THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)

Discipline ou spécialité :
Domaine mathématiques – Mathématiques appliquées

Présentée et soutenue par
Guillaume MIJOULE
le : 4 juin 2013

Titre :
Modélisation du processus d'inclusion de patients dans un essai clinique multicentrique.

École doctorale :
Mathématiques Informatique Télécommunications (MITT)

Unité de recherche :
UMR 5219

Directeurs de thèse :
Pr Laure Coutin - Université de Toulouse 3
Dr Nicolas Savy - Université de Toulouse 3

Rapporteurs :
Pr Stephen Senn - CCMS Luxembourg
Dr-HDR Adeline Samson - Université Paris 5

Autres membres du jury :
Pr Antoine Chambaz - Université Paris 10
Pr Aurélien Garivier - Université de Toulouse 3
Pr Sandrine Andrieu - Université de Toulouse 3 - INSERM unité 1027
Pr Vladimir Anisimov - Université de Glasgow

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse, Nicolas Savy et Laure Coutin, qui ont accepté de m'encadrer pendant ces quatre années. Je les remercie pour la confiance qu'ils m'ont toujours témoignée, la motivation qu'ils ont su me redonner quand elle faisait défaut, et pour toute l'énergie dépensée afin de m'aider à mener à bien cette thèse. Je dois beaucoup à leur soutien permanent et leur suis infiniment reconnaissant.

Je remercie Adeline Samson et Stephen Senn d'avoir accepté d'être les rapporteurs de cette thèse, ainsi que Sandrine Andrieu, Vladimir Anisimov, Aurélien Garivier et Antoine Chambaz d'avoir accepté de participer à mon jury de thèse.

J'adresse en particulier mes remerciements à Sandrine Andrieu et à l'unité 1027 de l'INSERM pour les données qu'ils ont accepté de nous prêter, ainsi qu'à Vladimir Anisimov pour le rôle qu'il a joué dans mon travail à travers notre collaboration. Je suis également reconnaissant des conditions de travail optimales qui m'ont été offertes au sein de l'Institut Mathématique de Toulouse.

Je remercie sincèrement mes amis et ma famille pour m'avoir soutenu durant ces années.

Enfin, merci à Laury-Anne, pour avoir vécu et partagé cette thèse avec moi, et pour la patience dont elle a fait preuve dans les moments difficiles.

Table des matières

Introduction

0.1 Les essais cliniques

Un essai thérapeutique est une expérience menée *in vivo* chez l'animal ou chez l'homme. Pour mettre au point un nouveau traitement, de multiples essais sont nécessaires avec différents objectifs pouvant être classés en deux grands axes : l'**efficacité** et l'**innocuité**.

- **efficacité.** Le traitement étudié doit agir efficacement sur la pathologie qu'il est censé soigner et, idéalement, plus efficacement que les traitements déjà utilisés. L'efficacité est une notion moyenne, quelques individus pouvant être "non-répondeurs" au traitement. Comme on parle d'effet moyen, on comprend qu'il faudra essayer le traitement sur un échantillon de malades, échantillon dont la taille sera souvent élevée pour assurer la validité des résultats obtenus.
- **innocuité.** Le proverbe est bien connu : "Le remède ne doit pas être pire que le mal". L'histoire est pourtant parsemée d'exemples d'accidents (pathologies iatrogènes) où l'issue d'une thérapeutique s'est révélée nocive (purges et saignées au moyen âge, Thalidomide, Distilbène, Cérivastatine, Vioxx). Pour l'innocuité, malgré les grandes tailles d'échantillons, il se peut que des effets nocifs dont la probabilité d'apparition est très faible ne soient pas révélés au cours de l'étude. Contrairement à la notion moyenne d'efficacité, l'innocuité est une notion individuelle et l'on notera chez chaque individu la survenue d'événements dits indésirables.

Les essais cliniques constituent une étape capitale dans le processus de mise au point d'un médicament. Les résultats de ces essais cliniques constituent la base du dossier d'enregistrement du médicament soumis à la validation des autorités de santé. C'est sur la base de ce dossier que sera accordée, ou non, l'autorisation de mise sur le marché du médicament (A.M.M.). La réalisation d'un essai clinique se fait dans le respect de l'éthique médicale et selon des normes scientifiques et réglementaires très strictes portées par l'ICH (International Conference on Harmonisation). Les statistiques ont un rôle central dans cette méthodologie.

0.2 Etapes du développement d'un médicament

Le développement d'un médicament se fait en deux temps : la phase préclinique et la phase clinique, elle même divisée en quatre phases.

0.2.1 La phase préclinique

Elle commence par une expérimentation *in vitro* pour vérifier les caractéristiques physiques et chimiques de la molécule candidate à devenir un médicament. Elle se poursuit par une expérimentation *in vivo* **sur l'animal** pour explorer :

- La pharmacocinétique (PK) : décrit le devenir dans l'organisme d'une substance active contenue dans un médicament.
- La pharmacodynamique (PD) : décrit les effets qu'un principe actif produit sur l'organisme : c'est l'étude détaillée de l'interaction récepteur/substance active.
- La toxicologie : étudie et analyse expérimentalement la toxicité des produits.
- La tératologie : étudie les anomalies du développement et les malformations congénitales.

Les phases d'essais cliniques impliquant des personnes ne peuvent être entreprises que si les résultats de l'expérimentation animale ont été jugés prometteurs et non dangereux.

0.2.2 Les phases cliniques

On distingue les trois phases successives pré-AMM de la phase IV dite post-AMM. Chaque phase permet d'obtenir des informations précises. Il est impératif qu'une phase soit terminée et que ses résultats soient jugés concluants pour pouvoir passer à la phase suivante. Un essai de phase IV est réalisé après la commercialisation du produit et continue à évaluer la balance bénéfice/risque du médicament tout au long de sa vie.

- **phase I.** Ces essais sont souvent menés sur quelques dizaines (20 à 100) volontaires sains, lorsque la toxicité escomptée du médicament est limitée. Les objectifs sont :
  - la détermination des conditions de tolérance humaine avec la dose maximale tolérée (DMT). La détermination de la dose initiale est aléatoire mais déduite des paramètres du dossier pharmaco-toxicologique animal.
  - les études initiales de paramètres pharmacocinétiques humains.
  - la détermination des doses entraînant les premiers effets pharmacodynamiques souhaités (l'effet pharmacologique principal et les effets pharmacologiques secondaires).
- **phase II.** Représente les premières administrations chez la population cible à des petits groupes de sujets (100 à 200 sujets) présentant la pathologie en question. La phase II se subdivise en deux phases selon l'objectif :
  - **phase IIa :** établir que la molécule a bien un effet de traitement de la maladie cible chez l'homme ("proof of concept"),
  - **phase IIb :** déterminer la ou les doses à utiliser préférentiellement ainsi que les conditions optimales de prescription (posologie, voie, rythme, durée, effets indésirables, interactions médicamenteuses, forme galénique).
- **phase III.** Les essais cliniques de phase III sont les éléments essentiels de l'étude de l'efficacité thérapeutique. L'objectif est essentiellement de déterminer l'efficacité du traitement dans une pathologie ou une indication donnée. Les essais concernent un plus grand nombre de patients (500 à 3000) qui sont des malades volontaires. Ces malades sont sélectionnés selon des critères d'inclusion et d'exclusion (affection, sexe, âge, forme clinique, degré d'évolution, etc...), de manière à ne pas être atteints d'autres pathologies.

– **phase IV.** Les effectifs des échantillons observés sont variables et les objectifs nombreux :
  – Déceler et tenter de quantifier les effets indésirables en particulier rares et imprévisibles induits par le médicament : la pharmacovigilance.
  – Evaluer l'efficacité réelle du médicament en utilisation au long cours ou en association à d'autres traitements.
  – Améliorer la comparaison avec d'autres molécules.
  – Cerner les populations ayant la probabilité la plus grande d'être sensibles au produit ou de présenter plus fréquemment les effets indésirables.
  – Développer de nouvelles indications thérapeutiques ou de nouvelles présentations du produit (extension d'AMM).
  – Réévaluer régulièrement l'intérêt du produit en fonction de la découverte de nouveaux médicaments.
  – Réaliser une étude de pharmaco-épidémiologie et de pharmaco-économie.

0.3 Les essais de phase III

0.3.1 Le trépied méthodologique

Les essais cliniques de phase III interviennent avant la mise sur le marché et sont indispensables et décisifs pour la demande d'AMM. Ce sont les éléments essentiels de l'étude de l'efficacité thérapeutique. Accessoirement, ils permettent de détecter les effets indésirables, d'affiner la posologie et le choix des meilleurs voies d'administration.

Nous nous focaliserons sur ces essais car, les tests étant effectués sur un grand nombre de volontaires présentant la pathologie étudiée (500 à 3000), les statistiques y jouent un rôle prépondérant. La méthodologie des essais de phase III est très rigoureusement établie et repose sur un protocole très strict dont la trame est maintenant normalisée (voir les travaux du Consort Group [36, 45] et de l'International Conference on Harmonisation [38]). L'analyse statistique sur laquelle s'appuie l'essai repose sur le trépied méthodologique suivant :

– **Comparaison à un groupe contrôle.** Une partie des patients reçoit soit un placebo (ce qui permet de déterminer l'efficacité absolue du médicament par rapport à l'évolution spontanée de l'affection) soit un traitement de référence, lorsqu'il en existe un dont l'intérêt même partiel est reconnu, ce qui permet de déterminer l'efficacité relative du nouveau médicament et son apport thérapeutique.

– **La randomisation.** Pour que la comparaison statistique ait un sens, il importe que les deux lots de malades (lot traité et lot témoin) soient identiques vis-à-vis de tous les paramètres pouvant influencer les résultats. Le seul moyen d'être certain que ces facteurs se répartissent de manière équivalente entre les deux lots, est de les constituer par tirage au sort.

– **L'aveugle.** Il faut éliminer les facteurs subjectifs (effet placebo) qui pourraient perturber les résultats. Ils peuvent naître de la connaissance par le médecin ou le malade appartenant au lot traité ou au lot témoin. Donc on opère en simple ou

double aveugle (insu) : les deux (médecin et malade) sont laissés dans l'ignorance des résultats du tirage au sort.

0.3.2 Le nombre de sujets nécessaire

La comparaison entre les groupes s'effectue au moyen d'un test statistique adapté au critère de jugement principal (quantitatif, qualitatif, données de survie, ...) et à l'objectif de l'essai (essai de supériorité, de non-infériorité, d'équivalence, ...). Une fois le test choisi et les hypothèses fixées, il est possible, étant donnés les risques de première et de seconde espèce ainsi que l'amplitude de l'essai que l'on souhaite observer, de calculer **le nombre de sujets nécessaire** pour observer cette amplitude avec une puissance (risque de seconde espèce) fixée *a priori*. Plus la différence d'efficacité que l'on souhaite observer entre les deux traitements est faible, plus le nombre de sujets à inclure est important pour montrer une différence significative. A l'opposé, si le traitement est très efficace, un petit nombre de patients suffit pour un résultat statistiquement significatif.

0.4 La phase de recrutement des patients

0.4.1 Les étapes de l'inclusion d'un patient dans les essais cliniques

1. Le malade volontaire fait connaissance avec l'essai que le médecin lui propose de suivre et signe un formulaire reprenant les grandes lignes de l'essai, notamment les risques encourus. On parle de **consentement éclairé**. Le candidat est alors *screené*.
2. On vérifie que le candidat satisfait les conditions d'inclusion et les critères de non-inclusion des patients dans l'essai thérapeutique. Le respect de ces critères permet de constituer des groupes homogènes de patients. Cette homogénéité entre les groupes est importante pour deux raisons :
   - La première afin d'éviter de mettre en doute les résultats d'une étude après constat d'un non respect des critères d'éligibilité entraînant la non comparabilité des groupes étudiés.
   - La seconde repose sur des considérations d'ordre statistique. Dans la mesure où plus la population éligible est homogène, moins la variabilité de la réponse est importante, plus la probabilité d'obtenir une différence statistiquement significative entre les groupes est grande.
3. S'il passe les tests inhérents à la visite d'inclusion avec succès, il entre alors dans l'étude proprement dite. Il lui est alors assigné **aléatoirement et à son insu** soit le traitement testé soit un placebo (ou un traitement de référence). On dira que le patient est *randomisé*.
4. S'il est rejeté à l'issue des tests ou se rétracte, le patient n'est pas inclus, on parle alors de *screening failure*.

L'analyse statistique finale portera sur des échantillons de patients dont la réunion est abusivement appelée "population". L'analyse de la population "en intention de traiter"

FIGURE 1 – Etapes de l'inclusion d'un patient dans un essai clinique.

signifie que chaque individu est affecté à son groupe de randomisation même s'il a subi un autre traitement que celui prévu par la randomisation. L'analyse "per protocole" ne conserve que les malades ayant subi strictement le protocole de traitement que leur avait affecté la randomisation.

0.4.2 Définition du recrutement

Une des principale difficultés dans un essai clinique est d'être capable de recruter un nombre fixé par le protocole de malades volontaires satisfaisant à des critères d'inclusion et d'exclusion très précis. Pour ce faire, dans la majeure partie des cas, on fait appel à plusieurs centres dit centres investigateurs pour se partager la tâche (on parle d'étude multicentrique par opposition aux études monocentriques où un seul centre est chargé du recrutement). Nous sommes alors en mesure de définir ce qui est au centre de la problématique de cette thèse, la phase de recrutement.

Définition 0.4.1. *La phase de recrutement est la période entre l'initiation du premier des C centres investigateurs et l'instant $T(N)$ où les N patients sont inclus.*

Les paramètres du recrutement sont donc :

- N, qui est fixé par le protocole.
- $T(N)$, qui a une valeur cible $T_R(N)$ fixée par les investigateurs (en pratique souvent sous-estimée).
- C, qui est a priori fixé par les investigateurs. On peut néanmoins s'autoriser à ouvrir de nouveaux centres ou en fermer en cours d'essai.

0.4.3 Contraintes sur le recrutement

La phase de recrutement dans un essai clinique est soumise à de nombreuses contraintes. On peut identifier trois contraintes principales :

- **Contraintes économiques.** Un essai clinique est très coûteux et d'autant plus coûteux qu'il est long. De plus, un essai clinique qui dure est un manque à gagner pour l'industrie pharmaceutique. En effet, la période d'exploitation (avant l'autorisation de développer un générique) est de 20 ans [49] et inclut la phase de développement clinique (voir figure ci dessous).
- **Contraintes éthiques.** Les patients randomisés sont exposés à des médicaments "expérimentaux" avec des risques potentiels. La randomisation est souvent mal perçue dans l'esprit public. Elle doit avoir lieu après l'inclusion des patients sélectionnés, donc le plus tardivement possible juste avant le début des traitements.

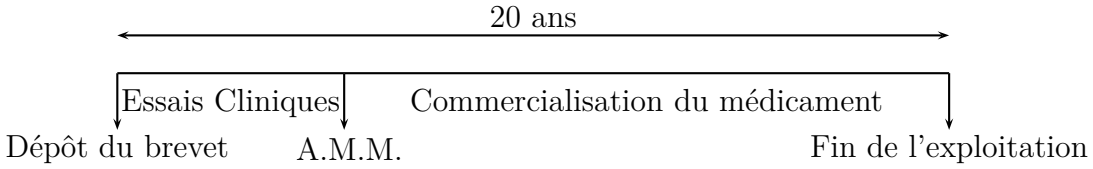

FIGURE 2 – Durée de vie d'un brevet de médicament.

– **Contraintes organisationnelles.** Une évaluation correcte de la date de fin d'essai plusieurs mois à l'avance est très utile pour le coordinateur de l'étude et pour les équipes cliniques pour organiser les services. De plus un modèle prédictif correct de la dynamique de l'essai est très important pour l'organisation des chaînes d'approvisionnement des médicaments [4].

0.4.4 Pourquoi modéliser le recrutement ?

La principale question d'intérêt est de prédire la date de fin de l'étude avec une précision acceptable. La méthode actuellement utilisée est principalement déterministe et basée sur la proportionnalité (j'ai recruté 100 patients en un an, je dois en recruter 200, donc je vais devoir encore attendre un an). Cependant, cette approche n'est pas satisfaisante car elle ne tient pas compte de la grande variabilité du processus de recrutement. Il est donc tout à fait naturel d'introduire un modèle de recrutement de patients. Une grande attention a été portée sur les méthodes de calcul du nombre de sujets nécessaires (NSN), avec le développement de logiciels dédiés (EAST par exemple). En revanche, la description du processus de recrutement n'a été que très peu étudiée et est très peu exploitée. Rojavin [41] exprime très bien cette situation en disant : "Patient recruitment and retention remains until now more of an art rather than a science".

Un tel modèle conduit notamment aux deux applications suivantes :

– La possibilité d'évaluer les principales caractéristiques de l'essai (durée, coût,...) en utilisant des estimations non seulement ponctuelles mais également par intervalle de confiance, le tout basé sur des données recueillies lors d'études intermédiaires.

– Le développement d'outils d'aide à la décision à n'importe quelle étape de l'essai clinique. Cela signifie qu'en utilisant les paramètres de l'essai (donnés par les investigateurs ou estimés à partir d'une étude intermédiaire), il est possible de proposer des stratégies optimales pour la performance de l'essai (évaluer le nombre optimal de centres à ouvrir pour compléter l'étude dans les temps à un risque fixé à l'avance près, par exemple).

Remarque 0.4.2. *Il est important de noter que seules les données de recrutement sont utilisées par ces modèles et les études intermédiaires en question ne nécessitent pas de levée d'aveugle.*

L'introduction d'un modèle de recrutement permet de répondre aux principales problématiques du recrutement de patients

– **au niveau éthique.** Il apporte aux patients des garanties de la faisabilité de l'essai et une visibilité sur les résultats en terme de recrutement.

– **au niveau économique.** Il aide les coordinateurs à prendre des décisions objectives sur l'essai (ouverture ou fermeture de centres, arrêt de l'essai).

– **au niveau organisationnel.** Il permet au coordinateur de manager son équipe et d'optimiser la chaîne d'approvisionnement des médicaments.

Le modèle le plus abouti est celui d'Anisimov [11]. Nous l'introduirons par la suite.

0.4.5 Comment modéliser le recrutement ?

Le premier travail sur la modélisation d'essais cliniques date des années 1980 avec les travaux de Lee [34], Williford et al. [50] et Morgan [37]. Le point de vue de Morgan [37] est heuristique et basé sur l'étude des études complètes ; cependant, les résultats de Williford et al [50] et Morgan [37] sont proches du notre par l'introduction de modèles basés sur des processus de Poisson.

L'utilisation de processus de Poisson est pertinent car, comme le souligne Stephen Senn dans [47], la dynamique du recrutement de patients satisfait les principales propriétés du processus de Poisson : il est à valeurs entières, ses accroissements sont indépendants et stationnaires. De plus, la propriété d'additivité des processus de Poisson est particulièrement utile pour considérer les essais multicentriques. En effet, si l'on modélise le processus de recrutement de chaque centre par un processus de Poisson, alors le processus de recrutement global est encore un processus de Poisson [46].

La principale faiblesse du modèle de Poisson est qu'il ne dépend que d'un seul paramètre (que l'on appelle taux et ne prend pas en compte les nombreuses sources de variabilité), et donc est très peu flexible. Cette difficulté a été soulignée par Carter et al. [21, 20] : ils recommandent l'introduction d'aléatoire dans le taux en utilisant des distributions uniformes ; les modèles suggérés sont donc des modèles doublement stochastiques (ou processus de Cox).

Anisimov et co-auteurs, dans une série d'articles [2] [4] [5] [11], proposent d'utiliser un modèle particulièrement flexible et réaliste, le modèle Gamma-Poisson où le taux de recrutement est modélisé par une distribution Gamma. Ce modèle sera introduit au chapitre 1 puisqu'à la base des travaux que j'ai réalisés dans le cadre de cette thèse. Ce modèle a été appliqué pour de nombreux essais. Il a donné de bons résultats pour la prédiction du temps d'inclusion total à partir de données recueillies à des études intermédiaires.

Remarque 0.4.3. *Bien sûr, les techniques développées ici peuvent s'étendre à un cadre plus large que les essais de phase III, notamment la constitution de cohortes de patients.*

Pour finir, on citera [16] pour une revue complète des modèles de recrutement de patients.

0.5 Autour de modèles Bayésiens de recrutements

L'intérêt d'une approche bayésienne pour modéliser les intensités d'inclusion est multiple. Elle permet d'abord d'éviter une surparamétrisation du modèle : au lieu d'estimer les C intensités des centres séparément, on estime seulement les paramètres de la loi des intensités (2 paramètres dans le cas Gamma). De plus, mettre de l'aléatoire dans l'intensité d'inclusion empêche certaines aberrations : par exemple, si un centre ne recrute pas

sur une période donnée, l'estimateur de son intensité d'inclusion est 0, donc le modèle non bayésien prédit que le centre ne recrutera plus dans le futur. Ce n'est pas le cas dans un modèle bayésien.

0.5.1 Extensions et compléments aux modèles d'Anisimov [35].

Dans les chapitres 1 et 2, nous ne regardons que les patients inclus dans l'étude, c'est-à-dire les patients randomisés.

Le chapitre 1 présente différents modèles bayésiens pour le recrutement, en commençant par le modèle Gamma-Poisson où les intensités d'inclusion des centres sont supposées distribuées suivant une loi Gamma. Nous introduisons une variante de ce modèle où les intensités suivent une loi de Pareto : en effet, la distribution de Pareto modélise bien le fait, observé dans un grand nombre de cas, que 20% des centres recrutent environ 80% des patients. Il est également possible de modéliser une période de "mise en route" des centres (l'intensité d'inclusion n'est pas maximale dès l'ouverture), ou un ralentissement du recrutement dû à un effet de saturation des centres : nous proposons dans ces deux cas une extension du modèle Gamma-Poisson où les taux d'inclusion dépendent du temps. Enfin, à partir d'un jeu de données prêté par l'unité INSERM 1027, où les dates d'ouverture des centres sont inconnues, nous avons introduit un modèle, que nous appellerons Gamma-Poisson uniforme, permettant de contourner ce problème : la date d'ouverture d'un centre est supposée uniformément distribuée entre l'instant initial et l'instant de première inclusion de ce centre. Pour chaque modèle, nous estimons les paramètres par la méthode du maximum de vraisemblance à partir des données recueillies lors d'une étude intermédiaire, et calculons leur matrice de variance-covariance asymptotique, ce qui nous permet de mesurer l'erreur entre les estimateurs et les paramètres réels.

Dans le chapitre 2 est étudié le problème de la prédiction du recrutement. Les données recueillies à un instant intermédiaire t_1 sont utilisées pour recalculer les lois des intensités d'inclusion de chaque centre par ré-estimation bayésienne. Ceci nous permet d'obtenir la loi du processus de recrutement après t_1, donc celle du temps total de recrutement T. Ainsi, il est possible de calculer le temps de recrutement moyen (l'espérance de T) ou un quantile d'ordre p, c'est-à-dire un temps t tel que la probabilité de finir le recrutement avant t soit plus grande que p (en pratique, nous prendrons $p = 95\%$). Ce chapitre est également dédié à une étude de la sensibilité du modèle aux paramètres. En effet, une erreur est commise lors de l'estimation des paramètres, et elle est d'autant plus grande que la quantité d'information est petite (c'est-à-dire, dans notre cas, quand le nombre de centres C est petit ou quand on observe le recrutement sur un temps t_1 petit). Comme la prédiction du recrutement est faite avec les paramètres estimés, cette erreur se répercute sur cette prédiction – par exemple, sur le calcul du temps moyen du recrutement. Etre capable de mesurer l'erreur commise sur les estimateurs peut donc être primordial dans les cas limites, c'est-à-dire lorsque le nombre de centres est petit ($C \leq 20$) ou lorsque l'étude intermédiaire se fait à un instant t_1 petit.

0.5.2 Les modèles avec prise en compte des screening failures [12]

Une autre extension possible du modèle est d'introduire un processus de sortie d'étude pendant la phase de screening. En effet, on ne sait pas si un patient arrivant dans l'essai sera apte à recevoir le traitement médical testé (résultats positifs aux tests d'inclusion). Usuellement, on contourne le problème en augmentant arbitrairement le nombre de sujets nécessaires (de 20% en général). Cet arbitraire a des conséquences économiques qu'il serait intéressant de contrôler.

Dans le chapitre 3 sera proposé un modèle joint pour les patients screenés, randomisés et perdus. La première idée est de supposer qu'un patient a une certaine probabilité r de réussir aux tests de la phase de screening. Afin de gagner en flexibilité, nous proposerons dans un autre modèle une approche bayésienne où la probabilité r dépend du centre et est distribuée suivant une loi Beta. Une autre approche est d'associer à chaque patient un temps exponentiel τ représentant le temps qu'il passe dans la phase de screening. Le patient n'est alors considéré comme recruté que si $\tau \geq R$ où R est le temps nécessaire à la phase de screening. Suivant le modèle, le paramètre θ de cette loi exponentielle sera supposé constant, ou dépendant du centre et distribué selon une loi Gamma.

Il sera également intéressant d'évaluer l'importance des paramètres dirigeant le processus de sortie d'étude sur la fonction de coût en considérant le coût engendré par un patient sorti d'étude sur la période où il a été suivi.

0.5.3 Un modèle pour le coût d'un essai clinique [13]

L'évaluation du coût d'un essai clinique n'est pas une tâche aisée, car ces essais sont longs et nécessitent une logistique complexe. Par exemple, un centre investigateur s'occupe en général de plusieurs essais en même temps, et il est difficile de connaitre la répartition du coût de fonctionnement d'un centre entre les différents essais. Bien souvent, le coût n'est connu que grossièrement et à la fin de l'essai.

Nous proposons néanmoins un modèle pouvant servir de base à une étude du coût. En effet, on peut catégoriser ces différents coûts de la manière suivante :

- **coût fixe d'un centre** qui peut inclure le coût de mise en route du centre, de la fourniture en médicaments (en général, la quantité de médicaments fournie à un centre est déterminée au début de l'essai et ne dépend donc pas du processus d'inclusion du centre), etc,
- **coût d'un centre proportionnel à sa durée d'activité** incluant une partie du coût du personnel (par exemple, un médecin est requis en permanence dans un centre ouvert, que le centre recrute ou non), les coûts en énergie, etc,
- **coût fixe par patient screené**
- **coût fixe par patient randomisé** incluant par exemple le coût du traitement
- **coût par patient randomisé dépendant du temps passé dans l'essai** : si les patients ne restent pas tous le même temps dans l'essai (à cause par exemple de l'évolution de leur état de santé ; ce sera en particulier le cas dans le cadre des données de survie).

Précisons comment modéliser le coût du dernier point. Considérons les instants d'arrivée $(T_i)_{i=1,\dots,n}$ des patients randomisés. L'arrivée du $i$ème patient engendre des coûts à partir de cet instant selon la relation

$$t \to g(t, T_i) \quad t \geq T_i,$$

où g est une fonction à déterminer. Le coût de l'essai pour le centre i à l'instant $t \geq 0$ peut donc s'écrire :

$$\mathcal{C}_i(t) = C_1 N_i^R(t) + C_2 N_i(t) + \sum_{0 \leq T_n^i \leq t} g(t, T_n^i) + F_i + G_i t, \tag{1}$$

où $(T_n^i)_{n \geq 0}$ sont les instants de randomisation dans le centre i, $N_i^R(t)$ est le nombre de patients ayant été randomisés au temps t, $N_i(t)$ le nombre de patients screenés, F_i est le coût fixe du centre, G_i le coût par année d'activité du centre, C_1 le coût fixe d'un patient randomisé, C_2 le coût de screening d'un patient.

Dans le chapitre 4, nous calculons le coût moyen de l'essai, c'est-à-dire l'espérance de $\sum_{i=1}^{C} \mathcal{C}_i(T)$ où T est l'instant (aléatoire) de fin de recrutement. Le modèle régissant la dynamique de l'inclusion est un modèle bayésien présenté au chapitre 3 prenant en compte la perte de patients en phase de screening. Notamment, nous montrons que le différentiel du coût moyen lors de la fermeture d'un centre peut être aisément calculé par simulations, et l'on peut ainsi adapter le recrutement pour minimiser le coût de l'essai.

Chapitre 1

Modèles bayésiens de l'enrôlement de patients dans les essais cliniques multicentriques

Nous proposons en premier lieu l'article [35] publié en collaboration avec Nicolas Savy et Stéphanie Savy dans Statistics in Medicine. Son contenu recoupe en grande partie celui des chapitres 1 et 2 de cette thèse.

