	0.439
nTTP-0.28	-0.097	**0.00**	*+0.079*	+0.129	+0.152	+0.159
p(DLT)	0.195	**0.330**	*0.447*	0.512	0.557	0.558
Scenario B						
nTTP	0.100	*0.198*	**0.310**	0.390	0.441	0.481
nTTP-0.28	-0.180	*-0.082*	**+0.029**	+0.110	+0.161	+0.201
p(DLT)	0.057	0.182	**0.379**	*0.491*	0.592	0.656
Scenario C						
nTTP	0.108	0.183	**0.280**	*0.359*	0.409	0.432
nTTP-0.28	-0.172	-0.097	**0.00**	*+0.079*	+0.129	+0.152
p(DLT)	0.065	0.195	**0.330**	*0.447*	0.512	0.557
Scenario D						
nTTP	0.051	0.119	**0.270**	*0.370*	0.404	0.460
nTTP-0.28	-0.229	-0.161	**-0.010**	*+0.090*	+0.124	+0.180
p(DLT)	0.002	0.014	**0.169**	*0.319*	0.417	0.554
Scenario E						
nTTP	0.051	0.096	*0.188*	**0.312**	0.418	0.446
nTTP-0.28	-0.229	-0.184	*-0.092*	**+0.032**	+0.138	+0.166
p(DLT)	0.002	0.014	*0.186*	**0.320**	0.506	0.554
Scenario F						
nTTP	0.054	0.108	0.183	**0.280**	*0.359*	0.409
nTTP-0.28	-0.226	-0.172	-0.097	**0.00**	*+0.079*	+0.129
p(DLT)	0.011	0.065	0.195	**0.330**	*0.447*	0.512
Scenario G						
nTTP	0.045	0.054	0.108	0.183	**0.280**	*0.359*
nTTP-0.28	-0.235	-0.226	-0.172	-0.097	**0.00**	*+0.079*
p(DLT)	0.008	0.011	0.065	0.195	**0.330**	*0.447*
Scenario H						
nTTP	0.051	0.111	0.141	0.189	**0.253**	*0.352*
nTTP-0.28	-0.229	-0.169	-0.139	-0.091	**-0.027**	*+0.072*
p(DLT)	0.001	0.006	0.015	0.032	**0.097**	*0.196*

Les valeurs en gras correspondent aux valeurs à la vraie dose à recommander (définie sur la base du score-cible). Les valeurs en italique correspondent aux valeurs à la dose suivante la plus proche de la vraie dose à recommander.

Les différents scénarios sont également décrits dans la figure ci-dessous :

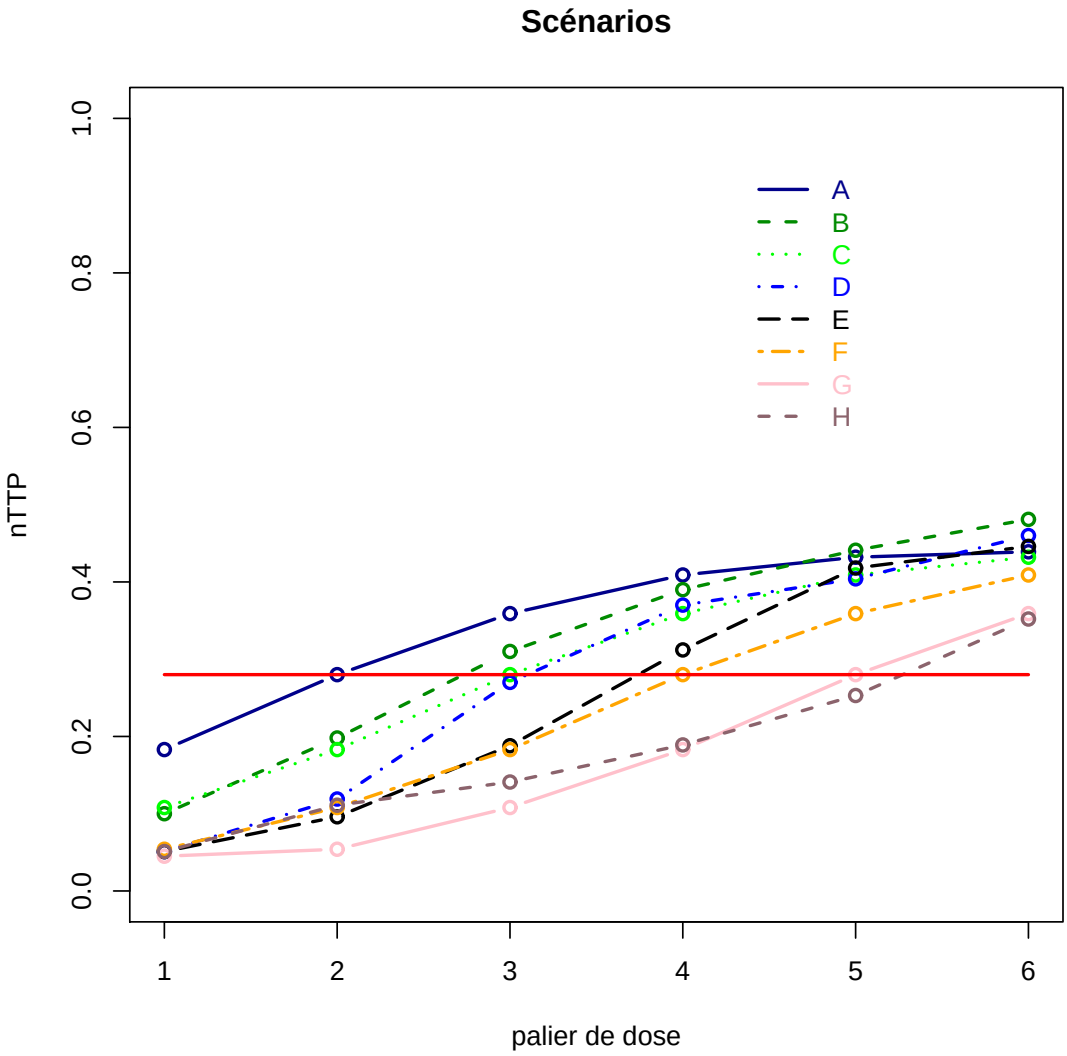


FIGURE 4.3 – Représentation graphique des principaux scénarios utilisés

4.3.2 Scénarios pour comparer les méthodes basées sur un score de toxicité à celles basées sur le critère DLT

Parmi les nombreux scénarios générés pour comparer les méthodes basées sur le score de toxicité nTTP, nous avons observé que la dose à recommander par le score est toujours inférieure ou égale à la dose à recommander par le critère binaire DLT (cf. tableau 4.1).

Afin de faciliter la comparaison de la performance des méthodes utilisant le score nTTP et les méthodes CRM classiques basées sur la DLT, nous avons choisi de travailler

avec les scénarios pour lesquels la vraie dose à recommander est la même pour les deux critères de toxicité. Les scénarios utilisés sont donc les scénarios A, C, F et G décrits dans le tableau 4.1. Etant donné que chaque scénario est la translation du scénario A, ces scénarios présentent la même pente autour de la dose à recommander, tant pour la relation dose-score que pour la relation dose-probabilité de DLT. Ceci permet également de faciliter la comparaison des différentes méthodes. Nous avons exploré des scénarios avec d'autres pentes autour de la vraie dose à recommander, ainsi qu'avec un score et une probabilité d'observer une DLT légèrement au-dessus ou au-dessous des valeurs cibles (ces scénarios sont détaillés dans l'annexe 8.4).

4.4 Génération des mesures de toxicité (scores TTP et DLT)

Une fois les scénarios définis, nous avons généré les données de toxicité (G_R, G_N et G_H) pour chacun des 36 patients de l'essai par un tirage au sort aléatoire avec remise. A chaque palier de dose et pour chaque patient, le grade de chaque type de toxicité prend la valeur 0, 1, 2, 3 ou 4 selon la matrice de probabilité P_R, P_N ou P_H.

A chaque patient traité à une dose donnée sont alors associés un score (cf. formule 3.1) et une variable DLT (cf. formule 4.3).

4.5 Etude de sensibilité des méthodes aux erreurs d'observation des grades

La première partie de l'étude de simulation a porté sur l'évaluation de la performance des différentes méthodes en supposant l'absence d'erreur d'observation quant aux grades de toxicité, donc l'absence d'erreur sur le score et la variable DLT mesurés. Cependant, dans un essai de phase I, ces valeurs peuvent être mal rapportées. Dans cette partie de travail, nous avons étudié la sensibilité des méthodes aux erreurs de classement des grades. Nous avons étudié successivement l'impact successivement des erreurs de sous-classement (*under-grading*) puis des erreurs de sur-classement (*over-grading*). Nous avons considéré, séparément, deux types d'erreurs :

- Les erreurs qui affectent tous les grades d'une manière indépendante et égale (erreurs indépendantes du grade).
- Les erreurs qui affectent seulement les grades définissant la DLT.

Notations

Pour un type donné de toxicité, soient :

T le grade de toxicité observé chez un patient donné ;

p_{Tj} la probabilité de survenue du grade j, avec $j \in \{0, 1, 2, 3, 4\}$, avec $\sum_{j=0}^{4} p_{Tj} = 1$;

Y le grade rapporté chez un patient donné ;

$p_{lm} = P(Y = G_l | T = G_m)$ la probabilité qu'une toxicité de grade l soit rapportée, sachant le grade réel m. Par exemple, p_{23} est la probabilité de rapporter une toxicité

de grade 2 sachant que la toxicité est de grade 3 (erreur de sous-classement), et p_{32} est la probabilité de rapporter une toxicité de grade 3 sachant que le vrai grade est égal à 2, (erreur de sur-classement). Soient p_O et p_U, respectivement, les probabilités de sur-classement et sous-classement des grades.

4.5.1 Hypothèses

Nous avons considéré les hypothèses suivantes :

Hypothèses	**Erreur de sous-classement**	**Erreur de sur-classement**
1. Les erreurs de sous et sur-classement sont mutuellement exclusives	$p_{lm} > 0$ si $l <$ m $p_{lm} = 0$ si $l >$ m	$p_{lm} > 0$ si $l >$ m $p_{lm} = 0$ si $l <$ m
2. Les erreurs touchent seulement les grades adjacents	$p_{lm} > 0$ si $l = m - 1$ $p_{lm} = 0$ si $l \leq m - 2$	$p_{lm} > 0$ si $l = m + 1$ $p_{lm} = 0$ si $l \geq m + 2$
3. Deux types possibles d'erreur		
3.1 Erreurs affectant seulement la DLT	Si la DLT est définie par le grade 3 ou 4 $p_U = p_{23}$ et $p_{01} = p_{12} = p_{34} = 0$	Si la DLT est définie par le grade 3 ou 4 $p_O = p_{32}$ et $p_{10} = p_{21} = p_{43} = 0$
	Si la DLT est seulement définie par le grade 4 $p_U = p_{34}$ et $p_{01} = p_{12} = p_{23} = 0$	Si la DLT est seulement définie par le grade 4 $p_O = p_{43}$ et $p_{10} = p_{21} = p_{32} = 0$
3.2 Erreurs affectant tous les grades (erreurs indépendantes du grade)	$p_U = p_{01} = p_{12} = p_{23} = p_{34}$	$p_O = p_{10} = p_{21} = p_{32} = p_{43}$
4. La probabilité de sous-classement ou sur-classement des grades est indépendante du type de toxicité		

4.5.2 Génération des mesures de toxicité observées (score TTP et DLT)

Pour étudier l'impact de l'erreur de classement des événements toxiques sur les méthodes utilisant le score nTTP ou la DLT, nous avons considéré les quatre scénarios A, C, F et G décrits dans le tableau 4.1.

Nous avons généré séparément les deux types d'erreur selon les hypothèses suivantes :

- l'erreur de sous classer un grade donné, notée par la variable E_U, est générée selon une loi de Bernoulli du paramètre P_U.
- l'erreur de sur-classer un grade donné, notée par la variable E_O, est générée selon une loi de Bernoulli du paramètre P_O.

Où P_U et P_O sont respectivement les probabilités de sous-classer et de sur-classer un grade donné. Nous avons fait varier ces probabilités d'erreur de 5 à 25%.

Pour chaque type de toxicité, à partir des "vrais" grades de toxicité, G_m, générés selon les différents scénarios étudiés (cf. paragraphe 4.3.2), nous avons ensuite généré les grades "rapportés" G_l :

1. Si les grades sont sous-classés :

$$G_l = \begin{cases} G_m & \text{si } E_U = 0 \\ G_m - 1 & \text{si } E_U = 1 \end{cases}$$

2. Si les grades sont sur-classés :

$$G_l = \begin{cases} G_m & \text{si } E_O = 0 \\ G_m + 1 & \text{si } E_O = 1 \end{cases}$$

Le score de toxicité ainsi que la DLT rapportés pour chaque patient sont obtenus respectivement selon les formules 3.1, 3.2 et 4.3. Dans chaque essai simulé, les mesures de toxicité pour l'escalade de dose et l'identification de la dose à recommander sont basées sur les grades rapportés, éventuellement erronés.

4.6 Critères de jugement pour l'évaluation des différentes méthodes

Rappelons qu'une "bonne" méthode doit permettre de :

- maximiser le nombre de patients traités à la vraie dose à recommander,
- minimiser le nombre de patients traités à des doses infra-thérapeutiques,
- minimiser le nombre de patients traités à des doses trop toxiques.

Les métriques suivantes ont été utilisées pour comparer les différentes méthodes :

- Le pourcentage de sélection de chaque dose comme dose à recommander. Nous nous sommes intéressés en particulier au pourcentage de sélection correcte de la dose à recommander (en anglais, *Percentage of Correct Selection*, PCS).
- La distribution des doses allouées aux patients au cours des essais.
- Le nombre de toxicités dose-limitantes, DLT, observées durant l'essai.
- La convergence des méthodes en termes de PCS, en faisant varier le nombre de sujets inclus dans l'essai de 15 à 99. Cette étude permet d'évaluer le comportement des méthodes dans un cas "réaliste" où un nombre faible de sujets est inclus $(15 - 40)$ ainsi que dans un cas "idéal" où un nombre plus important de sujets serait inclus.

Pour la suite du travail, notons DR la dose à recommander, DR+1 la dose au-dessus de la dose à recommander, et DR+2 la dose au-dessus de la dose DR+1. De la même manière, nous définissons les doses DR-1 et DR-2.

Chapitre 5

Résultats

5.1 Distribution des scores de toxicité nTTP

Nous avons utilisé un diagramme en bâtons pour visualiser un exemple de la distribution des scores normalisés, nTTP. Pour cet exemple, nous avons considéré les scores de toxicité dérivés de la matrice de poids W (cf. paragraphe 4.2.1), en supposant une équiprobabilité de chaque profil de toxicité. Au total, 30 valeurs de score normalisé sont possibles.
Cette répartition ne ressemble à aucune distribution de loi de probabilité connue.

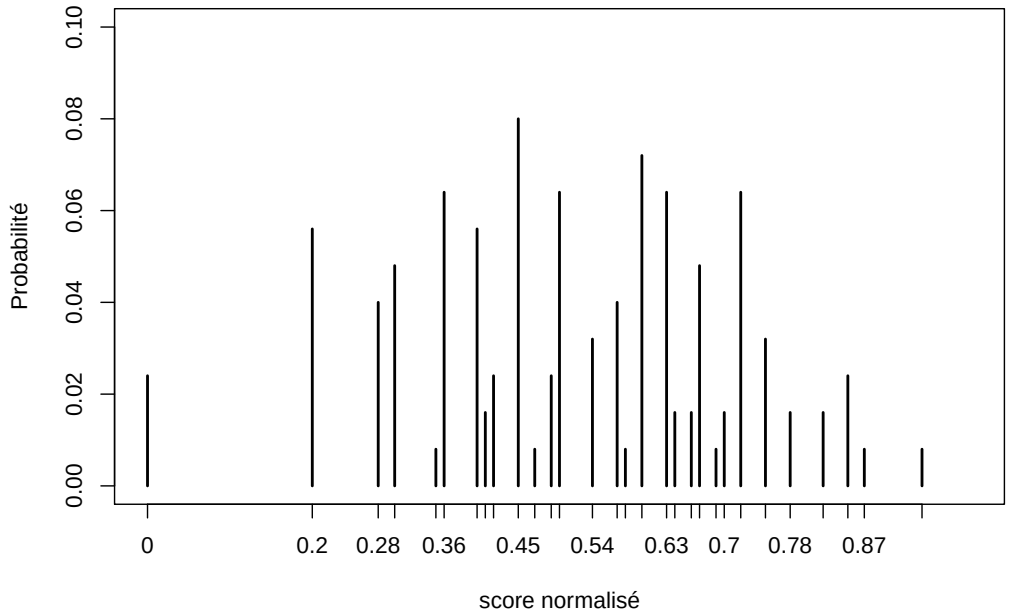


FIGURE 5.1 – Distribution des scores normalisés

5.2 Evaluation comparative de la méthode QLCRM

5.2.1 Choix de l'ordonnée à l'origine pour la méthode QLCRM

Pour évaluer l'impact de la valeur de l'ordonnée à l'origine sur la méthode QLCRM, nous avons évalué la performance de cette méthode pour trois valeurs de l'ordonnée à l'origine (2, 3 et 5). La performance de la méthode QLCRM avec une ordonnée à l'origine égale à 3, valeur recommandée pour la CRM classique, est supérieure à celle de la méthode QLCRM avec une ordonnée à l'origine égale à 2 ou à 5. Ces dernières valeurs conduisent toujours à une identification erronée de la dose à recommander. La DR est toujours sur-estimée avec une ordonnée à l'origine égale à 2, tandis qu'elle est toujours sous-estimée avec une ordonnée à l'origine égale à 5. A titre d'exemple, les résultats des scénarios A à D sont présentés dans le tableau 5.1. La méthode QLCRM (avec une ordonnée à l'origine égale à 3) est retenue pour la suite de l'étude de comparaison des différentes méthodes étudiées dans ce travail. Nous détaillons ultérieurement la performance de cette méthode.

TABLE 5.1 – Performance de la méthode QLCRM pour différentes ordonnées à l'origine (3, 2 et 5), pour $n = 36$ patients

	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Sc A												
QLCRM	3.4	**85.9**	*10.7*	0.0	0.0	0.0	19.4	**62.6**	*17.4*	0.6	0.0	0.0
$QLCRM^2$	0.0	**0.0**	*0.0*	10.8	72.6	16.6	8.4	**8.4**	*9.6*	28.2	41.0	4.6
$QLCRM^5$	100	**0.0**	*0.0*	0.0	0.0	0.0	99.4	**0.6**	*0.0*	0.0	0.0	0.0
Sc B												
QLCRM	0.0	*12.7*	**85.3**	2.0	0.0	0.0	9.1	*21.1*	**60.8**	8.9	0.2	0.0
$QLCRM^2$	0.0	*0.0*	**0.0**	0.5	50.0	49.5	8.3	*8.3*	**8.4**	11.3	43.0	20.6
$QLCRM^5$	99.9	*0.1*	**0.0**	0.0	0.0	0.0	96.6	*3.1*	**0.3**	0.0	0.0	0.0
Sc C												
QLCRM	0.0	3.0	**83.8**	*13.3*	0.0	0.0	9.2	14.9	**57.0**	*18.3*	0.7	0.0
$QLCRM^2$	0.0	0.0	0.0	**0.0**	*19.0*	81.0	8.3	8.3	**8.4**	*9.7*	30.2	35.1
$QLCRM^5$	99.9	0.1	0.0	**0.0**	*0.0*	0.0	96.7	2.9	**0.4**	*0.0*	0.0	0.0
Sc D												
QLCRM	0.0	0.1	**83.0**	*17.0*	0.0	0.0	8.4	8.6	**51.4**	*30.3*	1.4	0.0
$QLCRM^2$	0.0	0.0	**0.0**	*0.0*	3.32	96.68	8.33	8.33	**8.33**	*8.39*	17.07	49.54
$QLCRM^5$	95.6	4.4	**0.0**	*0.0*	0.0	0.0	85.5	12.7	**1.8**	*0.0*	0.0	0.0

Sc : Scénario.

QLCRM : méthode QLCRM avec une ordonnée à l'origine égale à 3.

$QLCRM^2$: méthode QLCRM avec une ordonnée à l'origine égale à 2.

$QLCRM^5$: méthode QLCRM avec une ordonnée à l'origine égale à 5.

Les valeurs en gras correspondent aux résultats obtenus à la dose cible.

Les valeurs en italiques correspondent aux résultats obtenus à la dose la plus proche de la dose cible.

5.2.2 Choix de la fonction de la variance pour la méthode QLCRM

Nous avons ensuite comparé le comportement de la méthode QLCRM avec une variance de Bernoulli à celui de la méthode QLCRMW avec une variance de Wedderburn (avec une ordonnée à l'origine égale à 3 pour les deux cas).

Hormis le scénario B, la méthode QLCRM présente globalement une performance supérieure à celle de la méthode QLCRMW. La différence de PCS est importante, variant de +2.5 à 18% en faveur de la méthode QLCRM.
La méthode QLCRM avec variance de Bernoulli est retenue pour la suite de l'évaluation des différentes méthodes.

Table 5.2 – Performance de la méthode QLCRM en utilisant la variance de Bernoulli et la variance de Wedderburn, pour $n = 36$ patients

	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Sc A												
QLCRM	3.4	**85.9**	*10.7*	0.0	0.0	0.0	19.4	**62.6**	*17.4*	0.6	0.0	0.0
QLCRMW	3.1	**83.4**	*13.4*	0.1	0.0	0.0	18.6	**60.5**	*19.7*	1.2	0.0	0.0
Sc B												
QLCRM	0.0	*12.7*	**85.3**	2.0	0.0	0.0	9.1	*21.1*	**60.8**	8.9	0.2	0.0
QLCRMW	0.0	*9.9*	**85.5**	4.6	0.0	0.0	9.0	*18.6*	**58.6**	13.0	0.7	0.0
Sc C												
QLCRM	0.0	3.0	**83.8**	*13.3*	0.0	0.0	9.2	14.9	**57.0**	18.3	0.7	0.0
QLCRMW	0.0	2.5	**77.2**	*19.9*	0.4	0.0	9.1	14.0	**52.5**	22.3	1.9	0.0
Sc D												
QLCRM	0.0	0.1	**83.0**	*17.0*	0.0	0.0	8.4	8.6	**51.4**	*30.3*	1.4	0.0
QLCRMW	0.0	0.0	**65.0**	*33.7*	1.2	0.0	8.3	8.5	**38.4**	*38.5*	6.1	0.1
Sc E												
QLCRM	0.0	0.0	*8.8*	**90.5**	0.6	0.0	8.4	8.4	*15.0*	**60.4**	7.8	0.1
QLCRMW	0.0	0.0	*4.9*	**88.7**	6.3	0.1	8.3	8.4	*12.4*	**53.5**	15.9	1.4
Sc F												
QLCRM	0.0	0.0	2.7	**80.7**	*16.5*	0.0	8.4	8.5	13.1	**50.9**	*18.5*	0.6
QLCRMW	0.0	0.0	1.8	**67.9**	*28.7*	1.6	8.4	8.5	11.9	**43.7**	*24.2*	3.4
Sc G												
QLCRM	0.0	0.0	0.0	2.6	**79.6**	*17.8*	8.4	8.3	8.4	12.3	**45.0**	*17.6*
QLCRMW	0.0	0.0	0.0	2.8	**61.8**	*35.4*	8.3	8.3	8.5	12.7	**36.7**	*25.4*
Sc H												
QLCRM	0.0	0.0	0.0	1.8	**82.5**	*15.7*	8.3	8.4	8.7	15.0	**48.0**	*11.5*
QLCRMW	0.0	0.0	0.0	6.0	**73.2**	*20.8*	8.3	8.4	8.9	18.6	**41.6**	*14.1*

Sc : Scénario.

QLCRM : méthode QLCRM basée sur la variance de Bernoulli (ces résultats correspondent aux mêmes résultats détaillés dans le tableau 5.1).

QLCRMW : méthode QLCRM basée sur la variance de Wedderburn et la fonction de quasi-vraisemblance correspondante.

Les valeurs en gras correspondent aux résultats obtenus à la dose cible.

Les valeurs en italiques correspondent aux résultats obtenus à la dose la plus proche de la dose cible.

5.2.3 Choix de la fonction de lien et de l'inférence d'estimation

Comme détaillé précédemment, la QLCRM conduit à modéliser le score nTTP en utilisant un modèle logistique dans un cadre fréquentiste. Cependant, nous avons également étendu la QLCRM en utilisant une modélisation logistique dans un cadre bayésien, une modélisation empirique dans un cadre fréquentiste et bayésien et une modélisation utilisant une fonction de lien cloglog dans un cadre fréquentiste. Ces cinq

différentes variantes présentent, en moyenne, des résultats très similaires pour les huit scénarios étudiés (résultats détaillés dans le tableau 5.3).
Pour la suite de l'évaluation, nous avons choisi de retenir, d'une part, notre méthode QLCRM car elle est basée sur un modèle logistique présentant un bon ajustement des relations dose-réponse, la méthode QCRM car c'est une méthode publiée. Ces deux méthodes sont analogues aux versions les plus utilisées de la CRM pour critère binaire. La performance de ces deux méthodes est commentée, dans le paragraphe suivant, en comparaison aux autres méthodes existantes.