Research Article

Statistics in Medicine

Received 22 September 2010, Accepted 8 November 2011 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4495

Models for patients' recruitment in clinical trials and sensitivity analysis

Guillaume Mijoule,[a,b] Stéphanie Savy[a,c] and Nicolas Savy[a,b*†]

Taking a decision on the feasibility and estimating the duration of patients' recruitment in a clinical trial are very important but very hard questions to answer, mainly because of the huge variability of the system. The more elaborated works on this topic are those of Anisimov and co-authors, where they investigate modelling of the enrolment period by using Gamma–Poisson processes, which allows to develop statistical tools that can help the manager of the clinical trial to answer these questions and thus help him to plan the trial. The main idea is to consider an ongoing study at an intermediate time, denoted t_1. Data collected on $[0, t_1]$ allow to calibrate the parameters of the model, which are then used to make predictions on what will happen after t_1. This method allows us to estimate the probability of ending the trial on time and give possible corrective actions to the trial manager especially regarding how many centres have to be open to finish on time. In this paper, we investigate a Pareto–Poisson model, which we compare with the Gamma–Poisson one. We will discuss the accuracy of the estimation of the parameters and compare the models on a set of real case data. We make the comparison on various criteria : the expected recruitment duration, the quality of fitting to the data and its sensitivity to parameter errors. We discuss the influence of the centres opening dates on the estimation of the duration. This is a very important question to deal with in the setting of our data set. In fact, these dates are not known. For this discussion, we consider a uniformly distributed approach. Finally, we study the sensitivity of the expected duration of the trial with respect to the parameters of the model : we calculate to what extent an error on the estimation of the parameters generates an error in the prediction of the duration.

Keywords: clinical trials; recruitment time; Bayesian statistics; Poisson process; maximum likelihood estimation; sensitivity analysis

The aim of this paper is to enrich the collection of models we can consider for dealing with the enrolment of patients in clinical trials. This is a question of paramount importance for the trialists, because clinical trials are very expensive and the methods usually used to manage these studies are mainly deterministic while the environment is obviously random. Section 1 of this paper is state of the art on the topic of the modelling of patients' recruitment in clinical trials. We aim to enlighten three questions: Why model patients' recruitment? How to model patients' recruitment and what is a model of patients recruitment for? Section 2 investigates the so-called Pareto–Poisson model. This model is defined in the same way as the Gamma–Poisson model introduced and studied in a series of papers of Anisimov and co-authors [1–6] with a Pareto distribution of the rate instead of a Gamma one. We compare these models and investigate a sensitivity analysis of these models, which defines how the model reacts to a small modification of the parameters. Section 3 inspired by a real data set is devoted to the question of the centres opening. In fact, what can we do if the centres opening dates are unknown? Anisimov [6] partially gave the answer, but here, we give a more precise study of a model in which the opening dates are uniformly distributed. We investigate a sensitivity analysis too. Finally, in Section 4, we summarize the methodology and give some of the main lines for a routine application of these tools. We relegate most of the proofs in Appendix A.

[a] *University of Toulouse III, F-31073, Toulouse, France*
[b] *Toulouse Institute of Mathematics, UMR C5583, F-31062, Toulouse, France*
[c] *INSERM, U1027, F-31073 Toulouse, France*
**Correspondence to: Nicolas Savy, University of Toulouse III, Toulouse Institute of Mathematics, UMR C5583, F-31062, Toulouse, France.*
[†] *E-mail: nicolas.savy@math.univ-toulouse.fr*

1. General considerations

1.1. Why model patients' recruitment?

To get marketing authorization, a new product has to succeed in clinical trials. A clinical trial is based on statistical considerations to show the product efficiency, taking into account the variability of the environment. It is a well-known fact that the power of this test is linked to the number of patients we deal with. If an inadequate number of enrolled patients is used, then the study may fail to reject the null hypothesis because of lack of power. So, the number of patients to include is a fixed parameter of the trial. There has been much effort in computing the sample size for clinical trials. Its computation is now standard and mandatory in trial protocol (see Consort Group works [7, 8]). On the other side, relatively little attention is focused on improving the predictions of the recruitment process.

Patients' recruitment is a question of paramount importance in every clinical trial because it is a quite long process and one has no time to waste for mainly two reasons. First, there is an economical reason. A clinical trial is an expensive study in itself, and as the duration of the trial is included in the duration of the licence of exploitation of the product (20 years) [9], a delay generates an enormous loss of income. Second, there is an ethical reason. Patients included in a trial that does not end due to lack of recruitment have followed a trial protocol for nothing. Stopping or continuing a trial is thus a decision with huge consequences, and it will be useful to have some objective tools based on scientific criterion to take it.

In practice, many clinical trials end after the deadline or are prematurely stopped. The main reason of such delays is the contrast between the difficulty of the question we deal with and the simplicity of the tools used to follow the recruitment. In fact, the recruitment process is complex because it involves many sources of randomness (inclusion of patients in each centre, rate of recruitment in each centre, delays in initiation of each centre and others), whereas the tools used to estimate the recruitment time are mainly deterministic and based on proportionality. For instance, consider a clinical trial that has run for 1 year and that 100 patients have been recruited. It is a basic reasoning to say that 2 years later, we will have included 200 more patients. The estimation is very rough, because it does not take into account the randomness of the trial and it tells us nothing about the confidence we can make in this estimation. The weakness comes from the rate of inclusion, which is assumed fixed, while it represents actually an average rate. Thus, attention has to be directed to the variance of the estimated inclusion rate. Rojavin [10] gives a good summary of the situation: 'Patient recruitment and retention remains until now more of an art rather than a science'. To take into account this variability, the idea is to develop a model of the recruitment period involving a random component.

1.2. How to model patients' recruitment?

1.2.1. State of the art. The aim of this paper is to develop tools to make recruitment a science rather than an art. Few authors have considered this problem. The reader can refer to [11] for a systematic review of the existing models for recruitment. As far as we know, the pioneer works were those of Morgan [12] where an estimation of the total study duration is proposed as a function of inclusion duration and based on data from previous clinical trials. Let us cite Lee [13] and Williford *et al.* [14] for a model of the recruitment by Poisson processes. Let us give the very first definition of this process.

Definition 1
Let λ be an integrable function. A stochastic process $\mathcal{N}$ is a Poisson process with rate λ if it is a process with independent and stationary increments and if the increments have a Poisson distribution:

$$\forall s < t, \qquad \mathcal{N}_t - \mathcal{N}_s \overset{\mathcal{L}}{=} \mathcal{P}\left(\int_s^t \lambda_u \mathrm{d}u\right).$$

This definition can be extended to random λ, and in this setting, the process is usually called a conditional Poisson process or Cox process.

Inspired by the queueing theory, this point of view has been widely developed. In fact, the Poisson process appears as a natural assumption in literature [15]. Indeed, the inclusion of a new patient is a random phenomenon, and several arrivals can be assumed independent. For a review of techniques for models based on Poisson processes or more generally models based on punctual processes, see [16, 17]. Senn [18] introduced a model of a multicentric trial based on Poisson process. Poisson processes depend on only one parameter, which is the rate of enrolment in our case. The rate of this process can be a

constant, a random variable, a function of time or even a stochastic process. This large choice of rate makes the model very rich. Piantadosi and Patterson [19] introduced a Poisson model with increasing rate up to a maximal value in the following form:

$$\lambda(t) = \zeta(1 - e^{-\kappa t}),$$

where ζ and κ being parameters. In [20, 21], Carter and co-authors have developed models based on Poisson processes and have noticed that the use of the historic mean is a too simple model, so we have to take into consideration the variability of the rate. The most advanced works in Poisson modelling are those of Anisimov and co-authors [1–3]. In these papers, the authors have taken into account the variability in the rates considering Gamma random variables in a very elegant way. They have validated the model using several sets of real data. These works widely inspired our paper.

Let us point out that, when we have a long-term trial with a lot of patients to be included, we can involve Donsker's theorem and evoke a Brownian model even a fractional Brownian model [22–24] to obtain results on recruitment time. Finally, Abbas *et al.* [25] have studied these questions by using Monte Carlo simulations.

1.2.2. Main notations. Consider now a clinical trial. The parameters of this trial are the following:

- N the number of patients we have to enrol. This parameter is fixed and related to the statistical analysis. Rojavin [10, 26] highlighted different metrics of this parameter (patients screened, patients randomised and patients who have completed the study per protocol). Here, we do not discuss the definition of N and suppose it has been correctly chosen by the investigator.
- T_R the time expected for the inclusion of these N patients. The subscript R means 'reference'.
- C the number of centres. This parameter is not random but can be modified if necessary.

We will model the enrolment in each centre by a Cox process denoted by $(\mathcal{N}^i, i = 1, \ldots, C)$. The distribution of the rate λ_i will be denoted by $\mathcal{L}(i, \theta)$. The actual end time of the study is denoted by T_N; it is a random variable.

Remark 1
λ may depend on the time t. It is especially the case if all centres do not start to enrol at the same time. In this case, we denote by τ_i the time centre i has been active.

The following property is very important, because it allows us to consider easily the multicentric case. In fact, if we consider a multicentric trial with C centres, then the global enrolment is still a Cox process whose rate is the sum of the intensities of each centre, and it is a function (more or less complicated) of the number C of centres. The global enrolment process will be denoted by $\tilde{\mathcal{N}}^C$.

Proposition 1 ([17])
The sum of two Cox processes of rate λ and μ is still a Cox process of rate $\lambda + \mu$.

1.2.3. Discussion on feasibility of a study. We can consider that the value of the rate is known by considering the value of the enrolment rate given by the investigator of the trial, but very often, this value is overestimated because of the optimism of the investigator. This is known as the Lasagna's law: 'Investigators overestimate, manifold, the pool of available patients who meet the inclusion and would be willing to enrol in a particular trial' [20], and references therein. Meanwhile, this point of view can be of interest when testing the feasibility of a clinical trial.

Instead, one should use the observed data of the trial, at some intermediate time, to reach a more consistent value of the rate. This study at an intermediate time will be called an ongoing study. We will discuss the instant of this ongoing study in the Section 4. In the following discussion, we consider that the rate of each centre is unknown and have to be estimated by an ongoing study. We will base the estimation procedure on a maximum likelihood (ML) procedure and detail it in Section 2.

1.2.4. Discussion on the number of centres.

Monocentric studies. It is a clinical trial where there is only one investigator centre. It is the easier case to handle. The easiest model we can deal with is when the rate λ is a constant, but in [20, 21], the authors showed that this assumption is too strong. To take into account the variability of the rate, we can model λ as a random variable; for instance, a gamma distribution with parameters (α, β) or a Pareto distribution with parameters (γ, δ) (see Section 2 for details).

Multicentric studies: low number of centres. When C is small, we can model the enrolment of each centre and apply property 1 to get the complete model. This is relevant because when C is small, the number of patients included to each centre is 'large' and estimating the rate of each centre separately is meaningful. When C increases and N fixed, obviously the number of patients to be included by each centre decreases and the ML estimators tend to be meaningless. In fact, a centre that has included no patient at ongoing study has a rate estimated by 0, and the model tells us that this centre will not include any patient until the end.

Multicentric studies: large number of centres. When the number of centres becomes large, the ML estimation of the λ_i's becomes meaningless. The good idea to get around this problem is to develop a Bayesian approach. We consider the intensities $(\lambda_1, \ldots, \lambda_C)$ as a sample of a distribution denoted by $\mathcal{L}(i, \theta)$. By this way, instead of estimating the C values of the λ_i's, we just have to estimate the parameters of the distribution. In Section 2, we investigate two examples of such a setting, the one introduced by Anisimov and co-authors in the series of articles [1–6] with a $\Gamma(\alpha, \beta)$-distribution and another one we develop in the following text with a Pareto $\Pi(\gamma, \delta)$-distribution.

1.3. What is a model of patients' recruitment for?

Assume we deal with a clinical trial where N is the number of patients to include, C is the number of centres and T_R the expected duration of the trial. We have to keep in mind that N cannot be modified, that C can be modified and that T_N is a random variable. One considers a model for the enrolment process whose parameters will be denoted by θ (θ may be a vector). θ may represent the global rate estimated without wondering about centres, or the rate of each centre, or the parameters of the distribution of the centres in the Bayesian setting.

1.3.1. θ known: feasibility of the trial. Assume that the recruitment process is modelled by a Poisson process or a Cox process and that the parameter θ describing the rate is known. The model is thus perfectly defined. The key point of the method is the fact that we are able to calculate the probability of finishing on time and the expectation of the duration of the trial. In fact, when all centres are initiated initially,

Theorem 2
For any θ, $C \geqslant 0$, $t \geqslant 0$ and $N \geqslant 0$, we have

$$\mathbb{P}\left[\tilde{\mathcal{N}}^C(t) \geqslant N\right] = 1 - \mathbb{E}_{\lambda_1,\ldots,\lambda_C}\left[\sum_{k=0}^{N-1} e^{-(\lambda_1+\cdots+\lambda_C)t}\frac{[(\lambda_1+\cdots+\lambda_C)t]^k}{k!}\right] \tag{1}$$

$$= 1 - \sum_{k=0}^{N-1}\frac{t^k}{k!}\int_{\mathbb{R}^C}(x_1+\cdots+x_C)^k e^{-t(x_1+\cdots+x_C)}\prod_{i=1}^{C} p_{\theta,i}(x_i)\,\mathrm{d}x_i, \tag{2}$$

$$\mathbb{E}\left[T_n\right] = \mathbb{E}_{\lambda_1,\ldots,\lambda_C}\left[\frac{N}{\lambda_1+\cdots+\lambda_C}\right] \tag{3}$$

$$= N\int_{\mathbb{R}^C}\frac{\prod_{i=1}^{C} p_{\theta,i}(x_i)}{x_1+\cdots+x_C}\mathrm{d}x_1,\ldots,\mathrm{d}x_C. \tag{4}$$

Remark 2
To lighten the notations, for any continuously distributed random variable X, p_X will denote its density.

Proof
Conditioning on the value of the λ_i's gives a straightforward proof. □

If the centres are not initiated at $t = 0$ but at different times, such quantities can still be evaluated by means of a Monte Carlo method.

Now, as already pointed out in [1, 3], we are able to consider some tools to monitor the clinical trial (we provide some details in Section 1.3.2).

- We can investigate the **feasibility of the trial**, which is given by
$$\mathbb{P}\left[\tilde{\mathcal{N}}^{\mathcal{C}}(T_R) \geqslant N\right].$$
- Given a fixed probability (say, 80% for instance), we can calculate an **estimation of the duration** of the trial up to this probability, that is the time
$$T \ \ s.t. \ \mathbb{P}\left[\tilde{\mathcal{N}}^{\mathcal{C}}(T) \geqslant N\right] = 0.80.$$
- Given a fixed probability (say, 80% for instance), we can calculate an **estimation of the number of centres** necessary for ending the trial on time up to this probability by
$$C_N \ \ s.t. \ \mathbb{P}\left[\tilde{\mathcal{N}}^{\mathcal{C}_N}(T_R) \geqslant N\right] \geqslant 0.80.$$

1.3.2. θ known: ongoing study. Consider now an ongoing study at time t_1. During the period $[0, t_1]$, N_1 patients are assumed to be included. We will denote by $\mathcal{F}_{t_1}$ the history (filtration) of the enrolment process until time t_1, this means the σ-algebra generated by the different events (inclusions of patients) up to t_1. The key point that makes this approach of paramount interest is that we can reach the probability of including the $N - N_1$ remaining patients before the deadline and to estimate the duration of the trial. In fact, when all centres are initiated initially, one can apply Theorem 2, where the expectations are taken with respect to the forward laws of the λ_i's, that is, the predictive laws based on using interim data.

Like in Theorem 2, if the centres are not initiated at $t = 0$ but at different times, such quantities can still be evaluated by means of a Monte Carlo method.

Remark 3
We have two choices to evaluate the integral (2): calculate it explicitly when a closed form of the integral is available or use Monte Carlo simulations.

Remark 4
In the setting of the models developed in Section 2, we can calculate the forward laws; see Propositions 4 and 5.

Now, we can use the same kind of tools as those introduced in the previous section.
We are also able to introduce corrective actions on the trial.

- We are able to **estimate the value of the recruitment rate** to reach the deadline. When the rate is constant or when the expected rate is easily linked to the parameters θ (not possible for Pareto but realistic for Gamma, see Remark 9), we can calculate an estimation of the rate necessary to reach the deadline, that is, the value $\tilde{\theta}$ such that
$$\mathbb{P}_{\tilde{\theta}}\left[\tilde{\mathcal{N}}^{\mathcal{C}}(T) \geqslant N \mid \mathcal{F}_{t_1}\right] = 0,80.$$
For example, in the case of constant rate λ, because the function $\lambda \mapsto \mathbb{P}_{\lambda}\left[\tilde{\mathcal{N}}^{\mathcal{C}}(T) \geqslant N \mid \mathcal{F}_{t_1}\right]$ is increasing and tends to 1 as λ grows to infinity, then there exists a particular λ for which this probability is 0.80. This action is quite artificial because, in practice, it is quite hard to change the rate of recruitment. Meanwhile, it is a useful tool for taking a decision on the continuation of a clinical trial, especially for institutional studies.
- We are able to **estimate the number of centres to open** to reach the deadline. In the Bayesian setting, the overall rate is directly linked to the number of centres, so given a fixed probability (say, 80% for instance), given N and T_R, we can calculate an estimation of the number C_N of centres needed to include these N patients in the time T_R. Indeed, the overall rate of inclusion is a sum of the two random variables
$$\Lambda = \Lambda_A + \Lambda_B,$$
where Λ_A is the contribution of the already opened centres (their distribution is thus the forward one) and Λ_B is the contribution of the new centres, the distribution does not depend of the history $\mathcal{F}_{t_1}$. Under some condition (for instance that $\exists \epsilon > 0, \delta > 0, \forall i, \mathbb{P}\left[\lambda_i > \epsilon\right] > \delta$), each increment of C_N increases the probability to end the trial on time, and this probability tends to 1 as C_N grows to infinity. To calculate the smallest number C_N of centres to open, a simple procedure consists in incrementing C_N until reaching the desired probability (here 0.80).

By an easy modification of this reasoning, one can calculate the number of centres to close if the rate is too high. This is not a problem in practice, but it remains an interesting theoretical question.

1.3.3. θ unknown: ongoing study. In most cases, θ is unknown. The classic idea is to replace the real parameter θ by an estimation $\hat{\theta}$ in each relationship. For this, we use the data collected on $[0, t_1]$ to estimate θ conformally to Section 2.1.3. In most situations, θ is unknown or given by the investigator and often overestimated. We will discuss the error made on predictions (that is, on $\mathbb{E}_{\hat{\theta}}[T_N]$ and $\mathbb{P}_{\hat{\theta}}[\tilde{\mathcal{N}}^C(T_R) \geqslant N]$) when replacing the true parameters θ_0 by the estimated parameters $\hat{\theta}$ in Sections 2.2 and 3.2. This is a key point in the method. Anisimov [3, p. 4959] tells us that when $C > 20$, the precision is good enough for the method to be relevant. The method of estimation is by maximisation of the likelihood function. The likelihood of a Poisson process with deterministic rate is a consequence of the Girsanov theorem:

Theorem 3 (Girsanov, [17])
The probability density of a Poisson process with a rate λ, which is a deterministic function of time, with respect to a standard Poisson process, is given by the following:

$$L_T = \prod_{n \geqslant 1} (\lambda(s_n)\mathbf{1}_{s_n \leqslant T})\, e^{\int_0^T [1-\lambda(s)]\mathrm{d}s}, \tag{5}$$

where $(s_n,\ n \in \mathbb{N})$ denote the jump times.

Remark 5
When λ is not deterministic, one replaces L_T by $\mathbb{E}_\lambda[L_T]$.

2. Models for patients recruitment: centre opening dates known

2.1. Presentation of the models

2.1.1. Definitions. In this section, we investigate two models. First is the model introduced by Anisimov and co-authors [1–3], which is the model of reference. It is called the Γ-Poisson model; this means the rate of the Poisson process is $\Gamma(\alpha, \beta)$-distributed whose density is given by the following:

$$f_{(\alpha,\beta)}(\lambda) = \lambda^{\alpha-1}\frac{\beta^\alpha\, e^{-\beta\,\lambda}}{\Gamma(\alpha)}\mathbf{1}_{\{\lambda \geqslant 0\}} \quad \text{with} \quad \Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1}\, e^{-t}\, \mathrm{d}t.$$

Second, we investigate the Π-Poisson model this means the rate of the Poisson process is distributed from a Pareto-decentred to avoid problems at 0-denoted by $\Pi(\gamma, \delta)$, whose density is given by the following:

$$f_{(\gamma,\delta)}(\lambda) = \gamma\frac{\delta^\gamma}{\lambda^{\gamma+1}}\,\mathbf{1}_{\{\lambda \geqslant \delta\}}.$$

Remark 6
One could argue that the condition $\lambda \geqslant \delta$ for some $\delta > 0$ in the Pareto model is not realistic and use a parametrization of the form $\gamma\delta^\gamma/(\lambda+\delta)^{\gamma+1}$ instead. However, when estimating the parameters (see next sections), δ has no constraints: it can become very close to zero. Moreover, we wanted to stick to the definition of $f_{(\gamma,\delta)}$ given previously, because it is the density of the well-known Pareto distribution.

Remark 7
The introduction of the Gamma distribution is very convenient here because it is easy to handle with the Poisson process. The Pareto distribution is computationally harder to handle and forces us to use Monte Carlo techniques, but it is the probabilistic formulation of the well-known fact that 'around 80% of enrolments is made by 20% of centres', and it seems relevant for the application. Despite the fact that Anisimov has provided evidence in [4] that the Gamma–Poisson model is more relevant than the Pareto–Poisson, we think that a comparative analysis of these models is important.

For the investigations we make, we have to introduce some notations. We denote by $k_{i,t}$ the number of patients included by centre i in the period $[0, t]$ and by $K_t = \sum_{i=1}^C k_{i,t}$ the total number of patients included on $[0, t]$. When doing an ongoing study at time t_1, when no confusion is possible, we will denote by $k_i = k_{i,t_1}$ and $K = K_{t_1}$. We denote by τ_i the duration of activity of the ith centre in $[0, t_1]$ (which means τ_i may depend on t_1). Finally, we denote by $p^{t_1}_{\theta,i}(x)$ the probability density of the forward law of λ_i on the basis of using interim data.

2.1.2. θ known, analysis from an ongoing study: the forward laws. The key point to use these models from an ongoing study is the forward laws. The results are the following:

Proposition 4 ([3] Relation (13), p. 4969)
In the Gamma case, $p_{\theta,i}^{t_1}(x)$ is the density of a $\Gamma(\alpha + k_i, \beta + \tau_i)$ distribution.

Proposition 5
In the Pareto case, the density $p_{\theta,i}^{t_1}$ is

$$p_{\theta,i}^{t_1}(x) = \frac{\tau_i^{k_i-\gamma}}{\Gamma_{\text{inc}}(k_i-\gamma, \delta\tau_i)} e^{-x\tau_i} x^{k_i-\gamma-1} \mathbf{1}_{\{x\geqslant\delta\}}, \quad \text{where} \quad \Gamma_{\text{inc}}(y,x) = \int_x^\infty e^{-t} t^{y-1} \mathrm{d}t$$

is usually called (upper) incomplete Gamma function and exists in many mathematical software packages.

Proof
Section A.1. □

Remark 8
In the Pareto case, one has to simulate random variables with an incomplete Gamma distribution and calculate the expected value of the end time of the study T_N by means of a Monte Carlo simulation.

2.1.3. θ unknown: estimation of parameters from an ongoing study. The first step to achieve an interim prediction is to estimate the parameters of the model. To this end, we use the data available at the interim time t_1. In our setting, the ML estimators of the parameters of the models are given by the following propositions:

Proposition 6 ([3] p. 4968)
For the Γ-Poisson's model, the ML estimation of (α, β) is obtained by the maximisation of

$$\ln L = C\alpha \ln\beta - C\ln\Gamma(\alpha) + \sum_{i=1}^{C} [\ln\Gamma(\alpha + k_i) - (\alpha + k_i)\ln(\beta + \tau_i) + \tau_i].$$

Proposition 7
For the Π-Poisson's model, the ML estimation of (γ, δ) is obtained by the maximisation of

$$\ln L = C\gamma \ln\delta + C\ln\gamma + \sum_{i=1}^{C} [\ln\Gamma_{\text{inc}}(k_i - \gamma, \delta\tau_i) - (k_i - \gamma)\ln\tau_i + \tau_i].$$

Proof
Section A.2. □

Remark 9
In the case of $\tau_i = t_1$ for all i, then, if $(\hat{\alpha}, \hat{\beta})$ is the ML estimation in the Gamma model, we have $\frac{\hat{\alpha}}{\hat{\beta}} = \frac{1}{C}\frac{K}{t_1}$. Indeed, the derivative of $\ln L$ with respect to β is

$$\frac{\partial \ln L}{\partial\beta} = C\frac{\alpha}{\beta} - \frac{C\alpha + K}{\beta + t_1},$$

which cancels out if and only if $\frac{\alpha}{\beta} = \frac{1}{C}\frac{K}{t_1}$. One just has to find the extremum of a single variable function. Moreover,

$$\mathbb{E}_{\hat{\lambda}}\left[\sum_{i=1}^{C}\lambda_i\right] = \frac{K}{t_1}.$$

This means that the mean value of the overall rate is the classical estimator of the rate in a constant rate model.

2.2. Error in the estimation of the parameters

In this section, we investigate the statistical properties of the parameter estimators used in the empirical Bayesian setting. To take into account the error made by the replacement of the sample by its distribution, we have to insure the quality of the estimation of this one. This way, we show the consistency and the asymptotic normality of these estimators by means of the delta method. To deepen this analysis, we investigate the calculation of sensitivity parameters. These parameters known in mathematical finance as Greeks [27] allows us to estimate the impact of the error made on a parameter on a function of interest (here, the expectation of the end of the trial).

2.2.1. Properties of the estimators. In the following discussion, we recall some remarks made in [3], which remain true in the Pareto setting. Here, the k_is are seen as random variables. In both models, the likelihood takes the form

$$\ln L = \sum_{i=1}^{C} f(\theta, k_i, \tau_i).$$

We denote the Fischer information matrix by $I_i(\theta_0, k_i, \tau_i) = -\mathbb{E}_{\theta_0}\left[\nabla^2 f(\theta_0, k_i, \tau_i)\right]$ and the total Fischer information matrix by

$$I(\theta_0) = \sum_{i=1}^{C} I_i(\theta_0, k_i, \tau_i).$$

If the τ_i's satisfy $\tau_i \geqslant \xi$ for any i and for some $\xi > 0$ and that, in probability, the limit matrix $I_0 = \frac{1}{C} I(\theta_0)$ exists, then, ML estimation theory ensures that the estimated parameters are consistent and that $\hat{\theta}$ has an asymptotic normal distribution with mean θ_0 and covariance matrix $\frac{1}{C} I_0^{-1}$.

Proposition 8
For the $\Gamma(\alpha, \beta)$-Poisson model, the 2×2 matrix I_i is given by

$$\begin{aligned}
(I_i)_{11} &= -\mathbb{E}_{\alpha_0,\beta_0}\left[\psi^{(3)}(\alpha_0 + k_i)\right] + \psi^{(3)}(\alpha_0),\\
(I_i)_{12} = (I_i)_{21} &= -\frac{1}{\beta_0} + \frac{1}{\beta_0 + \tau_i},\\
(I_i)_{22} &= \frac{\alpha_0}{\beta_0^2} - \alpha_0 \frac{1 + \frac{\tau_i}{\beta_0}}{(\beta_0 + \tau_i)^2},
\end{aligned}$$

where $\psi^{(3)}$ is the trigamma function and the aforementioned expectation is calculated by means of Monte Carlo simulations.

Proof
Section A.3 □

Proposition 9
For the $\Pi(\gamma, \delta)$-Poisson model, the 2×2 matrix I_i is given by

$$\begin{aligned}
(I_i)_{11} &= -\frac{1}{\gamma_0^2} - \mathbb{E}_{\gamma_0,\delta_0}\left[\frac{\partial^2}{\partial \gamma_0^2}\Gamma_{\text{inc}}(k_i - \gamma_0, \delta_0 \tau_i)\right],\\
(I_i)_{12} = (I_i)_{21} &= -\frac{1}{\delta_0} + \tau_i e^{-\delta_0 \tau_i} \ln(\delta_0 \tau_i)\mathbb{E}_{\gamma_0,\delta_0}\left[(\delta_0 \tau_i)^{k_i - \gamma_0 - 1}\right],\\
(I_i)_{22} &= \frac{\gamma_0}{\delta_0^2} + \tau_i^2 e^{-\delta_0 \tau_i}\mathbb{E}_{\gamma_0,\delta_0}\left[(\delta_0 \tau_i)^{k_i - \gamma_0 - 2}(\delta_0 \tau_i + k_i - \gamma_0 - 1)\right],
\end{aligned}$$

where the aforementioned means are calculated using Monte Carlo simulations.

Proof
Section A.3 □

2.2.2. Sensitivity analysis for the expectation of the trial duration. We discuss here the impact of error on the estimation of the set of parameter θ on the computation of the expected duration of the study. This is characterised by the following:

Definition 2
We call sensitivity parameter the quantities

$$\mathcal{SP}(\theta_i) = \frac{\partial}{\partial \theta_i} \mathbb{E}_\lambda [T_N]. \tag{6}$$

Remark 10
In practice, (6) can be interpreted by

$$\Delta \mathbb{E}_\lambda [T_N] = \mathcal{SP}(\theta) \Delta \theta \tag{7}$$

and gives the error on the estimation of $\mathbb{E}_\lambda [T_N]$ induced by an error on the parameter. For example, if the sensitivity parameter is 0.75, this means that a variation of the parameter of 1% induces a variation of the expectation of the trial duration of $0.075\,\theta$.

Proposition 10
For the Γ-Poisson model, we denote

$$m = \sum_{i=1}^{C} \frac{\alpha + k_i}{\beta + \tau_i} \qquad \text{and} \qquad v = \sum_{i=1}^{C} \frac{\alpha + k_i}{(\beta + \tau_i)^2},$$

and we have the approximate sensitivity parameters:

$$\mathcal{SP}(\alpha) = \partial_\alpha \mathbb{E}_\theta [T_N] \approx -N \frac{\partial_\alpha (m - \frac{v}{m})}{(m - \frac{v}{m})^2},$$
$$\mathcal{SP}(\beta) = \partial_\beta \mathbb{E}_\theta [T_N] \approx -N \frac{\partial_\beta (m - \frac{v}{m})}{(m - \frac{v}{m})^2},$$

where the expectations are calculated with the posterior rates (in particular, the k_i's are observed and thus fixed here). They can be easily numerically computed.

Proof
Section A.4. □

For the $\Pi(\gamma, \delta)$-Poisson model, the sensitivity parameters are calculated by means of discretised derivatives and Monte Carlo simulations.

2.3. Validation of the Π-Poisson model and comparison with the Γ-Poisson model

To validate the Π-Poisson model and to make the comparison properly, we use the data of Anisimov [6, pp. 28–29]. There are four data sets denoted by A,B,C and D, and we keep the same notations here. In [6], Anisimov and Fedorov did not give the duration of the studies: in fact, the likelihood is calculated conditionally on the total number of patients recruited so that it does not depend on the total duration and β. In the Π-Poisson model, such a conditional likelihood is not reachable in closed form, so we arbitrarily set the total duration of the study to be 1 year. This has no effect on the estimation of α in the Γ-Poisson model or on the occupancy curves, even though it has one on the parameters in the Π-Poisson model.

2.3.1. Validation and comparison. The first two columns of Table I indicate the estimations of the parameters for each study together with their confidence region (see next section) for both models. To validate the model, we plot in Figure 1 the mean occupancy (see [3, p. 4961] for details), calculated with estimated parameters for the Π-Poisson model (dot-dash lines), for the Γ-Poisson model (dotted lines) and for the real data (solid lines). We have closed-form formulae for the mean occupancy in the Γ-Poisson model and use Monte Carlo simulations to compute it in the Π-Poisson model.

Figure 1 shows that the Π-Poisson model fits quite well the real data, but not as well as the Γ-Poisson one. Nevertheless, it may work better for studies with a smaller number of 'big' centres compared with the number of 'small' ones, which conform to the Pareto idea evoked in Remark 7.

Table I. Estimated parameters and confidence region for Gamma–Poisson and Pareto–Poisson models.

Model	$\hat{\theta}_1$	$\hat{\theta}_2$	$l, \overrightarrow{V_l}$	$L, \overrightarrow{V_L}$
Γ-Poisson (A)	2.89	0.42	0.32, $\begin{bmatrix} 0.99 \\ 0.14 \end{bmatrix}$	0.016, $\begin{bmatrix} -0.14 \\ 0.99 \end{bmatrix}$
Π-Poisson (A)	1.74	3.38	0.036, $\begin{bmatrix} 0.72 \\ 0.69 \end{bmatrix}$	0.091 $\begin{bmatrix} -0.69 \\ 0.72 \end{bmatrix}$
Γ-Poisson (B)	2.89	0.26	0.47, $\begin{bmatrix} 1 \\ 0.09 \end{bmatrix}$	0.016 $\begin{bmatrix} -0.09 \\ 1 \end{bmatrix}$
Π-Poisson (B)	1.58	5.07	0.046, $\begin{bmatrix} 0.72 \\ 0.69 \end{bmatrix}$	0.17 $\begin{bmatrix} -0.69 \\ 0.72 \end{bmatrix}$
Γ-Poisson (C)	4.59	031	0.64, $\begin{bmatrix} 1 \\ 0.07 \end{bmatrix}$	0.012 $\begin{bmatrix} -0.07 \\ 1 \end{bmatrix}$
Π-Poisson (C)	2.04	8.34	0.043, $\begin{bmatrix} 0.72 \\ 0.69 \end{bmatrix}$	0.18 $\begin{bmatrix} -0.69 \\ 0.72 \end{bmatrix}$
Γ-Poisson (D)	3.46	0.43	0.59, $\begin{bmatrix} 1 \\ 0.12 \end{bmatrix}$	0.023 $\begin{bmatrix} -0.12 \\ 1 \end{bmatrix}$
Π-Poisson (D)	2.05	4.42	0.045, $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	0.16 $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Gamma–Poisson: $\theta_1 \equiv \alpha, \theta_2 \equiv \beta$. Pareto–Poisson: $\theta_1 \equiv \gamma, \theta_2 \equiv \delta$. In each of the last two columns are the length of the axis of the ellipse and the corresponding vector.

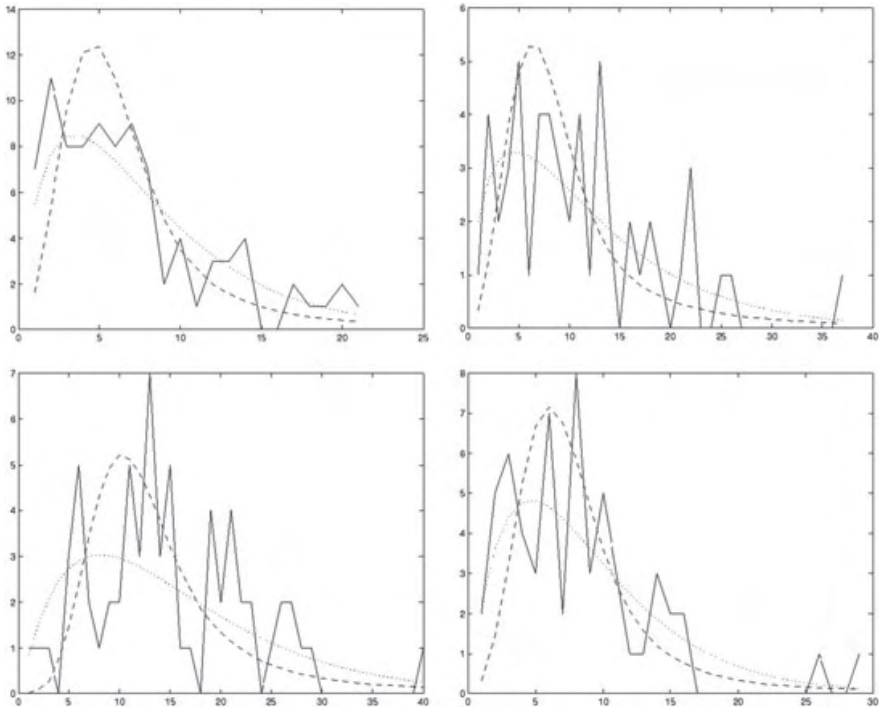

Figure 1. Occupancy curves for studies A (top left), B (top right), C (bottom left) and D (bottom right). Solid line, real data; dashed line, Π-Poisson model; dotted line, Γ-Poisson model.

2.3.2. Errors in estimated parameters. The error in estimated parameters is of the order of $1/\sqrt{C}$, as C grows to infinity, and the asymptotic distribution is normal with an inverse variance matrix given by Propositions 8 and 9. The 95% confidence region is then an ellipse defined by

$$\{X \in \mathbb{R}^2 : X^T I_0^{-1} X \leqslant p/C\},$$

where $p = 5.99$ is the 95% quantile of a Chi-square distribution with 2 degrees of freedom. In Table I, we give a vector for each axis of the ellipse and the corresponding length of the ellipse in this direction.

Statistics in Medicine

Note that in the Γ-Poisson model, the correlation between $\hat{\alpha}$ and $\hat{\beta}$ is almost zero, whereas it is not in the Π-Poisson one. Figure 2 plots the ellipse for the first data set for both models.

The errors in estimated parameters are not negligible because we have a 95% probability to make between 4% and 10% errors, depending on the study but not of the model, the ones with the least number of centres having the biggest errors (B and D).

2.4. *Comparison of predictions on real data*

We took the data from a clinical trial. The target was to enrol 680 patients in 5 years. The study actually ended after 4.25 years. Ninety centres were opened. Two ongoing studies were planned, at $t_1 = 3.32$ and $t_2 = 4.02$ years.

2.4.1. Estimation of the parameters of the model. We estimate the parameters at time t_1 and t_2, and results appear in Table II.

2.4.2. Errors in estimated parameters. As earlier, we show in Table II the lengths of the axis of the confidence region, which is an ellipse. The error made on parameters is less than 5%–7% with probability 95%.

2.4.3. On the duration of the trial. Table III shows the expected duration of the trial. We also include the error induced by parameter estimation. Indeed, let $g : \mathbb{R}^2 \mapsto \mathbb{R}$ be some numerical function (e.g. $g(\theta) = \mathbb{E}_\theta[T_N]$), which is approximated by a linear function for the values of θ close to θ_0 (in our case, this means in the confidence region). Also consider the smallest rectangle containing the ellipse

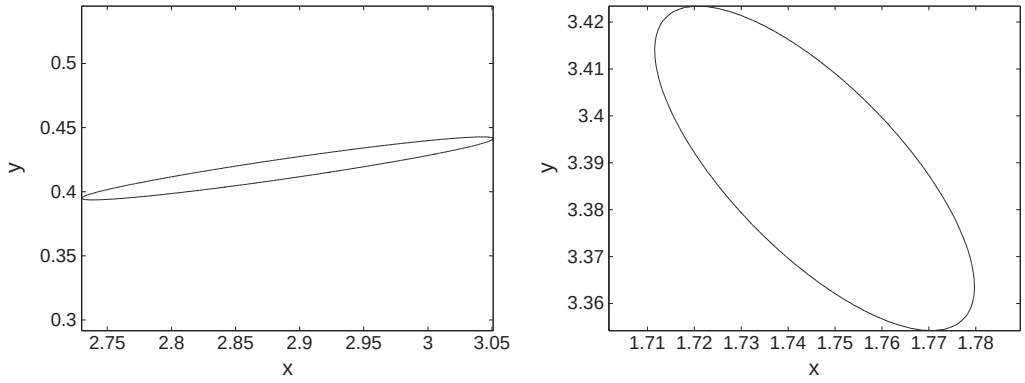

Figure 2. Confidence region in study A for the Γ-Poisson model (left—x-axis, α; y-axis, β) and the Π-Poisson model (right—x-axis, γ; y-axis, δ).

Table II. Estimated parameters and confidence region for Γ-Poisson and Π-Poisson models.

Model	$\hat{\theta}_1$	$\hat{\theta}_2$	$l, \overrightarrow{V_l}$	$L, \overrightarrow{V_L}$
Γ-Poisson (t_1)	1.06	0.44	$0.12, \begin{bmatrix} 0.92 \\ 0.40 \end{bmatrix}$	$0.025, \begin{bmatrix} -0.40 \\ 0.92 \end{bmatrix}$
Γ-Poisson (t_2)	0.91	0.41	$0.097, \begin{bmatrix} 0.90 \\ 0.43 \end{bmatrix}$	$0.024, \begin{bmatrix} -0.43 \\ 0.90 \end{bmatrix}$
Π-Poisson (t_1)	1.02	0.61	$0.061, \begin{bmatrix} 0.98 \\ 0.19 \end{bmatrix}$	$0.034, \begin{bmatrix} -0.19 \\ 0.98 \end{bmatrix}$
Π-Poisson (t_2)	0.96	0.49	$0.066, \begin{bmatrix} 0.98 \\ 0.21 \end{bmatrix}$	$0.028, \begin{bmatrix} -0.21 \\ 0.98 \end{bmatrix}$

Γ-Poisson: $\theta_1 \equiv \alpha, \theta_2 \equiv \beta$. Π-Poisson: $\theta_1 \equiv \gamma, \theta_2 \equiv \delta$. In each of the last two columns are the length of the axis of the ellipse and the corresponding vector.

Table III. Estimation of the trial duration: the expectation on the first line and the 95% confidence interval on the second one.

Model	Time t_1	Time t_2
Γ-Poisson	4.135	4.258
	[4.130, 4.140]	[4.257, 4.259]
Π-Poisson	4.074	4.254
	[4.068, 4.080]	[4.253, 4.255]

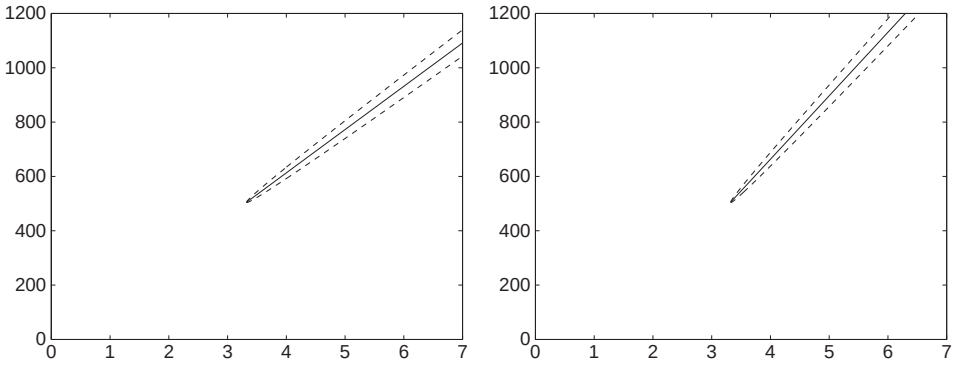

Figure 3. Predictive curves, calculated at t_1: Γ-Poisson model (left); Π-Poisson model (right).

and whose edges are parallel to the axis of the ellipse. Then, $|g(\theta) - g(\theta_0)|$ has its maximum in a corner of the rectangle, which depends on the direction of the vector $\nabla g(\theta_0)$. We use this value to compute the ($\approx 95\%$) confidence interval of $\mathbb{E}[T_N]$.

Note that, even though the error on estimated parameters is of order 5%, the error on the prediction of the end time of the study is very small. This can be explained by the fact that we are very close to the end time of the study, so the sensitivity parameters of $\mathbb{E}[T_N]$ are very small. Notice that there is a very little difference between the durations estimated by the two models at t_2. At t_1, the difference is larger and the Pareto model seems to underestimate the duration. We illustrate this in Figure 3, which shows the prediction of enrolment with parameters estimated at time t_1. The 95% confidence interval is also drawn.

3. Models for patients recruitment: centre opening times unknown

In this section, we investigate the problem of the nonknowledge of the centres opening dates. The problem comes from the analysis of a data set lent to us by the 'Intergroupe Francophone du Myélome' where the date of the centre opening is not specified.

3.1. Presentation of the data and the model

3.1.1. The data set. The INSERM unit 1027 lent to us the data used in this section. They are the results from a clinical trial involving 77 centres and aiming to recruit 610 patients over 3 years. The particularity of this trial is that the number of patients was recruited in 2.31 years. Obviously, there were too many centres. The dates at which the centres opened are not specified, but we know the times of the first inclusion of each centre. Figure 4 is the derivative of the interpolation of the global inclusion process by a piecewise quadratic function. It shows that a constant rate model is not realistic. The models presented in Section 2 are not usable because of the nonknowledge of the centres' activity durations τ_i. To overpass this difficulty, we can replace the opening centre dates by the following:

- 0, but this will underestimate the rate of the Poisson process and then overestimate the duration of the trial; and
- the first inclusion time, but this will overestimate the rate of the Poisson process and then underestimate the duration of the trial.

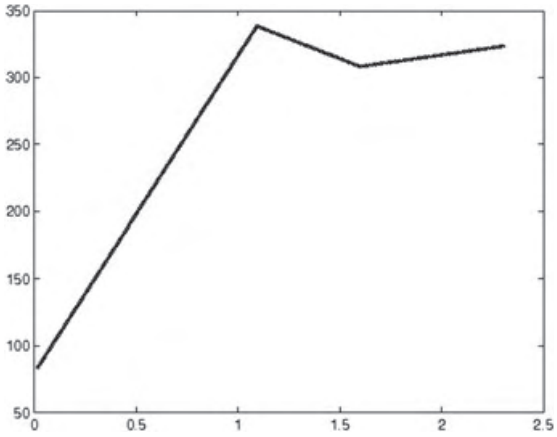

Figure 4. Estimation of the rate of the Poisson process.

To reach a more relevant estimation, we will introduce the so-called uniform Γ-Poisson model ($\mathcal{U}\Gamma$-Poisson model for short).

Remark 11
Another model can be introduced if, for each centre, the number of inclusions is high. In this setting, you can fit an exponential distribution on the sequence of inclusion times and consider the centre opening to be the first inclusion times minus an exponentially distributed number (or 0 if this quantity is negative).

3.1.2. The uniform Γ*-Poisson model.* Because the centres' opening dates are unknown to us, we introduce an extension of the Γ-Poisson model where the opening durations are uniformly distributed. We allow different duration intervals for each centre; therefore, τ_i is a uniform random variable in $[S_i', S_i]$.

In our particular study, because we observe the time of the first inclusion of each centre, but not the opening time of the centre, we will assume that the opening time is a random variable uniformly distributed in $[0, \rho_i]$, where ρ_i is the time of the first inclusion of centre i. Thus, when doing an ongoing study at time t_1, we will have $S_i' = t_1 - \rho_i$ and $S_i = t_1$; except when the first inclusion has not occurred, in which case we put $S_i' = 0$. Thus, S_i' and S_i do depend on t_1.

3.1.3. θ known, analysis from an ongoing study: the forward laws. To estimate the end time of the study, we need the forward distributions of the rates. To this end, we approximate the overall rate of inclusion by a Γ distribution, by matching the mean and variance of both random variables.

Proposition 11
If we put

$$m = \sum_{i=1}^{C} \frac{\alpha + k_i}{S_i - S_i'} \ln\left(\frac{\beta + S_i}{\beta + S_i'}\right) \qquad \text{and} \qquad v = \sum_{i=1}^{C} (\alpha + k_i) \frac{1}{(\beta + S_i')(\beta + S_i)},$$

we can approximate the overall forward rate by a Gamma distribution by matching the first two moments $\Lambda \overset{d}{\approx} \Gamma(A, B)$ with

$$A = \frac{m^2}{v} \qquad \text{and} \qquad B = \frac{m}{v},$$

and we have

$$\mathbb{E}\,[T_N] \approx N \frac{m}{m^2 - v}.$$

Proof
Section A.5. □

3.1.4. θ unknown: estimation of parameters from an ongoing study.

Proposition 12
The ML estimator of (α, β) is obtained by the maximisation of the following:

$$\ln L = C\alpha \ln \beta + C \ln \Gamma(\alpha) + \sum_{i=1}^{C} \left[\ln \Gamma(\alpha + k_i) - \ln(S_i - S_i') + \ln(J(\alpha, \beta, k_i, S_i', S_i)\right],$$

by putting $J(\alpha, \beta, k, S', S) = \int_{S'+\beta}^{S+\beta} t^{-k-\alpha}(t-\beta)^k \mathrm{d}t$.

Proof
Section A.5. □

3.2. Error in the estimation of the parameters

3.2.1. Properties of the estimators. We can make the same remarks as in the previous section. If the (S_i, S_i')'s satisfy $S_i - S_i' \geqslant \xi$ for any i and for some $\xi > 0$ and that, in probability, the limit matrix $I_0 = \frac{1}{C} I(\theta_0)$ exists, then ML estimation theory ensures that the estimated parameters are consistent and that $\hat{\theta}$ have an asymptotic normal distribution with mean θ_0 and covariance matrix I_0^{-1}.

3.2.2. Sensitivity analysis for the expectation of the trial duration. Using the approximation in Proposition 11 and recalling that m and v are functions of α and β, we can easily calculate (numerically or in closed form) the sensitivity parameters:

$$\partial_\theta \mathbb{E}_\theta[T_N] = -N \frac{\partial_\theta (m - \frac{v}{m})}{(m - \frac{v}{m})^2}, \quad \theta = \alpha, \beta.$$

3.3. Validation and Comparison using real data

3.3.1. Estimation of parameters. Table IV contains the values of the parameters estimations from ongoing studies at times 1, 1.5 and 2 years and at the end of the trial for the $\mathcal{U}\Gamma$-Poisson model.

3.3.2. Validation of the model. Let ν_j be the number of centres that recruited j patients. Then,

$$\mathbb{E}\left[\nu_j\right] = \sum_{i=1}^{C} \mathbb{P}\left[k_i = j\right] = \beta^\alpha \Gamma(\alpha)^{-1} \Gamma(j+\alpha) \sum_{i=1}^{C} \frac{1}{S_i - S_i'} J(\alpha, \beta, j, S_i', S_i).$$

Figure 5 compares the expected occupancy of centres with real data. The fitting is good.

3.3.3. Errors in estimated parameters. As in the previous section, the error in estimated parameters is of the order of $1/\sqrt{C}$, as C grows to infinity, and the asymptotic distribution is normal with inverse variance matrix given by Propositions 8 and 9. The 95% confidence region is then an ellipse defined by

$$\{X \in \mathbb{R}^2 : X^T I_0^{-1} X \leqslant p/C\},$$

Table IV. Estimated parameters and confidence region for $\mathcal{U}\Gamma$-Poisson model.

Model	Times	$\hat{\alpha}$	$\hat{\beta}$	$l, \overrightarrow{V_l}$		$L, \overrightarrow{V_L}$	
$\mathcal{U}\Gamma$-Poisson	$t_1 = 1$	0.824	0.233	0.135,	$\begin{bmatrix} 0.96 \\ 0.29 \end{bmatrix}$	0.020,	$\begin{bmatrix} -0.29 \\ 0.96 \end{bmatrix}$
$\mathcal{U}\Gamma$-Poisson	$t_1 = 1.5$	1.369	0.329	0.194,	$\begin{bmatrix} 0.97 \\ 0.24 \end{bmatrix}$	0.021	$\begin{bmatrix} -0.24 \\ 0.97 \end{bmatrix}$
$\mathcal{U}\Gamma$-Poisson	$t_1 = 2$	1.514	0.370	0.192,	$\begin{bmatrix} 0.97 \\ 0.24 \end{bmatrix}$	0.022	$\begin{bmatrix} -0.24 \\ 0.97 \end{bmatrix}$
$\mathcal{U}\Gamma$-Poisson	$T_f = 2.32$	1.506	0.365	0.178,	$\begin{bmatrix} 0.97 \\ 0.24 \end{bmatrix}$	0.021	$\begin{bmatrix} -0.24 \\ 0.97 \end{bmatrix}$

In each of the last two columns are the length of the axis of the ellipse and the corresponding vector.

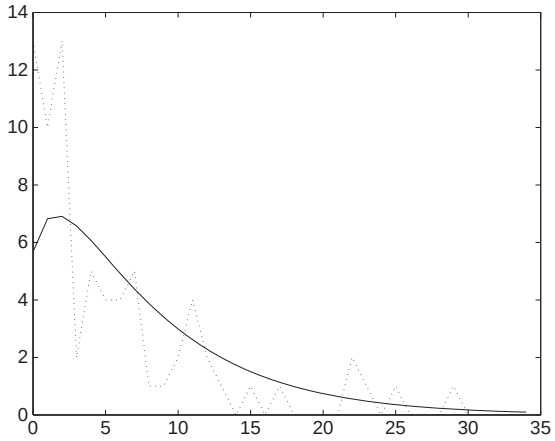

Figure 5. Occupancy.

Table V. Estimation of the trial duration: the expectation on the first line and the 95% confidence interval on the second one.

Model	$t_1 = 1$	$t_1 = 1.5$	$t_1 = 2$
$\mathcal{U}\Gamma$-Poisson	2.602	2.342	2.355
	[2.581, 2.623]	[2.334, 2.350]	[2.352, 2.358]

where $p = 5.99$ is the 95% quantile of a Chi-square distribution with 2 degrees of freedom. In Table IV, we give a vector for each axis of the ellipse and the corresponding length of the ellipse in this direction.

The correlation between $\hat{\alpha}$ and $\hat{\beta}$ is almost zero.

3.3.4. On the duration of the trial. Here, we use the same tools as in the previous section to compute the expected duration of the trial, which is shown in Table V. Recall that the actual end of the trial was 2.31 years. We also include the error induced by parameters estimation (that is, the $\approx$ 95% confidence interval). We refer to Section 2.4.3 for details on the calculation of this confidence interval.

The prediction is quite accurate at $t_1 = 1$ year and very accurate at $t_1 = 1.5$ years (only 1% error). Note that the error made in estimated parameters has only a limited impact on the error made in the prediction: indeed, even at $t_1 = 1$ year, the confidence interval of $\mathbb{E}[T_N]$ has a relative width of only 1%.

3.4. Conclusions

What appears clearly on this particular study is the good capacity of these models to capture the behaviour of the enrolment process. The $\mathcal{U}\Gamma$-Poisson model is satisfying because the parameters are not very variable from one estimation to another, it fits the data, and the predicted duration is close to the real one even at early ongoing studies. For these reasons, we suggest to use this model in Section 4. Figure 6 illustrates this. Note that at $t_1 = 1$ year, real data show that the prediction has underestimated the rate of enrolment: the curve of real data (dots) is above the upper bound of the confidence interval (Figure 6). This can be explained by an unusual high rate of enrolment at times 1.2–1.3 years, which was not due to an opening of centres and was therefore not statistically predictable.

4. A methodology for routine procedure

In this section, we propose a method for applying the tools developed properly in the following text in a routine procedure. It is important to keep in mind that it is a statistical procedure, and thus the data have to be collected accurately. Moreover, we have to take the closure of the centres (for holidays for instance) into consideration. These periods must be put aside, and we have to consider only worked days to model the rates as constants.

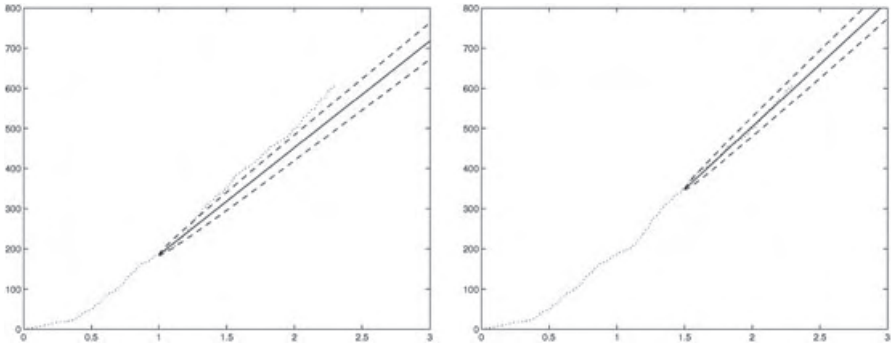

Figure 6. Predicted enrolment behaviour at time 1 (left) and at time 1.5 (right) years.

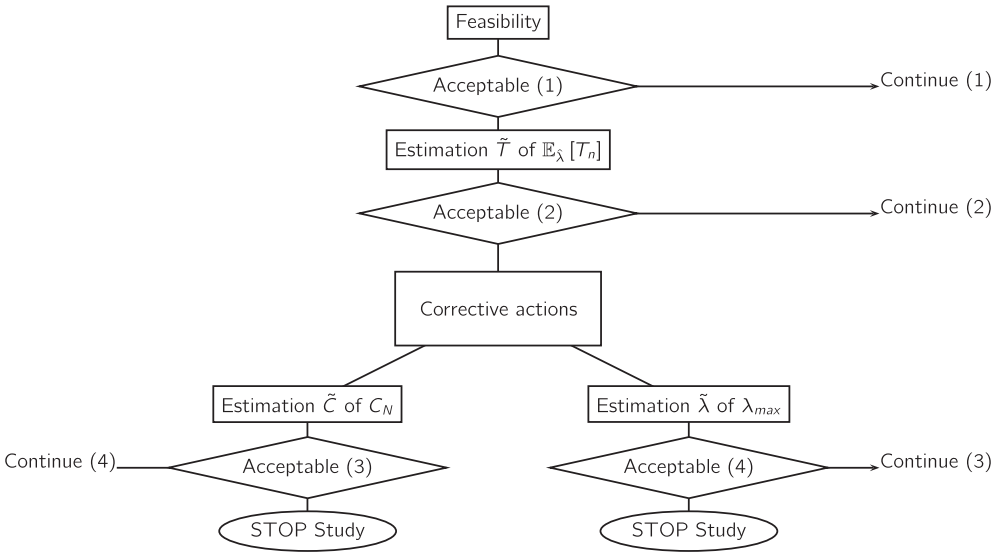

Figure 7. Steps for using the statistical tools.

4.1. Step 1: choice of the model

The method is split in two cases: C small ($C < 20$) and C large. If C is small, we have to consider all the rates for the C centres, and if C is large, we use a Bayesian approach. The conclusion of this paper is that the more relevant is the Γ-Poisson model if the date of centres opening are known; if not, we recommend to use a $\mathcal{U}\Gamma$-Poisson model. Nevertheless, in some setting, the Π-Poisson model fits better the real data in the occupancy curves and can be used, but the expectation of the duration is not very different of those given by the Γ-Poisson model.

4.2. Step 2: using the model at time $T = 0$

If the parameters are known or given by the investigator, then we check the feasibility of the trials. In fact, results of Section 1.3.1 tell you if the parameters of the trial and the parameters of the model are coherent. If they are not, one has to deal with the investigator to make this possible. We propose to follow the steps described on Figure 7. In this figure, a horizontal arrow means YES and a vertical one means NO.

What do we mean by acceptable?