Table 5.3 – Performance des méthodes selon la fonction de lien et la méthode d'estimation (fréquentiste ou bayésienne), pour $n = 36$ patients

	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Sc A												
QLCRM	3.4	**85.9**	*10.7*	0.0	0.0	0.0	19.4	**62.6**	*17.4*	0.6	0.0	0.0
QCRM-LB	4.5	**86.3**	*9.2*	0.0	0.0	0.0	22.9	**62.3**	*14.3*	0.5	0.0	0.0
QCRM	2.3	**83.6**	*14.1*	0.0	0.0	0.0	14.2	**63.1**	*22.0*	0.7	0.0	0.0
QCRM-EF	3.3	**86.2**	*10.5*	0.0	0.0	0.0	19.3	**62.7**	*17.4*	0.6	0.0	0.0
QLCRM-cl	3.4	**83.0**	*13.5*	0.0	0.0	0.0	18.6	**60.7**	*20.1*	0.6	0.0	0.0
Sc B												
QLCRM	0.0	*12.7*	**85.3**	2.0	0.0	0.0	9.1	*21.1*	**60.8**	8.9	0.2	0.0
QCRM-LB	0.0	*15.5*	**83.3**	1.2	0.0	0.0	9.7	*23.5*	**59.3**	7.4	0.1	0.0
QCRM	0.0	*10.5*	**87.4**	2.0	0.0	0.0	8.5	*18.7*	**63.1**	9.6	0.1	0.0
QCRM-EF	0.0	*12.6*	**85.5**	1.9	0.0	0.0	9.1	*21.0*	**61.1**	8.7	0.1	0.0
QLCRM-cl	0.0	*11.1*	**86.4**	2.5	0.0	0.0	9.0	*19.0*	**62.5**	9.5	0.0	0.0
Sc C												
QLCRM	0.0	3.0	**83.8**	*13.3*	0.0	0.0	9.2	14.9	**57.0**	*18.3*	0.7	0.0
QCRM-LB	0.0	3.7	**85.7**	*10.6*	0.0	0.0	9.8	16.4	**57.5**	*15.7*	0.6	0.0
QCRM	0.0	2.3	**84.1**	*13.6*	0.0	0.0	8.5	13.2	**58.5**	*19.3*	0.4	0.0
QCRM-EF	0.0	2.9	**83.8**	*13.3*	0.0	0.0	9.2	14.8	**57.2**	*18.2*	0.6	0.0
QLCRM-cl	0.0	2.5	**82.6**	*14.9*	0.0	0.0	9.1	13.8	**58.1**	*18.8*	0.2	0.0
Sc D												
QLCRM	0.0	0.1	**83.0**	*17.0*	0.0	0.0	8.4	8.6	**51.4**	*30.3*	1.4	0.0
QCRM-LB	0.0	0.2	**86.9**	*12.9*	0.0	0.0	8.4	8.9	**55.1**	*26.4*	1.2	0.0
QCRM	0.0	0.0	**82.5**	*17.5*	0.0	0.0	8.3	8.5	**50.2**	*32.1*	1.0	0.0
QCRM-EF	0.0	0.1	**82.4**	*17.5*	0.0	0.0	8.3	8.6	**51.0**	*30.7*	1.4	0.0
QLCRM-cl	0.0	0.1	**78.3**	*21.6*	0.0	0.0	8.3	8.5	**47.8**	*34.8*	0.5	0.0
Sc E												
QLCRM	0.0	0.0	*8.8*	**90.5**	0.6	0.0	8.4	8.4	*15.0*	**60.4**	7.8	0.1
QCRM-LB	0.0	0.0	*11.1*	**88.5**	0.4	0.0	8.4	8.4	*16.8*	**59.3**	7.0	0.1
QCRM	0.0	0.0	*8.2*	**91.4**	0.4	0.0	8.3	8.4	*14.6*	**62.4**	6.4	0.0
QCRM-EF	0.0	0.0	*8.5*	**90.8**	0.7	0.0	8.3	8.4	*15.0*	**60.7**	7.6	0.1
QLCRM-cl	0.0	0.0	*6.2*	**93.3**	0.5	0.0	8.3	8.4	*14.3*	**64.5**	4.5	0.0
Sc F												
QLCRM	0.0	0.0	2.7	**80.7**	*16.5*	0.0	8.4	8.5	13.1	**50.9**	*18.5*	0.6
QCRM-LB	0.0	0.0	3.5	**83.8**	*12.7*	0.0	8.4	8.6	14.1	**51.9**	*16.4*	0.6
QCRM	0.0	0.0	2.6	**84.7**	*12.7*	0.0	8.3	8.4	12.9	**54.9**	*15.3*	0.2

Suite page suivante

	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
QCRM-EF	0.0	0.0	2.7	**81.4**	*15.9*	0.0	8.4	8.5	13.2	**51.8**	*17.7*	0.5
QLCRM-cl	0.0	0.0	2.4	**86.5**	*11.1*	0.0	8.4	8.5	13.4	**58.8**	*10.9*	0.0
Sc G												
QLCRM	0.0	0.0	0.0	2.6	**79.6**	*17.8*	8.4	8.3	8.4	12.3	**45.0**	*17.6*
QCRM-LB	0.0	0.0	0.0	3.6	**81.3**	*15.1*	8.4	8.3	8.4	13.0	**45.8**	*16.1*
QCRM	0.0	0.0	0.0	3.8	**85.6**	*10.6*	8.3	8.3	8.4	13.9	**50.8**	*10.2*
QCRM-EF	0.0	0.0	0.0	2.9	**80.3**	*16.8*	8.3	8.3	8.4	12.9	**46.1**	*15.9*
QLCRM-cl	0.0	0.0	0.0	6.8	**90.6**	*2.6*	8.3	8.3	8.5	18.8	**54.1**	*1.9*
Sc H												
QLCRM	0.0	0.0	0.0	1.8	**82.5**	*15.7*	8.3	8.4	8.7	15.0	**48.0**	*11.5*
QCRM-LB	0.0	0.0	0.0	2.5	**85.3**	*12.2*	8.4	8.4	8.8	15.8	**48.6**	*10.1*
QCRM	0.0	0.0	0.0	3.6	**88.7**	*7.7*	8.3	8.4	8.7	18.4	**50.9**	*5.3*
QCRM-EF	0.0	0.0	0.0	2.6	**84.1**	*13.3*	8.3	8.4	8.8	16.5	**48.7**	*9.3*
QLCRM-cl	0.0	0.0	0.0	15.8	**83.5**	*0.7*	8.3	8.4	9.1	30.8	**43.0**	*0.4*

Sc : Scénario.
QLCRM : la méthode Quasi-LCRM avec un modèle **logistique** dans un cadre **fréquentiste**.
QCRM-LB : la méthode Quasi-LCRM avec un modèle **logistique**, avec une ordonnée à l'origine fixée à 3, dans un cadre **bayésien**. La distribution *a priori* de la pente est une distribution exponentielle du paramètre égal à 1.
QCRM : la méthode Quasi-CRM de Yuan et al avec un modèle **empirique** dans un cadre **bayésien**.
QCRM-EF : la méthode Quasi-CRM avec un modèle **empirique** dans un cadre **fréquentiste**.
QLCRM-cl : la méthode Quasi-LCRM avec une fonction de lien **cloglog** dans un cadre **fréquentiste**.
Les valeurs en gras correspondent aux résultats obtenus à la dose cible.

Les valeurs en italiques correspondent aux résultats obtenus à la dose la plus proche de la dose cible.

5.2.4 Comparaison de la méthode QLCRM aux méthodes existantes

Performance des méthodes avec une taille fixe d'échantillon

La distribution du pourcentage de sélection de chaque dose comme dose à recommander (DR) obtenue avec la QLCRM est très étroite autour de la vraie DR. Le PCS est très élevé, variant de 80% à 90% selon les scénarios étudiés. Dans tous les cas, plus de 90% des recommandations correspondent à la vraie DR ou à la dose la plus proche de la DR. Le risque de recommander des doses trop toxiques est faible avec la méthode QLCRM, avec 0% de recommandations à la dose DR+2 dans tous les scénarios. Le risque de recommander des doses trop basses est également très faible : la méthode n'a jamais recommandé la dose RD-2.

Avec la méthode QLCRM, plus de 45% des 36 patients inclus dans l'essai sont traités à la vraie DR. Le processus d'escalade de dose est efficace présentant un pourcentage très proche de 8.33% (=3/36) de patients traités à des doses inférieures à la dose *RD-1* dans tous les scénarios où la vraie DR est au-dessus de la deuxième dose (d_2). Ceci signifie que la dose est escaladée après l'inclusion de chaque cohorte de trois patients.

Les méthodes QCRM, UA et EID présentent également une très bonne performance pour tous les scénarios étudiés. Les PCS varient de 82% à 91% pour la méthode QCRM, de 78% à 93% pour la méthode UA, et de 69% à 85% pour la méthode EID. Les PCS des méthodes QLCRM et QCRM sont très similaires pour les différents scénarios considérés, avec une différence de PCS variant de −6% à +2%. La méthode QCRM

présente des résultats légèrement meilleurs que la méthode QLCRM pour six des huit scénarios étudiés.

La performance de la méthode QLCRM est supérieure, de plus de 9% en termes de PCS, à celle de la méthode EID dans sept des huit scénarios, atteignant une différence de PCS de 16% pour les scénarios B et E.

Le comportement de la méthode UA, en comparaison de la QLCRM, est très variable en fonction des scénarios, avec des différences de PCS allant de -10% à $+7\%$. Dans les quatre scénarios A, C, F et G pour lesquels les deux méthodes sont très proches en termes de PCS (différence inférieure à 1%), on remarque que la DR-1 est plus fréquemment recommandée avec UA qu'avec QLCRM, tandis que l'inverse est observé pour la dose DR+1. A noter que ces quatre scénarios correspondent à des scénarios pour lesquels le score associé à la vraie dose à recommander est égal exactement au score-cible. Ces scénarios se distinguent par la position de la dose à recommander. Au vu de ces résultats, il semble que la position de la dose à recommander a peu d'impact sur la performance des méthodes.

La distribution des doses allouées est similaire entre les méthodes QLCRM et QCRM. Avec ces méthodes, dans six des huit scénarios étudiés, plus de la moitié des patients sont traités à la vraie dose à recommander. Dans tous les scénarios sauf un (scénario D), les méthodes paramétriques (QLCRM et QCRM) allouent plus de patients à la vraie DR que les méthodes non paramétriques (UA et EID). La méthode UA semble être une méthode conservatrice dans le processus d'escalade de dose : elle inclut plus de patients à la DR-1, même si cette dose n'est pas la dose la plus proche de la dose cible. Pour la méthode EID, la distribution des doses allouées est plus étalée au tour de la DR. Comparée aux trois autres méthodes, la méthode EID inclut plus de patients à la dose DR+2, et ceci pour tous les scénarios étudiés.

Table 5.4 – Performance des différentes méthodes existantes, pour $n = 36$ patients

	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Sc A												
QLCRM	3.4	**85.9**	*10.7*	0.0	0.0	0.0	19.4	**62.6**	*17.4*	0.6	0.0	0.0
QCRM	2.3	**83.6**	*14.1*	0.0	0.0	0.0	14.2	**63.1**	*22.0*	0.7	0.0	0.0
UA	4.4	**86.4**	*9.1*	0.1	0.0	0.0	29.3	**59.2**	*11.3*	0.3	0.0	0.0
EID	11.3	**73.4**	*15.0*	0.3	0.0	0.0	21.5	**58.7**	*18.1*	1.7	0.1	0.0
Sc B												
QLCRM	0.0	*12.7*	**85.3**	2.0	0.0	0.0	9.1	*21.1*	**60.8**	8.9	0.2	0.0
QCRM	0.0	*10.5*	**87.4**	2.0	0.0	0.0	8.5	*18.7*	**63.1**	9.6	0.1	0.0
UA	0.0	*18.9*	**79.6**	1.5	0.0	0.0	10.8	*36.7*	**48.7**	3.8	0.1	0.0
EID	0.3	*26.0*	**69.2**	4.4	0.1	0.0	9.1	*30.4*	**51.5**	8.4	0.5	0.0
Sc C												
QLCRM	0.0	3.0	**83.8**	*13.3*	0.0	0.0	9.2	14.9	**57.0**	*18.3*	0.7	0.0
QCRM	0.0	2.3	**84.1**	*13.6*	0.0	0.0	8.5	13.2	**58.5**	*19.3*	0.4	0.0
UA	0.0	6.2	**84.4**	*9.4*	0.1	0.0	10.8	27.5	**51.5**	*10.0*	0.3	0.0
EID	0.3	12.6	**71.3**	*15.5*	0.4	0.0	9.1	21.2	**50.9**	*17.1*	1.7	0.1
Sc D												
QLCRM	0.0	0.1	**83.0**	*17.0*	0.0	0.0	8.4	8.6	**51.4**	*30.3*	1.4	0.0
QCRM	0.0	0.0	**82.5**	*17.5*	0.0	0.0	8.3	8.5	**50.2**	*32.1*	1.0	0.0
UA	0.0	0.0	**93.4**	*6.5*	0.0	0.0	8.5	17.2	**61.0**	*13.2*	0.1	0.0
EID	0.0	0.9	**84.6**	*14.2*	0.2	0.0	8.3	10.2	**62.8**	*17.5*	1.1	0.0
Sc E												
QLCRM	0.0	0.0	*8.8*	**90.5**	0.6	0.0	8.4	8.4	*15.0*	**60.4**	7.8	0.1
QCRM	0.0	0.0	*8.2*	**91.4**	0.4	0.0	8.3	8.4	*14.6*	**62.4**	6.4	0.0
UA	0.0	0.0	*16.0*	**83.6**	0.4	0.0	8.4	9.7	*34.4*	**45.0**	2.5	0.0
EID	0.0	0.1	*23.3*	**74.7**	1.8	0.1	8.3	8.6	*26.9*	**50.4**	5.6	0.2
Sc F												
QLCRM	0.0	0.0	2.7	**80.7**	*16.5*	0.0	8.4	8.5	13.1	**50.9**	*18.5*	0.6
QCRM	0.0	0.0	2.6	**84.7**	*12.7*	0.0	8.3	8.4	12.9	**54.9**	*15.3*	0.2
UA	0.0	0.0	8.1	**81.4**	*10.4*	0.1	8.6	11.0	26.9	**44.9**	*8.5*	0.2
EID	0.0	0.4	13.4	**69.8**	*16.0*	0.5	8.3	9.2	20.6	**45.8**	*14.6*	1.6
Sc G												
QLCRM	0.0	0.0	0.0	2.6	**79.6**	*17.8*	8.4	8.3	8.4	12.3	**45.0**	*17.6*
QCRM	0.0	0.0	0.0	3.8	**85.6**	*10.6*	8.3	8.3	8.4	13.9	**50.8**	*10.2*
UA	0.0	0.0	0.0	10.0	**79.1**	*10.9*	8.5	8.6	11.0	25.3	**39.3**	*7.4*
EID	0.0	0.0	0.3	13.3	**70.5**	*15.8*	8.3	8.4	9.1	19.0	**40.7**	*14.6*
Sc H												
QLCRM	0.0	0.0	0.0	1.8	**82.5**	*15.7*	8.3	8.4	8.7	15.0	**48.0**	*11.5*
QCRM	0.0	0.0	0.0	3.6	**88.7**	*7.7*	8.3	8.4	8.7	18.4	**50.9**	*5.3*
UA	0.0	0.0	0.0	1.9	**78.2**	*19.9*	8.4	9.4	11.0	18.4	**39.5**	*13.3*
EID	0.0	0.0	0.1	4.9	**72.4**	*22.6*	8.3	8.4	8.8	13.7	**41.1**	*19.7*

Sc : Scénario.
QLCRM : méthode Quasi-LCRM (notre proposition). QCRM : méthode Quasi-CRM de Yuan et al.
UA : Unified Algorithm (méthode d'Ivanova et al.). EID : Extended Isotonic Design (méthode de Chen et al.).

Les valeurs en gras correspondent aux résultats obtenus à la dose cible.

Les valeurs en italiques correspondent aux résultats obtenus à la dose la plus proche de la dose cible.

Distribution des scores normalisés et du nombre de DLTs

Les distributions des scores normalisés et du nombre de DLTs en cours d'essai permettent de quantifier la toxicité observée selon les différentes méthodes étudiées. A titre d'exemple, la figure 5.2 illustre la boîte à moustaches de ces distributions pour les scénarios A, C et G, le scénario G étant moins toxique que les deux autres.

Les distributions des scores nTTP et du nombre de DLTs observés dans les essais

simulés sont très similaires pour les différentes méthodes pour les huit scénarios étudiés, à l'exception de la méthode UA qui expose les patients à moins de DLTs, ce qui confirme le caractère conservateur de cette méthode.

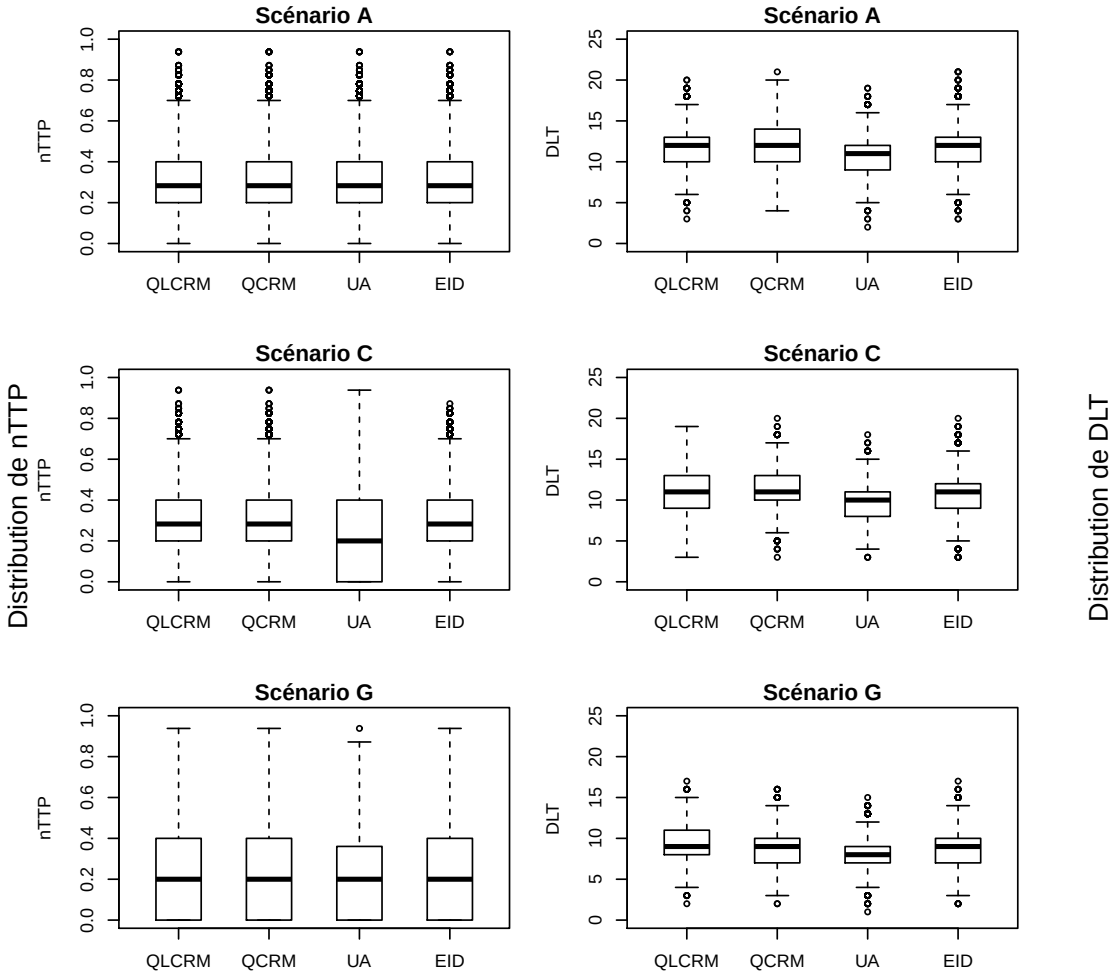


Figure 5.2 – Boîte à moustaches de la distribution des scores nTTP et du nombre de DLTs selon les méthodes QLCRM, QCRM, EID et UA

L'extrémité inférieure de la boite à moustache est définie par $Q_1 - 1.5 * IQR$, et l'extrémité maximale est définie par $Q_3 + 1.5 * IQR$. Où Q_1 et Q_3 sont respectivement le premier et le troisième quartile. $IR = Q_3 - Q_1$.
QLCRM : méthode Quasi-LCRM (notre proposition).
QCRM : méthode Quasi-CRM de Yuan et al., [3].
UA : Unified Algorithm (méthode d'Ivanova et al., [5]).
EID : Extended Isotonic Design (méthode de Chen et al., [2]).

Convergence des méthodes

La figure 5.3 illustre la convergence des méthodes en termes de pourcentage de sélection correcte, en faisant varier le nombre de sujets inclus dans l'essai de $n = 15$ à $n = 99$. Comme attendu, les méthodes paramétriques QLCRM et QCRM convergent vers la vraie DR pour tous les scénarios étudiés. La méthode UA converge également vers la vraie DR. Pour $n = 99$, le PCS des ces trois méthodes est supérieur à 95% dans sept des huit scénarios : pour le scénario B, le PCS de la méthode QCRM est égal à 85% quand n=99, tandis qu'il est supérieur à 93% pour les méthodes QLCRM et UA. Contrastant avec les excellents résultats observés avec ces trois méthodes, la méthode EID converge lentement, présentant le plus faible PCS quand n est grand.

Etant donné que les essais de phase I incluent un nombre faible de patients, généralement moins de 40 patients, le comportement du PCS des différentes méthodes a été également étudié à partir de $n = 15$. Les méthodes QLCRM et QCRM présentent des résultats très similaires. Ces résultats se distinguent de la méthode UA dans trois des huit scénarios étudiés pour $n < 40$: la méthode UA présente une performance supérieure à celle des méthodes paramétriques dans le scénario D, et elle présente une performance inférieure à celle de ces méthodes pour les scénarios B et E. Pour ces trois méthodes, le PCS est supérieur à 60% lorsque la taille de l'échantillon est supérieure à 24, quel que soit le scénario. D'une façon générale, la méthode EID présente une moins bonne performance que les autres méthodes. Elle apparaît meilleure que la QLCRM et QCRM pour $n < 40$, seulement dans un scénario sur huit (scénario D).

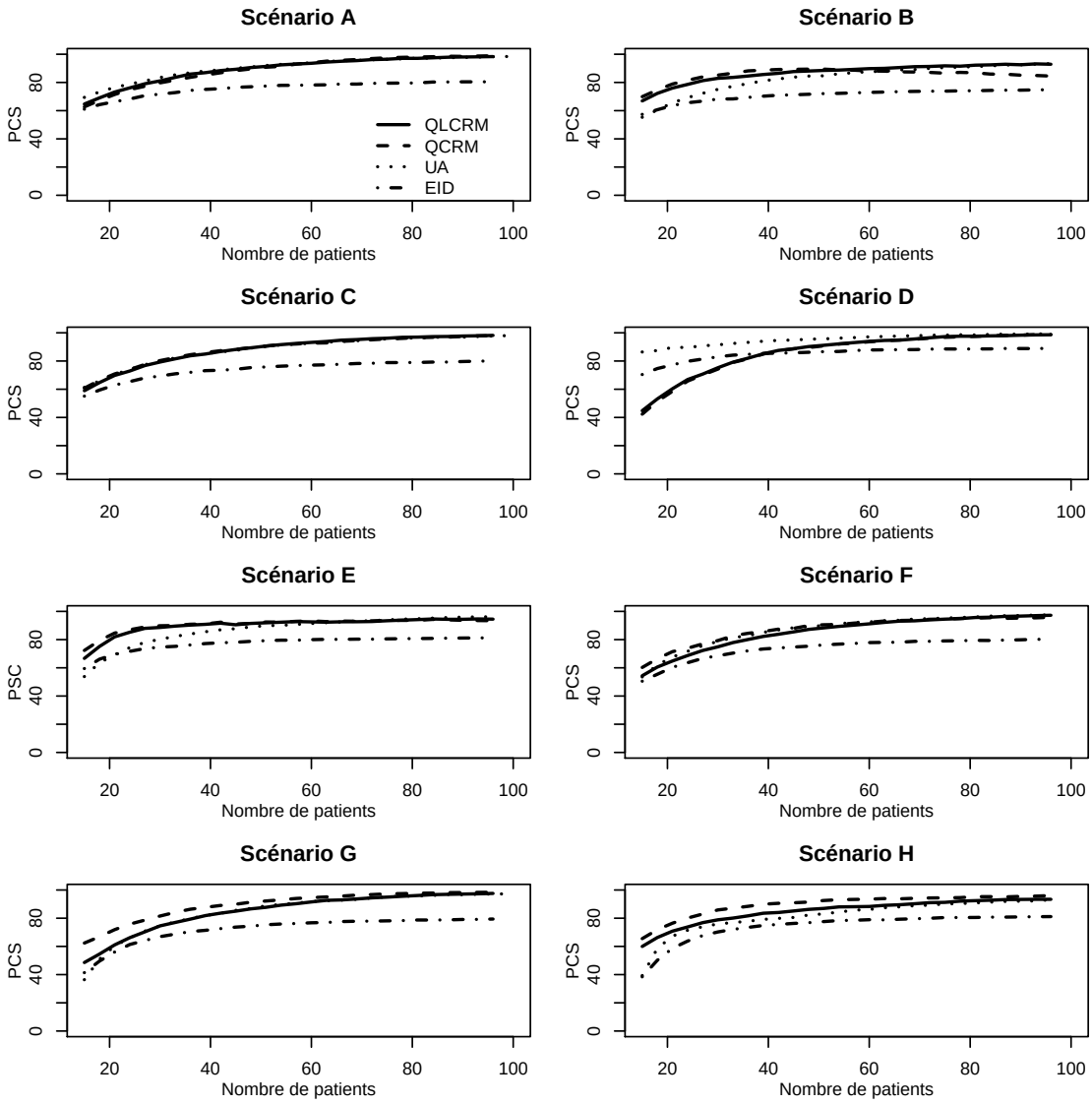


FIGURE 5.3 – Convergence des différentes méthodes QLCRM, QCRM, EID et UA

QLCRM : méthode Quasi-LCRM (notre proposition).
QCRM : méthode Quasi-CRM de Yuan et al., [3].
UA : Unified Algorithm (méthode d'Ivanova et al., [5]).
EID : Extended Isotonic Design (méthode de Chen et al., [2]).