(1) The probability of ending the study on time is high—say, higher than 80 %.
(2) It depends on the sponsor (time is money).
(3) It depends on the sponsor (centres are money).
(4) It depends on the different centres.

The conditions for continuing are as follows:

(1) No condition; the study will almost surely finish before the deadline.
(2) The study has to be prolonged of $\tilde{T} - T_R$ days.
(3) The different centres have to increase their rate of inclusion up to $\tilde{\lambda}$. This is obviously difficult in practice but indicates the effort we need to provide.
(4) $\tilde{C} - C$ centres have to be opened to end on time.

Remark 12
In many cases, we can mix both corrective actions: open $C < \tilde{C}$ centres and estimate the $\lambda_{\max}$ necessary to reach the deadline.

4.3. Step 3: dynamic using of the model

Now, consider an ongoing study at time t_1. First, we have to choose t_1. We want to have enough information for the parameters' estimation to be relevant, but we also want to do the ongoing study not too close to the real end date of the study. One considers that $t_1 = T_R/3$ is a good compromise. With the data collected at t_1, we estimate the parameters of the model and use the methodology described by Figure 7 to take a decision on the study. Another ongoing study can be made at time $2T_R/3$: the estimations would be better and the estimation of the end of the trial more relevant and more useful to organise the end of the trial.

Remark 13
If the parameters are given by the investigator, these can be compared with the estimated ones by the ongoing study and it is possible to take a decision between keeping the known parameters or using the estimated ones.

4.4. Application to the case study

Let us recall the setting of the case study. We plan to include 610 patients in 3 years. Seventy-seven centres are devoted to this trial. In the protocol of this trial, we would plan an ongoing study at the end of the first year. The opening dates of the centres are not known, so we use a $\mathcal{U}\Gamma$-Poisson model.

At the end of the first year, we are able to say that the probability of ending the study on time is very close to 100%. The expected time of the end of the enrolment is 2.60 and, with a probability of 80%, lower than 2.72.

Another ongoing study will be planned at 1.5 years. We are able to say that the probability of ending the study on time is very close to 100%. The expected time at the end of the enrolment is 2.34 and, with a probability of 80%, lower than 2.40.

In Figure 6, the dots are the dates of the inclusion and the line is the 80% confidence interval for the behaviour of the enrolment. If you are over the line, it is alright; if you are under the line, you have to verify the behaviour of the trial by another ongoing study. This is a useful figure for trialists to monitor the clinical trial.

Finally, in reality, the study ended in 2.31 years; this means that 10 months before the end, the model has predicted the end of the trial with an error of 15 days.

5. Conclusions

We have investigated models based on the Poisson process. Each of them gives very good results, and whatever the model you choose, the expected duration is not very different. In some setting, the Π-Poisson model fits better the real data in the occupancy curves, which is a good indicator to choose the model. Nevertheless, we recommend the use of the Γ-Poisson, which is easier to handle. Moreover, the use of a uniform distribution is a good tool when the opening dates of the centres are not known precisely. The sensitivity analysis that we have investigated is a good indicator of the performances of the model.

APPENDIX A. Proofs of the results

In the proofs, $x \propto y$ denotes two quantities equal up to a multiplicative constant.

A.1. Proof of Proposition 5

Proof
The conditional probability of λ when k events occurred between 0 and T is

$$\begin{aligned}\mathbb{P}[\lambda = x|k] &= \frac{\mathbb{P}[k|\lambda = x]\mathbb{P}[\lambda = x]}{\mathbb{P}[k]},\\ &\propto e^{-xT}x^k e^{-\beta x}x^{\alpha-1},\\ &\propto e^{-x(T+\beta)}x^{\alpha+k-1},\end{aligned}$$

and one recognizes the density of a $\Gamma(\alpha + k, \beta + T)$ distribution. In the same manner, in the Pareto setting,

$$\mathbb{P}[\lambda = x|k] \propto e^{-xT}x^k x^{-\gamma-1}\mathbf{1}_{x\geqslant\delta}.$$

□

A.2. Proofs of the expression of $\ln L$

Proof of Proposition 7
In the Bayesian setting, (5) leads to

$$\mathbb{E}_\lambda\left[\lambda^K e^{T(1-\lambda)}\right] = \int_{\mathbb{R}} \mathbb{E}_\lambda\left[\lambda^K e^{T(1-\lambda)} \mid \lambda = x\right] f_\lambda(x)\mathrm{d}x, \tag{8}$$

and (8) leads to

$$\begin{aligned}L &= \mathbb{E}_\lambda\left[\prod_{i=1}^{C}\lambda_i^{k_i}e^{T(1-\lambda_i)}\right],\\ &= e^{CT}\delta^{C\gamma}\gamma^C\prod_{i=1}^{C}\int_\delta^\infty x^{k_i}e^{-Tx}x^{-1-\gamma}\mathrm{d}x,\\ &= e^{CT}\delta^{C\gamma}\gamma^C\prod_{i=1}^{C}\frac{\Gamma_{\text{inc}}(k_i-\gamma,\delta T)}{T^{k_i-\gamma}},\end{aligned}$$

and thus

$$\ln L = CT + C\gamma\ln\delta + C\ln\gamma - (K - C\gamma)\ln T + \sum_{i=1}^{C}\ln\Gamma_{\text{inc}}(k_i-\gamma,\delta T),$$

where, for any $a > 0, b > 0, t > 0$,

$$\Gamma_{\text{inc}}(b, at) = a^b\int_t^\infty e^{-ax}x^{b-1}\mathrm{d}x.$$

□

A.3. Proofs of Propositions 8 and 9

These propositions come from

$$(I_i)_{rs}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 f(\theta;k_i)}{\partial\theta_r\partial\theta_s}\right],$$

where $r, s \in \{1, 2\}$, $\theta_1 = \alpha$ and $\theta_2 = \beta$ are straightforward calculations.

A.4. Proof of Proposition 10

Proof
For the $\Gamma(\alpha,\beta)$-Poisson model, recall that the overall rate is a sum of independent Gamma distributions $\Lambda \approx \sum_{i=1}^{C} \Gamma(\alpha+k_i, \beta+\tau_i)$. As Anisimov and Fedorov [3] pointed out, one can approximate this distribution by a $\Gamma(A,B)$ distribution, where $A=\frac{m^2}{v}$, $B=\frac{m}{v}$ with $m=\sum_{i=1}^{C}\frac{\alpha+k_i}{\beta+\tau_i}$ and $v=\sum_{i=1}^{C}\frac{\alpha+k_i}{(\beta+\tau_i)^2}$. The fact that the forward distribution of λ_i is a $\Gamma(\alpha+k_i,\beta+\tau_i)$ leads directly to the values m and v of the mean and variance of the overall rate of inclusion. The prediction of the end time of the trial reads

$$\mathbb{E}\left[T_N\right] = N\frac{m}{m^2-v}.$$

Matching the first two moments of a Γ distribution gives the corresponding values of A and B. The rest of the proposition is straightforward. □

A.5. Proof of Propositions 11 and 12

Proof of Proposition 11
Let us calculate

$$\begin{aligned}
\mathbb{E}\left[\lambda_i \,|\, k_i \text{ inclusions}\right] &= \mathbb{E}\left[\mathbb{E}\left(\lambda_i \,|\, k_i \text{ inclusions}, \tau_i\right) \,|\, k_i \text{ inclusions}\right], \\
&= \mathbb{E}\left[\frac{\alpha+k_i}{\beta+\tau_i} \,|\, k_i \text{ inclusions}\right], \\
&= (\alpha+k_i)\frac{1}{S_i-S_i'}\int_{S_i'}^{S_i}\frac{\mathrm{d}t}{\beta+t}, \\
&= \frac{\alpha+k_i}{S_i-S_i'}\ln\left(\frac{\beta+S_i}{\beta+S_i'}\right),
\end{aligned}$$

which gives the value of $m=\mathbb{E}\left[\Lambda \,|\, k_1,\ldots,k_C\right]$, the mean of the (forward) overall rate of inclusion. A similar calculus leads to the value of its variance v. The end of the proof is straightforward recalling that if $X \overset{d}{=} \Gamma(A,B)$, then $\mathbb{E}[X]=\frac{A}{B}$ and $\mathrm{var}[X]=\frac{A}{B^2}$. □

Proof of Proposition 12
We have,

$$\begin{aligned}
\mathbb{P}\left[k_i \text{ inclusions in centre } i\right] &\propto \mathbb{E}\left[e^{-\lambda\tau_i}\lambda^k\tau_i^k\right], \\
&\propto \beta^\alpha\Gamma(\alpha)^{-1}\Gamma(k_i+\alpha)\mathbb{E}\left[(\tau_i+\beta)^{-k_i-\alpha}\tau_i^{k_i}\right], \\
&\propto \beta^\alpha\Gamma(\alpha)^{-1}\Gamma(k_i+\alpha)\frac{1}{S_i-S_i'}\int_{S_i'}^{S_i}(t+\beta)^{-k_i-\alpha}t^{k_i}\,\mathrm{d}t, \\
&\propto \beta^\alpha\Gamma(\alpha)^{-1}\Gamma(k_i+\alpha)\frac{1}{S_i-S_i'}\int_{S_i'+\beta}^{\beta+S_i}t^{-k_i-\alpha}(t-\beta)^{k_i}\,\mathrm{d}t, \\
&\propto \beta^\alpha\Gamma(\alpha)^{-1}\Gamma(k_i+\alpha)\frac{1}{S_i-S_i'}J(\alpha,\beta,k_i,S_i',S_i),
\end{aligned}$$

and the developed form of J is obtained by developing the term $(t-\beta)^{k_i}$. □

Acknowledgements

The authors would like to thank the editors and the reviewers of this paper for their valuable comments especially Prof. Vladimir Anisimov, the very detailed report of the first version of the manuscript and the fruitful discussions have considerably improved our paper. The authors are also grateful to Toulouse INSERM Unit 1027 especially Prof. Sandrine Andrieu, Prof. Thierry Lang and Dr Valérie Lauwers-Cancès for the interest on this topic, the valuable comments during the investigations and for the careful reading of the manuscript. Finally, the authors thank IFM (Intergroupe Francophone du Myélome) for lending us the data used for evaluating the models in Section 3.3.

References

1. Anisimov VV. Using mixed Poisson models in patient recruit in multicentre clinical trials. *Proceedings of the World Congress on Engineering, Vol. II*, London, United Kingdom, 2008.
2. Anisimov VV, Downing D, Fedorov VV. *Recruitment in Multicentre Trials: Prediction and Adjustment. 8th International Workshop in Model-Oriented Design and Analysis*. Physica-Verlag/Springer, Heidelberg: Almagro, Spain, 2007. 1–8.
3. Anisimov VV, Fedorov VV. Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in Medicine* 2007; **26**(27):4958–4975. DOI: 10.1002/sim.2956.
4. Anisimov VV. Predictive modelling of recruitment and drug supply in multicenter clinical trials. *Proceedings of the Joint Statistical Meeting, ASA*, Washington, USA, 2009; 1248–1259.
5. Anisimov VV. Recruitment modeling and predicting in clinical trials. *Pharmaceutical Outsourcing* 2009; **10**(1):44–48.
6. Anisimov VV, Fedorov VV. Modeling of enrolment and estimation of parameters in multicentre trials. *Technical Report*, GSK BDS Technical Report, 2005.
7. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux P, Elbourne D, Egger M, Altman DG. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology* August 2010; **63**(8):e1–e37. DOI: 10.1016/j.jclinepi.2010.03.004. In Press, Corrected Proof.
8. Schulz KF, Altman DG, Moher D. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology* August 2010; **63**(8):834-840. DOI: 10.1016/j.jclinepi.2010.02.005. In Press, Corrected Proof.
9. Vlasto AP. Brevets et médicament en france. pourquoi l'application du droit des brevets au médicament est-elle autant critiquée ? *Médecine & Droit* 2007; **2007**(82):25–32. DOI: 10.1016/j.meddro.2006.12.001.
10. Rojavin M. Patient recruitment and retention: from art to science. *Contemporary Clinical Trials* 2009; **30**(5):387–387. DOI: 10.1016/j.cct.2009.06.002.
11. Barnard KD, Dent L, Cook A. A systematic review of models to predict recruitment to multicentre trials. *BMC Medical Research Methodology* 2010; **63**(10). DOI: 10.1186/1471-2288-10-63.
12. Morgan TM. Nonparametric estimation of duration of accrual and total study length for clinical trials. *Biometrics* 1987; **43**(4):903–912. DOI: 10.2307/2531544.
13. Lee YJ. Interim recruitment goals in clinical trials. *Journal of Chronic Diseases* 1983; **36**(5):379–389. DOI: 10.1016/0021-9681(83)90170-4.
14. Williford W, Bingham S, Weiss D, Collins J, Rains K, Krol W. The "constant intake rate" assumption in interim recruitment goal methodology for multicenter clinical trials. *Journal of chronic diseases* 1987; **40**(4):297–307. DOI: 10.1016/0021-9681(87)90045-2.
15. Senn S. *Statistical Issues in Drug Development*. John Wiley & Sons: Chichester, 1997. doi:10.1002/9780470723586.
16. Yagouti A, Abi-Zeid I, Ouarta TBMJ, Bobée B. Revue de processus ponctuels et synthse de tests statistiques pour le choix d'un type de processus (review and classification of statistical tests applied to point processes). *Revue des sciences de l'eau* 2001; **14**(3):323–361.
17. Brémaud P. *Point Processes and Queues*. Springer-Verlag: New York, 1981. doi:10.1002/bimj.4710300220. Martingale dynamics, Springer Series in Statistics.
18. Senn S. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine* 1998; **17**:1753–1765. DOI: 10.1002/(SICI)1097-0258(19980815/30)17:15/16h1753::AID-SIM977i3.0.CO;2-X.
19. Piantadosi S, Patterson B. A method for predicting accrual, cost, and paper flow in clinical trials. *Controlled Clinical Trials* 1987; **8**(3):202–215. DOI: 10.1016/0197-2456(87)90045-6.
20. Carter RE. Application of stochastic processes to participant recruitment in clinical trials. *Controlled Clinical Trials* 2004; **25**(5):429–436. DOI: 10.1016/j.cct.2004.07.002.
21. Carter RE, Sonne SC, Brady KT. Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC Medical Research Methodology* 2005; **5**(11):1–5. DOI: 10.1186/1471-2288-5-11.
22. Lai D, Moyé LA, Davis BR, Brown LE, Sacks FM. Brownian motion and long-term clinical trial recruitment. *Journal of Statistical Planning and Inference* 2001; **93**(1–2):239–246. DOI: 10.1016/S0378-3758(00)00203-2.
23. Lai D, Barry RD, Robert JH. Fractional Brownian motion and clinical trials. *Journal of Applied Statistics* 2000; **27**(1):103–108. DOI: 10.1080/02664760021853.
24. Lai D. Estimating the Hurst effect and its application in monitoring clinical trials. *Computational Statistics and Data Analysis* 2004; **45**(3):549–562. DOI: 10.1016/S0167-9473(03)00085-9.
25. Abbas I, Rovira J, Casanovas J. Clinical trial optimization: Monte Carlo simulation Markov model for planning clinical trials recruitment. *Contemporary clinical trials* 2007; **28**:220–231. DOI: 10.1016/j.cct.2006.08.002.
26. Rojavin MA. Recruitment index as a measure of patient recruitment activity in clinical trials. *Contemporary Clinical Trials* 2005; **26**(5):552–556. DOI: 10.1016/j.cct.2005.05.001.
27. Haug EG. *The Complete Guide to Option Pricing Formulas*. McGraw-Hill: New-York, 1997. doi:10.1036/0786312408.

Nous étudions une approche bayésienne pour la modélisation de l'enrôlement de patients dans un essai clinique multicentrique. Les premières investigations en ce sens ont été menées par Anisimov et al.[11], qui proposent un modèle doublement stochastique dans lequel les intensités d'inclusion sont distribuées selon une loi Gamma. Le modèle ainsi défini porte le nom de modèle Gamma-Poisson (ou Γ-Poisson).

Après avoir décrit le modèle d'Anisimov, nous proposerons une extension où les intensités suivent une loi de Pareto, qui rend bien compte du fait que 80% des patients sont inclus par 20% des centres. Puis, nous nous intéresserons au cas où les dates d'ouverture des centres sont inconnues.

L'objet de ce chapitre est, d'une part, la description des différents modèles ainsi définis, et d'autre part le problème de l'estimation des paramètres des modèles à partir des données recueillies à un instant intermédiaire.

Remarque 1.0.1. *Par abus de langage, nous appelons méthodes bayésiennes ce qui sont en réalité des* **méthodes bayésiennes empiriques**. *En effet, les approches bayésiennes proposées supposent que les intensités de recrutement $\lambda_1, \dots \lambda_C$ des centres sont aléatoires, indépendantes et distribuées suivant une loi a priori p_θ appartenant à une famille paramétrique $\{p_\theta \ ; \ \theta \in \Theta\}$ où $\Theta \in \mathbb{R}^d$. Le paramètre θ sera estimé via la méthode du maximum de vraisemblance. Puis, la prédiction utilisera les lois a posteriori des variables $\lambda_1, \dots, \lambda_C$, où θ sera remplacé par son estimateur $\widehat{\theta}$. Nous sommes donc bien dans le cadre de méthodes bayésiennes empiriques, et non strictes. Une démarche strictement bayésienne aurait supposé une loi a priori sur le paramètre θ, ce que nous ne faisons pas.*

1.1 Dates d'ouverture des centres connues

Dans cette section, les dates d'ouverture des C centres $\{u_i \ ; \ 1 \leq i \leq C\}$ sont supposées connues.

Le processus d'inclusion du centre i est un processus de Poisson doublement stochastique (ou processus de Cox), dont l'intensité λ_i suit une certaine loi de densité p_θ où $\theta \in \mathbb{R}^d$ est un paramètre. Les processus des différents centres sont supposés indépendants.

Si p_θ est la densité d'une loi Gamma de paramètres (α, β), nous sommes dans le cadre du modèle Gamma-Poisson (ou Γ-Poisson).

Nous étudierons également le cas où p_θ est la densité d'une loi de Pareto. En effet, cette loi modélise bien le fait, observé en pratique, que 20% des centres recrutent 80% des patients. On dénomme ce modèle par Pareto-Poisson (ou Π-Poisson).

L'objet de cette partie est de décrire ces deux modèles, et de montrer comment estimer les paramètres par maximum de vraisemblance en utilisant les données à un certain instant intermédiaire $t_1 > 0$.

1.1.1 Préliminaires

Nous supposons d'abord que la densité p_θ est quelconque. Dans la suite, si $a, b \in \mathbb{R}$, on notera $a \vee b = \max(a, b)$.

Notations 1.1.1. *Nous noterons $(N_i^R(t))_{t \geq 0}$ le processus d'inclusion du centre i, et $N^R = \sum_{i=1}^C N_i^R$ le processus global d'inclusion.*

Soit $(\lambda_1, \ldots, \lambda_C)$ un échantillon i.i.d. d'une loi p_θ sur $\mathbb{R}_+$. Sachant λ_i, le processus d'inclusion N_i^R du centre i est un processus de Poisson d'intensité $t \mapsto \lambda_i \mathbf{1}_{t \geq u_i}$.

Proposition 1.1.2. *Soit $t_1 \geq 0$. Posons $\tau_i = (t_1 - u_i) \vee 0,\ 1 \leq i \leq C$. Alors*

$$\mathbb{P}\left[N_1^R(t_1) = n_1, \ldots, N_C^R(t_1) = n_C\right] = \prod_{i=1}^{C} \frac{\tau_i^{n_i}}{n_i!} \int_0^{+\infty} e^{-x\tau_i} x^{n_i} p_\theta(x) dx.$$

Démonstration. Sachant λ_i, $N_i^R(t_1)$ suit une loi de Poisson de paramètre $\lambda_i \tau_i$, d'où

$$\mathbb{P}\left[N_i^R(t_1) = n_i\right] = \mathbb{E}\left[\exp(-\lambda_i \tau_i) \frac{(\lambda_i \tau_i)^{n_i}}{n_i!}\right] = \frac{\tau_i^{n_i}}{n_i!} \int_0^{+\infty} e^{-x\tau_i} x^{n_i} p_\theta(x) dx,$$

et le résultat suit par indépendance des variables $N_1^R(t_1), \ldots, N_C^R(t_1)$. □

Dans toute la suite de ce chapitre, nous fixons $t_1 > 0$ un instant intermédiaire. Pour alléger les notations, nous noterons $k_i = N_i^R(t_1)$ le nombre d'inclusions par le centre i jusqu'à t_1, et $\tau_i = (t_1 - u_i) \vee 0$.

1.1.2 Modèle Gamma-Poisson

D'après la proposition 1.1.2, appliquée à distribution Γ de paramètres (α, β) dont la densité $p_{(\alpha,\beta)}$ est donnée par

$$p_{(\alpha,\beta)}(x) = \mathbf{1}_{x>0} \beta^\alpha \Gamma(\alpha)^{-1} e^{-\beta x} x^{\alpha-1}, \tag{1.1}$$

la loi du vecteur $(k_1 = N_1^R(t_1), \ldots, k_C = N_C^R(t_1))$ est [11]

$$\begin{aligned}\mathbb{P}\left[k_1 = n_1, \ldots, k_C = n_C\right] &= \prod_{i=1}^{C} \frac{\tau_i^{n_i}}{n_i!} \int_0^{+\infty} e^{-x\tau_i} x^{n_i} p_{\alpha,\beta}(x) dx \\ &= \prod_{i=1}^{C} \frac{\Gamma(\alpha + n_i)}{n_i! \Gamma(\alpha)} \frac{\beta^\alpha \tau_i^{n_i}}{(\beta + \tau_i)^{\alpha+n_i}},\end{aligned}$$

où $\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt, x > 0$.

Dans la suite, on pose

$$\mu = \alpha/\beta.$$

A la date intermédiaire t_1, on estime les paramètres (α, μ). La méthode utilisée est celle du maximum de vraisemblance.

Dans le cadre du modèle Gamma-Poisson, travailler avec les paramètres (α, μ) permet d'obtenir une matrice d'information de Fisher diagonale (voir partie 2.3). D'où la proposition suivante :

Proposition 1.1.3 ([9], page 4968)**.** *L'estimateur du maximum de vraisemblance $(\hat{\alpha}_C, \hat{\mu}_C)$ est obtenu en maximisant :*

$$M_C^\Gamma(\alpha, \mu) = \alpha \ln(\alpha/\mu) - \ln \Gamma(\alpha) + \frac{1}{C} \sum_{i=1}^{C} \left[\ln \Gamma(\alpha + k_i) - (\alpha + k_i) \ln(\alpha/\mu + \tau_i)\right].$$

On peut obtenir une relation simple entre $\hat{\alpha}_C$ et $\hat{\mu}_C$.

Proposition 1.1.4. *Soit* $(\hat{\alpha}_C, \hat{\mu}_C) = \mathrm{argmax}_{(\alpha,\mu)\in(\mathbb{R}_+^*)^2} M_C^\Gamma(\alpha,\mu)$. *On a alors la relation :*

$$\hat{\mu}_C = \left(\sum_{i=1}^{C} \frac{k_i}{\hat{\alpha}_C + \hat{\mu}_C \tau_i}\right) \times \left(\sum_{i=1}^{C} \frac{\tau_i}{\hat{\alpha}_C + \hat{\mu}_C \tau_i}\right)^{-1}$$

Démonstration. Le maximum de M_C^Γ est atteint en un point où le gradient s'annule. La proposition découle de la condition $\partial_\mu M_C^\Gamma(\hat{\alpha}_C, \hat{\mu}_C) = 0$. □

Intéressons-nous aux propriétés asymptotiques de $(\hat{\alpha}_C, \hat{\mu}_C)$. A une constante additive près, la log-vraisemblance s'écrit

$$M_C^\Gamma = \frac{1}{C}\sum_{i=1}^{C} f^\Gamma(\theta, k_i, \tau_i),$$

où

$$\theta \equiv (\alpha, \mu),$$

et

$$f^\Gamma(\theta, k_i, \tau_i) = \alpha \ln(\alpha/\mu) - \ln\Gamma(\alpha) + \ln\Gamma(\alpha + k_i) - (\alpha + k_i)\ln(\alpha/\mu + \tau_i).$$

La i-ème matrice d'information de Fisher vaut $J_i(\theta, \tau_i) = -\mathbb{E}_\theta\left[\nabla^2 f^\Gamma(\theta, k_i, \tau_i)\right]$, et la matrice d'information globale

$$I_C(\theta) = \frac{1}{C}\sum_{i=1}^{C} J_i(\theta, \tau_i).$$

Soit $\theta_0 \equiv (\alpha_0, \mu_0)$ la vraie valeur du paramètre θ. Si $\tau_i \geq \xi$ pour tout i et un certain $\xi > 0$, et si la matrice limite $I(\theta_0) = \lim_{C\to\infty} I_C(\theta_0)$ existe et est définie positive, alors l'estimateur du maximum de vraisemblance est consistant [48], et $\hat{\theta}_C = (\hat{\alpha}_C, \hat{\mu}_C)$ possède une distribution asymptotique normale de moyenne (α_0, μ_0) et de matrice de covariance $\frac{1}{C} I(\theta_0)^{-1}$.

Proposition 1.1.5. *Dans le modèle* Γ*-Poisson, la matrice* 2×2 $J_i(\theta, \tau_i)$ *est*

$$\begin{aligned}
(J_i)_{11} &= -\frac{\mu\tau_i}{\alpha(\alpha + \mu\tau_i)} + \mathbb{E}_\theta\left[\psi^{(1)}(\alpha) - \psi^{(1)}(\alpha + k_i)\right], \\
(J_i)_{12} &= (J_i)_{21} = 0, \\
(J_i)_{22} &= \frac{\alpha\tau_i}{\mu(\alpha + \mu\tau_i)},
\end{aligned}$$

où $\psi^{(1)}$ *est la fonction trigamma :* $\psi^{(1)}(x) = \frac{d^2}{dx^2}\ln\Gamma(x)$.

Remarque 1.1.6. *Une condition suffisante pour que* $\lim_{C\to\infty} I_C(\theta_0)$ *existe et soit définie positive est donc que la suite* $(\tau_i)_{i\geq 1}$ *soit uniformément distribuée dans un intervalle* $[\tau_m, \tau_M]$ *où* $\tau_m \geq 0$ *et* $\tau_M > 0$. *En effet, dans ce cas, on a, pour toute fonction Riemann-intégrable* g, $\frac{1}{C}\sum_{i=1}^{C} g(\tau_i) \underset{C\to+\infty}{\to} \int_{\tau_m}^{\tau_M} g(\tau)d\tau$. *Notons* $J(\tau)$ *la matrice définie dans la proposition 1.1.5, où l'on a substitué* τ_i *par* τ. *Alors :*

$$I(\theta_0) = \lim_{C\to\infty} I_C(\theta_0) = \int_{\tau_m}^{\tau_M} J(\tau)d\tau,$$

et comme pour tout $\tau > 0, J(\tau)$ est définie positive et que $\tau \mapsto J(\tau)$ est continue, alors $I(\theta_0)$ est définie positive.

Un cas particulier est lorsque pour tout $i \geq 1$, $\tau_i = \tau > 0$, ou bien lorsque $\tau_i \underset{i \to +\infty}{\longrightarrow} \tau > 0$.

Remarque 1.1.7. *Anisimov et al. ont étudié dans [10] d'autres méthodes d'estimation : la méthode des moindres carrés et la méthode des moments. Ils montrent que ces deux méthodes donnent des résultats similaires à celle du maximum de vraisemblance.*

La variance asymptotique de l'estimateur se calcule via la vraisemblance, c'est pourquoi nous préférons utiliser la méthode du maximum de vraisemblance aux deux autres.

Néanmoins, dans le cas particulier où le nombre de centres n'ayant pas recruté n'est pas disponible, il apparait que l'estimateur le moins biaisé est celui des moindres carrés [10].

1.1.3 Modèle Pareto-Poisson

Nous supposons maintenant que les taux d'inclusion $(\lambda_i)_{1 \leq i \leq C}$ sont un échantillon i.i.d. d'une loi de Pareto de paramètres γ et δ, dont la densité est

$$x \mapsto p_{(\gamma,\delta)}(x) = \gamma \frac{\delta^\gamma}{x^{\gamma+1}} \, \mathbf{1}_{\{x \geq \delta\}}.$$

Les notations restent les mêmes que dans la section précédente : on note toujours $k_i = N_i^R(t_1)$ le nombre d'inclusions par le centre i jusqu'à t_1, et $\tau_i = (t_1 - u_i) \vee 0$. D'après la proposition 1.1.2, appliquée à la densité $p_\theta := p_{(\gamma,\delta)}$, la loi du vecteur $(k_1, \ldots, k_C)$ est

$$\begin{aligned}
\mathbb{P}\left[k_1 = n_1, \ldots, k_C = n_C\right] &= \prod_{i=1}^{C} \delta^\gamma \gamma \frac{\tau_i^\gamma}{n_i!} \Gamma_{inc}(n_i - \gamma, \delta\tau_i) \\
&= \delta^{C\gamma} \gamma^C \prod_{i=1}^{C} \frac{\tau_i^\gamma}{n_i!} \Gamma_{inc}(n_i - \gamma, \delta\tau_i),
\end{aligned}$$

où $\Gamma_{inc}(x, y) = \int_y^{+\infty} e^{-t} t^{x-1} dt, x \in \mathbb{R}, y > 0$. On en déduit la proposition suivante :

Proposition 1.1.8 ([35])**.** *L'estimateur du maximum de vraisemblance $(\hat{\gamma}_C, \hat{\delta}_C)$ est obtenu en maximisant :*

$$M_C^\Pi(\gamma, \delta) = \gamma \ln \delta + \ln \gamma + \frac{1}{C} \sum_{i=1}^{C} \left[\ln \Gamma_{inc}(k_i - \gamma, \delta\tau_i) + \gamma \ln \tau_i\right].$$

Intéressons-nous aux propriétés asymptotiques de $(\hat{\gamma}_C, \hat{\delta}_C)$. Notons $\theta \equiv (\gamma, \delta)$. La log-vraisemblance (à une constante additive près) est de la forme

$$M_C^\Pi = \frac{1}{C} \sum_{i=1}^{C} f^\Pi(\theta, k_i, \tau_i),$$

où

$$f^\Pi(\theta, k_i, \tau_i) = \gamma \ln \delta + \ln \gamma + \ln \Gamma_{inc}(k_i - \gamma, \delta\tau_i) + \gamma \ln \tau_i.$$

La i-ème matrice d'information de Fisher vaut $J_i(\theta, k_i, \tau_i) = -\mathbb{E}_\theta\left[\nabla^2 f^{\Pi}(\theta, k_i, \tau_i)\right]$, et la matrice d'information globale

$$I_C(\theta) = \frac{1}{C}\sum_{i=1}^{C} J_i(\theta, \tau_i).$$

Soit $\theta_0 \equiv (\gamma_0, \delta_0)$ la vraie valeur du paramètre θ. De même qu'à la section précédente, si $\tau_i \geq \xi$ pour tout i et un certain $\xi > 0$, et si la matrice limite $I(\theta_0) = \lim_{C\to\infty} I_C(\theta_0)$ existe et est définie positive, alors l'estimateur du maximum de vraisemblance est consistant [48], et $\hat{\theta}_C = (\hat{\gamma}_C, \hat{\delta}_C)$ possède une distribution asymptotique normale de moyenne (γ_0, δ_0) et de matrice de covariance $\frac{1}{C} I(\theta_0)^{-1}$. La proposition suivante donne la matrice J_i :

Proposition 1.1.9. *Dans le modèle* Π*-Poisson, la matrice* J_i *est donnée par*

$$\begin{aligned}
(J_i)_{11} &= -\frac{1}{\gamma^2} - \mathbb{E}_{(\gamma,\delta)}\left[\frac{\partial^2}{\partial\gamma^2}\Gamma_{inc}(k_i - \gamma, \delta\tau_i)\right], \\
(J_i)_{12} = (J_i)_{21} &= -\frac{1}{\delta} + \tau_i e^{-\delta\tau_i}\ln(\delta\tau_i)\mathbb{E}_{(\gamma,\delta)}\left[(\delta\tau_i)^{k_i-\gamma-1}\right], \\
(J_i)_{22} &= \frac{\gamma}{\delta^2} + \tau_i^2 e^{-\delta\tau_i}\mathbb{E}_{(\gamma,\delta)}\left[(\delta\tau_i)^{k_i-\gamma-2}(\delta\tau_i + k_i - \gamma - 1)\right].
\end{aligned}$$

Remarque 1.1.10. *Les espérances de la proposition précédente sont calculées par simulations de Monte-Carlo.*

Remarque 1.1.11. *Une condition suffisante pour que* $\lim_{C\to\infty} I_C(\theta_0)$ *existe et soit définie positive est donc que la suite* $(\tau_i)_{i\geq 1}$ *soit uniformément distribuée dans un intervalle* $[\tau_m, \tau_M]$ *où* $\tau_m \geq 0$ *et* $\tau_M > 0$*. Un cas particulier est lorsque pour tout* $i \geq 1$, $\tau_i = \tau > 0$ *ou bien lorsque* $\tau_i \underset{i\to+\infty}{\to} \tau > 0$.

1.1.4 Application aux données

Dans la table 1.1 sont données les estimations des paramètres, à l'instant final de l'essai, pour les quatre jeux de données tirés de [10], qui sont rappelés en annexe.

Modèle	$\widehat{\alpha}$	$\widehat{\mu}$
Γ-Poisson (A)	2.89	0.42
Γ-Poisson (B)	2.89	0.26
Γ-Poisson (C)	4.59	0.31
Γ-Poisson (D)	3.46	0.43

Modèle	$\widehat{\gamma}$	$\widehat{\delta}$
Π-Poisson (A)	1.74	3.38
Π-Poisson (B)	1.58	5.07
Π-Poisson (C)	2.04	8.34
Π-Poisson (D)	2.05	4.42

TABLE 1.1 – Estimateurs du maximum de vraisemblance pour les modèles Gamma-Poisson et Pareto-Poisson.

1.1.5 Validation des modèles

Courbes d'occupation

La validation des modèles se fait *via* les courbes d'occupation, qui est la méthode utilisée dans [10].

On se place à l'instant final T_f, et on note $\ell_i = N_i^R(T_f)$. Définissons pour $j \in \mathbb{N}$,

$$\nu_j = \text{Card}\{i \,:\, \ell_i = j\} = \sum_{i=1}^{C} \mathbf{1}_{\ell_i = j}. \tag{1.2}$$

ν_j est le nombre de centres ayant recruté exactement j patients.

Définition 1.1.12. *On appelle courbe d'occupation le graphe* $\{(j, \nu_j)\,;\; j \in \mathbb{N}\}$.

Remarque 1.1.13. *Dans le cas où pour tout* $1 \leq i \leq C$, $\tau = \tau_i$, *alors les variables aléatoires* $(\ell_i)_{1 \leq i \leq C}$ *ont toutes même loi. La fonction* $j \mapsto \nu_j$ *représente donc la distribution empirique de* ℓ_i.

Nous comparons graphiquement ν_j avec sa valeur moyenne $\mathbb{E}[\nu_j]$, qui est :

$$\mathbb{E}[\nu_j] = \sum_{i=1}^{C} \mathbb{P}[\ell_i = j].$$

Nous utilisons les données de [10, pages 28-29]. Les résultats sont présentés dans la figure 1.1. Les deux modèles sont en bonne adéquation avec les données. Le modèle Gamma-Poisson parait néanmoins plus proche des données réelles que le modèle Pareto-Poisson.

Test du χ^2

Dans cette partie on suppose en outre que $\tau_i \equiv \tau$, pour tout $1 \leq i \leq C$.

Chaque centre i possède une probabilité $p_j = \mathbb{P}[\ell_i = j]$ de recruter exactement j patients. Comme $\tau_i \equiv \tau$, les p_j ne dépendent pas de i. Ainsi, $(\nu_1, \ldots, \nu_N, \ldots)$ suit une loi multinomiale à C épreuves avec un nombre de classes infini (à la classe i étant associée une probabilité p_i).

On utilise alors un test du χ^2 d'adéquation aux données.

Si le nombre de classe N est fini, de probabilités $(p_1, \ldots, p_N)$, alors la variable

$$\sum_{i=1}^{N} \frac{(\nu_i - Cp_i)^2}{Cp_i}$$

suit approximativement une loi du χ^2 à $N-1$ degrés de liberté [28]. On suppose que l'approximation est valable dès que plus de 80% des classes contiennent, en moyenne, au moins 5 occurrences (i.e. $Cp_i \geq 5$ pour 80% des indices i) [28].

Dans notre cas, il convient donc de regrouper des classes (adjacentes) pour atteindre la condition de 80% des classes contenant au moins 5 éléments.

On souhaite tester l'hypothèse H_0 contre H_1 :

- $H_0 : \tilde{p}_j = \tilde{p}_j^0,\ \forall j \geq 0.$
- $H_1 : \exists j \geq 0$ tel que $\tilde{p}_j \neq \tilde{p}_j^0$,

où $(\tilde{p}_j)_{j \geq 1}$ sont les probabilités des classes regroupées induites par le modèle (Gamma-Poisson ou Pareto-Poisson), et $(\tilde{p}_j^0)_{j \geq 1}$ sont les vraies valeurs des probabilités (pour les classes regroupées).

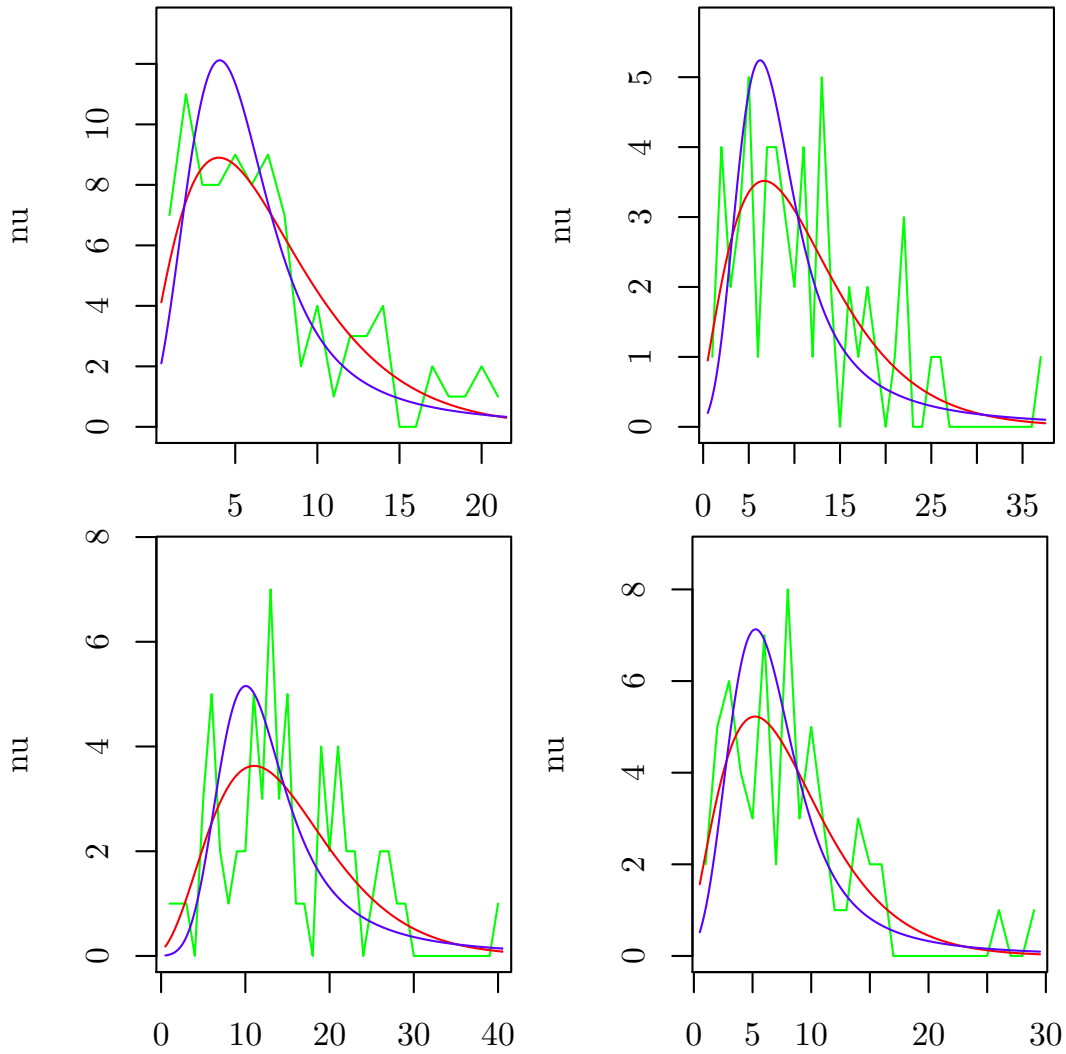

FIGURE 1.1 – Courbes d'occupation. Vert : données réelles, rouge : Gamma-Poisson, bleu : Pareto-Poisson. Haut : études A et B. Bas : études C et D.

Le regroupement est fait de la manière suivante :

Soit n_1 le plus grand entier tel que $\sum_{i=n_1}^{+\infty} Cp_i \geq 5$. On regroupe les classes $n_1, n_1+1, \ldots$ en posant $\tilde{\nu}_{n_1} = \sum_{i=n_1}^{+\infty} \nu_i$ et $\tilde{p}_{n_1} = \sum_{i=n_1}^{+\infty} p_i$. Puis, si l'objectif des 80% n'est pas atteint, on regroupe les classes précédentes $n_1 - 1, n_1 - 2, \ldots, n_2$ où n_2 est le plus grand entier tel que $\sum_{i=n_2}^{n_1-1} Cp_i \geq 5$. On recommence l'opération jusqu'à obtenir 80% de classes ayant plus de 5 occurrences.

Le test est effectué pour les probabilités $(p_j)_{j\geq 1}$ induites par le modèle Γ-Poisson puis par le modèle Π-Poisson. Les résultats du test, pour les quatre études de la partie précédente, sont donnés dans la table 1.2.

Dans le cas du modèle Γ-Poisson, l'hypothèse H_0 est acceptée pour les quatre études, tandis que dans le modèle Π-Poisson, elle est rejetée pour l'étude C. Le modèle Γ-Poisson est donc celui qui est le plus souvent en adéquation avec les données.

Remarque 1.1.14. *Le résultat des tests peut dépendre de la méthodologie employée pour regrouper les classes. Néanmoins, si le regroupement des classes se fait en commençant par la gauche, c'est-à-dire en regroupant d'abord les indices* $1, \ldots, n_1$ *où* n_1 *est choisi pour*

	A	B	C	D
Gamma	8.57	5.91	15.03	3.98
Pareto	17.62	16.99	70.40	10.44

TABLE 1.2 – Statistique du χ^2 d'adéquation des paramètres aux données pour les modèles Gamma-Poisson et Pareto-Poisson, appliqué à quatre jeux de données.

que $C(p_0 + \ldots + p_{n_1}) \geq 5$*, et ainsi de suite jusqu'à atteindre* 80% *des classes* j *vérifiant* $C\tilde{p}_j \geq 5$*, alors les résultats des tests sont les mêmes qu'avec l'autre méthode.*

1.2 Intensités d'inclusion dépendant du temps

Dans certains essais, on constate que l'intensité d'inclusion peut dépendre du temps. En particulier, deux situations peuvent se présenter :

- un centre n'atteint pas son intensité maximale d'inclusion dès l'ouverture : il est courant d'observer une période de "mise en route" de quelques semaines, avant que l'intensité maximale soit atteinte,
- dans certains essais, on constate au contraire que le taux d'inclusion diminue au fil du temps, à cause soit d'un effet de saturation de la capacité d'accueil des centres, soit d'un tarissement de la population cible.

Nous étendons le modèle Γ-Poisson à des intensités dépendant du temps pour prendre en compte ces deux phénomènes.

Les modèles suivants n'ont pas été testés sur données réelles, car ils nécessitent les instants d'arrivée de chaque patient et les dates d'ouverture des centres, et nous n'avons pas de tel jeu de données.

1.2.1 Prise en compte d'un temps de "mise en route"

Notons S la durée de la période de "mise en route" d'un centre. Cette durée S est inconnue mais supposée identique pour tous les centres. L'intensité du centre i peut être représentée par une fonction linéaire par morceaux

$$f_i(t) = \begin{cases} \lambda_i & \text{si } t \geq u_i + S, \\ \frac{\lambda_i}{S}(t - u_i) & \text{si } u_i \leq t \leq u_i + S, \\ 0 & \text{sinon,} \end{cases}$$

où u_i désigne la date d'ouverture du centre i et les intensités $(\lambda_i)_{1 \leq i \leq C}$ sont distribuées selon une loi Gamma de paramètres (α, β).

Plaçons nous à l'instant intermédiaire t_1. Alors

$$\mathbb{P}\left[k_i = n\right] = \mathbb{E}\left[e^{-\int_{u_i}^{t_1} f_i(t)dt} \frac{\left(\int_{u_i}^{t_1} f_i(t)dt\right)^n}{n!}\right],$$

où

$$\int_{u_i}^{t_1} f_i(t)dt = \begin{cases} 0 & \text{si } t_1 \leq u_i \\ \frac{(t_1-u_i)^2}{2S}\lambda_i & \text{si } u_i \leq t_1 \leq u_i + S \\ \lambda_i(t_1 - u_i - S/2) & \text{si } t_1 \geq u_i + S. \end{cases}$$

Proposition 1.2.1. *Soit $\tau_i = t_1 - u_i$. Posons*

$$g_i(S) = \begin{cases} 0 & \text{si } t_1 \leq u_i \\ \frac{\tau_i^2}{2S} & \text{si } u_i \leq t_1 \leq u_i + S \\ \tau_i - S/2 & \text{si } t_1 \geq u_i + S. \end{cases}$$

L'estimateur du maximum de vraisemblance $(\widehat{\alpha}, \widehat{\mu}, \widehat{S})$ est obtenu en maximisant :

$$M_C^{\Gamma-mr}(\alpha, \mu, S) = \alpha \ln(\alpha/\mu) - \ln\Gamma(\alpha) + \frac{1}{C}\sum_{i=1}^{C} \left[\ln\Gamma(\alpha + k_i) - (\alpha + k_i)\ln(\alpha/\mu + g_i(S))\right].$$

1.2.2 Décroissance exponentielle

L'intensité du centre i peut être représentée par une fonction

$$f_i(t) = \lambda_i \left(1 + \delta \exp(-(t - u_i)/t^*)\right) \mathbf{1}_{t \geq u_i},$$

où u_i désigne toujours la date d'ouverture du centre i et les constantes $\delta > 0, t^* > 0$ sont supposées identiques pour tous les centres. Les intensités limites $(\lambda_i)_{1 \leq i \leq C}$ sont distribuées suivant une loi Gamma de paramètres (α, β).

Un exemple de fonction f_i est donné dans la figure 1.2.

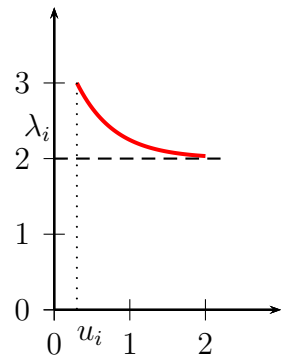

FIGURE 1.2 – Exemple d'intensité à décroissance exponentielle. $\lambda_i = 2, \delta = 0.5, t^* = 0.5, u_i = 0.3$.

On pose $\mu = \alpha/\beta$. A un instant intermédiaire t_1, les paramètres $(\alpha, \mu, \delta, t^*)$ sont estimés par maximum de vraisemblance. Soit $\tau_i = t_1 - u_i$. Par un calcul similaire à celui de la section précédente, on montre la proposition suivante :

Proposition 1.2.2. *Posons*

$$g_i(\delta, t^*) = \tau_i + \delta t^* \left(1 - e^{-\tau_i/t^*}\right).$$

L'estimateur du maximum de vraisemblance $(\widehat{\alpha}, \widehat{\mu}, \widehat{\delta}, \widehat{t^*})$ *est obtenu en maximisant :*

$$M_C^{\Gamma-e}(\alpha, \mu, \delta, t^*) = \alpha \ln(\alpha/\mu) - \ln \Gamma(\alpha) + \frac{1}{C} \sum_{i=1}^{C} \left[\ln \Gamma(\alpha + k_i) - (\alpha + k_i) \ln(\alpha/\mu + g_i(\delta, t^*)) \right].$$

1.3 Dates d'ouverture des centres inconnues

Les données utilisées dans cette partie nous ont été prêtées par l'unité 1027 de l'INSERM. L'essai clinique comprenait 77 centres et avait pour but de recruter 610 patients en 3 ans. En réalité, l'inclusion de patients s'est arrêtée au bout de 2.31 années. La particularité de ce jeu de données est que la date d'ouverture des centres n'a pas pu nous être fournie ; en revanche, nous connaissons les dates d'inclusion de chaque patient. Les modèles présentés dans la section précédente ne sont pas utilisables car à un instant t_1, nous ne connaissons pas la durée d'activité τ_i d'un centre. Une première possibilité serait de ne considérer le processus d'inclusion du centre i qu'à partir de la date de première inclusion de ce centre. Néanmoins, pour les centres recrutant peu de patients, cette méthode fait perdre beaucoup d'information.

Plutôt que de chercher à estimer la date d'ouverture de chaque centre, nous introduisons un modèle où la date d'ouverture est supposée uniformément distribuée dans l'intervalle $[0, v_i]$ où v_i est la date de première inclusion du centre i [35]. Les intensités d'inclusion sont supposées distribuées selon une loi Gamma. Nous appelons ce modèle le modèle Γ-Poisson uniforme (en abrégé modèle $\mathcal{U}\Gamma$-Poisson).

1.3.1 Modèle Gamma-Poisson uniforme

Dans le modèle Γ-Poisson uniforme, les dates d'ouvertures sont supposées uniformément distribuées sur $[0, v_i]$ où v_i est la date de première inclusion du centre i.

Soit $t_1 > 0$. Si le centre i n'a pas recruté sur $[0, t_1]$, on pose $v_i = t_1$. $\tau_i = t_1 - u_i$ est donc uniformément distribué sur $[t_1 - v_i, t_1]$. On pose $S_i' = t_1 - v_i$ et $S_i = t_1$.

Notons que v_i, S_i et S_i' dépendent de t_1.

Les intensités d'inclusion sont distribuées selon une loi Gamma de paramètres (α, β).

Proposition 1.3.1. *La loi de* $k_i = N_i^R(t_1)$ *est donnée par*

$$\mathbb{P}[k_i = n] = \beta^\alpha \frac{\Gamma(\alpha + n)}{\Gamma(\alpha) n!} \int_{S_i'}^{S_i} \frac{t^n}{(\beta + t)^{\alpha+n}} \frac{dt}{S_i - S_i'}.$$

Démonstration. Sachant l'intensité λ_i et la date d'ouverture du centre u_i, k_i suit une loi de Poisson de paramètre $\lambda_i \tau_i$ où $\tau_i = t_1 - u_i$. τ_i est uniformément distribué sur $[S_i', S_i]$. On en déduit

$$\begin{aligned}
\mathbb{P}[k_i = n] &= \int_0^{+\infty} \beta^\alpha \Gamma(\alpha)^{-1} dx \int_{S_i'}^{S_i} \frac{dt}{S_i - S_i'} e^{-xt} \frac{(xt)^n}{n!} e^{-\beta x} x^{\alpha-1} \\
&= \beta^\alpha \frac{\Gamma(\alpha + n)}{\Gamma(\alpha) n!} \int_{S_i'}^{S_i} \frac{t^n}{(\beta + t)^{\alpha+n}} \frac{dt}{S_i - S_i'}.
\end{aligned}$$

□

Estimation par maximum de vraisemblance

On déduit de la proposition 1.3.1 le corollaire :

Corollaire 1.3.2. *L'estimateur du maximum de vraisemblance* $(\widehat{\alpha}_C, \widehat{\mu}_C)$ *est obtenu en maximisant :*

$$M_C^{\mathcal{U}\Gamma}(\alpha,\mu) = \frac{1}{C}\sum_{i=1}^{C}\left[\alpha\ln(\alpha/\mu) - \ln\Gamma(\alpha) + \ln\Gamma(\alpha+k_i) + \ln(B(\alpha,\alpha/\mu,k_i,S_i',S_i)\right],$$

où $B(a,b,k,S',S) = \int_{S'+b}^{S+b} t^{-k-a}(t-b)^k dt$.

Remarque 1.3.3. *Un calcul direct montre que si* $a > 1$,

$$B(a,b,k,S',S) = b^{1-a}\left[\beta_{\text{inc}}\left(\frac{b}{S'+b};\, a-1, k+1\right) - \beta_{\text{inc}}\left(\frac{b}{S+b};\, a-1,k+1\right)\right],$$

où β_{inc} *est la fonction bêta incomplète : pour tous* $x \in [0,1]$, $a_1 > 0$, $a_2 > 0$,

$$\beta_{\text{inc}}(x;a_1,a_2) = \int_0^x t^{a_1-1}(1-t)^{a_2-1}dt. \tag{1.3}$$

Si $a \neq 1$ *et* $k > 0$,

$$\begin{aligned} B(a,b,k,S',S) = \frac{1}{a-1}\Bigg[&\frac{S'^k}{(S'+b)^{a+k-1}} - \frac{S^k}{(S+b)^{a+k-1}} \\ &+ kb^{1-a}\left(\beta_{\text{inc}}\left(\frac{b}{S'+b};\, a,k\right) - \beta_{\text{inc}}\left(\frac{b}{S+b};\, a,k\right)\right)\Bigg]. \end{aligned}$$

Propriétés de l'estimateur

Soit $\theta \equiv (\alpha,\mu)$, et $\theta_0 \equiv (\alpha_0,\mu_0)$ la vraie valeur des paramètres.
Posons $J_i(\theta,S_i,S_i') = -\mathbb{E}_\theta\left[\nabla^2 f(\theta,k_i,S_i,S_i')\right]$ où

$$f(\theta,k_i,S_i,S_i') = \alpha\ln(\alpha/\mu) - \ln\Gamma(\alpha) + \ln\Gamma(\alpha+k_i) + \ln(B(\alpha,\alpha/\mu,k_i,S_i',S_i).$$

Si la suite $(S_i,S_i')_{i\geq 1}$ vérifie $S_i - S_i' \geq \xi$ pour un certain $\xi > 0$, et si $\frac{1}{C}\sum_{i=1}^{C} J_i(\theta_0,S_i,S_i')$ converge vers une matrice définie positive I_0, alors $\widehat{\theta}_C$ suit asymptotiquement une loi normale de moyenne θ_0 et de matrice de covariance $\frac{1}{C}I_0^{-1}$.

Ceci nous permet de calculer une région de confiance pour l'estimateur $(\widehat{\alpha}_C,\widehat{\mu}_C)$: la région de confiance à 95% est une ellipse centrée en $(\widehat{\alpha}_C,\widehat{\mu}_C)$, définie par

$$\{X \in \mathbb{R}^2 : (X-\theta_0)^T I(\theta_0)(X-\theta_0) \leq p/C\},$$

où $p \simeq 5.99$ est le quantile d'ordre 95% d'une loi du χ^2 à deux degrés de liberté.

Date intermédiaire	$\hat{\alpha}$	$\hat{\mu}$
$t_1 = 1$	0.82	3.53
$t_1 = 1.5$	1.37	4.17
$t_1 = 2$	1.51	4.09
$T_f = 2.32$	1.51	4.13

TABLE 1.3 – Estimateurs de $\widehat{\alpha}_C$ et $\widehat{\mu}_C$ dans le modèle $\mathcal{U}\Gamma$-Poisson.

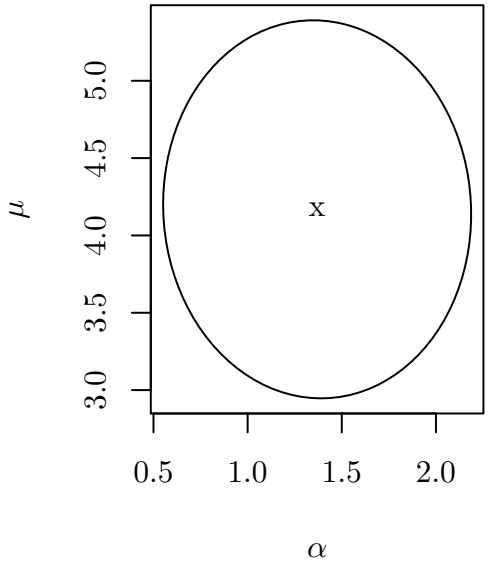

FIGURE 1.3 – Région de confiance à 95% pour $(\widehat{\alpha}_C, \widehat{\mu}_C)$ à $t_1 = 1.5$ années.

1.3.2 Application aux données

Dans la table 1.3 sont donnés les résultats des estimations des paramètres à différents instants intermédiaires t_1. Le temps réel d'inclusion de l'essai est $T_f = 2.32$ années.

Dans la figure 1.3, nous représentons la région de confiance à 95% pour $(\widehat{\alpha}_C, \widehat{\mu}_C)$ à $t_1 = 1.5$ années. Empiriquement, la corrélation entre $\widehat{\alpha}_C$ et $\widehat{\mu}_C$ est proche de 0. La région de confiance est assez large : la distance entre les estimateurs et les paramètres réels peut être grande. Nous verrons dans le chapitre 2 l'impact de cette erreur sur la prédiction.

Enfin, dans la figure 1.4 est comparée la densité empirique de $\widehat{\alpha}_C$ et $\widehat{\mu}_C$ à $t_1 = 1.5$ années à la loi normale asymptotique correspondante. Nous remarquons que l'estimateur de α est légèrement biaisé.