Analyse de sensibilité de la convergence des différentes méthodes à la règle de décision de la dose à recommander

Comme énoncé précédemment, la comparaison des différentes méthodes repose jusqu'à maintenant sur la même définition de la dose à recommander, à savoir la dose qui pourrait être administrée à la prochaine cohorte hypothétique, une fois le nombre maximum de sujets atteint. Pour les méthodes QLCRM, QCRM et EID, cette dose à recommander peut donc être une dose qui n'a jamais été allouée en cours d'essai.

Nous avons effectué une analyse de sensibilité pour évaluer l'impact de cette définition sur ces trois méthodes. Dans cette analyse, la DR est définie parmi les doses allouées dans l'essai. Nous n'avons pas observé d'effet significatif de cette règle de décision sur la performance des méthodes en termes de PCS en variant n de 15 à 99 (cf. figure 5.4). La convergence des différentes méthodes est la même à partir de $n = 20$. Pour les scénarios G et H pour lesquels la vraie DR est la cinquième dose, le PCS est très élevé avec les méthodes paramétriques pour $n = 15$, soit après avoir traité cinq cohortes.

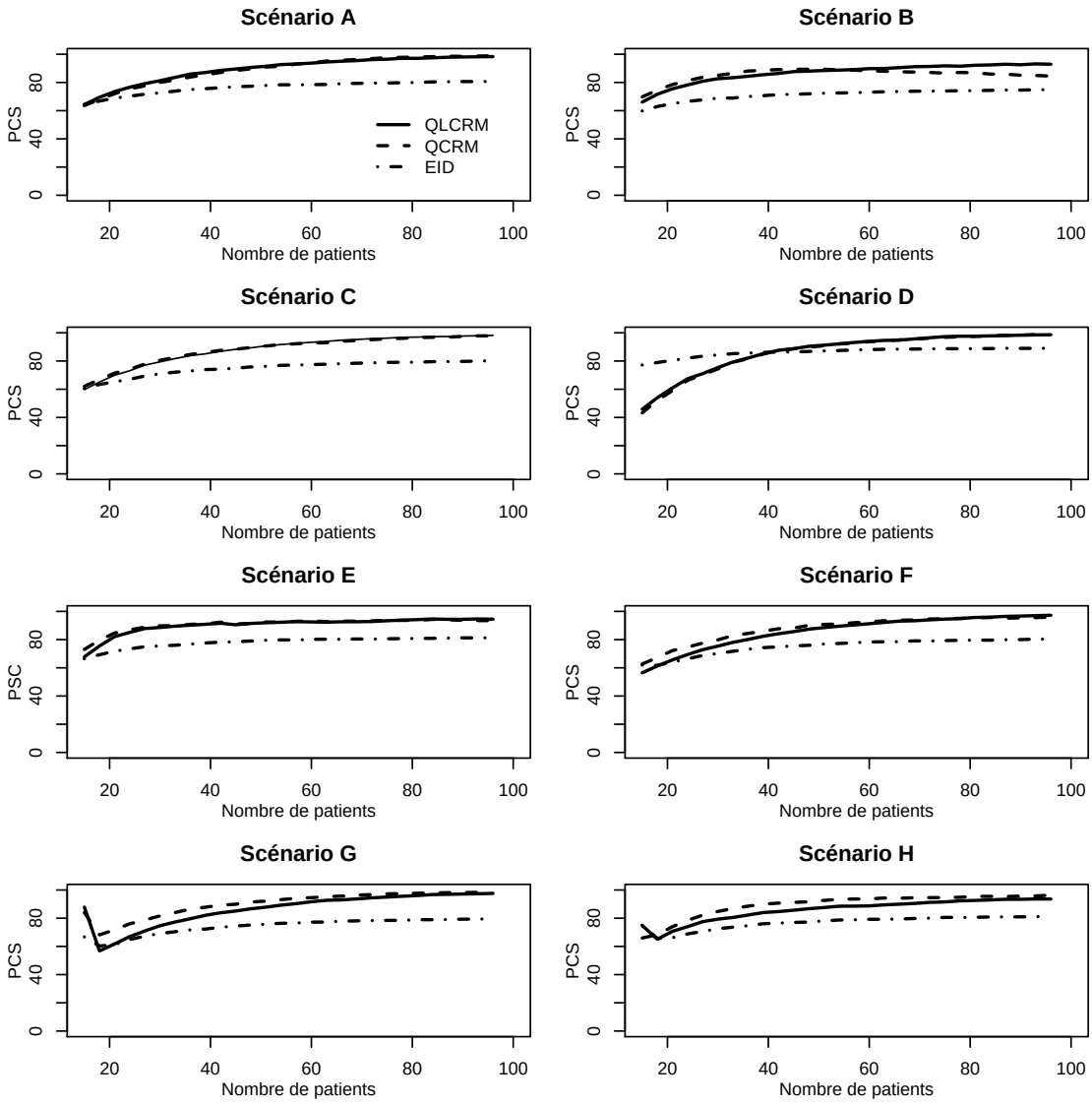


FIGURE 5.4 – Analyse de sensibilité des différentes méthodes à la règle de décision de la dose à recommander

QLCRM : méthode Quasi-LCRM (notre proposition).
QCRM : méthode Quasi-CRM de Yuan et al., [3].
EID : Extended Isotonic Design (méthode de Chen et al., [2]).
Cette analyse de sensibilité n'inclut pas la méthode UA car cette méthode ne peut pas recommander une dose qui n'a jamais été explorée.

Choix du modèle de travail pour les méthodes paramétriques

Dans les principaux résultats, les méthodes QLCRM et QCRM utilisent les modèles de travail obtenus par la fonction "getprior" du package *dfcrm* du logiciel *R* [41]. Nous avons comparé ces résultats avec ceux obtenus en utilisant le modèle de travail publié par Yuan et al. [3]. Le choix du modèle de travail entre ces deux options a peu d'impact

sur les résultats de deux méthodes, sauf pour le scénario H où la dose à recommander était sur-estimée en utilisant le modèle de travail de Yuan (les résultats sont détaillés dans l'annexe 8.3).

5.2.5 Evaluation des méthodes basées sur le score de toxicité en comparaison avec les méthodes de type CRM basées sur le critère binaire DLT

Dans cette partie, nous avons considéré les scénarios A, C, F et G, décrits dans le tableau 4.1, pour lesquels la vraie DR est la même pour les deux critères de toxicité (score et DLT). Le tableau 5.5 montre que les méthodes basées sur le score nTTP présentent une meilleure performance comparée aux méthodes CRM basées sur le critère DLT, et ceci pour tous les scénarios étudiés. Le PCS des méthodes LCRM et CRM varie, en fonction des scénarios, de 50.9% à 55.9%, tandis que celui des méthodes guidées par le score varie de 69.8% (pour la méthode EID) à 86.4%.

Le risque de recommander des doses supérieures à la DR est mieux contrôlé avec les méthodes guidées par le score nTTP qu'avec les méthodes basées sur la DLT. Les méthodes basées sur le score présentent moins de 17% de recommandations à des doses supérieures à la vraie DR *versus* 30.3 à 38.5% pour les méthodes LCRM et CRM. En particulier, il y a moins de 0.5% de recommandations à la dose DR+2, avec les méthodes basées sur le score (scénarios A, C et F) et aucune recommandation à des doses plus élevées (scénarios A et C). Par contre, les méthodes classiques LCRM et CRM présentent 4.3 à 7% de recommandations à la dose DR+2 et 0.3 à 1.1% de recommandations à des doses plus élevées.

Les méthodes QLCRM et QCRM incluent plus de patients à la vraie DR ($\geq 45\%$) que les méthodes classiques de type CRM (29.8 et 41.7%). Les méthodes guidées par le critère DLT incluent également plus de patients à des doses supérieures à la DR. Par exemple, elles incluent entre 21.9 et 29% de patients à la dose DR+1 *versus* un pourcentage variant de 7.4 à 19.3% avec les méthodes basées sur le score (cf. table 5.5). Avec les méthodes LCRM et CRM, ce pourcentage reste non négligeable à la dose DR+2, variant de 6.1 à 8.9% *versus* 0.2 à 1.7% pour les méthodes basées sur le score.

Comme attendu, un des avantages à utiliser un score de toxicité pour la méthode QLCRM développé dans un cadre fréquentiste est que l'estimation du paramètre du modèle logistique a été possible dès la première cohorte.

Table 5.5 – Performance des méthodes basées sur le score de toxicité TTP et des méthodes CRM basées sur la DLT, pour $n = 36$

	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
Méthode	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Sc A												
QLCRM	3.4	**85.9**	*10.7*	0.0	0.0	0.0	19.4	**62.5**	*17.4*	0.6	0.0	0.0
QCRM	2.3	**83.6**	*14.1*	0.0	0.0	0.0	14.1	**63.1**	*22.0*	0.7	0.0	0.0
UA	4.4	**86.4**	*9.1*	0.1	0.0	0.0	29.3	**59.1**	*11.3*	0.3	0.0	0.0
EID	11.3	**73.4**	*15.0*	0.3	0.0	0.0	21.5	**58.7**	*18.1*	1.7	0.1	0.0
LCRM	17.1	**52.7**	*25.2*	4.4	0.6	0.1	30.8	**38.5**	*21.9*	6.6	1.4	0.2
CRM	15.2	**53.5**	*26.7*	4.3	0.3	0.0	25.1	**41.7**	*25.2*	6.9	1.1	0.1
Sc C												
QLCRM	0.0	3.0	**83.8**	*13.3*	0.0	0.0	9.2	14.8	**57.0**	*18.3*	0.7	0.0
QCRM	0.0	2.3	**84.1**	*13.6*	0.0	0.0	8.5	13.2	**58.5**	*19.3*	0.4	0.0
UA	0.0	6.2	**84.4**	*9.4*	0.1	0.0	10.8	27.5	**51.5**	*10.0*	0.2	0.0
EID	0.3	12.6	**71.3**	*15.5*	0.4	0.0	9.1	21.2	**50.9**	*17.1*	1.7	0.1
LCRM	0.1	13.2	**51.9**	*28.0*	5.6	1.1	11.2	20.8	**35.9**	*22.2*	7.9	2.0
CRM	0.0	12.8	**53.9**	*28.2*	4.6	0.4	9.4	20.4	**38.5**	*24.0*	6.6	1.0
Sc F												
QLCRM	0.0	0.0	2.7	**80.7**	*16.5*	0.0	8.4	8.5	13.1	**50.9**	*18.5*	0.6
QCRM	0.0	0.0	2.6	**84.7**	*12.7*	0.0	8.3	8.4	12.9	**54.9**	*15.3*	0.2
UA	0.0	0.0	8.1	**81.4**	*10.4*	0.1	8.6	11.0	26.9	**44.9**	*8.5*	0.2
EID	0.0	0.4	13.4	**69.8**	*16.0*	0.5	8.3	9.2	20.6	**45.8**	*14.6*	1.6
LCRM	0.0	0.1	11.9	**51.4**	*29.6*	7.0	8.6	9.2	17.6	**32.4**	*23.3*	8.9
CRM	0.0	0.1	12.2	**54.2**	*28.6*	4.9	8.3	9.0	18.0	**35.8**	*22.7*	6.1
Sc G												
QLCRM	0.0	0.0	0.0	2.6	**79.6**	*17.8*	8.3	8.3	8.4	12.3	**45.0**	*17.6*
QCRM	0.0	0.0	0.0	3.8	**85.6**	*10.6*	8.3	8.3	8.4	13.9	**50.8**	*10.2*
UA	0.0	0.0	0.0	10.0	**79.1**	*10.9*	8.5	8.6	11.0	25.3	**39.3**	*7.4*
EID	0.0	0.0	0.3	13.3	**70.5**	*15.8*	8.3	8.4	9.1	19.0	**40.7**	*14.6*
LCRM	0.0	0.0	0.1	10.6	**50.9**	*38.5*	8.5	8.3	8.7	14.7	**29.8**	29.8
CRM	0	0.0	0.1	12.0	**55.9**	*32.0*	8.3	8.3	8.8	16.3	**33.5**	*24.8*

Sc : Scénario.
Méthodes basées sur le score :
QLCRM : méthode Quasi-LCRM (notre proposition).
QCRM : méthode Quasi-CRM de Yuan et al., [3].
UA : Unified Algorithm (méthode d'Ivanova et al., [5]).
EID : Extended Isotonic Design (méthode de Chen et al., [2]).
Méthodes basées sur la DLT :
LCRM : méthode CRM avec un modèle logistique dans un cadre fréquentiste.
CRM : méthode CRM avec un modèle empirique dans un cadre bayésien.
Les valeurs en gras correspondent aux résultats obtenus à la dose cible.

Les valeurs en italiques correspondent aux résultats obtenus à la dose la plus proche de la dose cible.

5.3 Etude de sensibilité des méthodes aux erreurs d'observation des grades

Dans cette partie, nous avons étudié la sensibilité des méthodes basées sur les deux critères de toxicité (score et DLT) aux erreurs d'observation des grades. Pour les méthodes basées sur le score, la méthode EID n'a pas été prise en compte car elle présente des résultats moins bons que les autres méthodes (QLCRM, QCRM et UA) en absence d'erreur d'observation.

5.3.1 Impact du sous-classement des grades sur les différentes méthodes étudiées

La figure 5.5 illustre la sensibilité des différentes méthodes aux erreurs de sous-classement des grades (*under-grading*). Il y a peu d'impact des erreurs de sous-classement sur les méthodes basées sur le critère nTTP lorsque ces erreurs touchent seulement les grades définissant la DLT. Pour les différentes méthodes basées sur le score nTTP, la diminution absolue du PCS est inférieure à 13% lorsque la probabilité d'erreur de sous-classement des grades, P_U, varie de 0 à 25%. Comme détaillé dans le tableau 5.6 pour le scénario F, le PCS reste supérieur à 69% pour toutes les méthodes guidées par le score nTTP, pour des valeurs d'erreur P_U inférieures ou égales à 25%.

Lorsque les erreurs touchent tous les grades de manière égale et indépendante, la performance des méthodes basées sur le critère nTTP diminue de façon monotone avec l'augmentation de la valeur de probabilité P_U. Pour toutes ces méthodes à l'exception de la méthode UA, le PCS est égal à environ la moitié de celui observé dans une situation idéale lorsque l'erreur $P_U = 0$. Par exemple pour le scénario F, le PCS diminue de 80.7% à 35.9% avec la méthode QLCRM (cf. tables 5.5 et 5.6). L'impact de l'erreur de sous-classement des grades sur la performance des différentes méthodes varie selon les scénarios. Ceci est particulièrement visible pour la méthode QLCRM, où la diminution de la performance est plus importante lorsque la vraie dose à recommander est parmi les doses les plus élevées de l'ensemble de doses à explorer (scénarios F et G). La baisse de performance de sous-classement des grades est moins importante avec la méthode UA et ceci pour les quatre scénarios considérés.

Les erreurs de sous-classement des grades ont également un impact important sur les méthodes classiques de type CRM guidées par la DLT, avec une diminution monotone du PCS lorsque les valeurs de P_U augmente. De manière similaire aux méthodes basées sur le critère nTTP, le PCS observé avec une probabilité de $P_U = 25\%$ est égal à environ la moitié de celui observé dans la situation idéale pour laquelle la probabilité de $p_U = 0$. Par exemple, le PCS diminue de 51.4% à 24.8% avec la méthode LCRM pour le scénario F (cf. les tableaux 5.5 et 5.6).

Le PCS des méthodes basées sur le critère nTTP reste supérieur à celui observé avec les méthodes guidées par la DLT, et ceci pour tous les scénarios considérés et pour les valeurs de P_U étudiées ($p_U \leq 0.25$).

Le détail des résultats pour le scénario F (cf. table 5.6) illustre le sens de biais de recommandation des doses pour une probabilité croissante d'erreur de sous-classement. Lorsque ces erreurs touchent tous les grades de manière indépendante et égale, la diminution du PCS est liée à un biais de recommandation vers les doses plus élevées que la DR. Il s'agit d'une sur-estimation de la DR. Cependant, il est important de noter que le risque de recommander des doses plus élevées que la DR est moins important avec les méthodes basées sur le score qu'avec les méthodes classiques de type CRM basées sur la DLT. Par exemple, pour le scénario F, lorsque la probabilité de $p_U = 25\%$, le pourcentage de recommandation de la dose DR+2 est respectivement de 3.1% et 1.1% pour les méthodes QLCRM et QCRM, contre 31.8% et 24.7% pour les méthodes LCRM et CRM.

La distribution des doses attribuées au cours de l'essai est également différente entre les méthodes comparées. Il y a plus de patients traités à la dose DR+2 avec les méthodes LCRM et CRM guidées par la DLT qu'avec les méthodes guidées par le score, et ceci pour les différentes valeurs de P_U. Par exemple, lorsque la probabilité $p_U = 25\%$, ce pourcentage est égal, respectivement, à 22% et 16.9% pour les méthodes LCRM et CRM, *versus* moins de 5% pour les méthodes basées sur le score (cf. 5.6).

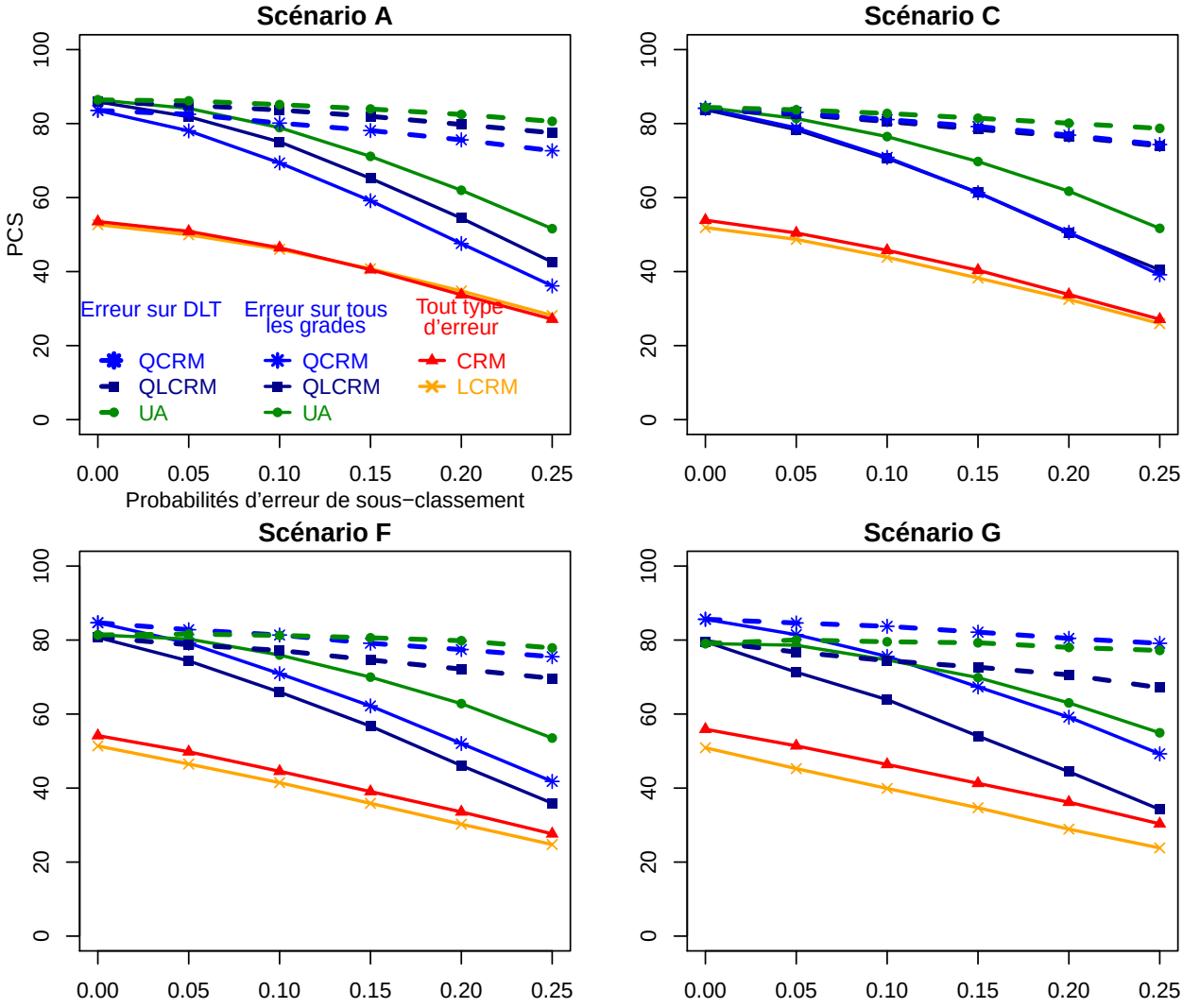


FIGURE 5.5 – Impact du sous-classement des grades sur les différentes méthodes étudiées

QCRM : méthode Quasi-CRM de Yuan et al., [3].
QLCRM : méthode Quasi-LCRM (notre proposition).
UA : Unified Algorithm (méthode d'Ivanova et al., [5]).