1.3.3 Validation du modèle

Comme dans la première section, nous traçons la courbe d'occupation des données, ainsi que la moyenne attendue dans le modèle $\mathcal{U}\Gamma$-Poisson. A l'instant T_f, soit ν_j le nombre de centres ayant recruté exactement j patients. Alors

$$\mathbb{E}\left[\nu_j\right] = \sum_{i=1}^{C} \mathbb{P}\left[k_i = j\right] = (\alpha/\mu)^{\alpha}\Gamma(\alpha)^{-1}\Gamma(j+\alpha)\sum_{i=1}^{C}\frac{1}{S_i - S_i'}B(\alpha, \alpha/\mu, j, S_i', S_i).$$

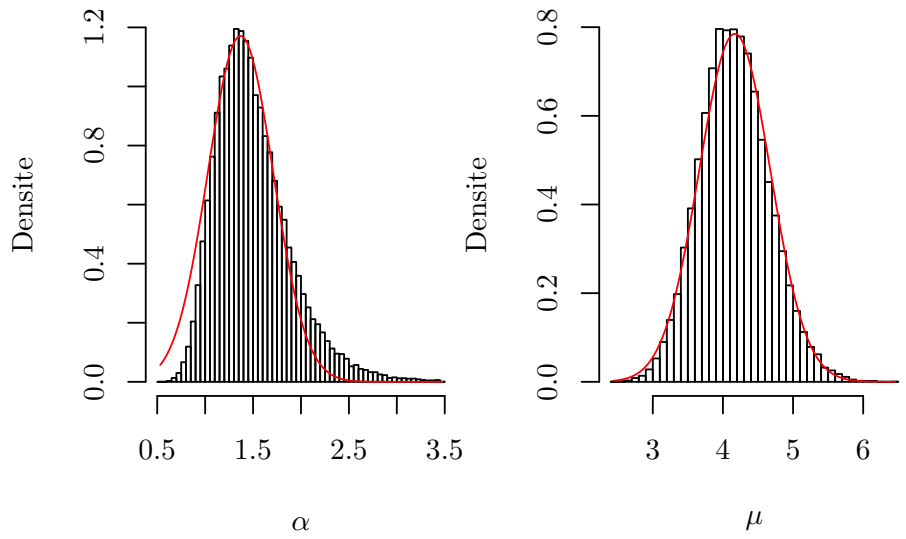

FIGURE 1.4 – Densité empirique et approximation normale pour $\widehat{\alpha}_C$ (gauche) et $\widehat{\mu}_C$ (droite) à $t_1 = 1.5$ années.

La courbe d'occupation est donnée dans la figure 1.5. Le modèle est en bonne adéquation avec les données.

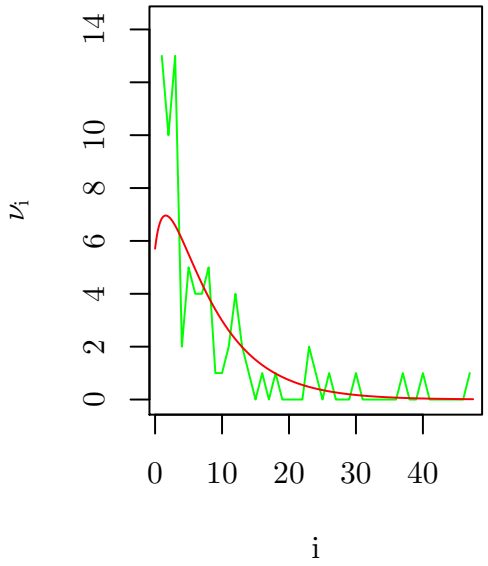

FIGURE 1.5 – Courbe d'occupation à $T_f = 2.31$ années.

Remarque 1.3.4. *Nous n'effectuons pas de test statistique pour ces données ; en effet, les probabilités que nous notons $(p_j)_{j\geq 1}$ qui représentent $\mathbb{P}[k_i = j]$ dépendent maintenant du centre i.*

Chapitre 2

Prédiction du processus d'inclusion

Dans ce chapitre, nous ne considérerons que les modèles Gamma-Poisson, Pareto-Poisson et Gamma-Poisson uniforme.

Nous fixons un instant intermédiaire $t_1 > 0$. Nous observons le processus d'inclusion jusqu'à t_1 et cherchons à prédire ce processus après t_1 sachant ces observations.

Comme dans le chapitre précédent, on notera, pour tout $1 \leq i \leq C$:

- u_i la date d'ouverture du centre i,
- $\tau_i = t_1 - u_i$ la durée d'activité du centre i entre 0 et t_1.

2.1 Ré-estimation bayésienne des intensités d'inclusion

Dans tous les modèles considérés, la variable aléatoire λ_i n'est pas indépendante du processus N_i^R : λ_i détermine l'intensité de N_i^R. Ainsi, la loi de λ_i sachant l'information à t_1 diffère de la loi a priori de λ_i. L'objet de ce chapitre est de caractériser les nouvelles lois par ré-estimation bayésienne, puis de décrire le processus d'inclusion après t_1.

2.1.1 Dates d'ouverture des centres connues

On suppose que les dates d'ouvertures des centres $(u_i)_{1\leq i\leq C}$ sont connues, et que les intensités d'inclusion sont distribuées suivant une loi de densité p_θ, où $\theta \in \mathbb{R}^d$.

Proposition 2.1.1. *La densité de λ_i sachant $N_i^R(t_1) = k_i$, notée $p_\theta^{t_1}(.)$ peut être évaluée par la formule de Bayes*

$$\begin{aligned} p_\theta^{t_1}(x) &= \frac{\mathbb{P}\left[N_i^R(t_1) = k_i \,|\, \lambda_i = x\right] p_\theta(x)}{\mathbb{P}\left[N_i^R(t_1) = k_i\right]} \\ &= e^{-\tau_i x} x^{k_i} p_\theta(x)\, \frac{\tau_i^{k_i}}{k_i! \mathbb{P}\left[N_i^R(t_1) = k_i\right]}, \end{aligned} \tag{2.1}$$

où p_θ est la densité de la loi a priori de λ_i et $\tau_i = (t_1 - u_i) \vee 0$.

Modèle Gamma-Poisson

Dans le cas du modèle Γ-Poisson, la densité est $x \mapsto p_{(\alpha,\beta)}(x) = \beta^\alpha \Gamma(\alpha)^{-1} e^{-\beta x} x^{\alpha-1} \mathbf{1}_{x\geq 0}$, et on déduit de (2.1) la proposition suivante tirée de [9].

Proposition 2.1.2 (Loi prédictive dans le modèle Γ-Poisson)**.** *Dans le modèle Gamma-Poisson, la densité de λ_i sachant $N_i^R(t_1) = k_i$ est celle d'une loi Gamma de paramètres $(\alpha + k_i, \beta + \tau_i)$.*

Démonstration. La proposition 2.1.1 appliquée à la densité $p_{(\alpha,\beta)} : x \mapsto \beta^\alpha \Gamma(\alpha)^{-1} e^{-\beta x} x^{\alpha-1} \mathbf{1}_{x>0}$ implique que la densité $p^{t_1}_{(\alpha,\beta)}$ de λ_i sachant k_i est de la forme

$$p^{t_1}_{(\alpha,\beta)}(x) = M e^{-\tau_i x} x^{k_i} e^{-\beta x} x^{\alpha-1} \mathbf{1}_{x>0} = M e^{-(\beta+\tau_i)x} x^{k_i+\alpha-1} \mathbf{1}_{x>0},$$

où M est une quantité ne dépendant pas de x. On en déduit le résultat en normalisant. □

Modèle Pareto-Poisson

Proposition 2.1.3 (Loi prédictive dans le modèle Π-Poisson)**.** *Dans le modèle Pareto-Poisson, la densité de λ_i sachant $N_i^R(t_1) = k_i$ admet une densité qui est*

$$x \mapsto \frac{\tau_i^{k_i-\gamma}}{\Gamma_{inc}(k_i - \gamma, \delta\tau_i)} e^{-x\tau_i} x^{k_i-\gamma-1} \mathbf{1}_{x\geq\delta},$$

où

$$\forall x \in \mathbb{R}, \forall y > 0, \quad \Gamma_{inc}(x, y) = \int_y^{+\infty} e^{-t} t^{x-1} dt. \tag{2.2}$$

Démonstration. La proposition 2.1.1 appliquée à la densité $p_{(\gamma,\delta)} : x \mapsto \gamma \frac{\delta^\gamma}{x^{\gamma+1}} \mathbf{1}_{\{x\geq\delta\}}$ implique que la densité $p^{t_1}_{(\gamma,\delta)}$ de λ_i sachant k_i est de la forme

$$p^{t_1}_{(\gamma,\delta)}(x) = M e^{-\tau_i x} x^{k_i-\gamma-1} \mathbf{1}_{x\geq\delta},$$

où M est une quantité ne dépendant pas de x. On déduit de la condition $\int_{\mathbb{R}} p^{t_1}_{(\gamma,\delta)}(x)dx = 1$ la valeur de M. □

2.1.2 Dates d'ouverture des centres inconnues

Lorsque les dates d'ouvertures des centres sont inconnues, on se place dans le cadre du modèle $\mathcal{U}\Gamma$-Poisson.

Si le centre i a recruté entre 0 et t_1, on note v_i l'instant de première inclusion. Sinon, on pose $v_i = t_1$. On suppose que la date d'ouverture u_i du centre i est uniformément distribuée sur $[0, v_i]$, et indépendante de λ_i.

Posons

$$S_i = t_1, \quad S_i' = t_1 - v_i.$$

Alors la durée d'activité τ_i du centre i sur $[0, t_1]$ est uniformément distribuée sur $[S_i', S_i]$.

Proposition 2.1.4 (Loi prédictive dans le modèle $\mathcal{U}\Gamma$-Poisson)**.** *Dans le modèle Gamma-Poisson uniforme, la loi de λ_i sachant $N_i^R(t_1) = k_i$ admet une densité qui est*

$$x \mapsto M\left(\Gamma_{inc}(xS_i', k_i + 1) - \Gamma_{inc}(xS_i, k_i + 1)\right) x^{-1} p_{\alpha,\beta}(x),$$

où M est une constante normalisante, $p_{\alpha,\beta}$ est définie par (1.1) *et Γ_{inc} par* (2.2)*.*

Démonstration. D'après la règle de Bayes, la densité de λ_i sachant k_i est

$$\begin{aligned} p(x) &= \frac{\mathbb{P}\left[N_i(t_1) = k_i \,|\, \lambda_i = x\right] p(x;\alpha,\beta)}{\mathbb{P}\left[N_i(t_1) = k_i\right]} \\ &= M\,\mathbb{E}\left[e^{-x\tau_i}(x\tau_i)^{k_i}\right] p(x;\alpha,\beta) \\ &= M\int_{S_i'}^{S_i} e^{-xt} t^{k_i} dt\, x^{k_i} p(x;\alpha,\beta) \\ &= M\left(\Gamma_{\text{inc}}(k_i+1, xS_i') - \Gamma_{\text{inc}}(k_i+1, xS_i)\right) x^{-1} p(x;\alpha,\beta), \end{aligned}$$

où M est une quantité indépendante de x pouvant varier d'une ligne à l'autre. □

2.2 Prédiction du processus d'inclusion

Les processus d'inclusion des centres (resp. processus global d'inclusion) après t_1 seront notés $\tilde{N}_i^R$; $1 \le i \le C$ (resp. $\tilde{N}^R$), et sont définis, pour tout $t \ge 0$, par

$$\tilde{N}_i^R(t) = N_i^R(t + t_1) - N_i^R(t_1) \quad \text{et} \quad \tilde{N}^R(t) = \sum_{i=1}^{C} \tilde{N}_i^R(t).$$

On définit également

$$\Lambda = \sum_{i=1}^{C} \lambda_i$$

que nous appellerons intensité globale du processus d'inclusion.

Dans les trois modèles considérés, les centres sont supposés ouverts après t_1. On en déduit la proposition :

Proposition 2.2.1.
- *Sachant λ_i, le processus $\tilde{N}_i^R$ est un processus de Poisson d'intensité homogène λ_i.*
- *Sachant $(\lambda_1, \ldots, \lambda_C)$, $\tilde{N}^R$ est processus de Poisson d'intensité homogène $\Lambda = \sum_{i=1}^{C} \lambda_i$.*

Corollaire 2.2.2. *Pour tout $t > 0$, l'espérance et la variance de $\tilde{N}_i^R(t)$ sachant $k_i = N_i^R(t_1)$ sont*

$$\begin{aligned} \mathbb{E}\left[\tilde{N}_i^R(t) \mid k_i\right] &= t\mathbb{E}\left[\lambda_i \mid k_i\right], \\ \text{Var}[\tilde{N}_i^R(t) \mid k_i] &= t^2 \,\text{Var}[\lambda_i \mid k_i] + t\mathbb{E}\left[\lambda_i \mid k_i\right]. \end{aligned}$$

Sachant l'information à t_1, les variables aléatoires $(\lambda_i)_{1\le i\le C}$ sont indépendantes, et suivent les lois prédictives dans le modèle considéré. Ces lois dépendent en général de l'indice i – les intensités ne sont plus identiquement distribuées. De plus, dans les modèles Pareto-Poisson et Gamma-Poisson uniforme, leur complexité ne se prête pas aux calculs. Il sera alors utile d'approcher l'intensité globale Λ par une loi Gamma, selon la proposition suivante.

Proposition 2.2.3 (Approximation du taux global par une loi Gamma)**.** *Soient $\lambda_1, \ldots, \lambda_C$ des variables aléatoires positives indépendantes, et supposons que, pour tout $i = 1, \ldots, C$,*

$$m_i = \mathbb{E}[\lambda_i] < \infty, \qquad \sigma_i^2 = \mathrm{Var}[\lambda_i] < \infty.$$

Posons

$$m = \sum_{i=1}^{C} m_i, \quad \sigma^2 = \sum_{i=1}^{C} \sigma_i^2, \quad A = \frac{m^2}{\sigma^2}, \quad B = \frac{m}{\sigma^2}. \tag{2.3}$$

Alors $\sum_{i=1}^{C} \lambda_i$ a même espérance et même variance qu'une loi Gamma de paramètres (A, B) dont la densité est donnée par (1.1)*.*

Nous nous intéresserons à deux quantités en particulier : la moyenne du temps d'inclusion et probabilité de terminer l'inclusion à temps. Donnons quelques notations :

- $K_1 = \sum_{i=1}^{C} k_i$ est le nombre de patients recrutés jusqu'au temps t_1,
- $\tilde{K} = K_f - K_1$ est le nombre de patients restant à recruter après t_1,
- $\tilde{T} = \inf_{t \geq 0}\{\tilde{N}^R(t) = \tilde{K}\}$ le temps d'inclusion restant.

On déduit de la proposition 2.2.1 :

Proposition 2.2.4 ([22, page 4])**.** *Conditionnellement à Λ, $\tilde{T}$ suit une loi $\Gamma(\tilde{K}, \Lambda)$.*

Corollaire 2.2.5. *Soit $\tilde{N}$ un processus doublement stochastique dont l'intensité homogène Λ suit une loi $\Gamma(A, B)$ et $\tilde{T} = \inf_{t \geq 0}\{\tilde{N}(t) = \tilde{K}\}$ où $\tilde{K} \in \mathbb{N}^*$. Alors la loi de $\tilde{T}$ admet une densité par rapport à la mesure de Lebesgue, qui vaut :*

$$p_{\tilde{T}} : t \mapsto \frac{B^A \Gamma(\tilde{K} + A)}{\Gamma(A)\Gamma(\tilde{K})} \frac{t^{\tilde{K}-1}}{(t+B)^{\tilde{K}+A}}.$$

En particulier,

$$\begin{aligned} \mathbb{E}\left[\tilde{T}\right] &= \tilde{K}\frac{B}{A-1} \quad \text{si } A > 1 \\ \mathbb{E}\left[\tilde{T}\right] &= +\infty \qquad\quad \text{si } 0 < A \leq 1, \end{aligned} \tag{2.4}$$

et

$$\mathbb{P}\left[\tilde{T} \leq x\right] = \frac{\Gamma(A + \tilde{K})}{\Gamma(A)(\tilde{K}-1)!} \beta_{inc}\left(\frac{x}{x+B}; \tilde{K}, A\right), \ \forall x \geq 0, \tag{2.5}$$

où β_{inc} est définie dans (1.3)*.*

Démonstration. Sachant Λ, la densité de $\tilde{T}$ est $t \mapsto \frac{\Lambda^{\tilde{K}}}{(\tilde{K}-1)!} e^{-\Lambda t} t^{\tilde{K}-1}$. En conditionnant, on en déduit la densité de $\tilde{T}$:

$$\begin{aligned} p_{\tilde{T}}(t) &= \frac{B^A}{\Gamma(A)(\tilde{K}-1)!} t^{\tilde{K}-1} \int_0^{+\infty} e^{-Bx} x^{A-1} x^{\tilde{K}} e^{-xt} dx \\ &= \frac{B^A \Gamma(A + \tilde{K})}{\Gamma(A)(\tilde{K}-1)!} \frac{t^{\tilde{K}-1}}{(t+B)^{\tilde{K}+A}}. \end{aligned}$$

Les relations en découlent. □

2.2.1 Dates d'ouverture des centres connues

Modèle Gamma-Poisson

D'après les propositions 2.1.2 et 2.2.1, sachant $k_i = N_i^R(t_1)$, $\tilde{N}_i^R$ est un processus doublement stochastique d'intensité homogène λ_i où $\lambda_i \overset{\mathcal{L}}{\sim} \Gamma(\alpha + k_i, \beta + \tau_i)$.

Nous déduisons de la proposition 2.2.3 le corollaire :

Corollaire 2.2.6. *Soient* $m_i = \mathbb{E}[\lambda_i \mid k_i]$ *et* $\sigma_i^2 = \mathrm{Var}[\lambda_i \mid k_i]$. *Alors*

$$m_i = \frac{\alpha + k_i}{\beta + \tau_i} \quad et \quad \sigma_i^2 = \frac{\alpha + k_i}{(\beta + \tau_i)^2}. \tag{2.6}$$

De plus, si A *et* B *sont définis par* (2.3), *alors conditionnellemment à* $(k_1, \ldots, k_C)$, Λ *a mêmes moments d'ordre 1 et 2 qu'une loi Gamma de paramètres* (A, B).

Remarque 2.2.7. *Si pour tout* $1 \leq i \leq C$, $\tau_i \equiv \tau$ *(c'est-à -dire si tous les centres ont ouvert en même temps), l'approximation précédente est exacte : sachant* $(k_1, \ldots, k_C)$,

$$\Lambda \overset{\mathcal{L}}{\sim} \Gamma\left(C\alpha + \sum_{i=1}^{C} k_i, \beta + \tau\right).$$

Une conséquence des corollaires 2.2.2 et 2.2.6 est la proposition suivante :

Proposition 2.2.8. *Pour tout* $t > 0$, *sachant* $(k_1, \ldots, k_C)$,

$$\begin{aligned} \mathbb{E}\left[\tilde{N}^R(t)\right] &= mt, \\ \mathrm{Var}[\tilde{N}^R(t)] &= \sigma^2 t^2 + mt, \end{aligned}$$

où $m = \sum_{i=1}^{C} m_i$, $\sigma^2 = \sum_{i=1}^{C} \sigma_i^2$, *et* m_i *et* σ_i^2 *sont définis dans* (2.6). *Si* C *est assez grand* ($C \geq 30$), $\tilde{N}^R(t)$ *suit approximativement une loi normale de moyenne* mt *et d'écart-type* $\sqrt{\sigma^2 t^2 + mt}$.

Modèle Pareto-Poisson

D'après les propositions 2.1.3 et 2.2.1, sachant $k_i = N_i^R(t_1)$, $\tilde{N}_i^R$ est un processus doublement stochastique d'intensité homogène λ_i où λ_i suit la loi donnée dans la proposition 2.1.3.

Corollaire 2.2.9. *Soient* $m_i = \mathbb{E}\left[\tilde{N}_i^R(t) \mid k_i\right]$ *et* $\sigma_i^2 = \mathrm{Var}[\tilde{N}_i^R(t) \mid k_i]$. *Alors*

$$\begin{aligned} m_i &= \frac{1}{\tau_i} \cdot \frac{\Gamma_{inc}(k_i - \gamma + 1, \delta\tau_i)}{\Gamma_{inc}(k_i - \gamma, \delta\tau_i)}, \\ \sigma_i^2 &= \frac{1}{\tau_i^2} \frac{\Gamma_{inc}(k_i - \gamma + 2, \delta\tau_i)}{\Gamma_{inc}(k_i - \gamma, \delta\tau_i)} - m_i^2, \end{aligned} \tag{2.7}$$

et soient A *et* B *définis par* (2.3). *Alors conditionnellemment à* $(k_1, \ldots, k_C)$, Λ *a mêmes moments d'ordre 1 et 2 qu'une loi Gamma de paramètres* (A, B).

Une conséquence des corollaires 2.2.2 et 2.2.9 est la proposition suivante :

Proposition 2.2.10. *Pour tout $t > 0$, sachant $(k_1, \ldots, k_C)$,*

$$\mathbb{E}\left[\tilde{N}^R(t)\right] = mt,$$
$$\mathrm{Var}[\tilde{N}^R(t)] = \sigma^2 t^2 + mt,$$

où $m = \sum_{i=1}^{C} m_i$, $\sigma^2 = \sum_{i=1}^{C} \sigma_i^2$, et m_i et σ_i^2 sont définis dans (2.7). Si C est assez grand ($C \geq 30$), $\tilde{N}^R(t)$ suit approximativement une loi normale de moyenne mt et d'écart-type $\sqrt{\sigma^2 t^2 + mt}$.

2.2.2 Dates d'ouverture des centres inconnues

Si les dates d'ouverture des centres sont inconnues, on utilise le modèle $\mathcal{U}\Gamma$-Poisson défini dans le chapitre 1.

D'après les propositions 2.1.4 et 2.2.1, sachant $k_i = N_i^R(t_1)$, $\tilde{N}_i^R$ est un processus doublement stochastique d'intensité homogène λ_i où λ_i suit la loi donnée dans la proposition 2.1.4.

Corollaire 2.2.11. *Soient $m_i = \mathbb{E}\left[\tilde{N}_i^R(t) \mid k_i\right]$ et $\sigma_i^2 = \mathrm{Var}[\tilde{N}_i^R(t) \mid k_i]$. Soient*

$$\begin{aligned} m_i &= \frac{\alpha + k_i}{S_i - S_i'} \ln\left(\frac{\beta + S_i}{\beta + S_i'}\right), \\ \sigma_i^2 &= \frac{\alpha + k_i}{(\beta + S_i')(\beta + S_i)}, \end{aligned} \tag{2.8}$$

et soient A et B définis par (2.3). Alors conditionnellemment à $(k_1, \ldots, k_C)$, Λ a mêmes moments d'ordre 1 et 2 qu'une loi Gamma de paramètres (A, B).

Une conséquence des corollaires 2.2.2 et 2.2.11 est la proposition suivante :

Proposition 2.2.12. *Pour tout $t > 0$, sachant $(k_1, \ldots, k_C)$,*

$$\mathbb{E}\left[\tilde{N}^R(t)\right] = m_i t,$$
$$\mathrm{Var}[\tilde{N}^R(t)] = \sigma_i^2 t^2 + m_i t,$$

où $m = \sum_{i=1}^{C} m_i$, $\sigma^2 = \sum_{i=1}^{C} \sigma_i^2$, et m_i et σ_i^2 sont définis dans (2.8). Si C est assez grand ($C \geq 30$), $\tilde{N}^R(t)$ suit approximativement une loi normale de moyenne mt et d'écart-type $\sqrt{\sigma^2 t^2 + mt}$.

2.3 Sensibilité aux paramètres

On rappelle que, sous de bonnes hypothèses, l'estimateur du maximum de vraisemblance $\hat{\theta}_C$ est consistant et asymptotiquement normal quand $C \to +\infty$ et la matrice d'information de Fischer associée est donnée dans chaque modèle. Ceci nous permet de calculer une région de confiance au quantile q (typiquement, $q = 0,95$), à l'aide de l'approximation

$$\sqrt{C}(\hat{\theta}_C - \theta_0) \overset{\mathcal{L}}{\simeq} \mathcal{N}(0, I_0^{-1}).$$

Considérons maintenant une quantité d'intérêt, c'est-à -dire une fonction f du paramètre θ – par exemple, $\theta \mapsto \mathbb{E}_\theta\left[\tilde{T} \,|\, \mathcal{F}_{t_1}\right]$ ou $\theta \mapsto \mathbb{P}_\theta\left[\tilde{T} \leq t \,|\, \mathcal{F}_{t_1}\right]$. La région de confiance pour $\hat{\theta}_C$ induit un intervalle de confiance pour $f(\theta)$: c'est une conséquence de la δ-méthode [48].

Proposition 2.3.1 (δ-méthode)**.** *Pour toute fonction $f : \mathbb{R}^2 \to \mathbb{R}$ de classe $\mathcal{C}^2$, on a*

$$\sqrt{C}(f(\hat{\theta}_C) - f(\theta_0)) \xrightarrow[C\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla f_{\theta_0}^T I_0^{-1} \nabla f_{\theta_0})$$

Les fonctions f auxquelles nous appliquerons la proposition précédente peuvent être calculées numériquement de manière efficace, donc leur gradient également (par différences finies). Néanmoins, dans le cas général où f est l'espérance sous θ d'une fonctionnelle des trajectoires, le théorème de Girsanov est un moyen de calculer son gradient. Nous énonçons donc un théorème de Girsanov pour les processus de Cox.

Proposition 2.3.2 (Girsanov)**.** *Sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, soit N un processus de Cox d'intensité λ où λ suit la loi p_{θ_0} sur $\mathbb{R}_+$. Soit pour tout $t \geq 0$, $\mathcal{F}_t^N = \sigma(N_s;\ 0 \leq s \leq t)$ la filtration engendrée par N.*

Supposons que $\forall \theta \in \Theta$, p_θ est absolument continue par rapport à p_{θ_0} et notons $L_\theta = \frac{dp_\theta}{dp_{\theta_0}}$.

Soit $Z_t^\theta = \mathbb{E}\left[L_\theta(\lambda)|\mathcal{F}_t^N\right]$ et $\mathbb{P}^\theta$ la probabilité définie par $\mathcal{F}_T^N$ par $\frac{d\mathbb{P}^\theta}{d\mathbb{P}} = Z_T^\theta$.

Alors, sous $\mathbb{P}^\theta$ et sur $[0, T]$, N est un processus de Cox dont l'intensité suit la loi p_θ.

Démonstration. Soit X une variable aléatoire $\mathcal{F}_T^N$-mesurable. $\mathbb{E}_\theta[X]$ désignera l'espérance de X sous $\mathbb{P}^\theta$ (sans indice, l'espérance s'entend sous θ_0).

$$\begin{aligned}
\mathbb{E}_\theta[X] &= \mathbb{E}\left[\mathbb{E}\left[L_\theta(\lambda)|\mathcal{F}_T^N\right] X\right] \\
&= \mathbb{E}[L_\theta(\lambda) X] \\
&= \int_{\mathbb{R}} \mathbb{E}[X|\lambda = x]\, L_\theta(x) p_{\theta_0}(dx) \\
&= \int_{\mathbb{R}} \mathbb{E}[X|\lambda = x]\, p_\theta(dx)
\end{aligned}$$

d'où le résultat. □

Nous appliquons ces résultats aux données de la partie 1.3. Les dates d'ouverture des centres étant inconnues, nous utilise le modèle Gamma-Poisson uniforme.

Nous utilisons l'approximation de l'intensité globale d'inclusion Λ de $\tilde{N}^R$ par une loi Γ de paramètres (A, B) où A et B sont données dans (2.3). Les moyenne et variance m et σ^2 de Λ, qui dépendent des paramètres du modèle α et μ, sont définies dans le corollaire 2.2.11. Puis, nous utilisons (2.4) et (2.5) pour calculer, sous les paramètres (α, μ), l'espérance et le quantile d'ordre $p > 0$ du temps d'inclusion restant, que nous noterons $\mu^{\tilde{T}}(\alpha, \mu)$ et $q_p^{\tilde{T}}(\alpha, \mu)$. Ainsi, on peut calculer numériquement le gradient de ces deux quantités par rapport aux paramètres (α, μ), et utiliser le lemme de la δ-méthode pour calculer un intervalle de confiance pour $\mu^{\tilde{T}}(\widehat{\alpha}_C, \widehat{\mu}_C)$ et $q_p^{\tilde{T}}(\widehat{\alpha}_C, \widehat{\mu}_C)$, où $(\widehat{\alpha}_C, \widehat{\mu}_C)$ est l'estimateur du maximum de vraisemblance du paramètre (α, μ) tel que défini dans la proposition 1.3.2.

On note $\widehat{\mu}^T = t_1 + \mu^{\tilde{T}}(\widehat{\alpha}_C, \widehat{\mu}_C)$ l'estimateur du temps moyen d'inclusion total, et $\widehat{q}^T = t_1 + q_{0.95}^{\tilde{T}}(\widehat{\alpha}_C, \widehat{\mu}_C)$ le quantile d'ordre 0.95 du temps d'inclusion total.

On rappelle que le temps d'inclusion réel est de 2.31 années. Les résultats de prédiction sont dans les tables 2.1 et 2.2.

t_1	$\widehat{\mu}_T$	Intervalle de confiance
1	2.60	[2.46,2.75]
1.5	2.34	[2.29,2.40]
2	2.36	[2.34,2.37]

TABLE 2.1 – Estimateur du temps moyen d'inclusion et intervalle de confiance à 95%.

t_1	$\widehat{q}_T$	Intervalle de confiance
1	2.83	[2.67,3.00]
1.5	2.47	[2.39,2.52]
2	2.42	[2.40,2.44]

TABLE 2.2 – Estimateur du quantile d'ordre 0.95 du temps d'inclusion et intervalle de confiance à 95%.

A une étude intermédiaire à 1.5 années, soit près de 10 mois avant la fin de l'inclusion, on est capable d'estimer la fin du processus de recrutement avec une précision de deux semaines.

La proposition 2.2.12 nous permet de calculer pour tout $t \geq 0$ l'espérance et un intervalle de confiance à 95% de $\tilde{N}^R(t) = N^R(t_1 + t)$ sachant $(k_1, \ldots, k_C)$. Ces quantités ainsi que le processus d'inclusion réel sont représentés dans la figure 2.1 pour différentes valeurs de t_1.

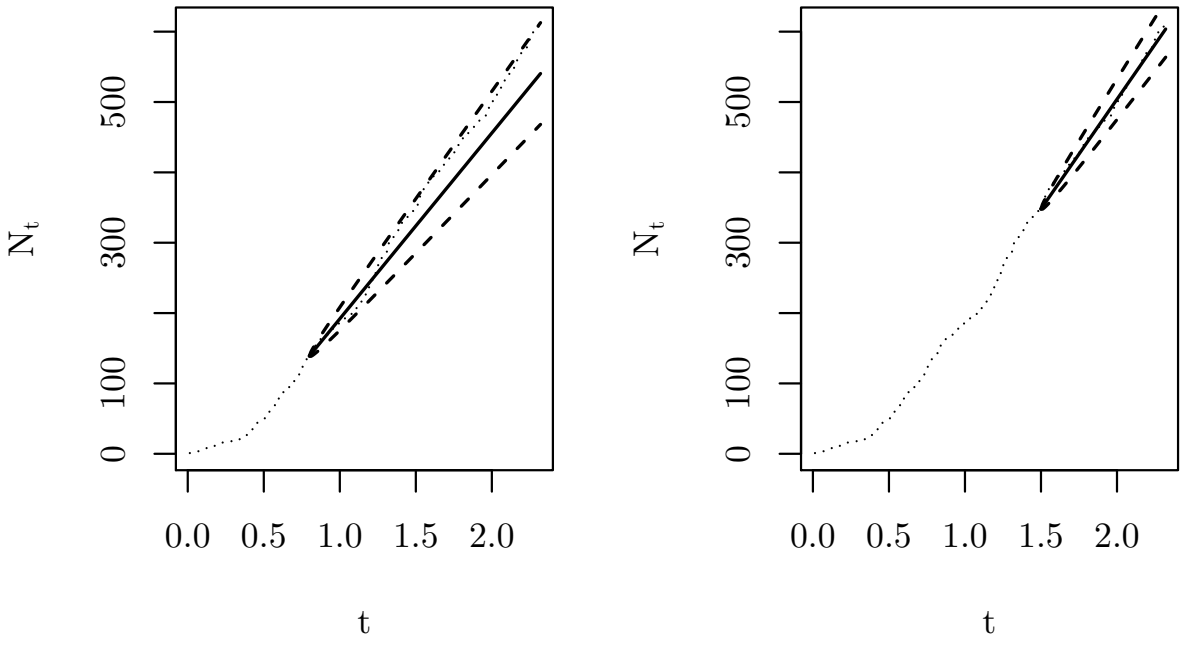

FIGURE 2.1 – Processus d'inclusion réel (points), moyenne de N^R (ligne) et bornes de l'intervalle de confiance à 95% (tirets) sachant l'information à t_1. Gauche : $t_1 = 0.8$ années, droite : $t_1 = 1.5$ années.

2.4 Ouverture et fermeture de centres

La flexibilité des modèles présentés permet de recalculer facilement la loi du temps d'inclusion si l'on décide de fermer ou ouvrir un centre. Si l'on ferme le centre j, l'intensité globale vaut $\Lambda_2 = \sum_{\substack{i=1 \\ i\neq j}}^{C} \lambda_i$, tandis que si l'on ouvre un nouveau centre, elle vaut $\Lambda_3 = \lambda + \sum_{i=1} \lambda_i$ où λ est l'intensité du nouveau centre.

Conditionnellement à $(k_1, \ldots, k_C)$, les intensités $(\lambda_i)_{1\leq i\leq C}$ suivent les lois conditionnelles du modèle considéré. Comme aucune information n'est disponible pour le nouveau centre, on suppose que son intensité λ suit la loi non conditionnée du modèle considéré (loi Gamma pour les modèles Γ-Poisson et $\mathcal{U}\Gamma$-Poisson, loi de Pareto pour le modèle Π-Poisson).

Ainsi, pour approcher la nouvelle intensité globale Λ par une loi Gamma, il suffit d'appliquer la proposition 2.2.3 avec :

- si l'on ferme le centre j, $m = \displaystyle\sum_{\substack{i=1 \\ i\neq j}}^{C} \mathbb{E}\left[\lambda_i \mid k_i\right],\ \sigma^2 = \sum_{\substack{i=1 \\ i\neq j}}^{C} \mathrm{Var}[\lambda_i \mid k_i]$,
- si l'on ouvre un centre, $m = \mathbb{E}\left[\lambda\right] + \displaystyle\sum_{i=1}^{C} \mathbb{E}\left[\lambda_i \mid k_i\right],\ \sigma^2 = \mathrm{Var}[\lambda] + \sum_{i=1}^{C} \mathrm{Var}[\lambda_i \mid k_i]$.

Appliquons cette méthode aux données de la section précédente. A $t_1 = 1.5$, l'estimateur du temps moyen de recrutement est $\widehat{\mu}^T \simeq 2.34$ années et la probabilité de finir l'essai dans les 3 ans est très proche de 1. En ne gardant ouverts que les 39 centres (sur 77) ayant recruté au moins 3 patients, l'estimateur du temps moyen de recrutement est 2.73 années et la probabilité de finir l'essai en moins de 3 ans est de 99.5%.

Chapitre 3

Perte de patients en phase de screening

Avant d'être inclus dans l'essai, un patient subit une phase de tests, appelée phase de screening, au cours de laquelle il est déterminé s'il est apte à recevoir le traitement médical testé dans l'essai. Si le patient ne passe pas la phase de tests, il quitte l'essai (on parle de "screening failure"). S'il la réussit, il entre à proprement parler dans l'essai. Un traitement parmi les traitements testés (par exemple, un médicament ou son placebo) lui est assigné aléatoirement : on dit qu'il est "randomisé" ou inclus.

Dans ce chapitre, nous proposons plusieurs modélisations de la perte de patients en phase de screening, et nous regardons son impact sur le processus d'inclusion de patients. Les patients arrivent dans l'essai suivant le modèle Gamma-Poisson de la partie 1. Le modèle de base (modèle 1) suppose qu'un patient a une probabilité p d'échouer à la phase de tests ; le modèle 2 utilise une approche bayésienne où cette probabilité dépend du centre et est distribuée suivant une loi Beta. Dans les modèles 3, 4, et 5, nous utilisons en plus le temps passé par chaque patient dans la phase de screening : ce temps est supposé être une variable aléatoire exponentielle de paramètre θ, où, suivant le modèle, θ peut être identique dans tous les centres ou distribué suivant une loi Gamma.

Pour chaque modèle, nous montrons comment estimer les paramètres par la méthode du maximum de vraisemblance, puis nous abordons le problème de la prédiction. L'applicabilité des modèles est testée sur données simulées.

Ce travail fait l'objet d'un article [12], en collaboration avec Vladimir Anisimov et Nicolas Savy, soumis à Statistics in Medecine en mars 2013. Cet article est proposé dans les pages suivantes. J'ai participé à l'élaboration des différents modèles (notamment les modèles 3-5) et ai programmé les simulations.

Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Statistical modelling of recruitment in multicentre clinical trials with patients' dropout

Vladimir V. Anisimov[a], Guillaume Mijoule[b,c], Nicolas Savy[b,d]*

This paper focuses on statistical modelling and prediction in clinical trials accounting for patients dropout. The recruitment model is based on a Poisson-gamma model introduced by Anisimov & Fedorov (2007), where the patients arrive at different centres according to Poisson processes with rates viewed as gamma-distributed random variables. Each patient can drop the study during some screening period. A few models of dropout are proposed. The technique for estimation of parameters and predicting the number of recruited patients over time and recruitment time is developed. Simulation results confirm the applicability of the technique. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: clinical trials; recruitment time; dropout; Bayesian statistics; Poisson process; maximum likelihood estimation

1. Introduction.

The problem of predicting patient recruitment and evaluating the recruitment time in clinical trials has been given much attention during the past years. However, till now the most of techniques used by pharma companies are still based on deterministic models and various *ad hoc* techniques. Using a Poisson process to describe the recruitment process is now an accepted approach (Senn, [1, 2], Carter et al. [3, 4]). However in real trials the recruitment rates in different centres vary and to mimic this variation it is natural to use a gamma distribution. Note that the use of Poisson-gamma mixtures for describing the variation of positive variables in modelling flows of various events has a long history, e.g. [5].

For modelling patient recruitment Anisimov and Fedorov [6] proposed to use a doubly stochastic Poisson process to take into consideration the variation in recruitment rates between different centres. This model, called as a Poisson-gamma model, assumes that the patients arrive at different centres according to Poisson processes with the rates viewed as independent gamma distributed random variables. In [6] the procedure of parameters estimation at interim stage and the technique for predicting future recruitment process using empirical Bayesian technique have been suggested. The model has been validated using data from a large number of real trials [7]. This model was developed further for predicting recruitment process at initial and interim stages [8], to account for the situations when the centres opening dates may not be known and assumed to be uniformly distributed in some intervals [9, 10], some centres can be closed or open in the future [11], for sensitivity analysis to parameter errors [10]. This model was also used as a basis for developing techniques for the analysis of the effects of unstratified and centre-stratified randomization [12], predictive event modelling [13] and predicting randomization process [11].

Here, we use a Poisson-gamma model as a starting point for the patient arrival process and develop technique further assuming that each patient can be lost during a screening process following patient arrival. Suppose that screening interval, which is the time that a patient has to stay initially in clinical centre to complete some preliminary tests for inclusion-exclusion criteria and to be randomized into the study, is a fixed positive number R which is the same for all patients. We assume that a patient can

[a] *Quintiles,Green Park, 500 Brook Drive, Reading, Berkshire, RG2 6UU, UK.*
[b] *University of Toulouse III, F-31073, Toulouse, France.*
[c] *University Paris Ouest Nanterre La Defense, 200 Avenue de la Republique, 92001 NANTERRE.*
[d] *Toulouse Institute of Mathematics, UMR C5583, F-31062, Toulouse, France.*

* *Correspondence to: University of Toulouse III, Toulouse Institute of Mathematics, UMR C5583, F-31062, Toulouse, France.*
E-mail: nicolas.savy@math.univ-toulouse.fr

be lost either at the start of the screening process with some probability or during the screening interval at some random time. We consider a few models for dropout.

In Section 2, we define the model. Section 3 provides the technique for estimating parameters at the interim stage, whereas Section 4 is devoted to the prediction of the recruitment time using the parameters estimated in Section 3. Finally Section 5 illustrates these results by simulation studies.

2. Models for recruitment with patients' dropout.

Consider a multicentre study with M clinical centres. Denote by u_i the opening date of centre i. The patients arrive at centres according to independent doubly stochastic Poisson processes $\left\{N_t^i,\, t\geq 0\,;\; 1\leq i\leq M\right\}$ with time-dependent rates of the form $\lambda_i(t)=\lambda_i\mathbf{1}_{t\geq u_i}$. The values $\{\lambda_i\,;\; 1\leq i\leq M\}$ are independent identically distributed random variables (i.i.d.r.v. for short) having a gamma distribution with some unknown parameters (α,β). Denote by $\mathrm{Ga}(\alpha,\beta)$ a gamma-distributed random variable with parameters (α,β) and probability density function

$$p_{\alpha,\beta}(x)=\beta^{\alpha}\Gamma(\alpha)^{-1}e^{-\beta x}x^{\alpha-1},\;\; x\geq 0.$$

Let $N_t=\sum_{i=1}^{M}N_t^i$ be the total number of patients arrived at all centres in time interval $[0,t]$ and let $\{t_j, j\geq 1\}$ be the increasing series of the jump times of N_t (respectively, $\{t_{i,j}, j\geq 1\}$ - jump times for $N_t^i,\; 1\leq i\leq M$).

Consider modelling of the dropout effect. Let us introduce the independent families of the i.i.d.r.v. $\{r_i\,;\; 1\leq i\leq M\}$ with values in $[0,1]$ and the parametric family of positive random variables $\{Z_{ij}(\theta_i), j\geq 1\,;\; 1\leq i\leq M\}$, where for each i and fixed θ_i the variables $\{Z_{ij}(\theta_i), j\geq 1\}$ have the same distribution, and the values $\{\theta_i\,;\; 1\leq i\leq M\}$ are i.i.d.r.v. with some distribution. Here $1-r_i$ stands for the probability of dropout upon arrival and $Z_{ij}(\theta_i)$ - for time of dropout in center i. Randomness in $\{r_i\}$ and $\{\theta_i\}$ reflects the variation in these values across different centres.

The patient j arriving at centre i at time $s=t_{i,j}$ may drop the study upon arrival with probability $1-r_i$ due to different initial tests. Otherwise, the patient drops the study at some random time $s+Z_{ij}(\theta_i)$ during the screening interval if $Z_{ij}(\theta_i)\leq R$. If neither one of these events happen, the patient is successfully randomized at time $s+R$ and registered to participate in the trial.

Consider centre i at some interim time t_1 and assume for simplicity that $R<t_1-u_i$. Suppose that the values (r_i,θ_i) are given. Then at time t_1 for patient j that has arrived at time $s=t_{i,j}<t_1$ there can be three events:

- patient is successfully screened and randomized at time $s+R$ with probability

$$p_i(s,t_1,r_i,\theta_i)=r_i\mathbb{P}\left(Z_{ij}(\theta_i)>R\right)\mathbf{1}_{s\leq t_1-R};$$

- patient is lost with probability

$$q_i(s,t_1,r_i,\theta_i)=1-r_i+r_i\mathbb{P}\left(Z_{ij}(\theta_i)\leq \min(R,t_1-s)\right);$$

- patient is still in screening process with probability

$$g_i(s,t_1,r_i,\theta_i)=r_i\mathbb{P}\left(Z_{ij}(\theta_i)>t_1-s\right)\mathbf{1}_{t_1-R\leq s<t_1}.$$

Let us define the independent families of indicators $\{\chi_{ij}(r_i), j\geq 1\,;\; 1\leq i\leq M\}$, where for a given r_i the variables $\{\chi_{ij}(r_i), j\geq 1\}$ are conditionally independent and for any $1\leq i\leq M$ and any $j\geq 1$,

$$\mathbb{P}\left(\chi_{ij}(r_i)=0\right)=1-\mathbb{P}\left(\chi_{ij}(r_i)=1\right)=r_i.$$

Now, for each centre i, at any time $t\geq 0$, we define three processes:

- **randomized patients:**
$$N_t^{i,R}=\mathrm{Card}\left\{j\;:\; u_i\leq t_{i,j}\leq t-R \text{ and } \chi_{ij}(r_i)=0,\;\; Z_{ij}(\theta_i)\geq R\right\},$$
- **lost patients:**
$$N_t^{i,L}=\mathrm{Card}\left\{j\;:\; u_i\leq t_{i,j}\leq t \text{ and } \left\{\chi_{ij}(r_i)=1\right\}\cup\left\{Z_{ij}(\theta_i)\leq \min(R,t-t_{i,j})\right\}\right\},$$
- **patients in screening process:**
$$N_t^{i,S}=N_t^i-N_t^{i,R}-N_t^{i,L}.$$

Finally, denote

$$N_t^X=\sum_{i=1}^{M}N_t^{i,X}\qquad \text{for } X:=R,L,S.$$

The trial stops as soon as the desired number of randomized patients N_R is reached, that is when $N_t^R = N_R$ - sample size. We consider several models for dropout.

Model 1. For all $1 \le i \le M$, $r_i = r$ where r is a fixed constant in $[0,1]$ and for all $j \ge 1$, $Z_{ij} = +\infty$. That means, we consider only dropout at the time upon arrival.

Model 2. The variables $\{r_i\ ;\ 1 \le i \le M\}$ are i.i.d.r.v. having a beta distribution with parameters (ψ_1, ψ_2), and $Z_{ij} = +\infty$ for all $1 \le i \le M$ and $j \ge 1$. This means, the variation in randomization probability between different centres is described using a beta distribution.

Model 3. For all $1 \le i \le M$, $r_i = r$ and the values $\{Z_{ij}(\cdot), j \ge 1\ ;\ 1 \le i \le M\}$ are i.i.d.r.v. having an exponential distribution with parameter θ (the same for all centres).

Model 4. For all $1 \le i \le M$, $r_i = r$ and the values $\{Z_{ij}(\cdot), j \ge 1\}$ given θ_i are i.i.d.r.v. having an exponential distribution with parameter θ_i, where the values $\{\theta_i\ ;\ 1 \le i \le M\}$ are i.i.d.r.v. having a gamma distribution with parameters (α_2, β_2).

Model 5. The variables $\{r_i\ ;\ 1 \le i \le M\}$ are i.i.d.r.v. having a beta distribution with some parameters (ψ_1, ψ_2). The values $\{Z_{ij}(\cdot), j \ge 1\ ;\ 1 \le i \le M\}$ given θ_i are i.i.d.r.v. having an exponential distribution with parameter θ_i, where the values $\{\theta_i\ ;\ 1 \le i \le M\}$ are i.i.d.r.v. having a gamma distribution with parameters (α_2, β_2).

As we see, model 5 is the most advanced model that accounts for the variation in the probability of dropout upon patient arrival and in the distribution of dropout time during screening process across clinical centres. For each model we consider the procedure of estimating unknown parameters and predicting in time the future process of randomized patients and the total recruitment time.

In models 1 and 2, the actual time of patient dropout during screening process is not taken into account. Thus, in the estimation procedure at some interim time t_1, it is enough to know for each centre i the number of recruited and randomized patients (that is, $\{N_{t_1}^i, N_{t_1}^{i,R}\}$). On the other hand, in models 3, 4 and 5, we assume that full data is available: for each patient it is known the arrival and dropout (or randomization) time. If the dropout time were unknown, it would be impossible to distinguish between a patient lost upon arrival or during screening process, and the distinction within the model would be irrelevant.

3. Parameters' estimation at interim time.

We use a Poisson-gamma recruitment model for modelling patients recruitment process [6, 8, 9, 11]. That means, recruitment rates λ_i are viewed as a sample from a gamma distributed population with some unknown parameters (α, β).

Let t_1 be some interim time and assume for simplicity that $\tau_i = t_1 - u_i \ge R$. In this case the number of recruited patients, $n_i = N_{t_1}^i$, as a random variable, has a negative binomial distribution with parameters $(\alpha, \tau_i/(\beta + \tau_i))$ [14, p.199] (or a Poisson-gamma distribution with parameters (α, β, τ_i) [15, p.119]). Denote by $\mu = \mathbb{E}[\lambda]$ the mean rate and $\sigma^2 = \mathbb{V}[\lambda]$ and notice a useful relation between parameters: $\mu = \alpha/\beta$, $\sigma^2 = \alpha/\beta^2$ and $\alpha = \mu^2/\sigma^2$, $\beta = \mu/\sigma^2$. Using the parametrization in terms of mean rate let us introduce a probability distribution for negative binomial distribution (Poisson-gamma) with parameters (α, π):

$$\mathrm{NegBin}(k; \alpha, \pi) = \frac{\Gamma(\alpha + k)}{k!\Gamma(\alpha)} \pi^k (1-\pi)^\alpha.$$

Then in centre i,

$$\mathbb{P}(n_i = k) = \mathbb{E}\left[e^{-\lambda\tau}\frac{(\lambda\tau)^k}{k!}\right] = \mathrm{NegBin}\left(k; \alpha, \frac{\mu\tau}{\alpha + \mu\tau}\right). \tag{1}$$

3.1. Model 1.

Let r_i be the probability of randomization in centre i. Denote by $\mathrm{Bin}(n, \pi)$ a binomial random variable with parameters (n, π):

$$\mathbb{P}(\mathrm{Bin}(n,\pi) = k) = \mathrm{Bin}(k; n, \pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k}, \quad 0 \le k \le n.$$

Assume for simplicity that there is no screening delay. Then in centre i the number of randomized patients k_i has a binomial distribution with parameters (n_i, r_i). Suppose that $r_i \equiv r$ (probability of randomization is the same). Then the log-likelihood function can be written in the form:

$$\mathcal{L}_1(\alpha, \mu, r) = \sum_{i=1}^M \ln\left[\mathrm{NegBin}\left(n_i; \alpha, \frac{\mu\tau_i}{\alpha + \mu\tau_i}\right)\right] + \sum_{i=1}^M \ln[\mathrm{Bin}(k_i; n_i, r)].$$

As we see, the parameter r is separated from (α, μ) and $\mathcal{L}_1(\alpha, \mu, r)$ can be re-written in the form: $\mathcal{L}_1(\alpha, \mu, r) = \mathcal{L}_{1,1}(\alpha, \mu) + \mathcal{L}_{1,2}(r)$, with

$$\mathcal{L}_{1,1}(\alpha, \mu) = \sum_{i=1}^M \ln\Gamma(n_i + \alpha) - M\ln\Gamma(\alpha) + N_1(\ln\mu - \ln\alpha) - \sum_{i=1}^M (n_i + \alpha)\ln(1 + \mu\tau_i/\alpha) + C, \tag{2}$$

and

$$\mathcal{L}_{1,2}(r) = \sum_{i=1}^{M} \Big[k_i \ln r + (n_i - k_i) \ln(1-r) \Big] + C.$$

where C is some generic constant independent of the parameters which varies in different lines, and $N_1 = \sum_{i=1}^{M} n_i$ is the total number of recruited patients up to time t_1.

Taking derivative in r it is easy to calculate that maximum likelihood estimator

$$\widehat{r} = \Big(\sum_{i=1}^{M} n_i \Big)^{-1} \sum_{i=1}^{M} k_i. \tag{3}$$

Note that if there is a screening delay, then such patients that entered screening process but the results of their screening procedure are unknown yet should be excluded in the calculations of probability of randomization, otherwise this probability will be underestimated. Therefore, instead of n_i we should count $\widetilde{n}_i$, the number of patients with known screening results. Parameters (α, μ) can be estimated using for function $\mathcal{L}(\alpha, \mu)$ a two-dimensional optimization procedure. Note that the estimator of the variance is $\widehat{\sigma}^2 = \widehat{\mu}^2/\widehat{\alpha}$.

3.2. Model 2.

Assume now that r_i can vary between different centres and we describe this variation using a beta distribution with some unknown parameters (ψ_1, ψ_2). Denote by $\mathrm{Beta}(\psi_1, \psi_2)$ a beta-distributed random variable with p.d.f.

$$p_\beta(x; \psi_1, \psi_2) = x^{\psi_1 - 1}(1-x)^{\psi_2 - 1} / \mathcal{B}(\psi_1, \psi_2), \quad x \in]0,1[,$$

where $\mathcal{B}(\psi_1, \psi_2) = \int_0^1 x^{\psi_1 - 1}(1-x)^{\psi_2 - 1}\, \mathrm{d}\, x$ is a beta function.
Notice the fact that if $r = \mathrm{Beta}(\psi_1, \psi_2)$, then a doubly stochastic binomial variable $\mathrm{Bin}(n, r)$ has a beta-binomial distribution:

$$\mathbb{P}\left(\mathrm{Bin}(n; \mathrm{Beta}(\psi_1, \psi_2)) = k\right) = \binom{n}{k} \frac{\mathcal{B}(k + \psi_1, n - k + \psi_2)}{\mathcal{B}(\psi_1, \psi_2)}.$$

Note that for beta distribution

$$\mathbb{E}\left[\mathrm{Beta}(\psi_1, \psi_2)\right] = \frac{\psi_1}{\psi_1 + \psi_2}, \qquad \mathbb{V}\left[\mathrm{Beta}(\psi_1, \psi_2)\right] = \frac{\psi_1 \psi_2}{(\psi_1 + \psi_2)^2 (1 + \psi_1 + \psi_2)},$$

and for beta-binomial

$$\mathbb{E}\left[\mathrm{Bin}(n; \mathrm{Beta}(\psi_1, \psi_2))\right] = \frac{n\psi_1}{\psi_1 + \psi_2}, \qquad \mathbb{V}\left[\mathrm{Bin}(n; \mathrm{Beta}(\psi_1, \psi_2))\right] = \frac{n\psi_1\psi_2(n + \psi_1 + \psi_2)}{(\psi_1 + \psi_2)^2 (1 + \psi_1 + \psi_2)}.$$

Given data $\{n_i, k_i, \tau_i\}$, the log-likelihood function can be written in the form:

$$\mathcal{L}_2(\alpha, \mu, \psi_1, \psi_2) = \mathcal{L}_{2,1}(\alpha, \mu) + \mathcal{L}_{2,2}(\psi_1, \psi_2),$$

where $\mathcal{L}_{2,1} = \mathcal{L}_{1,1}$ given by (2), and

$$\mathcal{L}_{2,2}(\psi_1, \psi_2) = \sum_{i=1}^{M} \ln \mathcal{B}(k_i + \psi_1, n_i - k_i + \psi_2) - M \ln \mathcal{B}(\psi_1, \psi_2) + C.$$

Parameters (ψ_1, ψ_2) can be estimated using for function $\mathcal{L}_2(\psi_1, \psi_2)$ a two-dimensional optimization procedure.

Denote the parameter estimators by $(\widehat{\alpha}, \widehat{\mu})$ and $(\widehat{\psi}_1, \widehat{\psi}_2)$. Consider now a Bayesian procedure of adjusting (re-estimating) parameters in each centre given the values n_i and k_i in this centre similar to [6, 11]. As λ_i has a prior gamma distribution with parameters (α, β), given data $\{n_i, \tau_i\}$ and using Bayesian formula, one can calculate that the posterior distribution of λ_i is also a gamma distribution with parameters $(\alpha + n_i, \beta + \tau_i)$. Correspondingly, if r_i has a prior beta distribution with parameters (ψ_1, ψ_2), then, given data $\{n_i, k_i\}$, one can calculate that the posterior distribution of r_i is also a beta distribution with parameters $(\psi_1 + k_i, \psi_2 + n_i - k_i)$.

Therefore, given data, we can represent the posterior estimators of the rates and the probabilities of randomization in each centre in the form:

$$\widehat{\lambda}_i = \mathrm{Ga}(\widehat{\alpha} + n_i, \widehat{\beta} + \tau_i) \quad \text{and} \quad \widehat{r}_i = \mathrm{Beta}(\widehat{\psi}_1 + k_i, \widehat{\psi}_2 + n_i - k_i), \tag{4}$$

where given data $\widehat{\lambda}_i$ and $\widehat{r}_i$, $i = 1, .., M$ are independent.

3.3. Models 3, 4, 5.

For calculation of the likelihood function we need to account for that the variables $\{\lambda_i, r_i, \theta_i\ ;\ 1 \le i \le M\}$ are independent and in general are some random functions. As we assume that the rates $\{\lambda_i\ ;\ 1 \le i \le M\}$ are viewed as i.i.d.r.v. having a gamma distribution with parameters (α, β), then for any $1 \le i \le M$ the variable N_t^i has a negative binomial distribution $\text{NegBin}\left(k; \alpha, \frac{\mu\tau_i}{\alpha+\mu\tau_i}\right)$ (see (1)). The types of distributions of r_i and θ_i are specified by the types of models 3-5. In these models, we assume that more information is available, so we can estimate parameters of dropout times $\{Z_{ij}, j \ge 1\ ;\ 1 \le i \le M\}$. Moreover, the calculation of the posterior distributions of parameters is not straightforward if we use data as in models 1 and 2. At time t_1, we observe patients arrival times $\{t_{i,j} \le t_1\ ;\ j \ge 1\ ,\ 1 \le i \le M\}$ and the last time $\{s_{i,j}, j \ge 1\ ;\ 1 \le i \le M\}$ they were in the screening process (i.e $t_{i,j} \le s_{i,j} \le t_1$) . This means we also observe $\{\min(Z_{ij}, R, t_1 - t_{i,j}) \vee 0, j \ge 1\ ;\ 1 \le i \le M\}$, where $a \vee b = \max(a, b)$.

Conditioning on parameters $\{\theta_i, r_i\ ;\ 1 \le i \le M\}$, we can write a general expression for the likelihood

$$\mathbf{L}\left[(t_{i,j}); (s_{i,j})\right] = \exp\left[\mathcal{L}_{1,1}(\alpha, \mu)\right] \times \prod_{i=1}^{M} \mathbb{E}\left[\mathbf{L}_2\left(\theta_i, r_i; (t_{i,j}), (s_{i,j})\right)\right] \tag{5}$$

where $\mathcal{L}_{1,1}$ is given in (2), and

$$\mathbf{L}_2\left(\theta_i, r_i; (t_{i,j}), (s_{i,j})\right) = \prod_{\mathcal{D}_1}(1 - r_i) \prod_{\mathcal{D}_2} r_i \exp(-\theta_i(s_{i,j} - t_{i,j})) \times \prod_{\mathcal{D}_3} r_i\theta_i \exp\left(-\theta_i(s_{i,j} - t_{i,j})\right),$$

where $\mathcal{D}_1 = \bigcup_{i=1}^M \mathcal{D}_1^i$, $\mathcal{D}_2 = \bigcup_{i=1}^M \mathcal{D}_2^i$, $\mathcal{D}_3 = \bigcup_{i=1}^M \mathcal{D}_3^i$ and for any $1 \le i \le M$,

$$\begin{aligned}
\mathcal{D}_1^i &= \{j \ge 1, \text{ s.t. } \quad s_{i,j} = t_{i,j}\},\\
\mathcal{D}_2^i &= \{j \ge 1, \text{ s.t. } \quad s_{i,j} = (t_{i,j} + R) \wedge t_1\},\\
\mathcal{D}_3^i &= \{j \ge 1, \text{ s.t. } \quad t_{i,j} < s_{i,j} < (t_{i,j} + R) \wedge t_1\}.
\end{aligned}$$

In (5), the expectation is taken when θ_i and r_i vary according to their respective distributions defined by models 3-5.