Table 5.6 – Impact du sous-classement des grades sur les différentes méthodes étudiées pour le scénario F (vraie DR=d_4), pour $n = 36$

Méthodes	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Méthodes guidées par le score, Erreur touchant les grades définissant la DLT												
QLCRM												
$p_U = 0.05$	0	0	2.7	**78.7**	*18.4*	0.1	8.4	8.5	12.7	**49.7**	*20.0*	0.7
$p_U = 0.10$	0	0	2.2	**77.2**	*20.5*	0.1	8.4	8.5	12.4	**48.9**	*21.0*	0.8
$p_U = 0.15$	0	0	1.9	**74.6**	*23.4*	0.1	8.4	8.5	12.1	**47.9**	*22.3*	0.9
$p_U = 0.20$	0	0	1.7	**72.2**	*26.1*	0.1	8.4	8.5	11.7	**47.0**	*23.5*	0.9
$p_U = 0.25$	0	0	1.4	**69.6**	*28.8*	0.2	8.4	8.5	11.5	**45.8**	*24.8*	1.1
QCRM												
$p_U = 0.05$	0	0	2.5	**82.8**	*14.6*	0.0	8.3	8.4	12.5	**53.8**	*16.7*	0.2
$p_U = 0.10$	0	0	2.1	**81.4**	*16.5*	0.0	8.3	8.4	12.2	**53.1**	*17.8*	0.2
$p_U = 0.15$	0	0	1.8	**79.1**	*19.0*	0.0	8.3	8.4	11.9	**52.3**	*18.9*	0.2
$p_U = 0.20$	0	0	1.5	**77.4**	*21.0*	0.0	8.3	8.4	11.6	**51.5**	*19.9*	0.3
$p_U = 0.25$	0	0	1.3	**75.5**	*23.2*	0.0	8.3	8.4	11.3	**50.6**	*21.0*	0.3
UA												
$p_U = 0.05$	0	0	7.3	**81.6**	*10.9*	0.1	8.6	11.0	25.9	**45.2**	*9.1*	0.2
$p_U = 0.10$	0	0	6.4	**81.3**	*12.2*	0.1	8.6	10.9	25.0	**45.4**	*9.9*	0.2
$p_U = 0.15$	0	0	5.5	**80.6**	*13.8*	0.1	8.6	10.8	24.2	**45.4**	*10.7*	0.3
$p_U = 0.20$	0	0	4.5	**79.8**	*15.4*	0.2	8.6	10.7	23.4	**45.4**	*11.6*	0.3
$p_U = 0.25$	0	0	4.0	**77.9**	*17.8*	0.3	8.6	10.6	22.7	**45.2**	*12.6*	0.4
Méthodes guidées par le score, Erreur touchant tous les grades												
QLCRM												
$p_U = 0.05$	0	0	2.0	**74.4**	*23.6*	0.1	8.4	8.5	11.9	**47.5**	*22.8*	1.0
$p_U = 0.10$	0	0	1.2	**65.9**	*32.6*	0.3	8.4	8.5	11.0	**43.7**	*27.1*	1.4
$p_U = 0.15$	0	0	0.5	**56.7**	*42.0*	0.7	8.4	8.4	10.2	**39.6**	*31.2*	2.1
$p_U = 0.20$	0	0	0.3	**46.1**	*52.1*	1.6	8.4	8.4	9.7	**35.2**	*35.2*	3.1
$p_U = 0.25$	0	0	0.2	**35.9**	*60.8*	3.1	8.3	8.4	9.4	**30.8**	*38.5*	4.5
QCRM												
$p_U = 0.05$	0	0	1.9	**79.2**	18.9	0.0	8.3	8.4	11.8	51.9	**19.3**	*0.3*
$p_U = 0.10$	0	0	1.1	**70.9**	*27.9*	0.1	8.3	8.4	10.8	**48.3**	*23.6*	0.5
$p_U = 0.15$	0	0	0.5	**62.2**	*37.2*	0.2	8.3	8.4	10.1	**44.3**	*28.1*	0.8
$p_U = 0.20$	0	0	0.2	**52.0**	*47.2*	0.5	8.3	8.4	9.6	**39.7**	*32.7*	1.3
$p_U = 0.25$	0	0	0.1	**41.8**	*56.9*	1.1	8.3	8.4	9.3	**35.1**	*36.9*	2.0
UA												
$p_U = 0.05$	0	0	5.8	**80.3**	*13.8*	0.1	8.6	10.8	24.5	**45.3**	*10.5*	0.3
$p_U = 0.10$	0	0	4.0	**76.0**	*19.7*	0.3	8.5	10.6	22.5	**44.9**	*13.0*	0.4
$p_U = 0.15$	0	0	2.7	**70.0**	*26.6*	0.7	8.5	10.4	20.6	**43.9**	*15.9*	0.8
$p_U = 0.20$	0	0	1.5	**62.8**	*34.5*	1.2	8.5	10.1	19.1	**42.2**	*18.9*	1.1
$p_U = 0.25$	0	0	1.0	**53.5**	*43.2*	2.4	8.5	10.0	17.6	**40.0**	*22.2*	1.8
Méthodes guidées par la DLT,												
LCRM												
$p_U = 0.05$	0	0	9.5	**46.5**	*34.0*	10.0	8.6	9.1	16.2	**30.5**	*24.5*	11.0
$p_U = 0.10$	0	0	6.9	**41.5**	*37.9*	13.7	8.6	9.0	14.9	**28.5**	*25.8*	13.3
$p_U = 0.15$	0	0	5.0	**35.9**	*40.7*	18.4	8.6	8.9	13.8	**26.4**	*26.6*	15.8
$p_U = 0.20$	0	0	3.4	**30.2**	*41.4*	25.0	8.6	8.8	12.8	**24.1**	*26.8*	18.9
$p_U = 0.25$	0	0	2.3	**24.8**	*41.1*	31.8	8.6	8.7	12.0	**22.0**	*26.7*	22.0
CRM												
$p_U = 0.05$	0	0	9.5	**49.8**	*33.7*	6.9	8.3	9.0	16.7	**33.9**	*24.5*	7.6

Suite page suivante

Méthodes	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
$p_U = 0.10$	0	0	6.8	**44.5**	*39.0*	9.6	8.3	8.8	15.2	**31.8**	*26.4*	9.4
$p_U = 0.15$	0	0	5.1	**39.1**	*42.4*	13.4	8.3	8.8	14.2	**29.6**	*27.7*	11.5
$p_U = 0.20$	0	0	3.4	**33.6**	*44.3*	18.8	8.3	8.7	13.1	**27.3**	*28.5*	14.1
$p_U = 0.25$	0	0	2.2	**27.6**	*45.5*	24.7	8.3	8.7	12.3	**25.2**	*28.6*	16.9

Méthodes basées sur le score :
QLCRM : méthode Quasi-LCRM (notre proposition).
QCRM : méthode Quasi-CRM de Yuan et al., [3].
UA : Unified Algorithm (méthode d'Ivanova et al., [5]).
EID : Extended Isotonic Design (méthode de Chen et al., [2]).
Méthodes basées sur la DLT :
LCRM : méthode CRM avec un modèle logistique dans un cadre fréquentiste.
CRM : méthode CRM avec un modèle empirique dans un cadre bayésien.
Les valeurs en gras correspondent aux résultats obtenus à la dose cible.

Les valeurs en italiques correspondent aux résultats obtenus à la dose la plus proche de la dose cible.

5.3.2 Impact du sur-classement des grades sur les différentes méthodes étudiées

La figure 5.6 illustre la sensibilité des différentes méthodes aux erreurs de sur-classement des grades. Lorsque l'erreur de sur-classement touche seulement les grades définissant la DLT, elle a peu d'impact sur les méthodes guidées par le critère nTTP, et ceci pour tous les scénarios étudiés. Il y a même une légère amélioration des valeurs de PCS avec l'augmentation des valeurs de p_O. Lorsque les erreurs de sur-classement touchent tous les grades de manière indépendante et égale, la performance des méthodes basées sur le critère nTTP augmente initialement avec des petites valeurs de p_O, puis elle décline fortement avec l'augmentation de ces valeurs. Lorsque $p_O = 25\%$, le PCS des différentes méthodes basées sur le critère nTTP est inférieur à la moitié de celui observé dans la situation idéale pour laquelle $p_O = 0$. Par exemple, pour le scénario F, le PCS diminue de 80.7% à 30.5% avec la méthode QLCRM, de 84.7% à 30.2% avec la méthode QCRM et de 81.4% à 28% avec la méthode UA (cf. les tableaux 5.5 et 5.7). Cependant, dans tous les scénarios, lorsque la probabilité d'erreur de sur-classement est inférieur ou égal à 15%, le PCS de ces méthodes reste supérieur à 50% et équivalent ou supérieur à celui des méthodes de type CRM guidées par la DLT.

En comparaison des figures 5.6 et 5.5, il s'avère que la méthode UA est beaucoup plus sensible aux erreurs de sur-classement qu'aux erreurs de sous-classement des grades.

Les erreurs de sur-classement ont un faible impact sur les méthodes guidées par la DLT, avec une légère augmentation de PCS pour les petites valeurs de probabilité d'erreur de sur-estimation des grades ($p_O < 10\%$), et une légère baisse pour les valeurs de p_O supérieures ou égales à 15%.

Comme l'illustre l'exemple détaillé du scénario F (cf. tableau 5.7), la diminution du PCS des méthodes basées sur le critère nTTP, lorsque les erreurs touchent tous les grades de manière indépendante et égale, est liée à un biais important de recommandation vers les doses plus faibles. La dose *RD-1* devient plus fréquemment recommandée que la vraie DR. Cependant la dose DR-2 est très rarement recommandée. En ce

qui concerne les méthodes de type CRM pour lesquelles le PCS semble stable même pour une probabilité élevée d'erreur ($P_O = 0.25$), la répartition des pourcentages de recommandation de dose varie grandement en fonction de la probabilité d'erreur de sur-classement : alors que la dose DR+1 est la plus fréquemment recommandée après la vraie DR en l'absence d'erreur, la dose DR-1 qui devient la plus fréquemment recommandée quand P_O augmente.

La distribution des doses attribuées au cours de l'essai est également différente entre les méthodes guidées par le critère nTTP et les méthodes basées sur le critère DLT, présentant les mêmes tendances que la distribution des doses à recommander, commentées ci-dessus.

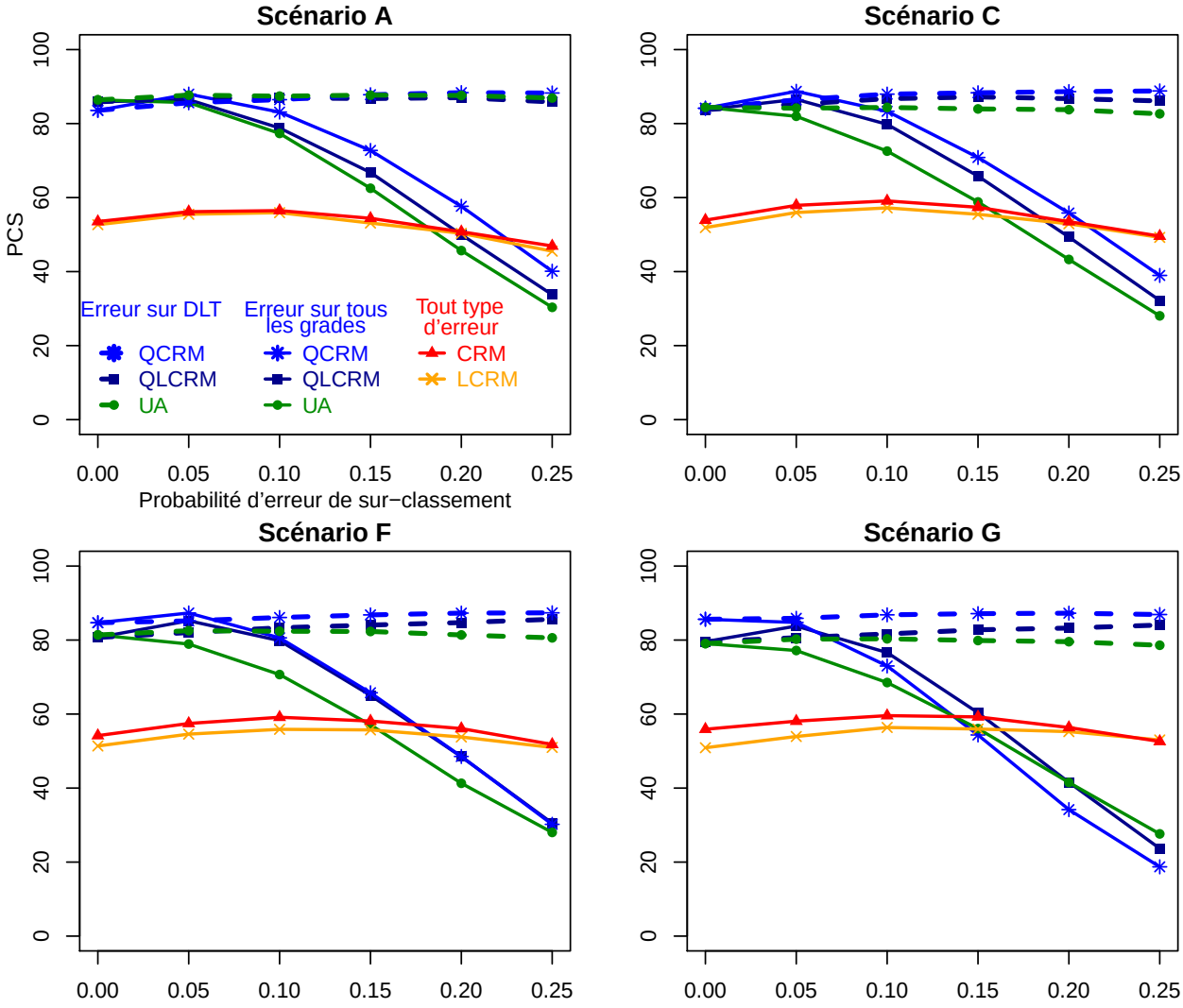


FIGURE 5.6 – Impact du sur-classement des grades sur les différentes méthodes étudiées

QCRM : méthode Quasi-CRM de Yuan et al., [3].
QLCRM : méthode Quasi-LCRM (notre proposition).
UA : Unified Algorithm (méthode d'Ivanova et al., [5]).

Table 5.7 – Impact du sur-classement des grades sur les différentes méthodes étudiées pour le scénario F (vraie DR=d_4), pour $n = 36$

Méthodes	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Méthodes guidées par le score, Erreur touchant les grades définissant la DLT												
QLCRM												
$p_O = 0.05$	0	0.0	4.0	**82.0**	*14.0*	0.0	8.4	8.5	13.6	**51.5**	*17.5*	0.6
$p_O = 0.10$	0	0.0	4.8	**83.3**	*11.8*	0.1	8.4	8.5	14.2	**52.1**	*16.3*	0.5
$p_O = 0.15$	0	0.0	5.7	**84.1**	*10.2*	0.1	8.4	8.5	14.8	**52.8**	*15.0*	0.5
$p_O = 0.20$	0	0.0	6.5	**84.7**	*8.7*	0.1	8.4	8.6	15.4	**53.2**	*14.1*	0.4
$p_O = 0.25$	0	0.0	7.4	**85.6**	*7.0*	0.1	8.4	8.6	16.0	**53.6**	*13.0*	0.4
QCRM												
$p_O = 0.05$	0	0.0	3.8	**85.2**	*10.9*	0.0	8.3	8.4	13.4	**55.1**	*14.6*	0.2
$p_O = 0.10$	0	0.0	4.4	**86.1**	*9.5*	0.0	8.3	8.4	13.9	**55.6**	*13.6*	0.1
$p_O = 0.15$	0	0.0	5.5	**86.8**	*7.7*	0.0	8.3	8.4	14.5	**55.9**	*12.7*	0.1
$p_O = 0.20$	0	0.0	6.3	**87.3**	*6.5*	0.0	8.3	8.4	15.1	**56.3**	*11.7*	0.1
$p_O = 0.25$	0	0.0	7.2	**87.4**	*5.5*	0.0	8.3	8.4	15.7	**56.6**	*10.8*	0.1
UA												
$p_O = 0.05$	0	0.0	9.3	**82.5**	*8.0*	0.1	8.6	11.1	27.6	**44.9**	*7.6*	0.1
$p_O = 0.10$	0	0.0	10.6	**82.4**	*7.0*	0.1	8.6	11.2	28.6	**44.7**	*6.7*	0.1
$p_O = 0.15$	0	0.0	12.1	**82.3**	*5.6*	0.1	8.6	11.3	29.6	**44.4**	*6.0*	0.1
$p_O = 0.20$	0	0.0	14.0	**81.4**	*4.6*	0.0	8.6	11.4	30.6	**43.8**	*5.5*	0.1
$p_O = 0.25$	0	0.0	15.5	**80.6**	*3.9*	0.0	8.6	11.5	31.6	**43.3**	*4.9*	0.1
Méthodes guidées par le score, Erreur touchant tous les grades												
QLCRM												
$p_O = 0.05$	0	0.0	8.0	**85.2**	*6.7*	0.0	8.4	8.7	17.8	**54.0**	*10.9*	0.2
$p_O = 0.10$	0	0.0	18.0	**79.8**	*2.1*	0.0	8.5	9.2	24.7	**51.7**	*5.8*	0.1
$p_O = 0.15$	0	0.0	34.1	**65.1**	*0.8*	0.0	8.6	10.0	34.1	**44.3**	*2.9*	0.0
$p_O = 0.20$	0	0.3	51.1	**48.5**	*0.1*	0.0	8.9	11.2	43.4	**35.1**	*1.4*	0.0
$p_O = 0.25$	0	1.0	68.4	**30.5**	*0.1*	0.0	9.2	13.5	51.6	**25.1**	*0.6*	0.0
QCRM												
$p_O = 0.05$	0	0.0	7.9	**87.3**	*4.8*	0.0	8.3	8.5	17.6	**56.9**	*8.5*	0.1
$p_O = 0.10$	0	0.0	17.9	**80.6**	*1.5*	0.0	8.3	8.8	24.7	**53.8**	*4.3*	0.0
$p_O = 0.15$	0	0.0	34.0	**65.8**	*0.3*	0.0	8.4	9.3	34.4	**45.9**	*2.0*	0.0
$p_O = 0.20$	0	0.2	51.3	**48.5**	*0.0*	0.0	8.4	10.1	44.4	**36.2**	*0.9*	0.0
$p_O = 0.25$	0	0.6	69.2	**30.2**	*0.0*	0.0	8.4	11.8	53.7	**25.7**	*0.4*	0.0
UA												
$p_O = 0.05$	0	0.0	16.2	**78.9**	*4.8*	0.0	8.7	11.9	31.6	**42.4**	*5.3*	0.1
$p_O = 0.10$	0	0.1	27.1	**70.7**	*2.1*	0.0	8.9	13.1	36.7	**37.9**	*3.4*	0.1
$p_O = 0.15$	0	0.1	41.8	**57.1**	*1.0*	0.0	9.2	14.5	42.0	**32.3**	*2.0*	0.0
$p_O = 0.20$	0	0.5	57.7	**41.3**	*0.4*	0.0	9.7	16.4	46.0	**26.7**	*1.2*	0.0
$p_O = 0.25$	0	1.4	70.3	**28.0**	*0.3*	0.0	10.2	19.0	49.0	**21.1**	*0.7*	0.0
Méthodes guidées par la DLT												
LCRM												
$p_O = 0.05$	0	0.2	17.2	**54.6**	*23.8*	4.3	8.7	9.5	20.0	**34.0**	*20.9*	7.0
$p_O = 0.10$	0	0.4	22.9	**55.9**	*18.2*	2.6	8.7	9.7	22.5	**35.2**	*18.5*	5.4
$p_O = 0.15$	0	0.8	28.9	**55.7**	*13.1*	1.5	8.8	10.0	25.2	**35.9**	*16.0*	4.1
$p_O = 0.20$	0	1.0	35.2	**53.8**	*9.2*	0.7	8.8	10.3	28.0	**35.7**	*14.0*	3.2
$p_O = 0.25$	0	1.5	40.8	**51.0**	*6.3*	0.4	8.9	10.7	30.4	**35.3**	*12.3*	2.4
CRM												
$p_O = 0.05$	0	0.1	17.3	**57.5**	*22.2*	2.9	8.4	9.2	20.6	**37.3**	*20.0*	4.6

Suite page suivante

Méthodes	Pourcentage de sélection de chaque dose comme dose à recommander						Pourcentage des doses allouées					
$p_O = 0.10$	0	0.3	22.9	**59.1**	*16.2*	1.5	8.4	9.4	23.1	**38.3**	*17.5*	3.3
$p_O = 0.15$	0	0.5	29.0	**58.1**	*11.6*	0.8	8.4	9.7	25.9	**38.7**	*14.8*	2.5
$p_O = 0.20$	0	0.9	35.0	**56.1**	*7.6*	0.4	8.4	10.0	28.7	**38.4**	*12.7*	1.8
$p_O = 0.25$	0	1.3	41.5	**51.8**	*5.2*	0.2	8.4	10.3	31.3	**37.6**	*11.1*	1.3

Méthodes basées sur le score :
QLCRM : méthode Quasi-LCRM (notre proposition).
QCRM : méthode Quasi-CRM de Yuan et al., [3].
UA : Unified Algorithm (méthode d'Ivanova et al., [5]).
EID : Extended Isotonic Design (méthode de Chen et al., [2]).
Méthodes basées sur la DLT :
LCRM : méthode CRM avec un modèle logistique dans un cadre fréquentiste.
CRM : méthode CRM avec un modèle empirique dans un cadre bayésien.
Les valeurs en gras correspondent aux résultats obtenus à la dose cible.

Les valeurs en italiques correspondent aux résultats obtenus à la dose la plus proche de la dose cible.

Chapitre 6

Illustration de la méthode QLCRM dans un essai réel

Nous avons rétrospectivement exploré l'approche de construction du score de toxicité ainsi que l'application de la méthode QLCRM dans le contexte clinique d'un essai de phase I pédiatrique évaluant l'erlotinib en combinaison avec la radiothérapie dans le traitement des gliomes du tronc cérébral de l'enfant (NCT00418327)[8]. La problématique de la prise en compte des toxicités modérées et multiples dans l'escalade de dose et l'identification de la dose à recommander a été soulevée par les cliniciens de l'essai.

6.1 Motivation

Vingt et un patients ont été inclus dans cet essai. L'escalade de dose comprenait quatre paliers de dose : 75, 100, 125 et 150 mg/m^2 ; le palier de 75 mg/m^2 était la dose de départ. Une dose inférieure (60 mg/m^2) était prévue si nécessaire. L'essai a été conduit en utilisant la méthode de réévaluation séquentielle (CRM) avec un modèle logistique à un paramètre, dans un cadre fréquentiste, pour estimer la relation entre la dose et la probabilité d'observer une DLT. Le critère principal de jugement considéré pour l'escalade de dose et pour l'identification de la dose à recommander était la toxicité dose-limitante, DLT. Elle a été définie comme tout événement indésirable de grade 4 hématologique ou de grade supérieur ou égal à 3 extra-hématologique possiblement lié à la drogue, survenant dans les six semaines après le début du traitement. La dose à recommander était définie comme dose associée à 20% de DLT.

L'erlotinib est un inhibiteur des récepteurs d'un facteur de croissance épithélial (EGF) déjà largement utilisé chez l'adulte. Cette molécule ciblée est connue pour induire une toxicité dermatologique. Une relation entre la dose d'erlotinib et la toxicité dermatologique a été rapportée dans les essais conduits chez l'adulte [60]. Plusieurs évaluations par un dermatologue étaient donc prévues dans l'essai pédiatrique : avant le début du traitement, puis au $7^{ème}$, $21^{ème}$ et $42^{ème}$ jour. Une surveillance par le pédiatre était prévue ensuite toutes les trois semaines avec consultation spécialisée en cas de toxicité dermatologique. Le recueil de données correspondant à chaque évaluation

dermatologique était standardisé sous forme d'un tableau décrivant l'intensité de différents symptômes : folliculite, érythème, prurit, autre toxicité cutanée (un champ texte permettant de détailler le symptôme), trouble des phanères (alopécie, trichomégalie, hypertrichose, paronychie...). La toxicité dermatologique, atteignant exceptionnellement un grade 3 ou 4, est le plus souvent ignorée dans le processus d'escalade de dose et dans l'estimation de la dose à recommander basés sur l'observation de la toxicité dose-limitante, DLT.

Parmi les 21 patients inclus dans cet essai, 20 patients ont été pris en compte dans l'escalade de dose ; un patient a été considéré comme non évaluable pour la toxicité car il avait reçu par erreur le quart de la dose prévue. Trois paliers de dose ont été explorés (75, 100 et 125 mg/m^2). Un patient a présenté un événement neurologique létal au premier palier, à type d'hémorragie intratumorale d'interprétation difficile compte tenu de la pathologie sous-jacente. Cet événement a finalement été classé par précaution comme DLT bien que l'imputabilité au traitement soit discutée. Au $3^{ème}$ palier de dose, un patient a présenté une toxicité dermatologique de grade 3, à type de folliculite. Au terme de l'essai, la dose 125 mg/m^2 a été identifiée comme dose à recommander, associée à une probabilité de DLT de 15.9% (l'intervalle de confiance à 95% est égal à [4 - 45%]).

Le tableau 6.1 décrit les toxicités dermatologiques observées pendant les six premières semaines après le début du traitement, telles qu'elles ont été rapportées par les dermatologues. Dans la catégorie "autre toxicité cutanée" était fréquemment rapportée l'existence d'une xérose (sécheresse cutanée). Nous avons donc considéré séparément ce type de toxicité. En revanche, les autres toxicités rapportées dans cette catégorie étaient très rares et n'étaient probablement pas liées au traitement étudié, d'après la dermatologue référent de l'essai. Elles ont donc été ignorées dans la suite de l'analyse. Les troubles des phanères étant peu fréquents au cours des six premières semaines, ils ont été considérés globalement en une seule catégorie.

Comme décrit dans le tableau suivant rapportant la fréquence des grades de toxicité observés chez les 20 patients évaluables, tous les patients ont expérimenté au moins une toxicité cutanée au cours des six premières semaines. La folliculite est le symptôme le plus fréquemment rapporté. La majorité des symptômes dermatologiques rapportés consistait en des symptômes de grade 1 ou de grade 2. Une folliculite de grade 3 associée à un prurit de grade 3 ont été observés. Bien qu'il ne soit pas pertinent de tester l'indépendance entre les différents symptômes compte tenu du nombre limité d'observations, il apparaît que les symptômes étaient fréquemment associés, le nombre moyen de symptômes rapportés étant de 2.7 par patient (médiane 2.5, maximum 5). Le détail des toxicités observées par patient sera rapporté ultérieurement (cf. tableau 6.5).

TABLE 6.1 – Fréquence des toxicités dermatologiques observées dans l'essai étudié (n=20)

Symptôme dermatologique	Grade 0 N	Grade 1 N	Grade 2 N	Grade 3 N	Grade > 0 %
Folliculite	5	8	6	1	75%
Erythème	9	7	4	0	55%
Prurit	12	7	0	1	40%
Xérose	7	12	1	0	65%
Autre lésion cutanée :	16	3	1	0	20%
- Rhagades	19	1	0	0	5%
- Impétigo, pyodermite	19	0	1	0	5%
- Hypopigmentation	19	1	0	0	5%
- Desquamation autour des oreilles	19	1	0	0	5%
Trouble des phanères :	13	7	0	0	35%
- Alopecie	15	5	0	0	25%
- Trichomegalie et anomalie des cils	19	1	0	0	5%
- Hypertrichose	18	2	0	0	10%
- Paronychie	19	1	0	0	5%

Face au constat du faible nombre de DLTs observées et des difficultés possibles d'interprétation de ces événements, d'une part, et de la non prise en compte de toxicités fréquentes liées à la molécule, d'autre part, nous avons proposé de construire un score de toxicité dermatologique et de réestimer la dose à recommander en utilisant la méthode QLCRM guidée par ce score sur les données de cet essai. En particulier, il nous a semblé important d'évaluer la faisabilité de cette approche qui nécessite une interaction étroite avec les cliniciens tant pour définir les poids associés aux différents types de toxicité que pour définir le score-cible.

6.2 Méthodes

6.2.1 Construction du score de toxicité nTTP et du score-cible

Dans cette partie, nous avons travaillé avec trois cliniciens experts onco-pédiatres impliqués dans l'essai erlotinib, notés X, Y et Z correspondant respectivement au Dr Vita Ridola de l'université Cattolica-A Gemelli, Italie, au Dr Birgit Geoerger de l'Institut Gustave Roussy, France et au Dr Darren Hargrave du Royal Marsden Hospital, UK.

Nous avons utilisé une variante de la méthode Delphi nous permettant d'étudier la variabilité des réponses entre les cliniciens interviewés : i) nous avons initialement interrogé indépendamment les cliniciens pour définir les poids associés à chaque grade de chaque type de toxicité, puis les décisions concernant les cohortes hypothétiques, ii) plusieurs semaines plus tard, après avoir réexpliqué la méthode, les mêmes experts sont interrogés de nouveau sur le même questionnaire concernant le classement des cohortes hypothétiques (sans concertation entre eux et sans revoir leurs premières réponses), iii) enfin une réunion commune a été organisée afin de discuter leurs propositions et d'aboutir à des décisions consensuelles. Pour valider ces dernières, nous avons interrogé un quatrième expert indépendant, qui n'a pas participé au processus de construction du score et du score-cible, pour nous donner son avis sur les différentes propositions.

Les experts ont considéré que la toxicité dermatologique était correctement décrite en retenant cinq types différents de toxicité : Folliculite, Erythème, Prurit, Xérose, et trouble des phanères, ignorant les autres types de toxicité hormis le décès toxique. Comme énoncé précédemment, la toxicité dermatologique atteint exceptionnellement un grade 3 ou 4 dans cet essai et plus généralement avec ce type de molécule. Nous avons donc limité notre travail aux grades variant de 0 à 3.

Dans une première étape, nous avons demandé aux cliniciens de définir un poids compris entre 0 et 10 pour chaque grade de chaque type de toxicité considéré (cinq types et trois grades possibles). Nous avons fixé initialement à 10 la valeur maximale de poids pour le décès possiblement lié à la drogue. Afin de définir le score-cible, nous avons ensuite présenté aux cliniciens 23 cohortes hypothétiques dans un ordre aléatoire. Chaque cohorte contient quatre patients avec des profils de toxicité différents. Les profils de toxicité de ces cohortes sont détaillés dans le tableau 6.3. Nous avons fait varier le grade de la folliculite de 0 à 3 comme cela a été observé dans l'essai, tandis que les troubles des phanères présentés dans les cohortes hypothétiques étaient limités au grade 1, les troubles des phanères plus sévères n'ayant jamais été observés dans les six premières semaines (toxicité retardée). Les profils de toxicité de ces cohortes ont été générés en se basant sur les toxicités dermatologiques observées dans l'essai. La pertinence clinique des profils a été validée par MC Le Deley. Pour chaque cohorte, nous avons calculé la valeur moyenne des mesures de toxicité en considérant successivement la somme des poids (TTB, correspondant au score de Bekele et Thall) et la norme des poids (TTP). Sur le plan statistique, nous avons veillé à ce que la moyenne et la variance des mesures soient cohérentes avec les données de l'essai.

6.2.2 Application rétrospective de la méthode QLCRM

Après avoir défini de façon consensuelle le vecteur de poids et la valeur cible de toxicité, nous avons appliqué la méthode QLCRM aux données de l'essai erlotinib, pour réestimer la dose qui aurait été recommandée en prenant en compte des toxicités dermatologiques et les DLTs observées durant l'essai. Nous avons fixé la valeur de l'ordonnée à l'origine du modèle logistique à 3. Le modèle de travail a été obtenu avec la fonction "getprior" du package dfcrm du logiciel R, en supposant la troisième dose comme dose à recommander et en fixant la demi-largeur de l'intervalle d'indifférence à 0.04. Cette analyse rétrospective inclut les 20 patients évaluables pour la toxicité sur une période de 6 semaines. Notons qu'il s'agit d'une application rétrospective des données observées sur les 20 patients en fin d'essai, l'escalade de dose étant figée.

6.3 Résultats

6.3.1 Définition des poids de toxicité

La figure 6.1 illustre les réponses des trois cliniciens. La première observation dans les réponses des différents cliniciens est que les poids proposés diffèrent des grades et varient d'un type de toxicité à l'autre. Ainsi, pour chaque expert, la folliculite semble avoir une importance clinique supérieure aux autres types de toxicité de grade égal. La démarche semble donc avoir été comprise par les investigateurs. Bien que les réponses entre les trois experts ne soient parfaitement concordantes que dans deux situations (Xérose de grade 1 et Trouble des phanères de grade 2), nous remarquons que les poids proposés par les trois experts ne diffèrent pas trop entre eux, la différence maximale étant de 2 points ; cette différence n'est observée que pour trois situations. Les poids proposés par l'expert Y sont toujours les plus élevés sauf dans un cas (prurit de grade 3), ce que reflète la moyenne de l'ensemble des poids proposés par cet expert (moyenne tous types et tous grades confondus = 4.13 *versus* 3.27 et 3.53 pour les experts X et Z respectivement). A la demande des experts, le poids associé au décès possiblement lié à la drogue a été réajusté à 20 pour le différencier plus d'une toxicité dermatologique la plus sévère possible. Les poids consensuels ont été obtenus facilement au cours de la réunion finale sans nécessité de réajustement secondaire. Les poids consensuels sont présentés dans le tableau 6.2.

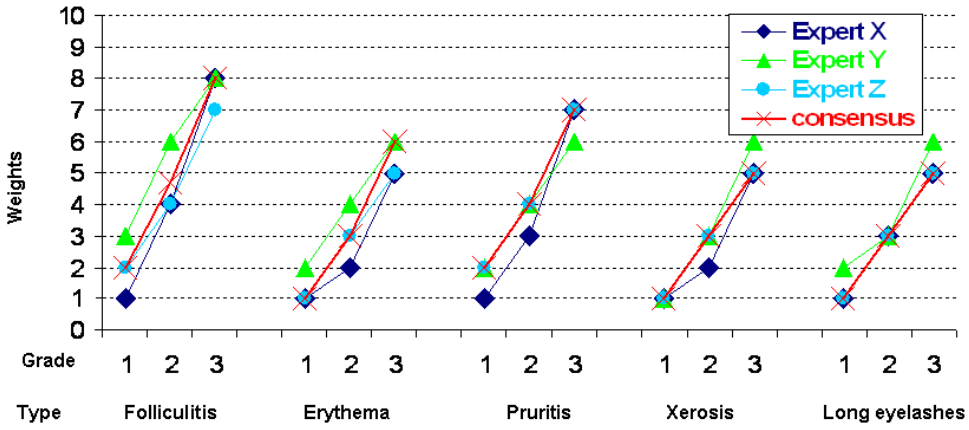


FIGURE 6.1 – Poids proposés par les experts pour les différents types et grades de toxicité dermatologique

TABLE 6.2 – Poids consensuels pour chaque grade de chaque type de toxicité

	grade-1	**grade-2**	**grade-3**
Folliculite	2	4.5	8
Erythème	1	3	6
Prurit	2	4	7
Xérose	1	3	6
Trouble des phanères	1	3	6

6.3.2 Construction des cohortes hypothétiques

Nous avons invité les cliniciens, tout d'abord, séparément, à donner leurs décisions (escalader, répéter la dose ou désescalader) pour les 23 cohortes hypothétiques de quatre

patients.

Les réponses $X1$, $Y1$ et $Z1$ correspondent aux premières réponses des experts, obtenues en les interrogeant séparément. Rappelons que les décisions seront jugées cohérentes si les décisions d'escalader correspondent aux scores TTP les plus faibles, et les décisions de désescalader correspondent aux scores TTP les plus élevés. Ceci n'est pas le cas avec les premières réponses des cliniciens.

Après avoir discuté la méthode et les premières réponses, les mêmes experts ont été interrogés de nouveau sur les mêmes cohortes, sans concertation entre eux et sans revoir leurs premières réponses. Les deuxièmes réponses sont présentées par les colonnes $X2$, $Y2$ et $Z2$. Même si ces réponses diffèrent des premières, elles ne correspondent pas à une classification cohérente, c'est-à-dire à trois blocs distincts de décision (escalader, répéter et désescalader la dose). Les décisions consensuelles aboutissant à une classification cohérente ont été obtenues facilement lors de la réunion commune sans nécessité de réajuster les poids.

Le score-cible de toxicité, θ_{nTTP}, est ainsi défini par la moyenne des scores TTP associés à des cohortes pour lesquelles la décision est de rester à la même dose. Il est égal à 4.66.

A noter que le classement des cohortes par ordre croissant de score TTP validé par les experts diffère de celui obtenu avec les scores TTB. Ceci est illustré par les cohortes 17 et 23 respectivement : de nombreuses toxicités minimes dans la cohorte 17 aboutissant à un score TTB supérieur à celui de la cohorte 23 dans laquelle les patients présentent moins de types de toxicité, mais d'intensité plus sévère.

Notons que le quatrième expert interrogé après l'ensemble du processus a jugé acceptable les poids consensuels ainsi que les décisions consensuelles.

Table 6.3 – Cohortes hypothétiques utilisées pour définir le score-cible dans l'essai erlotinib

N°	Outcomes	N°	Outcomes
1	Folliculite 1 + Erythème 1 + Xérose 1 + TPhanères 1 Folliculite 2 + Erythème 1 + Prurit 2 Folliculite 1 + Erythème 2 + Xérose 1 Folliculite 2 + Erythème 2	**2**	Erythème 2 + Xérose 1 + Prurit 1 Erythème 2 + Xérose 1 Erythème 2 Folliculite1 + Erythème 2 + Xérose 1
3	Folliculite 1 + TPhanères 1 Folliculite 2 + Erythème 2 + Xérose 1 + Prurit 1 Folliculite 1 + Prurit 1 Folliculite 2 + Erythème 1 + Prurit 1	**4**	Folliculite 2 + Erythème 1 + Xérose 2 + Prurit 1 Folliculite 1 + Xérose 2 + Prurit 1 + TPhanères 1 Erythème 1 + Xérose 1 + Prurit 1 + TPhanères 1 Folliculite 2 + Erythème 2 + Xérose 1 + Prurit 2 + TPhanères 1
5	Erythème 1 + Prurit 1 Folliculite 2 + Xérose 2 + Prurit 1 Folliculite 1 + Xérose 1 Folliculite 2 + TPhanères 1	**6**	Erythème 2 + Xérose 1 + TPhanères 1 Folliculite 1 + Erythème 1 Folliculite 2 Folliculite 1 + Erythème 1 + Xérose 1
7	Erythème 1 + Xérose 2 Xérose 1 Erythème 1 + Xérose 1 + Prurit 1 Erythème 1	**8**	Folliculite 1 + Xérose 2 Pas de toxicité Folliculite 2 + Xérose 2 Erythème 2 + Prurit 1
9	Erythème 2 + Xérose 1 Folliculite 1 + Erythème 1 + Xérose 2 Xérose 1 + Prurit 1 + TPhanères 1 Folliculite 1 + Erythème 2 + Xérose 1	**10**	Xérose 1 + Prurit 1 Folliculite 2 + Erythème 2 + Xérose 1 + Prurit 1 + TPhanères 1 Erythème 1 + Xérose 1 + Prurit 1 + TPhanères 1 Folliculite 2 + Erythème 2 + Prurit 1
11	Folliculite 1 Folliculite 1 + Xérose 1 + TPhanères 1 Pas de toxicité Erythème 2 + Xérose 2 + Prurit 1	**12**	Folliculite 2 + Erythème 2 Folliculite 1 Folliculite 3 + Erythème 2 + Prurit 3 Folliculite 3 + Erythème 1 + Prurit 2
13	Erythème 1 + Xérose 1 + Prurit 1 Folliculite 1 + Erythème 1 + Xérose 1 + Prurit 1 + TPhanères 1 Folliculite 2 + Prurit 2 + TPhanères 1 Xérose 2 + TPhanères 1	**14**	Folliculite 2 + Xérose 1 Folliculite 2 Folliculite 1 + Prurit 1 + TPhanères 1 Folliculite 3 + Erythème 2
15	Erythème 1 + Xérose 1 Folliculite 1 + Erythème 1 + Xérose 2 Erythème 2 Erythème 1 + Xérose 1 + Prurit 1	**16**	Folliculite 1 + Xérose 1 Folliculite 1 + Erythème 1 + Xérose 2 + Prurit 1 + TPhanères 1 Erythème 1 + Prurit 1 + TPhanères 1 Folliculite 2 + Prurit 1
17	Folliculite 1 + Erythème 1 + Xérose 2 + Prurit 1 Folliculite 1 + Xérose 1 + Prurit 1 + TPhanères 1 Erythème 1 + Xérose 1 + Prurit 1 + TPhanères 1 Folliculite 2 + Erythème 1 + Xérose 1 + Prurit 1 + TPhanères 1	**18**	Xérose 1 + Death Folliculite 2 + Prurit 1 Folliculite 1 + Erythème 1 + Xérose 1 + TPhanères 1 Folliculite 1
19	Erythème 1 + Xérose 1 + Prurit 1 + TPhanères 1 Folliculite 1 + Xérose 1 + Prurit 1 + Death Folliculite 2 + Erythème 1 + Xérose 2 + Prurit 1 Folliculite 1 + Erythème 2 + Xérose 1 + TPhanères 1	**20**	Folliculite 2 + Erythème 2 Folliculite 1 + Xérose 1 + Death Folliculite 1 + Erythème 1 + Prurit 1 Folliculite 1 + Erythème 2 + Prurit 3
21	Xérose 1 Folliculite 2 + Prurit 1 Folliculite 1 + Erythème 1 + Xérose 1 + TPhanères 1 Folliculite 1	**22**	Erythème 1 + Xérose 1 + Prurit 1 + TPhanères 1 Folliculite 1 + Xérose 1 + Prurit 1 Folliculite 2 + Erythème 1 + Xérose 2 + Prurit 1 Folliculite 1 + Erythème 2 + Xérose 1 + TPhanères 1
23	Folliculite 2 + Erythème 2 Folliculite 1 + Xérose 1 Folliculite 1 + Erythème 1 + Prurit 1 Folliculite 1 + Erythème 2 + Prurit 3		

Les grades 1-2-3 des différentes toxicités dermatologiques sont basées sur la classification NCI-CTC. TPhanères : Trouble des Phanères.

TABLE 6.4 – Description des cohortes triées par ordre croissant de score TTP et décisions des experts

Numéro de cohorte	$\overline{TTB}$	$\overline{TTP}$	X1	Y1	Z1	X2	Y2	Z2	Consensus
7	2.50	1.90	E	E	E	E	E	E	E
11	3.50	2.28	E	E	E	E	E	E	E
21	3.63	2.64	E	E	E	E	E	E	E
15	3.75	2.65	E	E	E	E	E	E	E
6	4.13	3.13	E	E	E	E	E	E	E
8	4.38	3.15	**R**	E	E	E	E	E	E
9	5.00	3.27	E	E	E	E	E	E	E
2	4.75	3.41	**R**	E	E	E	E	E	E
16	5.63	3.49	E	E	E	E	E	E	E
5	5.25	3.71	E	E	E	E	E	E	E
13	6.13	3.76	**R**	E	E	E	E	E	E
17	7.13	3.82	E	E	E	E	E	E	E
22	6.88	3.84	E	E	E	E	E	E	E
3	6.25	3.99	E	E	E	E	E	E	E
10	7.25	4.15	**R**	**R**	E	**R**	E	E	**R**
1	7.00	4.47	**R**	E	E	E	E	E	**R**
23	6.88	4.63	E	**R**	E	**R**	E	E	**R**
4	9.25	4.90	**R**	**R**	E	E	E	E	**R**
14	6.50	5.16	D	D	**R**	**R**	**R**	**R**	**R**
12	10.13	6.86	D	D	D	D	**R**	**R**	D
18	8.63	7.40	D	D	D	D	D	**R**	D
19	11.88	8.15	D	D	D	D	D	**R**	D
20	11.88	9.10	D	D	D	D	D	D	D

Numéro de cohorte : numéro attribué au hasard dans la présentation des cohortes hypothétiques aux cliniciens (cf. tableau 6.3).
$\overline{TTB}$ et $\overline{TTP}$: moyennes du score TTB (score de Bekele) et TTP (notre proposition), calculées pour chaque cohorte avec les poids consensuels. Ces valeurs n'ont pas été dévoilées lorsque les cohortes ont été présentées aux experts.
X1, Y1 et Z1 : décisions initiales des experts X, Y et Z obtenues indépendamment.
X2, Y2 et Z2 : décisions des mêmes experts obtenues quelques semaines plus tard sans discussion commune sur leurs décisions et sans en référer à leurs premières réponses.
Consensus : décisions consensuelles des experts, obtenues lors de la discussion face à face.
E : Escalader au palier de dose supérieur.
R : Répéter le même palier de dose.
D : Désescalader au palier de dose inférieur.

6.3.3 Application rétrospective de la méthode QLCRM

Une fois le score-cible de toxicité défini, nous avons appliqué la méthode QLCRM (modèle logistique avec une inférence fréquentiste) aux données de l'essai erlotinib. Dans ce contexte, le profil de toxicité le plus sévère TTP_{max} correspond au poids consensuel associé au décès toxique (= 20). Le score-cible normalisé est alors égal à 0.233. Le modèle de travail est égal au vecteur (0.10, 0.16, 0.23,0.32).

Le tableau 6.5 présente les symptômes dermatologiques observés chez les 20 patients inclus dans cet essai ainsi que leurs scores normalisés. A partir de ces derniers, la méthode QLCRM permet d'estimer la valeur du score nTTP à chaque palier de dose. Les scores nTTP estimés par la méthode QLCRM pour chaque palier de dose sont respectivement de 0.15, 0.22 et 0.30 aux paliers 75, 100 et 125mg/m^2. La deuxième dose (100 mg/m^2) est la dose associée au score estimé le plus proche du score cible et donc considérée comme dose à recommander quand la toxicité dermatologique observée au premier cycle est prise en compte, tandis qu'une dose de 120 mg/m^2 avait été recommandée par la CRM guidée par le critère DLT.

Table 6.5 – Valeurs du score de toxicité, nTTP, pour chaque patient avec les poids consensuels

N°	palier de dose	profil de toxicité	nTTP	nTTP moyen	nTTP estimé
1	75	Gr-1 Folliculite	0.100		
2	75	Gr-2 Folliculite	0.225		
3	75	Décès toxique	1.000	0.31	0.15
4	75	Gr-1 Erythème + Gr-1 Prurit + Gr-1 Xérose + Gr-1 Trouble des phanères	0.133		
5	75	Gr-1 Folliculite + Gr-1 Erythème	0.112		
6	75	Gr-2 Folliculite + Gr-2 Erythème	0.271		
7	100	Gr-2 Folliculite + Gr-1 Prurit	0.231		
8	100	Gr-1 Xérose	0.050		
9	100	Gr-1 Folliculite + Gr-1 Erythème + Gr-2 Xérose + Gr-1 Prurit + Gr-1 Trouble des phanères	0.218	0.15	0.22
10	100	Gr-1 Folliculite + Gr-1 Xérose + Gr-1 Trouble des phanères	0.123		
11	100	Gr-1 Folliculite + Gr-1 Erythème + Gr-1 Xérose + Gr-1 Trouble des phanères	0.133		
12	100	Gr-1 Folliculite + Gr-1 Erythème + Gr-1 Xérose + Gr-1 Trouble des phanères	0.123		
13	125	Gr-1 Folliculite + Gr-1 Erythème + Gr-1 Xérose	0.166		
14	125	Gr-1 Folliculite + Gr-1 Erythème + Gr-1 Xérose + Gr-1 Prurit + Gr-1 Trouble des phanères	0.158		
15	125	Gr-2 Erythème + Gr-1 Xérose	0.112		
16	125	Gr-1 Folliculite + Gr-1 Xérose	0.553	0.22	0.30
17	125	Gr-3 Folliculite + Gr-2 Erythème + Gr-3 Prurit	0.297		
18	125	Gr-2 Folliculite + Gr-2 Erythème + Gr-1 Xérose + Gr-1 Prurit + Gr-1 Trouble des phanères	0.123		
19	125	Gr-1 Xérose + Gr-1 Prurit + Gr-1 Trouble des phanères	0.112		
20	125	Gr-1 Prurit + Gr-1 Trouble des phanères	0.225		

Profil de toxicité : les profils de toxicité, parmi les cinq symptômes de toxicité dermatologique, observés sur la période de 6 semaines pour chaque patient.

nTTP : le score TTP normalisé.

Gr : Grade.

Chapitre 7

Discussion

Les traitements classiques en oncologie sont la chirurgie, la radiothérapie et la chimiothérapie cytotoxique. Même si la survie des patients traités pour un cancer s'est globalement améliorée au cours des dernières décennies, il reste des situations pour lesquelles les approches classiques sont insuffisantes. Par ailleurs, dans le contexte des cancers associés à un taux élevé de guérison, tels que la majorité des cancers de l'enfant par exemple, il est devenu essentiel de réduire le prix de la guérison en termes de toxicité immédiate et de séquelles tardives fonctionnelles, psychologiques et développementales.

La recherche de nouveaux traitements plus efficaces et moins toxiques est donc intense.

Une nouvelle génération de traitements, appelés thérapies ciblées, a émergé ces dernières années grâce à la meilleure compréhension des mécanismes de développement des tumeurs. Ces agents ont une action ciblée en intervenant à un niveau précis du développement de la cellule tumorale ou de son environnement. De par leur mécanisme d'action plus spécifique, ils diffèrent également des chimiothérapies cytotoxiques par leurs profils de toxicité.

L'exemple le plus frappant de ces thérapies est l'imatinib, inhibiteur de la tyrosine kinase, qui a transformé l'histoire de la maladie des tumeurs stromales gastro-intestinales et des leucémies myéloïdes chroniques, ou encore le trastuzumab dans le traitement adjuvant du cancer du sein $HER2^+$ [61, 62].

Actuellement, près de 800 molécules sont en cours de développement, le plus souvent obtenues par génie génétique, avec des mécanismes d'action très différents de ceux des chimiothérapies cytotoxiques classiques. Seulement 5 à 25% des molécules passant de l'évaluation pré-clinique à l'évaluation chez l'homme deviendront commercialisées [63]. L'échec dans les phases ultérieures de développement peut être en partie expliqué par une mauvaise sélection de la dose au terme de l'essai de phase I [64].

Bien que ces thérapies ciblées soient apparues il y a maintenant plus d'une décennie, elles sont encore évaluées par des méthodes développées pour des traitements cytotoxiques. Concernant les essais de phase I, le paradigme établi pour la chimiothérapie classique est basé sur une relation monotone et croissante entre la dose et la toxicité, d'une part, et entre la dose et l'efficacité, d'autre part, définissant le concept "Plus est mieux" [6, 65, 66]. Bien que de nombreux travaux soient en cours pour in-

tégrer des critères d'activité biologique ou d'efficacité afin de définir la dose optimale biologique [25, 26, 27, 30, 33, 34, 35, 36], la toxicité reste le critère de jugement principal des essais de phase I [28, 29, 67]. Quoiqu'elle soit mesurée pour les différents organes sur une échelle gradée [7], elle est généralement réduite à un indicateur binaire définissant la DLT. De fait, ce critère ne prend pas en compte la multiplicité éventuelle des événements toxiques, ni les toxicités modérées qui pourtant peuvent être le reflet de l'activité biologique de la molécule. Par ailleurs, une revue de littérature de 201 essais de phase I où la toxicité est le critère principal de jugement a montré que la DLT est moins fréquemment observée dans les essais évaluant les thérapies ciblées que dans ceux évaluant les traitements cytotoxiques [29]. Pourtant ces thérapies ne sont pas dépourvues de toxicité : une toxicité modérée chronique peut représenter un problème pour le patient et donc induire une diminution de l'observance (ou compliance) au traitement. Par exemple, une fatigue, des nausées ou une diarrhée de grade 2 pourraient être acceptables pendant quelques jours chez les patients recevant un traitement par voie intraveineuse toutes les trois semaines, mais ces toxicités peuvent devenir insupportables pour le patient lorsque le traitement est administré quotidiennement sans interruption pendant plusieurs mois. Le critère de DLT s'avère ainsi non adapté pour les thérapies ciblées.

Mon travail de thèse a été motivé par le souhait de développer une méthode de recherche de dose plus adaptée aux essais de phase I évaluant les thérapies ciblées. L'idée de ce projet est venue des cliniciens insatisfaits devant la réduction d'information en tout ou rien dans le contexte d'un essai clinique de phase I pédiatrique évaluant l'erlotinib, inhibiteur des récepteurs d'un facteur de croissance épithélial (NCT00418327) [8].

Ce travail de recherche a débuté par l'élaboration d'un critère de jugement alternatif à la DLT combinant l'ensemble des toxicités liées au traitement. Ce score, appelé TTP pour *Total Toxicity Profile*, est défini par la norme euclidienne des poids associés aux différentes toxicités observées chez un patient. Les poids reflètent l'importance clinique relative de chaque type et grade de toxicité. Notre approche est dérivée de celle de Bekele et Thall proposant un score de toxicité défini par la somme arithmétique des poids [38]. Deux autres scores de toxicité ont été récemment publiés : i) le score TBS (*Total Burden Score*) de Lee et al. défini également par la somme arithmétique des poids, la manière de définir les poids étant différente de celle de Bekele et Thall [40], et (ii) le score ETS (*Equivalent Toxicity Score*) de Chen et al. défini par le grade maximum observé, les autres toxicités de grade inférieur contribuant à la part décimale du score à l'aide d'une fonction logistique. Notons que les grades définissant le score ETS sont dérivés d'une nouvelle échelle qui différencie les grades 3 non-DLT des grades 3 DLT, d'une part, et les grades 4 non-DLT des grades 4 DLT, d'autre part [2].

Il est important de noter que ces scores conduisent à un classement relatif très différent des profils de toxicité. Afin d'illustrer ce point, prenons l'exemple suivant en supposant la même matrice de poids pour les scores TTB/TBS et pour le score TTP. Soit un patient A présentant deux toxicités de poids (ou de grade) égal à 2, un patient B avec trois toxicités de poids égal à 2 et un patient C avec une seule toxicité classé

DLT de poids égal à 3. Avec le score TTB ou TBS, le profil de toxicité du patient B apparaît beaucoup plus sévère (score =6) que celui des patients A et C (respectivement 4 et 3). De façon très différente, avec le score ETS, les patients A et B auraient des scores proches (= 1.38 et 1.44) et très inférieurs à celui du patient C (= 2). Le score TTP apparaît comme une alternative intermédiaire, avec un score de TTP égal à 2.83, 3.46 et 3 pour les patients A, B et C, respectivement. La figure 7.1 illustre le classement de ces trois patients selon le score utilisé.

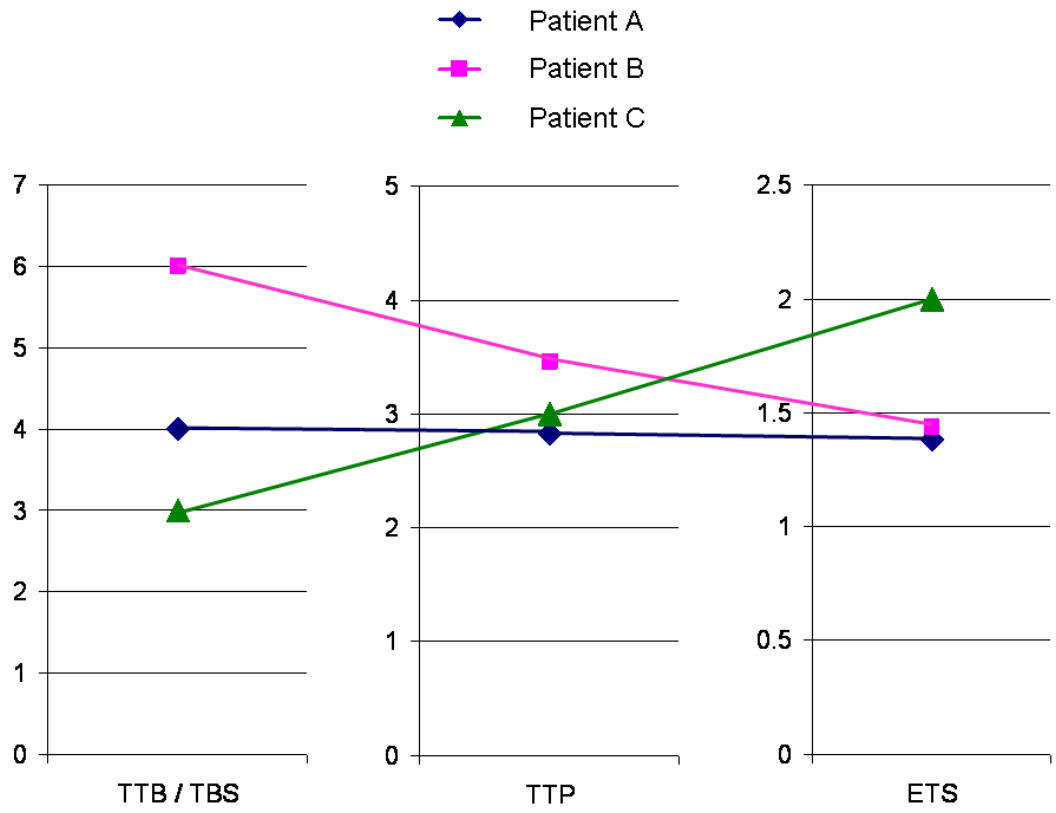


FIGURE 7.1 – Classement des différents profils de toxicité selon le score utilisé

TTB/TBS : Le score de Bekele et Thall (*Total Toxicity Burden*)/Le score de Lee et al. (*Total Burden Score*).
TTP : Notre proposition (*Total Toxicity Profile*).
ETS : Le score de Chen et al. (*Equivalent Toxicity Score*). Pour ce score, les paramètres de la fonction logistique définissant la partie décimale ont été fixés selon les recommandations des auteurs à -2 et 0.25 [2].

Le cumul de toxicités modérées peut donc aboutir à un score TTB ou TBS très élevé, possiblement très supérieur à celui d'un patient présentant une DLT isolée. Ce classement semble aberrant. De façon générale, même si le résultat dépend de la matrice de poids, la norme euclidienne (TTP) est plus appropriée que la somme arithmétique (TTB et TBS) pour mesurer la longueur d'un vecteur dans un espace multidimensionnel. Grâce à ses propriétés géométriques (inégalité triangulaire), le score d'un patient cumulant deux toxicités est plus faible que la somme des scores de deux patients différents éprouvant chacun une toxicité : les toxicités multiples de patients différents

apportent ainsi plus de poids que celles cumulées par un même patient. Contrairement au score ETS, le cumul de toxicités modérées chez un même patient peut cependant aboutir à un score TTP égal ou supérieur au score associé à une DLT isolée telle que définie classiquement.

Les méthodes classiques basées sur le critère binaire DLT n'étant pas directement applicables à notre score, nous avons donc proposé une extension de la méthode CRM utilisant un score de toxicité pour guider la recherche de dose. Cette méthode appelée QLCRM pour *Quasi-Likelihood Continual Reassessment Method* suppose que le score de toxicité, une fois normalisé (nTTP, *normalized Total Toxicity Profile*), peut être assimilé à une fraction d'événement qui suit une loi de quasi-Bernoulli. Elle utilise un modèle linéaire généralisé avec une fonction de lien *Logit* à un seul paramètre pour modéliser ce score dans un cadre fréquentiste.

Telle que nous l'avions définie, la méthode QLCRM suppose que les deux premiers moments de la variable à expliquer sont ceux d'une variable aléatoire distribuée selon une loi de Bernoulli, d'où le terme quasi-Bernoulli. Cependant, différentes variantes de cette méthode sont possibles, en termes de choix de la fonction de la variance, de choix de la fonction de lien et d'inférence d'estimation (bayésienne ou fréquentiste). La variance de Bernoulli est utilisée, par défaut, pour décrire la variabilité du score nTTP compris entre 0 et 1. Nous avons calculé la fonction de variance "analytique" du score normalisé à l'aide de la méthode Delta (cf. annexe 8.6). A noter qu'une modélisation conjointe des différentes composantes de toxicité serait nécessaire pour développer la méthode de recherche de dose associée à cette fonction de variance. Ce volet pourrait faire l'objet de nouveaux travaux de recherche.

Nous avons également calculé la variance de Wedderburn permettant de décrire la variabilité du score nTTP [55]. Contrairement à la variance de Bernoulli, la variance de Wedderburn présente des valeurs proches de la variance analytique, pour tous les scénarios étudiés (cf. annexe 8.5). Après avoir exprimé la fonction de quasi-vraisemblance associée à la variance de Wedderburn, nous avons comparé la performance de la méthode QLCRM avec une variance de Bernoulli à celle de la méthode QLCRM avec une variance de Wedderburn. Contrastant avec les estimations de la variance, la méthode QLCRM avec la variance de Bernoulli présente globalement une performance supérieure à celle de la méthode utilisant la variance de Wedderburn.

En ce qui concerne la fonction de lien et l'inférence d'estimation, notre choix initial s'est porté sur un modèle logistique dans un cadre fréquentiste. Nous avons ensuite étendu la QLCRM en utilisant différentes fonctions de lien et inférences. Les cinq variantes étudiées, toutes basées sur un seul paramètre, présentent en moyenne des résultats très similaires dans les différents scénarios étudiés. Le choix de la fonction de lien ainsi que l'inférence d'estimation ont ainsi un impact faible sur la performance de la méthode. Comme pour les différentes variantes possibles de la CRM classique avec le critère DLT, l'objectif de ces propositions est plus de généraliser le principe de cette approche que de sélectionner la meilleure méthode. Les paramètres de la méthode (choix de la fonction de lien et de l'inférence d'estimation) doivent ainsi être spécifiés, pour chaque nouvel essai, à l'aide d'une étude de simulation pour différents scénarios

plausibles dans le contexte de l'étude. A noter que l'influence du choix du modèle de travail aurait pu être étudiée de façon plus extensive puisque nous nous sommes limités à deux propositions de modèle de travail. Cependant, au vu des excellents résultats obtenus avec le modèle de travail défini par la fonction *getprior* sur un nombre de scénarios couvrant une gamme variée de situations, ce choix nous a semblé raisonnable comme point de départ.

Sans faire de recommandations dans l'absolu, nous avons choisi de travailler avec la fonction de lien *Logit* et l'inférence fréquentiste pour les raisons suivantes : i) la fonction de lien *Logit* présente un bon ajustement des relations dose-réponse, et elle est la plus utilisée dans le cas classique avec le critère binaire DLT ii) l'inférence fréquentiste est mieux comprise par les cliniciens car elle ne nécessite pas de choix *a priori* sur le paramètre du modèle à estimer. Un des avantages à utiliser un score de toxicité dans un cadre fréquentiste, par comparaison au critère binaire DLT, est que l'estimation du modèle est possible dès la première observation de toxicité. Ceci conduit à une première étape algorithmique courte.

Nous avons ensuite comparé les méthodes QLCRM et QCRM à deux autres méthodes de recherche de dose existantes susceptibles d'intégrer un score de toxicité : la méthode d'Ivanova et Kim et la méthode de Chen et al. [5, 2].

En utilisant le score nTTP pour guider l'escalade de dose et l'identification de la dose à recommander, les quatre méthodes présentent une bonne performance en termes de capacité à identifier correctement la dose à recommander et de contrôle du surdosage. Dans un essai incluant 36 patients, le pourcentage de sélection correcte de la dose à recommander (PCS) obtenu avec les méthodes QLCRM et QCRM varie de 80 à 91% en fonction des situations. Bien que la méthode UA présente une performance proche de celle des méthodes paramétriques, elle inclut globalement moins de patients à la vraie dose à recommander, confirmant son caractère conservateur. Ceci peut s'expliquer par le fait que la méthode UA répète la même dose si le score associé à cette dose est relativement proche du score-cible, contrairement aux méthodes QLCRM et QCRM qui escaladent la dose dans ce cas de figure. La méthode EID présente en moyenne une performance plus faible que les trois autres méthodes pour les différents scénarios étudiés, notamment en termes de convergence. La moindre performance de la méthode EID par rapport aux méthodes QLCRM et QCRM peut sembler surprenante si l'on considère que cette méthode est également basée sur une régression pendant l'étape d'escalade de dose. De plus, la règle de décision de la méthode EID est similaire à celle des méthodes paramétriques. Une des explications à la moindre performance de la méthode EID par rapport aux méthodes paramétriques est que les données collectées à des doses inférieures n'ont aucune influence sur les estimations des scores à des doses plus élevées, contrairement aux méthodes paramétriques. A noter que des études de simulation comparant différentes méthodes de recherche de dose basées sur une régression isotonique pour le critère de toxicité classique DLT ont montré que les méthodes utilisant la régression isotonique uniquement en fin d'essai pour l'identification de la dose à recommander ont une performance supérieure à celles utilisant la régression isotonique également pour l'escalade de dose [68, 69].

L'utilisation d'un score de toxicité augmente considérablement la performance de

la méthode de recherche de dose par rapport aux méthodes CRM guidées par le critère DLT : le pourcentage de sélection correcte varie de 51 à 56% en fonction du scénario avec les méthodes CRM. Ce pourcentage est cohérent avec les résultats précédemment obtenus par différents auteurs [13, 24, 50]. Cette performance est limitée par la simple variabilité binomiale du critère binaire DLT, en particulier pour les essais de petite taille [24, 45].

Cette différence importante de résultats selon que l'on utilise un critère binaire ou un critère quasi-continu est en accord avec les exemples donnés par Senn et Julious dans d'autres contextes d'essais cliniques : le fait de dichotomiser des variables continues ou ordinales dans l'analyse des essais entraîne une perte d'information préjudiciable [70]. Comme mentionné par les auteurs, le choix des critères de jugement dans les essais cliniques devrait être une préoccupation de tous et en particulier du statisticien.

Malgré les efforts établis pour assurer la qualité du suivi des essais de phase I, rapporter les événements indésirables reste un processus complexe et sensible aux erreurs d'observation. Brundage et al. avaient noté une forte variabilité des classifications d'événements indésirables entre différents évaluateurs (clinicien, attaché de recherche clinique, etc) suite à l'utilisation de deux échelles de toxicité [71]. Sackett a identifié un certain nombre de facteurs pouvant contribuer à ce désaccord liés au patient, à l'évaluateur et à leur interaction [72]. Ces erreurs de classement peuvent ainsi avoir un impact négatif sur les méthodes basées sur un score de toxicité, d'autant plus que le score inclut tous les grades des toxicités. Nous avons étudié la sensibilité de ces méthodes pour différents types d'erreur.

Etant donné la faible performance de la méthode EID comparée à celles des autres méthodes basées sur le score, cette méthode a été exclue de l'étude de robustesse. Les méthodes utilisant le score TTP sont relativement robustes aux erreurs de mesure. Leur performance reste excellente si l'erreur ne concerne que les grades définissant la classification DLT. Comme prévu, pour une probabilité d'erreur donnée, les erreurs affectant tous les grades ont un impact plus important sur la performance des différentes méthodes basées sur le score. Cependant, cette dernière reste supérieure à celle des méthodes basées sur le critère DLT pour des probabilités d'erreur $< 15\%$.

Les erreurs de sur-classement peuvent probablement être plus fréquemment observées que les erreurs de sous-classement, en particulier dans les études de première administration chez l'homme où les cliniciens voudraient protéger les patients d'un risque de toxicité potentiellement sévère [73]. Les méthodes basées sur le score semblent plus robustes à ce type d'erreur qu'aux erreurs de sous-classement.

Nous avons conscience que les hypothèses considérées pour étudier l'impact des erreurs d'observation sont peut-être simplistes. Il est probable que "dans la vraie vie" : i) les erreurs de sur-classement et de sous-classement soient associées, ii) les erreurs puissent porter sur un grade non adjacent, et iii) le risque d'erreur varie au cours de l'essai (phénomène d'apprentissage, principe de précaution en début d'essai ou à l'inverse quand les investigateurs pensent approcher la dose cible). De plus, nous avons considéré que les erreurs affectaient uniquement le classement du grade. Cependant, une question fréquente rencontrée dans les essais de phase I concerne l'imputabilité d'un

événement indésirable au traitement évalué [74, 75]. La FDA a récemment proposé des recommandations afin de mieux clarifier la définition des événements indésirables liés aux médicaments [76]. Iasonos et al. ont étudié l'impact des toxicités non-liées au traitement évalué sur les méthodes guidées par la DLT [73]. Cela pourrait faire l'objet de nouveaux travaux de recherche en utilisant les méthodes guidées par le score.

Toutes les méthodes comparées dans ce travail méthodologique sont basées sur le score TTP. Nous avons travaillé avec un clinicien pour définir une matrice de poids ainsi que la valeur de toxicité jugée acceptable (score-cible), indépendamment d'un contexte clinique particulier. Ces valeurs seraient à définir pour un essai clinique réel en prenant en compte le contexte de la maladie et de la population de l'étude ainsi que le profil de toxicité attendu du traitement évalué. En ce qui concerne la définition des poids, une étude de sensibilité de l'impact des erreurs associées à la définition des poids sur les différentes méthodes comparées pourrait ouvrir à des perspectives de recherche. Bien que l'élaboration du score-cible ait été guidée par un clinicien en analogie avec la définition du pourcentage-cible de DLT, cette valeur peut être assez critiquée. Comme pour la méthode CRM classique utilisant la DLT, on peut étendre ce travail en choisissant d'autres valeurs cibles de score de toxicité. Nous pouvons aussi définir un intervalle contenant des valeurs de toxicité jugées acceptables et non une valeur ponctuelle. La dose à recommander pourrait alors être définie comme dose associée à un score estimé appartenant à cet intervalle. Plusieurs doses pourraient ainsi être sélectionnées au terme de l'essai de phase I pour être comparées dans les phases ultérieures du développement du traitement, comme c'est le cas habituellement en dehors de la cancérologie.

Pour être cohérents dans la comparaison des différentes méthodes, nous avons choisi de travailler avec le même score pour les différentes méthodes. Ceci ne nous permet donc pas de séparer l'intérêt du score TTP de l'intérêt de la méthode de recherche de dose. Il nous a semblé difficile, voire discutable de comparer les méthodes en utilisant les différents scores possibles. En effet, comme mentionné précédemment, l'utilisation d'un autre score peut aboutir à un classement différent des patients (cf. figure 7.1). Dans l'étape de définition du score-cible, quand nous obtenons un classement cohérent des cohortes hypothétiques avec les scores TTP, le classement des mêmes cohortes avec le score TTB ne semble plus cohérent. Un classement cohérent des patients avec ces deux scores n'est pas possible en utilisant les mêmes poids. La comparaison des résultats obtenus en utilisant les scores TTP, TTB/TBS ou ETS aurait donc nécessité une démarche complète avec les cliniciens, incluant la définition du score cible pour chaque type de score. De plus, à partir des mêmes scénarios définis sur la base des probabilités des différentes toxicités élémentaires, l'utilisation des différents scores aurait peut-être conduit à définir des doses différentes comme vraie dose à recommander.

Loin d'être exhaustifs dans notre étude de comparaison, nous avons volontairement omis d'évaluer la méthode de Bekele et Thall, publiée en 2004 [38]. Bekele et Thall ont proposé d'identifier la dose à recommander à partir de la modélisation du score TTB construit sur l'ensemble des toxicités observées (cf. page 84). La méthode d'escalade de dose et d'estimation de la dose à recommander est une méthode paramétrique utilisant

des variables latentes pour décrire la distribution de probabilité conjointe des poids. Le modèle, de dimension égale au nombre de types de toxicité considérés, est très complexe conduisant à une méthode de calcul très lourde et difficile à appliquer dans la pratique.

A noter qu'à l'étape préalable de génération de scénarios, nous avons observé que la dose à recommander est plus faible, dans la plupart des cas, si l'on considère le score de toxicité plutôt que le critère DLT. Ceci dépend des valeurs cibles de toxicité définies pour les deux critères utilisés (DLT ou score). Le score TTP est plus conservateur que la DLT, puisqu'il inclut l'ensemble des grades des différentes toxicités et non pas seulement les grades définissant la DLT.

A notre connaissance, nous sommes les premiers à comparer différentes approches susceptibles d'intégrer un score de toxicité en simulant des données élémentaires de toxicité dans un espace multidimentionnel. Dans l'étude de simulation évaluant les différentes méthodes, nous avons supposé que la toxicité attendue portait essentiellement sur trois types de toxicité indépendants. Ainsi, nous avons défini chaque scénario par la distribution des probabilités d'observer chaque grade de chaque type de toxicité. Afin d'obtenir des scénarios plausibles, nous avons fixé différentes contraintes sur la distribution des probabilités d'observer les grades de toxicité élémentaire : par exemple, nous avons supposé que la distribution de la probabilité d'observer un grade 0 était strictement décroissante avec la dose. Yuan et Chen ont également proposé une méthode de recherche de dose (QCRM et EID) permettant de prendre en compte l'ensemble des toxicités en utilisant le score (TTB de Bekele et Thall pour la méthode de Yuan ; ETS pour la méthode de Chen). Cependant, dans leurs études de simulation, un seul type de toxicité était considéré. Par ailleurs, dans l'article de Yuan, certains scénarios semblaient peu plausibles cliniquement : par exemple considérer la probabilité de grade 3 égale à 0 quelle que soit la dose et la probabilité de grade 2 et de grade 4 différentes de 0.

S'il semble raisonnable de travailler avec des scénarios de relation dose-toxicité monotones et croissants comme point de départ, il serait intéressant de considérer d'autres scénarios peut-être plus adaptés aux essais de phase I évaluant les thérapies ciblées, en particulier en combinant des critères d'efficacité au critère de toxicité. En effet, l'hypothèse d'une relation monotone et croissante est remise en question avec les thérapies ciblées, notamment en ce qui concerne la relation dose-efficacité. Nous avons prévu d'intégrer des données d'efficacité, ou du moins des marqueurs d'activité, afin de définir la dose optimale biologique [1].

Nous sommes conscients que le sens clinique du score TTP et du score-cible est moins intuitif que la définition d'une probabilité de DLT. Par analogie au pourcentage-cible de DLT pour la CRM classique, ce dernier correspond au score moyen jugé acceptable permettant de définir la dose à recommander en fin d'essai. Le processus de construction du score TTP doit être élaboré préalablement pour chaque nouvel essai

1. Cette thématique fait partie d'un vaste projet de recherche soumis par l'équipe de la Mayo Clinic auprès du NCI pour un financement RO1.

clinique en interaction étroite avec les cliniciens. Nous avons illustré la construction de ce score, d'une manière rétrospective, dans l'essai de phase I évaluant l'erlotinib chez l'enfant. Les investigateurs interrogés dans le cadre de l'analyse rétrospective de l'essai erlotinib ont facilement adhéré à notre proposition de score, mais on peut s'interroger quant à leur représentativité. Nous sommes conscients que la méthode que nous proposons est complexe. A la complexité relative de la méthode CRM qui nécessite des interactions avec le statisticien en continu contrairement à la majorité des méthodes algorithmiques, s'ajoute la complexité liée à l'utilisation du score. En effet, elle demande plus d'efforts que l'utilisation simple du critère DLT. Avant de commencer l'essai, la difficulté réside dans la définition des toxicités liées au traitement et leurs poids relatifs ainsi que dans l'élaboration du score-cible. En cours d'essai, la collecte de données de toxicité peut être une étape complexe nécessitant beaucoup d'interactions entre cliniciens, ARC (attaché de recherche clinique), data manager et statisticien afin d'avoir les données en temps réel. Bien que ceci soit possible dans certains centres (par exemple à la Mayo Clinic, Rochester, USA), ce n'est probablement pas encore le cas dans de nombreux centres où même la simple information binaire DLT n'est pas obtenue en temps réel. A noter que tous les efforts supplémentaires pour l'élaboration du score et du score-cible seraient récompensés par l'excellente performance de ces méthodes.

Dans le contexte de l'essai erlotinib, nous avons interrogé trois cliniciens experts pour définir les poids et le score de toxicité jugé acceptable. Dans cette partie, nous nous sommes concentrés sur la toxicité dermatologique d'intensité inférieure ou égale à 3. Ce choix peut être critiqué puisque la prise en charge des patients est améliorée et que les toxicités cutanées sévères peuvent être partiellement évitées par une l'utilisation de crèmes adaptées. Il aurait pu être intéressant d'intégrer d'autres types de toxicité, digestive par exemple. De plus, les différents types de toxicité dermatologique considérés sont probablement corrélés, alors que le score de toxicité suppose l'indépendance des différents axes de toxicité. Notons que ce score permet d'introduire, si nécessaire en cours d'essai, une toxicité inattendue plus sévère que le profil de toxicité le plus grave calculé à partir de la matrice initialement définie. Pour une application prospective de notre approche, il est préférable de connaître au préalable le profil de toxicité attendu de la molécule afin de sélectionner les items de toxicité à prendre en compte dans la construction du score. Ceci est plus facile quand on évalue une molécule déjà étudiée dans une nouvelle population (par exemple chez l'enfant) ou quand on évalue une nouvelle molécule d'une classe thérapeutique déjà connue.

Dans le contexte de l'essai erlotinib, les poids consensuels ont été facilement obtenus. Ces poids ont conduit à une échelle de toxicité différente de celle de la classification NCI-CTC, ce qui confirme l'intérêt de l'approche. L'échelle finale de poids variait de 0 à 20. Cette échelle est arbitraire puisque la méthode est invariante à ce choix. Les poids doivent surtout refléter au mieux l'importance clinique relative des différentes toxicités. Dans notre application rétrospective, le poids associé au décès a été réajusté à 20 pour le différencier de la toxicité dermatologique la plus sévère possible. Dans ce travail, les poids reflètent l'importance clinique relative des toxicités. Cependant, on peut étendre cette approche pour prendre en compte d'autres paramètres dans la définition des poids, par exemple la durée d'une toxicité donnée, sa réversibilité ou non.

Comme énoncé précédemment, la définition du score de toxicité jugé acceptable est cruciale dans l'élaboration de la méthode. Cette étape repose sur le classement de cohortes hypothétiques par les cliniciens. Le nombre de cohortes hypothétiques à classer doit être limité pour que le processus reste acceptable. Il est important de définir ces cohortes de façon à bien identifier "les scores limites" correspondant aux décisions escalader/répéter d'une part et répéter/désescalader d'autre part. Sur l'ensemble des 23 cohortes générées dans notre application, les décisions des experts diffèrent pendant les étapes intermédiaires reflétant leurs différents points de vue. Suite à une discussion riche entre les cliniciens, les décisions consensuelles ont été facilement obtenues. A la fin de ce processus, un quatrième expert indépendant a été interrogé sur les conclusions des trois premiers experts. Ce clinicien a validé le choix consensuel. Travailler avec différents experts au cours du processus en utilisant la méthode Delphi et utiliser une étape finale de validation en interrogeant un expert externe permet d'augmenter la fiabilité de cette approche et de limiter la subjectivité des réponses. Il est probable qu'aucune règle de conversion standard entre les grades et les poids numériques ne puisse être définie pour une utilisation ultérieure prospective pour tous les essais de recherche de dose. Il s'agit donc d'une étape préalable à l'initiation d'un nouvel essai. Cependant, il est important de noter que dans les essais classiques de phase I, la définition de DLT peut également varier entre les essais parce que les évaluateurs peuvent être plus "conservateurs" ou plus "agressifs" dans leur définition de la toxicité acceptable et inacceptable selon le contexte de l'essai [77], plus particulièrement dans les essais évaluant les thérapies ciblées [78]. Un résumé de la démarche d'élaboration du score et du score-cible, ainsi que la méthode QLCRM est illustré par les figures 7.2 et 7.3

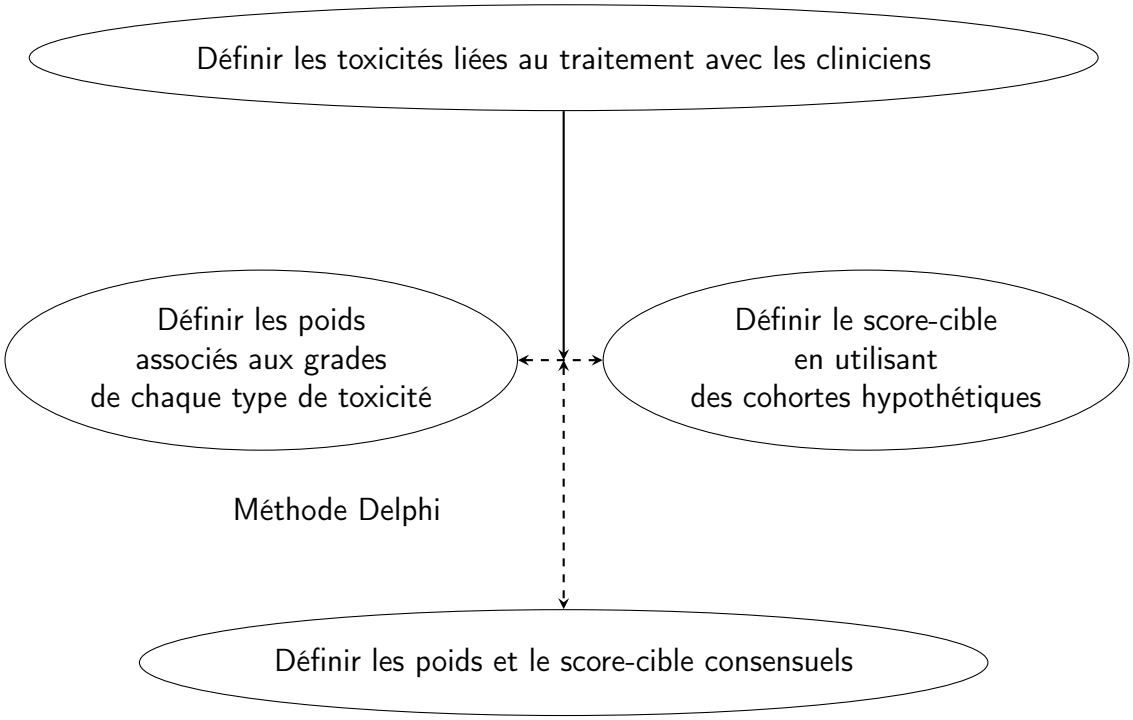


FIGURE 7.2 – Elaboration du score et du score-cible

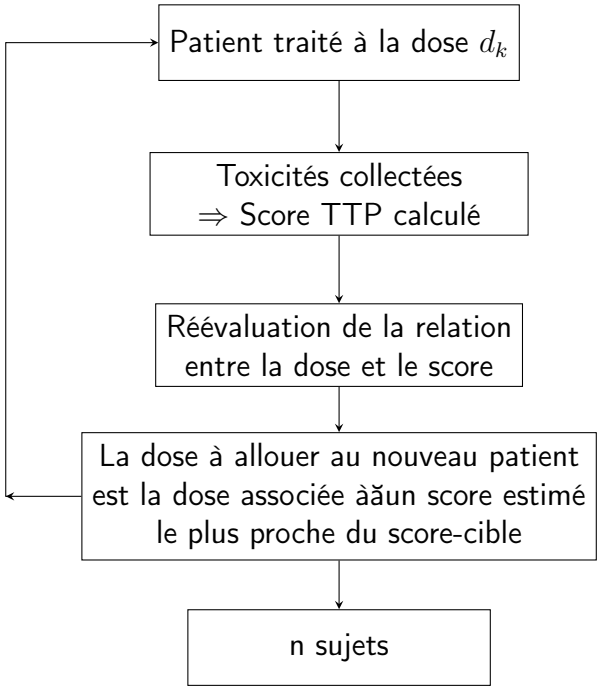


FIGURE 7.3 – Méthodes paramétriques QLCRM et QCRM (extension de la CRM) pour un score de toxicité

La dose à recommander estimée avec la méthode QLCRM était inférieure à celle identifiée par la méthode classique CRM basée sur le critère DLT (100 au lieu de 125 mg/m^2), ce qui corrobore le caractère conservateur de l'utilisation du score par rapport au critère DLT. L'application rétrospective a inclus les observations finales des 20 patients évalués au cours de la période de 6 semaines ainsi que leurs doses explorées. Le schéma d'escalade de dose aurait probablement été différent si la méthode avait été utilisée en amont. Dans cette étude, la dose à recommander d'erlotinib estimée avec la méthode CRM classique est plus élevée que celle de l'adulte (150 $mg \approx 87\ mg/m^2$) [79]. Cela peut être expliqué par la faible incidence des effets indésirables gastro-intestinaux dans l'étude pédiatrique. Une étude américaine publiée en 2008 a également évalué l'erlotinib chez l'enfant. Cette étude a recommandé la dose 85 mg/m^2 [80]. Celle-ci est également inférieure à celle recommandée par l'essai français (125 mg/m^2). Dans leur étude, la définition de DLT était plus conservatrice incluant, en partie, toute toxicité de grade 2 non-hematologique qui persiste pendant plus de 7 jours et qui conduit à l'interruption du traitement. Récemment, l'erlotinib a été évalué en combinaison avec la rapamycine chez les enfants (avec un gliome de bas grade) lors d'une étude de phase I/II où la toxicité et l'efficacité sont conjointement considérées [81]. Dans cette étude de combinaison, la dose d'erlotinib a été fixée à 65 mg/m^2 au vu des profils de toxicité. Les doses 100 et 125 mg/m^2 pourraient être comparées dans les phases ultérieures du développement du traitement par erlotinib chez l'enfant.

Le nombre de thérapies ciblées en développement clinique est en pleine augmentation. Il apparaît un besoin urgent de nouvelles approches mieux adaptées à ce contexte, en particulier pour les phases précoces du développement thérapeutique. L'utilisation du score de toxicité TTP permet d'améliorer considérablement la performance de la méthode de recherche de dose par rapport aux méthodes CRM guidées par le critère binaire DLT. Notre approche s'avère séduisante pour les essais de phase I évaluant des médicaments associés à peu de DLTs *a priori*, mais avec des toxicités multiples modérées probables. Dans ce travail, la période d'observation était limitée aux premières semaines du traitement. Afin de mieux évaluer la toxicité au long cours de ces traitements administrés de plus en plus de façon prolongée, une extension de cette méthode est prévue pour prendre en compte les données longitudinales de toxicité, au-delà du premier cycle de traitement. L'utilisation des données répétées de toxicité permettant d'évaluer la variation du score de toxicité au cours du temps devrait être très informative [2].

2. Cette thématique fait partie également d'un vaste projet de recherche soumis par l'équipe de la Mayo Clinic auprès du NCI pour un financement RO1.

Bibliographie

[1] J. O'Quigley, M. Pepe, and L. Fisher. Continual reassessment method : a practical design for phase 1 clinical trials in cancer. *Biometrics*, 46(1) :33–48, 1990.

[2] Z. Chen, M. D. Krailo, S. P. Azen, and M. Tighiouart. A novel toxicity scoring system treating toxicity response as a quasi-continuous variable in Phase I clinical trials. *Contemp Clin Trials*, 31(15S (May 20 Supplement)) :473–482, 2010.

[3] Z. Yuan, R. Chappell, and H. Bailey. The continual reassessment method for multiple toxicity grades : a Bayesian quasi-likelihood approach. *Biometrics*, 63 :173–179, 2007.

[4] M. Ezzalfani, S. Zohar, R. Qin, S. J. Mandrekar, and M. C. Deley. Dose-finding designs using a novel quasi-continuous endpoint for multiple toxicities. *Stat Med*, 32(16) :2728–2746, 2013.

[5] A. Ivanova and S. H. Kim. Dose finding for continuous and ordinal outcomes with a monotone objective function : a unified approach. *Biometrics*, 65 :307–315, 2009.

[6] S. Chevret. *Statistical Methods for dose-Finding Experiments*, pages 5–18. Statistics in Practice. John Wiley and Sons Ltd., Chichester, 2006.

[7] National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE) v4.0. *http ://www.fda.gov/cder/cancer/toxicityframe.htm*, 2009.

[8] B. Geoerger, D. Hargrave, F. Thomas, A. Ndiaye, D. Frappaz, F. Andreiuolo, P. Varlet, I. Aerts, R. Riccardi, T. Jaspan, E. Chatelut, M. C. Le Deley, X. Paoletti, C. Saint-Rose, P. Leblond, B. Morland, J. C. Gentet, V. Meresse, and G. Vassal. Innovative Therapies for Children with Cancer pediatric phase I study of erlotinib in brainstem glioma and relapsing/refractory brain tumors. *Neuro-oncology*, 13 :109–118, 2011.

[9] J. W. Dixon and A. M. Mood. A Method for Obtaining and Analyzing Sensitivity Data. *JASA*, 43 :109–126, 1948.

[10] E. Storer Barry. design and analysis of phase I clinical trials . *Biometrics*, 45 :925–937, 1989.

[11] J. O'Quigley and S. Zohar. Experimental designs for phase I and phase I/II dose-finding studies. *Br. J. Cancer*, 94(5) :609–613, 2006.

[12] J. O'Quigley and S. Chevret. Methods for dose finding studies in cancer clinical trials : a review and results of a Monte Carlo study. *Stat Med*, 10(11) :1647–64, 1991.

[13] S. Chevret. The continual reassessment method in cancer phase I clinical trials : a simulation study. *Stat Med*, 12(12) :1093–1108, 1993.

[14] D. Faries. Practical modifications of the continual reassessment method for phase I cancer clinical trials. *J Biopharm Stat*, 4(2) :147–164, 1994.

[15] S. N. Goodman, M. L. Zahurak, and S. Piantadosi. Some practical improvements in the continual reassessment method for phase I studies. *Stat Med*, 14(11) :1149–1161, 1995.

[16] S. Piantadosi, J. D. Fisher, and S. Grossman. Practical implementation of a modified continual reassessment method for dose-finding trials. *Cancer Chemother. Pharmacol.*, 41(6) :429–436, 1998.

[17] A. Iasonos, A. S. Wilton, E. R. Riedel, V. E. Seshan, and D. R. Spriggs. A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in Phase I dose-finding studies. *Clin Trials*, 5(5) :465–477, 2008.

[18] J. O'Quigley and L. Z. Shen. Continual reassessment method : a likelihood approach. *Biometrics*, 52(2) :673–684, 1996.

[19] S. Piantadosi and G. Liu. Improved designs for dose escalation studies using pharmacokinetic measurements. *Stat Med*, 15(15) :1605–1618, 1996.

[20] J. Babb, A. Rogatko, and S. Zacks. Cancer phase I clinical trials : efficient dose escalation with overdose control. *Stat Med*, 17(10) :1103–1120, 1998.

[21] Y. K. Cheung and R. Chappell. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*, 56(4) :1177–1182, 2000.

[22] A. Mauguen, M. C. Le Deley, and S. Zohar. Dose-finding approach for dose escalation with overdose control considering incomplete observations. *Stat Med*, 30(13) :1584–1594, 2011.

[23] M. Y. Polley. Practical modifications to the time-to-event continual reassessment method for phase I cancer trials with fast patient accrual and late-onset toxicities. *Stat Med*, 30(17) :2130–2143, 2011.

[24] X. Paoletti and A. Kramar. A comparison of model choices for the Continual Reassessment Method in phase I cancer trials. *Stat Med*, 28 :3012–3028, 2009.

[25] K. A. Gelmon, E. A. Eisenhauer, A. L. Harris, M. J. Ratain, and P. Workman. Anticancer agents targeting signaling molecules and cancer cell environment : challenges for drug development ? *J. Natl. Cancer Inst.*, 91 :1281–1287, 1999.

[26] E. L. Korn, S. G. Arbuck, J. M. Pluda, R. Simon, R. S. Kaplan, and M. C. Christian. Clinical trial designs for cytostatic agents : are new approaches needed ? *J. Clin. Oncol.*, 19 :265–272, 2001.

[27] W. R. Parulekar and E. A. Eisenhauer. Novel endpoints and design of early clinical trials. *Ann. Oncol.*, 13 Suppl 4 :139–143, 2002.

[28] W. R. Parulekar and E. A. Eisenhauer. Phase I trial design for solid tumor studies of targeted, non-cytotoxic agents : theory and practice. *J. Natl. Cancer Inst.*, 96 :990–997, 2004.

[29] N. Penel, A. Adenis, S. Clisant, and J. Bonneterre. Nature and subjectivity of dose-limiting toxicities in contemporary phase 1 trials : comparison of cytotoxic versus non-cytotoxic drugs. *Invest New Drugs*, 2010.

[30] C. Le Tourneau, V. Dieras, P. Tresca, W. Cacheux, and X. Paoletti. Current challenges for the early clinical development of anticancer drugs in the era of molecularly targeted agents. *Target Oncol*, 5 :65–72, 2010.

[31] N. S. Azad, E. M. Posadas, V. E. Kwitkowski, S. M. Steinberg, L. Jain, C. M. Annunziata, L. Minasian, G. Sarosy, H. L. Kotz, A. Premkumar, L. Cao, D. McNally, C. Chow, H. X. Chen, J. J. Wright, W. D. Figg, and E. C. Kohn. Combination targeted therapy with sorafenib and bevacizumab results in enhanced toxicity and antitumor activity. *J. Clin. Oncol.*, 26 :3709–3714, 2008.

[32] S. S. Ramalingam, C. P. Belani, P. C. Mack, E. E. Vokes, J. Longmate, R. Govindan, M. Koczywas, S. P. Ivy, and D. R. Gandara. Phase II study of Cediranib (AZD 2171), an inhibitor of the vascular endothelial growth factor receptor, for second-line therapy of small cell lung cancer (National Cancer Institute). *J Thorac Oncol*, 5 :1279–1284, 2010.

[33] P. M. LoRusso, S. A. Boerner, and L. Seymour. An overview of the optimal planning, design, and conduct of phase I studies of new therapeutics. *Clin. Cancer Res.*, 16 :1710–1718, 2010.

[34] S. P. Ivy, L. L. Siu, E. Garrett-Mayer, and L. Rubinstein. Approaches to phase 1 clinical trial design focused on safety, efficiency, and selected patient populations : a report from the clinical trial design task force of the national cancer institute investigational drug steering committee. *Clin. Cancer Res.*, 16 :1726–1736, 2010.

[35] E. L. Korn. Nontoxicity endpoints in phase I trial designs for targeted, non-cytotoxic agents. *J. Natl. Cancer Inst.*, 96 :977–978, 2004.

[36] V. J. Suman, A. Dueck, and D. J. Sargent. Clinical trials of novel and targeted therapies : endpoints, trial design, and analysis. *Cancer Invest.*, 26 :439–444, 2008.

[37] R. K. Paul, W. F. Rosenberger, and N. Flournoy. Quantile estimation following non-parametric phase I clinical trials with ordinal response. *Stat Med*, 23 :2483–2495, 2004.

[38] B. Nebiyou Bekele and Peter F Thall. Dose-Finding Based on Multiple Toxicities in a Soft Tissue Sarcoma Trial. *JASA*, 99(465) :26–35, 2004.

[39] E. M. Van Meter, E. Garrett-Mayer, and D. Bandyopadhyay. Proportional odds model for dose-finding clinical trial designs with ordinal toxicity grading. *Stat Med*, 30(17) :2070–2080, 2011.

[40] S. M. Lee, D. L. Hershman, P. Martin, J. P. Leonard, and Y. K. Cheung. Toxicity burden score : a novel approach to summarize multiple toxic effects. *Ann. Oncol.*, 23(2) :537–541, 2012.

[41] S. M. Lee and Y. K. Cheung. Model calibration in the continual reassessment method. *Stat Med*, 6 :227–238, 2011.

[42] Ying Kuen Cheung. *Dose Finding by the Continual Reassessment Method*, volume 41. Chapman & Hall, 2011.

[43] J. O'Quigley and E. Reiner. A stopping rule for the continual reassessment method. *Biometrika*, 85(3) :741–748, 1998.

[44] S. Zohar and S. Chevret. The continual reassessment method : comparison of Bayesian stopping rules for dose-ranging studies. *Stat Med*, 20(19) :2827–2843, 2001.

[45] J. O'Quigley. Continual reassessment designs with early termination. *Biostatistics*, 3(1) :87–99, 2002.

[46] J. O'Quigley. Theoretical study of the continual reassessment method. *J. of statistical planning and inference*, 136(6) :1765–1780, 2006.

[47] E. Garrett-Mayer. The continual reassessment method for dose-finding studies : a tutorial. *Clin Trials*, 3(1) :57–71, 2006.

[48] L. Z. Shen and J. O'Quigley. Consistency of continual reassessment method under model misspecification. *Biometrika*, 83(2) :395–405, 1996.

[49] J. O'Quigley. Another look at two phase I clinical trial designs. *Statistics in medicine*, 18(20) :2683–2690, 1999.

[50] A. Iasonos, A. S. Wilton, E. R. Riedel, V. E. Seshan, and D. R. Spriggs. A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in Phase I dose-finding studies. *Clin Trials*, 5(5) :465–477, 2008.

[51] E. L. Korn, D. Midthune, T. T. Chen, L. V. Rubinstein, M. C. Christian, and R. M. Simon. A comparison of two phase I trial designs. *Stat Med*, 13(18) :1799–1806, 1994.

[52] M. J. Ratain, R. Mick, R. L. Schilsky, and M. Siegler. Statistical and ethical issues in the design and conduct of phase I and II clinical trials of new anticancer agents. *J. of the N. Cancer I.*, 85(20) :1637–1643, 1993.

[53] C. Ahn. An evaluation of phase I cancer clinical trial designs. *Stat Med*, 17(14) :1537–1549, 1998.

[54] S. Moller. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Stat Med*, 14(9) :911–922, 1995.

[55] P. McCullagh and J.A. Nelder. *Generalized Linear Model*, chapter 9. Monographs on Statistics and Applied Probability. 2nd ed. New York, Chapman and Hall, 1989.

[56] J. Whitehead and H. Brunier. Bayesian decision procedures for dose determining experiments. *Stat Med*, 14(9-10) :885–893, 1995.

[57] C. Gourieroux, A. Monfort, and A. Trognon. Pseudo maximum likelihood methods : Theory. *Econometrica : J. E. S.*, pages 681–700, 1984.

[58] L. Papke and J. Wooldridge. Econometric Methods For Fractional Response Variables with an Application to 401 (K) Plan Participation Ratess. *Journal of Applied Econometrics*, 11 :619–632, 1996.

[59] R. W. Wedderburn. Quasi-likelihood functions, generalized linear models, and the GaussŮNewton method. *Biometrika*, 61(3) :439–447, 1974.

[60] B. Wacker, T. Nagrani, J. Weinberg, K. Witt, G. Clark, and P. J. Cagnoni. Correlation between development of rash and efficacy in patients treated with the epidermal growth factor receptor tyrosine kinase inhibitor erlotinib in two large phase III studies. *Clin. Cancer Res.*, 13 :3913–3921, 2007.

[61] S. G. O'Brien and M. W. Deininger. Imatinib in patients with newly diagnosed chronic-phase chronic myeloid leukemia. *Semin. Hematol.*, 40(2 Suppl 2) :26–30, 2003.

[62] M. J. Piccart-Gebhart, M. Procter, B. Leyland-Jones, A. Goldhirsch, M. Untch, I. Smith, L. Gianni, J. Baselga, R. Bell, C. Jackisch, D. Cameron, M. Dowsett, C. H. Barrios, G. Steger, C. S. Huang, M. Andersson, M. Inbar, M. Lichinitser, I. Lang, U. Nitz, H. Iwata, C. Thomssen, C. Lohrisch, T. M. Suter, J. Ruschoff, T. Suto, V. Greatorex, C. Ward, C. Straehle, E. McFadden, M. S. Dolci, and R. D. Gelber. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N. Engl. J. Med.*, 353(16) :1659–1672, 2005.

[63] J. A. DiMasi and H. G. Grabowski. Economics of new oncology drug development. *J. Clin. Oncol.*, 25(2) :209–216, 2007.

[64] I. Kola and J. Landis. Can the pharmaceutical industry reduce attrition rates ? *Nat Rev Drug Discov*, 3(8) :711–715, 2004.

[65] C. M. Booth, A. H. Calvert, G. Giaccone, M. W. Lobbezoo, L. K. Seymour, and E. A. Eisenhauer. Endpoints and other considerations in phase I studies of targeted anticancer therapy : recommendations from the task force on Methodology for the Development of Innovative Cancer Therapies (MDICT). *Eur. J. Cancer*, 44 :19–24, 2008.

[66] S. Sleijfer and E. Wiemer. Dose selection in phase I studies : why we should always go for the top. *J. Clin. Oncol.*, 26(10) :1576–1578, 2008.

[67] C. Le Tourneau, J. J. Lee, and L. L. Siu. Dose escalation methods in phase I cancer clinical trials. *J. Natl. Cancer Inst.*, 101 :708–720, 2009.

[68] A. Ivanova and N. Flournoy. Comparison of Isotonic Designs for Dose-Finding. *Stat Biopharm Res*, 1(1) :101, 2009.

[69] D. H. Leung and Y. Wang. Isotonic designs for phase I trials. *Control Clin Trials*, 22(2) :126–138, 2001.

[70] S. Senn and S. Julious. Measurement in clinical trials : a neglected issue for statisticians ? *Stat Med*, 28(26) :3189–3209, 2009.

[71] M. D. Brundage, J. L. Pater, and B. Zee. Assessing the reliability of two toxicity scales : implications for interpreting toxicity data. *J. Natl. Cancer Inst.*, 85(14) :1138–1148, 1993.

[72] D. L. Sackett. Clinical disagreement : I. How often it occurs and why. *Can Med Assoc J*, 123(6) :499–504, 1980.

[73] A. Iasonos, M. Gounder, D. R. Spriggs, J. F. Gerecitano, D. M. Hyman, S. Zohar, and J. O'Quigley. The impact of non-drug-related toxicities on the estimation of the maximum tolerated dose in phase I trials. *Clin. Cancer Res.*, 18(19) :5179–5187, 2012.

[74] S. D. Mukherjee, M. E. Coombes, M. Levine, J. Cosby, B. Kowaleski, and A. Arnold. A qualitative study evaluating causality attribution for serious adverse events during early phase oncology clinical trials. *Invest New Drugs*, 29(5) :1013–1020, 2011.

[75] D. J. Sargent and S. L. George. Clinical trials data collection : when less is more. *J. Clin. Oncol.*, 28(34) :5019–5021, 2010.

[76] FDA. Investigational new drug safety reporting requirements for human drug and biological products and safety reporting requirements for bioavailability and bioequivalence studies in humans. Final rule. *Fed Regist*, 75(188) :59935–59963, 2010.

[77] S. G. Arbuck. Workshop on phase I study design. Ninth NCI/EORTC New Drug Development Symposium, Amsterdam, March 12, 1996. *Ann. Oncol.*, 7 :567–573, 1996.

[78] C. Le Tourneau, A. R. Razak, H. K. Gan, S. Pop, V. Dieras, P. Tresca, and X. Paoletti. Heterogeneity in the definition of dose-limiting toxicity in phase I cancer clinical trials of molecularly targeted agents : a review of the literature. *Eur. J. Cancer*, 47(10) :1468–1475, 2011.

[79] M. D. Prados, S. M. Chang, N. Butowski, R. DeBoer, R. Parvataneni, H. Carliner, P. Kabuubi, J. Ayers-Ringler, J. Rabbitt, M. Page, A. Fedoroff, P. K. Sneed, M. S. Berger, M. W. McDermott, A. T. Parsa, S. Vandenberg, C. D. James, K. R. Lamborn, D. Stokoe, and D. A. Haas-Kogan. Phase II study of erlotinib plus temozolomide during and after radiation therapy in patients with newly diagnosed glioblastoma multiforme or gliosarcoma. *J. Clin. Oncol.*, 27(4) :579–584, 2009.

[80] R. I. Jakacki, M. Hamilton, R. J. Gilbertson, S. M. Blaney, J. Tersak, M. D. Krailo, A. M. Ingle, S. D. Voss, J. E. Dancey, and P. C. Adamson. Pediatric phase I and pharmacokinetic study of erlotinib followed by the combination of erlotinib and temozolomide : a Children's Oncology Group Phase I Consortium Study. *J. Clin. Oncol.*, 26(30) :4921–4927, 2008.

[81] M. Yalon, B. Rood, T. J. MacDonald, G. McCowage, R. Kane, S. Constantini, and R. J. Packer. A feasibility and efficacy study of rapamycin and erlotinib for recurrent pediatric low-grade glioma (LGG). *Pediatr Blood Cancer*, 60(1) :71–76, 2013.

Chapitre 8

Annexes

8.1 Modèle de travail

Dans les méthodes CRM guidées par le critère DLT, le modèle de travail définit les probabilités de toxicité initiales à chaque niveau de dose. Ces probabilités reflètent l'avis *a priori* des cliniciens sur la toxicité du traitement étudié. L'ensemble de ces probabilités, noté $\{\alpha_1, ..., \alpha_K\}$, forme la relation hypothétique considérée entre la dose et la probabilité d'observer une toxicité. Ces probabilités sont en général définies en collaboration étroite avec les cliniciens de l'essai.

Un grand nombre d'essais de phase I utilise le premier modèle de travail arbitrairement proposé par O'Quigley en 1990, sans aucune justification. Cependant, il a été montré que le choix de ce modèle avait un impact sur le comportement de la méthode CRM tant en termes de recommandation finale que d'attribution des doses en cours de l'essai [41]. Récemment, Lee et Cheung ont proposé une méthode pour définir ce vecteur de probabilités initiales de DLT à chaque niveau de dose [41]. Cette méthode est basée sur les éléments suivants : la forme de la relation dose-toxicité, le nombre de doses prévues dans l'essai, la position de la dose à recommander *a priori*, la probabilité de toxicité acceptable et un paramètre de calibration noté δ. Ce dernier définit un intervalle contenant des valeurs de toxicité jugées acceptables, noté intervalle d'indifférence $(\theta_{DLT} \pm \delta)$.

Les auteurs ont proposé une étude de simulation pour trouver la largeur de l'intervalle d'indifférence maximisant le PCS pour différents scénarios étudiés.

L'algorithme de Lee et Cheung permettant de définir le modèle de travail, appelé "getprior", est disponible dans le package dfcrm du logiciel R.

8.2 Fonction de quasi-vraisemblance

La fonction de quasi-vraisemblance est une fonction des paramètres évaluée à partir des observations, à l'instar de la vraisemblance. Cette fonction ne fait pas d'hypothèse sur la loi de la variable à expliquer mais seulement sur les deux premiers moments de la variable à expliquer (espérance et variance).

Une fois le premier et le deuxième ordre spécifiés, la fonction de quasi-vraisemblance, L_i, définie pour le sujet i est :

$$L_i = \int_{z_i}^{\mu} \frac{z_i - t}{Var(t)} dt \tag{8.1}$$

Où μ et $Var(t)$ sont respectivement le premier et le deuxième ordre de la variable à expliquer Z. Les observations z_i sont supposées indépendantes.

La fonction de quasi-vraisemblance présente ces trois propriétés :

1. $E\left(\frac{\partial L_i}{\partial \mu}\right) = 0$
2. $Var\left(\frac{\partial L_i}{\partial \mu}\right) = \frac{1}{Var(\mu)}$
3. $-E\left(\frac{\partial^2 L_i}{\partial \mu^2}\right) = \frac{1}{Var(\mu)}$

La fonction de quasi-vraisemblance (ou plus précisément la fonction de Log quasi-vraisemblance) possède donc les mêmes propriétés que la fonction de Log vraisemblance d'une loi exponentielle sur l'espérance des deux premières dérivées et sur la variance de la première dérivée [55, 59]. Ces propriétés permettent l'estimation du paramètre ainsi que la définition de ses propriétés de convergence et la vérification de la normalité asymptotique. Choisir une loi exponentielle pour le modèle linéaire généralisé ou une quasi-vraisemblance revient donc aux mêmes estimations.

La fonction de quasi-vraisemblance s'écrit, pour n patients, comme suit :

$$L = \sum_{i=1}^{n} L_i \tag{8.2}$$

Fonction de quasi-vraisemblance avec une variance de Bernoulli :
La variance de Bernoulli est définie par :

$$Var(t) = t * (1 - t)$$

la fonction de quasi-vraisemblance, L_i, est égale à :

$$\begin{aligned} L_i &= \int_{z_i}^{\mu} \frac{z_i - t}{t(1-t)} \\ &= z_i \int_{z_i}^{\mu} \frac{1}{1-t} dt + z_i \int_{z_i}^{\mu} \frac{1}{t} dt - \int_{z_i}^{\mu} \frac{1}{1-t} dt \\ &\propto z_i Ln\left(\frac{\mu}{1-\mu}\right) + Ln(1-\mu) \end{aligned} \tag{8.3}$$

Fonction de quasi-vraisemblance avec une variance de Wedderburn :
La variance de Wedderburn est définie par :

$$Var(t) = E(t)^2 * (1 - E(t))^2$$

La fonction de quasi-vraisemblance, L_i, est égale :

$$\begin{aligned} L_i &= \int_{z_i}^{\mu} \frac{z_i - t}{t^2(1-t)^2} \\ &= z_i \int_{z_i}^{\mu} \frac{1}{t^2(1-t)^2} dt - \int_{z_i}^{\mu} \frac{1}{t(1-t)^2} dt \end{aligned}$$

Or

$$\begin{cases} \int_{z_i}^{\mu} \frac{1}{t^2(1-t)^2} dt = \int_{z_i}^{\mu} \frac{2t+1}{t^2} dt + \int_{z_i}^{\mu} \frac{3-2t}{(1-t)^2} dt = \left[2\left(Ln(t) - Ln(1-t)\right) - \frac{1}{t(1-t)}\right]_{\mu}^{z_i} \\ \int_{z_i}^{\mu} \frac{1}{t(1-t)^2} dt = \int_{z_i}^{\mu} \frac{1}{t} dt + \int_{z_i}^{\mu} \frac{(2-t)}{(1-t)^2} dt = \left[Ln(t) - Ln(1-t) - \frac{1}{1-t}\right]_{\mu}^{z_i} \end{cases}$$

La fonction de quasi-vraisemblance utilisant la variance de Wedderburn s'écrit donc comme suit :

$$L_i \propto (2z_i - 1)Ln(\frac{\mu}{(1-\mu)}) - \frac{z_i}{\mu} - \frac{1-z_i}{1-\mu} \tag{8.4}$$

8.3 Performance des méthodes QLCRM et QCRM en utilisant le modèle de travail de Yuan

Le modèle de travail publié dans l'article de Yuan est égal à (0.11, 0.25, 0.40, 0.55, 0.75, 0.85). Les résultats de la performance des méthodes avec le modèle de travail issu de la fonction "getprior" (QLCRM et QCRM) ou défini par Yuan ($QLCRM^Y$ et $QCRM^Y$) sont présentés dans le tableau 8.1.

Le PCS des différentes méthodes varient selon les scénarios. Le PCS des méthodes QLCRM et QCRM est supérieur de celui des méthodes $QLCRM^Y$ et $QCRM^Y$ (avec le modèle de travail de Yuan) dans les scénarios B, E et H avec une différence variant de $+1.14$ à 27.54. Dans les cinq autres scénarios, la différence de PCS est en faveur des méthodes utilisant le modèle de travail de Yuan, mais les écarts de PCS sont minimes variant de -0.38 à -7.12.

Table 8.1 – Performance de la méthode QLCRM, pour $n = 36$ patients, selon le modèle de travail

	Pourcentage des doses à recommander						Pourcentage des doses à allouer					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Sc A												
QLCRM	3.4	**85.9**	*10.7*	0.0	0.0	0.0	19.4	**62.6**	*17.4*	0.6	0.0	0.0
QLCRM^Y	7.2	**86.3**	*6.6*	0.0	0.0	0.0	24.6	**65.4**	*9.9*	0.1	0.0	0.0
QCRM	2.3	**83.6**	*14.1*	0.0	0.0	0.0	14.2	**63.1**	*22.0*	0.7	0.0	0.0
QCRM^Y	2.2	**87.9**	*9.8*	0.0	0.0	0.0	13.5	**70.4**	*15.9*	0.2	0.0	0.0
Sc B												
QLCRM	0.0	*12.7*	**85.3**	2.0	0.0	0.0	9.1	*21.1*	**60.8**	8.9	0.2	0.0
QLCRM^Y	0.0	*16.0*	**83.2**	0.8	0.0	0.0	9.6	*28.5*	**57.9**	4.0	0.0	0.0
QCRM	0.0	*10.5*	**87.4**	2.0	0.0	0.0	8.5	*18.7*	**63.1**	9.6	0.1	0.0
QCRM^Y	0.0	*12.9*	**86.3**	0.8	0.0	0.0	8.5	*22.6*	**64.2**	4.7	0.0	0.0
Sc C												
QLCRM	0.0	3.0	**83.8**	*13.3*	0.0	0.0	9.2	14.9	**57.0**	*18.3*	0.7	0.0
QLCRM^Y	0.0	5.9	**86.3**	*7.7*	0.0	0.0	9.9	21.6	**58.7**	*9.8*	0.1	0.0
QCRM	0.0	2.3	**84.1**	*13.6*	0.0	0.0	8.5	13.2	**58.5**	*19.3*	0.4	0.0
QCRM^Y	0.0	4.1	**87.8**	*8.1*	0.0	0.0	8.6	16.8	**63.7**	*11.0*	0.0	0.0
Sc D												
QLCRM	0.0	0.1	**83.0**	*17.0*	0.0	0.0	8.4	8.6	**51.4**	*30.3*	1.4	0.0
QLCRM^Y	0.0	0.06	**89.82**	*10.12*	0	0	8.4	9.16	**62.86**	*19.43*	0.14	0.0
QCRM	0.0	0.0	**82.5**	*17.5*	0.0	0.0	8.3	8.5	**50.2**	*32.1*	1.0	0.0
QLCRM^Y	0.0	0.0	**88.6**	*11.4*	0.0	0.0	8.3	8.62	**60.9**	*22.1*	0.1	0.0
Sc E												
QLCRM	0.0	0.0	*8.8*	**90.5**	0.6	0.0	8.4	8.4	*15.0*	**60.4**	7.8	0.1
QLCRM^Y	0.0	0.0	*10.8*	**89.1**	0.1	0.0	8.4	8.6	*20.7*	**60.1**	2.2	0.0
QCRM	0.0	0.0	*8.2*	**91.4**	0.4	0.0	8.3	8.4	*14.6*	**62.4**	6.4	0.0
QCRM^Y	0.0	0.0	*10.5*	**89.5**	0.1	0	8.3	8.4	*20.0*	**62.0**	1.2	0.0
Sc F												
QLCRM	0.0	0.0	2.7	**80.7**	*16.5*	0.0	8.4	8.5	13.1	**50.9**	*18.5*	0.6
QLCRM^Y	0.0	0.0	5.0	**87.8**	*7.2*	0.0	8.4	9.0	18.7	**56.8**	*7.0*	0.1
QCRM	0.0	0.0	2.6	**84.7**	*12.7*	0.0	8.3	8.4	12.9	**54.9**	*15.3*	0.2
QCRM^Y	0.0	0.0	5.0	**90.7**	*4.3*	0.0	8.3	8.6	18.2	**60.4**	*4.4*	0.0
Sc G												
QLCRM	0.0	0.0	0.0	2.6	**79.6**	*17.8*	8.4	8.3	8.4	12.3	**45.0**	*17.6*
QLCRM^Y	0.0	0.0	0.0	11.0	**81.9**	*7.0*	8.4	8.41	8.9	22.5	**45.7**	*6.2*
QCRM	0.0	0.0	0.0	3.8	**85.6**	*10.6*	8.3	8.3	8.4	13.9	**50.8**	*10.2*
QCRM^Y	0.0	0.0	0.0	18.3	**79.1**	*2.6*	8.3	8.4	8.8	29.5	**43.2**	*1.8*
Sc H												
QLCRM	0.0	0.0	0.0	1.8	**82.5**	*15.7*	8.3	8.4	8.7	15.0	**48.0**	*11.5*
QLCRM^Y	0.0	0.0	0.1	25.0	**70.0**	*4.9*	8.4	8.7	10.7	34.6	**34.6**	*3.0*
QCRM	0.0	0.0	0.0	3.6	**88.7**	*7.7*	8.3	8.4	8.7	18.4	**50.9**	*5.3*
QCRM^Y	0.0	0.0	0.1	37.5	**61.2**	*1.3*	8.3	8.4	10.6	44.4	**27.6**	*0.6*

Sc : Scénario.
QLCRM : la méthode Quasi-LCRM avec le modèle de travail (0.14,0.20,0.28,0.36,0.44,0.52) [4].
QLCRM^Y : la méthode QLCRM avec le modèle de travail de Yuan (0.11, 0.25, 0.40, 0.55, 0.75, 0.85).
QCRM : la méthode Quasi-CRM de Yuan et al., avec le modèle de travail (0.14,0.20,0.28,0.36,0.44,0.52).
QCRM^Y : la méthode QCRM avec le modèle de travail de Yuan (0.11, 0.25, 0.40, 0.55, 0.75, 0.85).
Les valeurs en gras correspondent aux résultats obtenus à la dose cible.
Les valeurs en italiques correspondent aux résultats obtenus à la dose la plus proche de la dose cible.

8.4 Scénarios supplémentaires pour évaluer les méthodes basées sur un score de toxicité en comparaison avec celles basées sur le critère DLT

Nous avons exploré quatre autres scénarios pour évaluer la performance des méthodes utilisant le score nTTP en comparaison avec les méthodes CRM classiques basées sur la DLT. Ces scénarios présentent des pentes différentes autour de la vraie dose à recommander, avec un score et une probabilité d'observer une DLT légèrement au-dessus ou au-dessous des valeurs cibles (cf. tableau 8.2).

TABLE 8.2 – Description des scénarios (score nTTP et probabilité de DLT à chaque palier de dose)

	d_1	d_2	d_3	d_4	d_5	d_6
Scénario I						
nTTP	0.051	0.096	0.168	**0.279**	0.418	0.446
p(DLT)	0.002	0.014	0.186	**0.332**	0.506	0.554
Scénario J						
nTTP	0.051	0.118	0.195	**0.279**	0.404	0.461
p(DLT)	0.002	0.105	0.241	**0.332**	0.470	0.526
Scénario K						
nTTP	0.051	0.093	0.190	**0.307**	0.409	0.446
p(DLT)	0.003	0.024	0.238	**0.361**	0.489	0.554
Scénario L						
nTTP	0.043	0.081	0.132	**0.248**	0.389	0.470
p(DLT)	0.004	0.064	0.175	**0.298**	0.446	0.584

8.5 Spécification de la variance pour la méthode QLCRM

L'estimation du paramètre de dispersion ϕ a été conduite par les méthodes dite "des moments" proposées par McCullag et Nelder [55] :

$$\hat{\phi} = \sum_{i=1}^{n} \frac{(z_i - \hat{\mu}_i)^2}{(n-1)Var(\hat{\mu}_i)} = \sum_{i=1}^{n} \frac{r_i^2}{n-1} \tag{8.5}$$

Où r_i est le résidu de Pearson.

Ce paramètre a été estimé pour les fonctions de variance de Bernoulli et Wedderburn pour chaque scénario étudié. Comme détaillé dans le tableau 8.3, le paramètre de dispersion estimé avec la variance de Bernoulli est très inférieur à 1 pour tous les scénarios. La variabilité de la variable à expliquer supposée avec la variance de Bernoulli est donc plus importante que celle calculée empiriquement. On parle ainsi de sur-dispersion. Ceci est explicitement illustré dans le tableau 8.4 présentant les valeurs estimées de la variance empirique et la variance de Bernoulli.

Avec la variance de Wedderburn, le paramètre de dispersion est inférieur à 1 pour trois scénarios (A, B et C) , il est supérieur à 1 pour quatre scénarios (E, F, G et H) et il est égal à 1 pour le scénario D (cf. 8.3 et 8.4). La variance de Wedderburn est également estimée selon les scénarios étudiés dans le tableau 8.4.

Table 8.3 – Estimation du paramètre de dispersion avec la variance de Bernoulli et la variance de Wedderburn pour les différents scénarios étudiés

Scénario	Variance de Bernoulli	Variance de Wedderburn
A	0.12	0.68
B	0.13	0.85
C	0.15	0.92
D	0.12	1.00
E	0.13	1.33
F	0.17	1.50
G	0.19	2.37
H	0.13	1.34

Nous avons également comparé les différentes variances (empirique, Bernoulli et Wedderburn) à la variance analytique du score nTTP calculée explicitement à l'aide de la méthode Delta (cf. 8.6 pour le détail). Le tableau 8.4 illustre les différents résultats. Il est clair que la variance analytique reflète mieux la variabilité de la variable à expliquer. Cependant, un modèle basé sur cette fonction de variance analytique conduit à une approche alternative qui dépasse ce projet de recherche.

Table 8.4 – Valeurs de la variance selon les différents scénarios

	d_1	d_2	d_3	d_4	d_5	d_6
Scénario A						
variance empirique	0.0307	0.0282	0.0238	0.0193	0.0178	0.0172
variance de Bernoulli	0.1498	0.2012	0.2303	0.2414	0.2452	0.2461
variance de Wedderburn	0.0224	0.0405	0.0531	0.0583	0.0601	0.0606
variance analytique *	0.0174	0.0233	0.0234	0.0225	0.0217	0.0211
Scénario B						
variance empirique	0.0182	0.0290	0.0281	0.0219	0.0186	0.0165
variance de Bernoulli	0.0900	0.1589	0.2135	0.2381	0.2464	0.2496
variance de Wedderburn	0.0081	0.0253	0.0456	0.0567	0.0607	0.0623
variance analytique	0.0087	0.0171	0.0239	0.0241	0.0229	0.0205
Scénario C						
variance empirique	0.0195	0.0307	0.0282	0.0238	0.0193	0.0178
variance de Bernoulli	0.0961	0.1498	0.2012	0.2303	0.2414	0.2452
variance de Wedderburn	0.0092	0.0224	0.0405	0.0531	0.0583	0.0601
variance analytique	0.0106	0.0174	0.0233	0.0234	0.0225	0.0217
Scénario D						
variance empirique	0.0088	0.0159	0.0204	0.0172	0.0158	0.0160
variance de Bernoulli	0.0478	0.1043	0.1972	0.2332	0.2408	0.2484
variance de Wedderburn	0.0023	0.0109	0.0389	0.0544	0.0580	0.0617
variance analytique	0.0079	0.0108	0.0173	0.0192	0.0191	0.0196
Scénario E						
variance empirique	0.0088	0.0143	0.0274	0.0243	0.0165	0.0149
variance de Bernoulli	0.0478	0.0864	0.1525	0.2148	0.2432	0.2470
variance de Wedderburn	0.0023	0.0075	0.0233	0.0461	0.0591	0.0610
variance analytique	0.0079	0.0086	0.0175	0.0222	0.0208	0.0191
Scénario F						
variance empirique	0.0088	0.0159	0.0204	0.0172	0.0158	0.0160
variance de Bernoulli	0.0478	0.1043	0.1972	0.2332	0.2408	0.2484
variance de Wedderburn	0.0023	0.0109	0.0389	0.0544	0.0580	0.0617
variance analytique	0.0079	0.0108	0.0173	0.0192	0.0191	0.0196
Scénario G						
variance empirique	0.0086	0.0100	0.0195	0.0307	0.0282	0.0238
variance de Bernoulli	0.0435	0.0510	0.0961	0.1498	0.2012	0.2303
variance de Wedderburn	0.0019	0.0026	0.0092	0.0224	0.0405	0.0531
variance analytique	0.0071	0.0085	0.0106	0.0174	0.0233	0.0234
Scénario H						
variance empirique	0.0087	0.0150	0.0170	0.0179	0.0176	0.0167
variance de Bernoulli	0.0478	0.0981	0.1210	0.1533	0.1890	0.2282
variance de Wedderburn	0.0023	0.0096	0.0147	0.0235	0.0357	0.0521
variance analytique	0.0079	0.0107	0.0129	0.0152	0.0169	0.0179

* : variance analytique calculée à l'aide de la méthode Delta (cf. 8.6 pour le détail).

8.6 Variance analytique du score de toxicité

Etant donné que le score de toxicité est dérivé des différentes variables aléatoires indépendantes (les poids des différents types de toxicité), nous avons cherché à calculer la fonction de la variance analytique. Considérant la norme euclidienne, nous ne pouvons pas appliquer la propriété de la variance telle que la variance de la somme des variables aléatoires est égale à la somme des variances de chaque variable aléatoire. L'idée est ainsi de dériver une fonction linéaire qui se rapproche de la somme euclidienne afin de calculer explicitement la fonction de la variance. Pour ceci, nous avons utilisé les deux premiers termes d'un développement en série de Taylor. Nous notons que la variable aléatoire Z (le score) dépend de la variable aléatoire w.

Le développement en série de Taylor pour la variable Z est défini comme suit :

$$Z = f(w) = f(t) + (w - t)f^{'}(t)$$

Où f est une fonction indéfiniment dérivable et t une réalisation de la variable w pour laquelle la fonction f est définie. $f^{'}$ est la dérivée de la fonction f par rapport à w. Elle est définie comme suit :

$$f^{'}(t) = \frac{\partial f(w)}{w}|_{w=t}$$

La variance approximative de la variable Z est dérivée directement de l'équation ci-dessus comme suit :

$$Var(Z) = var(f(w)) = Var(f(t) + (w - t)f^{'}(t)) = Var(w)[f^{'}(t)]^2$$

Nous avons ensuite utilisé la méthode Delta pour obtenir une estimation de la variance :

$$\widehat{Var}(Z) = \widehat{Var}(w)[f^{'}(t)]^2$$

En appliquant cette formule, nous obtenons :

$$\widehat{Var(Z)} = \frac{\sum_{t=1}^{T} \widehat{w_t}^2 * \widehat{var}(\widehat{w_t})}{\nu^2 * \sum_{t=1}^{T} \widehat{w_t}^2}$$

$\widehat{w_T}$ est le poids estimé pour chaque type de toxicité comme suit :

$$\widehat{w_t} = \sum_{j=0}^{4} w_{t,j} * P(G_{t,j})$$

Où $P(G_{t,j})$ est la probabilité d'observer un grade j de toxicité t,

$\widehat{Var}(\widehat{w_t})$ est la variance de la variable w_t, calculée comme suit :

$$\widehat{Var}(\widehat{w_t}) = \sum_{j=0}^{4} \widehat{w_{t,j}}^2 * P(G_{t,j}) - \left(\sum_{j=0}^{4} \widehat{w_{t,j}} * P(G_{t,j}) \right)^2$$

Un modèle basé sur cette fonction de variance analytique conduit à une approche alternative qui dépasse ce projet de recherche.