Denote by $m_i = \text{Card}\left(\mathcal{D}_3^i\right)$ the number of patients lost in the middle of screening process in centre i, $T_i = \sum_j (s_{i,j} - t_{i,j})$ - sum of screening durations in centre i, and $\tilde{k}_i = \text{Card}\left(\mathcal{D}_2^i\right) + \text{Card}\left(\mathcal{D}_3^i\right)$ the number of patients that are not lost immediately upon arrival. Then $\mathbf{L}_2$ can be rewritten as

$$\mathbf{L}_2\left(\theta_i, r_i; (t_{i,j}), (s_{i,j})\right) = (1 - r_i)^{n_i - \tilde{k}_i} r_i^{\tilde{k}_i} \theta_i^{m_i} \exp\left(-\theta_i T_i\right). \tag{6}$$

3.3.1. Model 3. Given data $\{t_{i,j}, s_{i,j}, \tau_i, j \ge 1\ ;\ 1 \le i \le M\}$ and using (5) and (6), the log-likelihood function can be written in the form:

$$\mathcal{L}_3(\alpha, \mu, r, \theta) = \mathcal{L}_{3,1}(\alpha, \mu) + \sum_{i=1}^{M} \left[(n_i - \tilde{k}_i)\ln(1 - r) + \tilde{k}_i \ln r + m_i \ln\theta - \theta T_i\right],$$

where $\mathcal{L}_{3,1} = \mathcal{L}_{1,1}$ is given in (2) and $(\widehat{\alpha}, \widehat{\mu})$ can be calculated using a two-dimensional optimization procedure for function $\mathcal{L}_{1,1}(\alpha, \mu)$. Taking derivatives in r and θ we get the maximum likelihood estimators

$$\widehat{r} = \left(\sum_{i=1}^{M} n_i\right)^{-1} \sum_{i=1}^{M} \tilde{k}_i \quad \text{and} \quad \widehat{\theta} = \left(\sum_{i=1}^{M} T_i\right)^{-1} \sum_{i=1}^{M} m_i. \tag{7}$$

3.3.2. Model 4. Using (5) and (6), the log-likelihood function can be written in the form:

$$\mathcal{L}_4(\alpha, \mu, r, \alpha_2, \beta_2) = \mathcal{L}_{4,1}(\alpha, \mu) + \sum_{i=1}^{M} \left[(n_i - \tilde{k}_i)\ln(1 - r) + \tilde{k}_i \ln r\right] + \mathcal{L}_{4,2}(\alpha_2, \beta_2)$$

where

$$\mathcal{L}_{4,2}(\alpha_2, \beta_2) = \sum_{i=1}^{M} \left[\ln\Gamma(m_i + \alpha_2) - \ln\Gamma(\alpha_2) + \alpha_2 \ln\beta_2 - (m_i + \alpha_2)\ln(\beta_2 + T_i)\right] \tag{8}$$

and $\mathcal{L}_{4,1} = \mathcal{L}_{1,1}$ is given by (2). $(\widehat{\alpha}, \widehat{\mu})$ and $(\widehat{\alpha}_2, \widehat{\beta}_2)$ are calculated numerically, and

$$\widehat{r} = \left(\sum_{i=1}^{M} n_i\right)^{-1} \sum_{i=1}^{M} \tilde{k}_i$$

Consider the Bayesian procedure of re-estimating parameters in each centre given data $\{n_i, \tau_i, (t_{i,j}), (s_{i,j}), j \ge 1\ ;\ 1 \le i \le M\}$. Then θ_i has a prior gamma distribution with parameters (α_2, β_2). Given data $\{m_i, T_i\ ;\ 1 \le i \le M\}$ (the knowledge of $\{t_{i,j}\}$'s and

$\{s_{i,j}\}$'s is not necessary here) and using Bayesian formula, we obtain that the posterior distribution of θ_i is a gamma distribution with parameters $(\alpha_2 + m_i, \beta_2 + T_i)$. Thus,

$$\begin{aligned}\widehat{\lambda}_i &= \mathrm{Ga}(\widehat{\alpha} + n_i, \widehat{\beta} + \tau_i),\\ \widehat{r} &= \Big(\sum_{i=1}^{M} n_i\Big)^{-1} \sum_{i=1}^{M} \tilde{k}_i,\\ \widehat{\theta}_i &= \mathrm{Ga}(\widehat{\alpha}_2 + m_i, \widehat{\beta}_2 + T_i), \quad i = 1, \dots, M.\end{aligned} \tag{9}$$

3.3.3. Model 5. Using (5) and (6), the log-likelihood function can be written in the form:

$$\mathcal{L}_5(\alpha, \mu, \psi_1, \psi_2, \alpha_2, \beta_2) = \mathcal{L}_{5,1}(\alpha, \mu) + \mathcal{L}_{5,2}(\alpha_2, \beta_2) + \mathcal{L}_{5,3}(\psi_1, \psi_2)$$

$$\mathcal{L}_{5,3}(\psi_1, \psi_2) = \sum_{i=1}^{M} \Big[\ln \mathcal{B}(\tilde{k}_i + \psi_1, n_i - \tilde{k}_i + \psi_2) - \ln \mathcal{B}(\psi_1, \psi_2) \Big],$$

$\mathcal{L}_{5,1} = \mathcal{L}_{1,1}$ is given by (2) and $\mathcal{L}_{5,2} = \mathcal{L}_{4,2}$ is given by (8).

Parameters (α, μ), (ψ_1, ψ_2) and (α_2, β_2) can be estimated using two-dimensional optimization procedures for corresponding functions $\mathcal{L}(\cdot)$.

Then, similar to (9), as λ_i has a prior gamma distribution with parameters (α, β), given data $\{n_i, \tau_i \,;\, 1 \le i \le M\}$, the posterior distribution of λ_i is a gamma distribution with parameters $(\alpha + n_i, \beta + \tau_i)$. Correspondingly, θ_i has a prior gamma distribution with parameters (α_2, β_2), thus, given data $\{m_i, T_i \,;\, 1 \le i \le M\}$, the posterior distribution of θ_i is a gamma distribution with parameters $(\alpha_2 + m_i, \beta_2 + T_i)$. Furthermore, r_i has a prior beta distribution with parameters (ψ_1, ψ_2), thus, given $\{n_i, \tilde{k}_i \,;\, 1 \le i \le M\}$, the posterior distribution of r_i is a beta distribution with parameters $(\psi_1 + \tilde{k}_i, \psi_2 + n_i - \tilde{k}_i)$. To sum up, the posterior distributions of the rates of inclusion, probabilities of instantaneous dropout and rate of dropout are

$$\begin{aligned}\widehat{\lambda}_i &= \mathrm{Ga}(\widehat{\alpha} + n_i, \widehat{\beta} + \tau_i),\\ \widehat{r}_i &= \mathrm{Beta}(\widehat{\psi}_1 + \tilde{k}_i, \widehat{\psi}_2 + n_i - \tilde{k}_i),\\ \widehat{\theta}_i &= \mathrm{Ga}(\widehat{\alpha}_2 + m_i, \widehat{\beta}_2 + T_i), \quad i = 1, \dots, M.\end{aligned} \tag{10}$$

4. Prediction.

4.1. Model 1.

Given data $\{(n_i, \tau_i) \,;\, 1 \le i \le M\}$ at interim time t_1, the predicted number of recruited patients $\widehat{n}_i(t)$ in center i is a Poisson-gamma process with posterior rate $\widehat{\lambda}_i = \mathrm{Ga}(\widehat{\alpha} + n_i, \widehat{\beta} + \tau_i)$ (see (4) and [11]). Consider now predicting the number of patients that will be randomized. Denote by $(\widehat{\mu}, \widehat{\beta}, \widehat{r})$ the estimators of parameters (μ, β, r). Recall that R is a screening delay. Let ν_i be the number of patients entered screening stage at centre i in the interval $[t_1 - R, t_1]$ and k_i be the total number of randomized patients up to time t_1.

Then, given ν_i, the number of patients that will be randomized in the interval $[t_1, t_1 + R]$ is a binomial random variable $\mathrm{Bin}(\nu_i, r)$ and the times when these patients are randomized are uniformly distributed in $[t_1, t_1 + R]$. The predicted number of randomized patients in $[t_1, t_1 + R]$ is $\mathrm{Bin}(\nu_i, \widehat{r})$, where $\widehat{r}$ is defined in (3). The number of patients randomized after time $t_1 + R$ can be considered as thinning of the process $N^i_\cdot$ with probability r. Let Π_a stand for a Poisson process with rate a. Then, for any $t > t_1 + R$, the predicted process of the number of randomized patients in centre i, $\{\widehat{k}_i(t),\ t \ge t_1 + R\}$, is developing as a Poisson process with rate $\widehat{r}\widehat{\lambda}_i$. Thus,

$$\widehat{k}_i(t) = k_i + \mathrm{Bin}(\nu_i, \widehat{r}) + \Pi_{\widehat{r}\widehat{\lambda}_i}(t - t_1 - R). \tag{11}$$

Note that for a random rate λ, $\mathbb{E}\left[\Pi_\lambda(t)\right] = t\mathbb{E}\left[\lambda\right]$, $\mathbb{V}\left[\Pi_\lambda(t)\right] = t\mathbb{E}\left[\lambda\right] + t^2\mathbb{E}\left[\lambda\right]$. Therefore, given current data,

$$\mathbb{E}\left[\widehat{k}_i(t) \mid \text{data}\right] = k_i + \nu_i\widehat{r} + \widehat{r}\,(t - t_1 - R)\mathbb{E}\left[\widehat{\lambda}_i\right],$$

where $\widehat{r}$ is given by (3), and

$$\mathbb{E}\left[\widehat{\lambda}_i\right] = \frac{\alpha + n_i}{\beta + \tau_i}, \qquad \mathbb{V}\left[\widehat{\lambda}_i\right] = \frac{\alpha + n_i}{(\beta + \tau_i)^2}. \tag{12}$$

As for given $\{n_i, k_i\}$ the posterior predictors $\widehat{\lambda}_i$ are independent of $\widehat{r}$, then the variance is

$$\mathbb{V}\left[\widehat{k}_i(t) \mid \text{data}\right] = \nu_i\widehat{r}(1 - \widehat{r}) + \widehat{r}(t - t_1 - R)\mathbb{E}\left[\widehat{\lambda}_i\right] + \widehat{r}^2(t - t_1 - R)^2\mathbb{V}\left[\widehat{\lambda}_i\right]. \tag{13}$$

Denote by $\widehat{k}(t) = \sum_{i=1}^{M} \widehat{k}_i(t)$ the total number of randomized patients at time $t > t_1 + R$ and put $K_2 = \sum_{i=1}^{M} k_i$. Then

$$\mathbb{E}\left[\widehat{k}(t) \mid \text{data}\right] = K_2 + \widehat{r} \sum_{i=1}^{M} \nu_i + \widehat{r}\,(t - t_1 - R) \sum_{i=1}^{M} \frac{\alpha + n_i}{\beta + \tau_i},$$

and

$$\mathbb{V}\left[\widehat{k}(t) \mid \text{data}\right] = \widehat{r}(1-\widehat{r})\sum_{i=1}^{M}\nu_i + \widehat{r}\,(t-t_1-R)\sum_{i=1}^{M}\frac{\alpha+n_i}{\beta+\tau_i} + \widehat{r}^2(t-t_1-R)^2\sum_{i=1}^{M}\frac{\alpha+n_i}{(\beta+\tau_i)^2}.$$

At large enough M ($M > 10$) we can use these expressions to create $(1-\delta)$-predictive bounds for $\widehat{k}(t)$ using a normal approximation similar to [11]. It is also possible to consider a joint distribution of two-component process $\{(\widehat{n}_i(t), \widehat{k}_i(t)),\ t \geq t_1 + R\}$. Given data at time t_1, in interval $[t_1, t_1+R]$ these processes are independent, as for $t \in [t_1, t_1+R]$ the process $\widehat{k}_i$ depends only on the data before time t_1. For $t > t_1 + R$, the process $\{\widehat{k}_i(t),\ t \geq t_1 + R\}$ can be represented as thinning of the process $\{\widehat{n}_i(t-R),\ t \geq t_1\}$ with probability $\widehat{r}$.

4.2. Model 2.

In this case for $t > t_1 + R$ we can use for predictive process $\widehat{k}_i(t)$ in centre i a formula similar to (11),

$$\widehat{k}_i(t) = k_i + \text{Bin}(\nu_i, \widehat{r}_i) + \Pi_{\widehat{r}_i\widehat{\lambda}_i}(t-t_1-R),$$

where the posterior estimators $\widehat{\lambda}_i$ and $\widehat{r}_i$ are given by (4). The mean and variance of λ_i are calculated in (12), and

$$\mathbb{E}\left[\widehat{r}_i \mid \text{data}\right] = \frac{\psi_1+k_i}{\psi_1+\psi_2+n_i}, \qquad \mathbb{V}\left[\widehat{r}_i \mid \text{data}\right] = \frac{(\psi_1+k_i)(\psi_2+n_i-k_i)}{(\psi_1+\psi_2+n_i)^2(1+\psi_1+\psi_2+n_i)},$$

where instead of (ψ_1, ψ_2) we should substitute $(\widehat{\psi}_1, \widehat{\psi}_2)$. As data $\{n_i, k_i\,;\, 1 \leq i \leq M\}$ are given, the posterior predictors $\widehat{\lambda}_i$ and $\widehat{r}_i$ are independent. Note that for a random probability r, $\mathbb{E}\left[\text{Bin}(n,r)\right] = n\mathbb{E}\left[r\right]$, and

$$\mathbb{V}\left[\text{Bin}(n,r)\right] = \mathbb{E}\left[\mathbb{V}\left[\text{Bin}(n,r) \mid r\right]\right] + \mathbb{V}\left[\mathbb{E}\left[\text{Bin}(n,r) \mid r\right]\right] = n\mathbb{E}\left[r(1-r)\right] + n^2\mathbb{V}\left[r\right].$$

Therefore,

$$\mathbb{E}\left[\widehat{k}_i(t) \mid \text{data}\right] = k_i + \nu_i\mathbb{E}\left[\widehat{r}_i\right] + (t-t_1-R)\mathbb{E}\left[\widehat{r}_i\right]\mathbb{E}\left[\widehat{\lambda}_i\right],$$

and

$$\mathbb{V}\left[\widehat{k}_i(t) \mid \text{data}\right] = \nu_i\mathbb{E}\left[\widehat{r}_i(1-\widehat{r}_i)\right] + \nu_i^2\mathbb{V}\left[\widehat{r}_i\right] + (t-t_1-R)\mathbb{E}\left[\widehat{r}_i\right]\mathbb{E}\left[\widehat{\lambda}_i\right] + (t-t_1-R)^2\mathbb{V}\left[\widehat{r}_i\widehat{\lambda}_i\right],$$

where we can use the formula

$$\mathbb{V}\left[\widehat{r}_i\widehat{\lambda}_i\right] = \mathbb{E}\left[\widehat{\lambda}_i^2\right]\mathbb{V}\left[\widehat{r}_i\right] + \left(\mathbb{E}\left[\widehat{r}_i\right]\right)^2\mathbb{V}\left[\widehat{\lambda}_i\right].$$

Using relations (12), (13) we can easy calculate the mean and variance of the global process $\widehat{k}(t)$ and use these characteristics at large enough M to create $(1-\delta)$-predictive bounds based on the normal approximation.

4.3. Models 3, 4, 5.

For models 3, 4, and 5, since we have more information at time t_1, the patients with unknown screening outcome correspond to the case $t_1 - R < t_{i,j} \leq t_1$ and $Y_{ij} > t_1 - t_{i,j}$. For $1 \leq i \leq M$, let Ω_i be the corresponding set of indices

$$\Omega_i = \{j \in \mathbb{N} : t_1 - R < t_{i,j} \leq t_1 \text{ and } Y_{ij} > t_1 - t_{i,j}\}$$

and $\nu_i = \text{Card}(\Omega_i)$. Conditionally on θ_i, the probability of this patient to be randomized in $[t_1, t_1+R]$ is

$$\mathbb{P}\left[Y_{ij} \geq R | Y_{ij} > t_1 - t_{i,j}; \theta_i\right] = e^{-\theta_i(t_{i,j}+R-t_1)}$$

Given data at t_1 and θ_i, the number of randomized patients between t_1 and t_1+R in centre i is the sum of ν_i independent Bernoulli r.v. with probabilities $e^{-\theta_i(R-t_1+t_{i,j})}$ denoted as $\text{Ber}\left(e^{-\theta_i(t_{i,j}+R-t_1)}\right)$. Thus, the predictive process $\{\widehat{k}_i(t), t \geq t_1 + R\}$ for $t > t_1 + R$ can be written as

$$\widehat{k}_i(t) = k_i + \Pi_{\widehat{p}_i\widehat{\lambda}_i}(t-t_1-R) + \sum_{j\in\Omega_i}\text{Ber}\left(e^{-\theta_i(t_{i,j}+R-t_1)}\right). \tag{14}$$

The probability of non-dropout is $\widehat{p}_i = \widehat{r}_i\exp(-\widehat{\theta}_i R)$.

4.3.1. Model 3. In this case $\widehat{r}_i \equiv \widehat{r}$ and $\widehat{\theta}_i \equiv \widehat{\theta}$, where $(\widehat{r}, \widehat{\theta})$ are given in (7). Thus, using (14) we get

$$\widehat{k}_i(t) = k_i + \Pi_{\widehat{p}\widehat{\lambda}_i}(t-t_1-R) + \sum_{j\in\Omega_i}\text{Ber}\left(e^{-\widehat{\theta}(t_{i,j}+R-t_1)}\right),$$

where $\widehat{p} = \widehat{r}e^{-\widehat{\theta}R}$. Therefore,

$$\mathbb{E}\left[\widehat{k}_i(t) \mid \text{data}\right] = k_i + (t-t_1-R)\widehat{r}e^{-\widehat{\theta}R}\mathbb{E}\left[\widehat{\lambda}_i\right] + \sum_{j\in\Omega_i}e^{-\widehat{\theta}(t_{i,j}+R-t_1)},$$

$$\mathbb{V}\left[\widehat{k}_i(t) \mid \text{data}\right] = (t-t_1-R)^2\widehat{r}^2e^{-2\widehat{\theta}R}\mathbb{V}\left[\widehat{\lambda}_i\right] + (t-t_1-R)\widehat{r}e^{-\widehat{\theta}R}\mathbb{E}\left[\widehat{\lambda}_i\right] + \sum_{j\in\Omega_i}e^{-\widehat{\theta}(t_{i,j}+R-t_1)}(1-e^{-\widehat{\theta}(t_{i,j}+R-t_1)}),$$

where $\mathbb{E}\left[\widehat{\lambda}_i\right]$ and $\mathbb{V}\left[\widehat{\lambda}_i\right]$ are given in (12).

4.3.2. Model 4. The predictive process in centre i is

$$\widehat{k}_i(t) = k_i + \Pi_{\widehat{p}_i\,\widehat{\lambda}_i}(t - t_1 - R) + \sum_{j\in\Omega_i} \mathrm{Ber}\left(e^{-\widehat{\theta}_i(t_{i,j}+R-t_1)}\right),$$

where $\widehat{p}_i = \widehat{r}\exp(-\widehat{\theta}_i R)$, and $\widehat{\lambda}_i, \widehat{r}$ and $\widehat{\theta}_i$ are given by (9). Denote by $F_i(s)$ a Laplace transformation of $\widehat{\theta}_i$:

$$F_i(s) = \mathbb{E}\left[e^{-\widehat{\theta}_i s}\right] = \left[1 + s/(\widehat{\beta}_2 + T_i)\right]^{-\widehat{\alpha}_2 - m_i}, \qquad s \geq 0$$

Then,

$$\mathbb{E}\left[\widehat{k}_i(t) \mid \text{data}\right] = k_i + (t - t_1 - R)\widehat{r}F_i(R)\mathbb{E}\left[\widehat{\lambda}_i\right] + \sum_{j\in\Omega_i} F_i(t_{i,j} + R - t_1),$$

$$\mathbb{V}\left[\widehat{k}_i(t) \mid \text{data}\right] = (t - t_1 - R)^2\widehat{r}^2\mathbb{V}\left[e^{-\widehat{\theta}_i R}\widehat{\lambda}_i\right] + (t - t_1 - R)\widehat{r}\mathbb{E}\left[e^{-\widehat{\theta}_i R}\right]\mathbb{E}\left[\widehat{\lambda}_i\right] + \mathbb{V}\left[\sum_{j\in\Omega_i} \mathrm{Ber}\left(e^{-\widehat{\theta}_i(t_{i,j}+R-t_1)}\right)\right],$$

where $\mathbb{E}\left[\widehat{\lambda}_i\right]$ and $\mathbb{V}\left[\widehat{\lambda}_i\right]$ are given in (12), $\mathbb{V}\left[e^{-\widehat{\theta}_i R}\widehat{\lambda}_i\right] = F_i(2R)\mathbb{E}\left[\widehat{\lambda}_i^2\right] - (F_i(R)\mathbb{E}\left[\widehat{\lambda}_i\right])^2$, $\mathbb{E}\left[e^{-\widehat{\theta}_i R}\right] = F_i(R)$, and straightforward calculations show that, by denoting $\Delta_{ij} = R - t_1 - t_{i,j}$,

$$\mathbb{V}\left[\sum_{j\in\Omega_i} \mathrm{Ber}\left(e^{-\widehat{\theta}_i(t_{i,j}+R-t_1)}\right)\right] = \sum_{j\in\Omega_i} F_i(\Delta_{ij}) - F_i(2\Delta_{ij}) + \sum_{(j_1,j_2)\in\Omega_i^2} F_i(\Delta_{ij_1} + \Delta_{ij_2}) - F_i(\Delta_{ij_1})F_i(\Delta_{ij_2}).$$

4.3.3. Model 5. In this case $\widehat{r}_i$ is also random and given in (10). In a similar way as in model 4, we can write

$$\mathbb{E}\left[\widehat{k}_i(t) \mid \text{data}\right] = k_i + (t - t_1 - R)\mathbb{E}\left[\widehat{r}_i\right]F_i(R)\mathbb{E}\left[\widehat{\lambda}_i\right] + \sum_{j\in\Omega_i} F_i(t_{i,j} + R - t_1),$$

$$\mathbb{V}\left[\widehat{k}_i(t) \mid \text{data}\right] = (t - t_1 - R)^2\mathbb{V}\left[\widehat{r}_i e^{-\widehat{\theta}_i R}\widehat{\lambda}_i\right] + (t - t_1 - R)\mathbb{E}\left[\widehat{r}_i\right]\mathbb{E}\left[e^{-\widehat{\theta}_i R}\right]\mathbb{E}\left[\widehat{\lambda}_i\right] + \mathbb{V}\left[\sum_{j\in\Omega_i} \mathrm{Ber}\left(e^{-\widehat{\theta}_i(t_{i,j}+R-t_1)}\right)\right],$$

and use previous formulae for calculation the expectations and variances of different variables in this expression.

5. Simulation study.

5.1. Data generation procedure.

We simulate data according to model 5. One has to generate the rates λ_i of the centres according to a $\Gamma(\alpha, \beta)$ distribution, the probabilities of staying in trial at arrival r_i according to a Beta(ψ_1, ψ_2) distribution, and the rates θ_i of the exponential durations according to a $\Gamma(\alpha_2, \beta_2)$ distribution. Then, the inter-arrival times between patients entering at centre i are exponential variables with parameter λ_i. With each inclusion time $t_{i,j}$ in centre i, it is associated an exponential time Z_{ij} with rate θ_i and a Ber(r_i) random variable denoted χ_{ij}. In the following, $\mu = \alpha/\beta$ denotes mean annual rate per centre.

Recall models 1 and 2 on one hand, and models 3, 4 and 5 on the other hand, need not to be the same type of data. In models 3, 4 and 5, for each patient we need to know exactly at what time he left screening process. This is not necessary in models 1 and 2, for which it is sufficient to know whether this process was successful.

5.2. Simulation scenarios.

In order to be close to the data used in [10], we have chosen the expected value of the number of predicted patients at the end of the recruitment period $N = 500$, and the number of centres $M = 75$. A set of parameters coherent with this data is given in Table 1. These coefficients imply $\beta = \alpha/\mu \simeq 0.34$ and $\beta_2 = \alpha_2/\mu_2 = 0.5$ where $\mu_2 = \mathbb{E}[\theta]$.

We generate data as specified in section 5.1 and choose interim times $t_1 = 0.5$, 1 and 2 years. At each of these interim time one collects for each centre i the required data for each model. First, we estimate the different values of the parameters. Second, using point and interval estimators, we estimate the duration of the trial.

5.3. Results.

For the recruitment process, the estimated parameters at interim times $t_1 = 0.5, 1$ and 2 are given in Table 2. We can see that the estimated parameters are rather stable when t_1 varies.

In models 1 and 2, which do not take time of dropout into account, the results of estimation procedure are given in Table 3. We can see that the parameter $\widehat{r}$ (model 1) and $\widehat{\mathbb{E}[r_i]} = \frac{\widehat{\psi}_1}{\widehat{\psi}_1+\widehat{\psi}_2}$ (model 2) are also stable when t_1 varies. Parameters for models 3, 4 and 5 are given in Table 4. As expected, $\widehat{r}$ (model 3) and $\widehat{\mathbb{E}[r_i]} = \frac{\widehat{\psi}_1}{\widehat{\psi}_1+\widehat{\psi}_2}$ (model 5) behave similar. Correspondingly, $\widehat{\theta}$ (model 3) and $\widehat{\mathbb{E}[\theta_i]} = \widehat{\mu}_2$ (models 4 and 5) also behave similar.

We now turn to the prediction of the number of randomized patients. At interim time $t_1 = 0.5$, we estimate the parameters for each model and plot in Figure 1 the expected number of patients and predictive interval for $t > t_1 + R$. We see that all five models perform almost the same. It is not surprising as we plot the global screening process, which does not take into account variability in dropout probabilities and screening durations among different centres.

On the other hand, we expect that models will perform somehow differently when we close the centres that lost most patients compared to included patients. More precisely, at time $t_1 = 0.5$, we choose to close the centres that randomized less than 20% of recruited patients (that is $k_i/n_i \leq 0.2$). We plot corresponding predictive curves in Figure 2. Different models can predict recruitment very differently. As expected, models that in some way account for variability in probabilities of screening failure (models 2, 4 and 5) are much more precise that models that do not (models 1 and 3). Recall that model 5 is used to simulate data – it is no suprise it performs the best.

In Table 5 we calculate numerically $\mathbb{E}[\hat{T}]$, expected recruitment time, for different interim times. Five models perform roughly the same. We see that prediction is accurate at interim time $t_1 = 0.5$ years – only about 10% of patients were recruited at this time.

Remark 5.1 *For five models, the distribution of N_t, $t > t_1$, conditionally on data at interim time t_1 is approximated by a normal distribution. In figure 3 is plotted the 'real' distribution, evaluated by simulation, at time $t = 3$ and the normal approximation, for model 5. We can see that the approximation is accurate.*

5.4. Conclusion.

New methodology for predicting patient recruitment accounting for various types of drop-out is developed. Five different models for drop-out are suggested. Two of them do not use screening duration of each patient. The technique for estimation of parameters using interim data and predicting the number of recruited patients over time and recruitment time is developed.

Simulation results show that all five models provide very good prediction and perform similar. However, if we close at interim time the centres that randomize less proportion of patients, then the models that account for variability in probability of screening failures perform better. Thus, simulation results confirm the applicability of the technique and the opportunity to use the approach in real trials analysis.

Acknowledgement

Authors thank Sandrine Andrieu and Stéphanie Savy for valuable discussions on this topic. This research has received the help from IRESP during the call for proposals launched in 2012 as a part of Cancer Plan 2009-2013.

References

1. Senn S. *Statistical Issues in Drug Development*. John Wiley & Sons, Chichester, 1997, doi:10.1002/9780470723586.
2. Senn S. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine* 1998; 17:1753–1765, doi:10.1002/(SICI)1097-0258(19980815/30)17:15/16⟨1753::AID-SIM977⟩3.0.CO;2-X.
3. Carter RE, Sonne SC, Brady KT. Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC Medical Research Methodology* 2005; 5(11):1–5, doi:10.1186/1471-2288-5-11.
4. Carter RE. Application of stochastic processes to participant recruitment in clinical trials. *Controlled Clinical Trials* 2004; 25(5):429–436, doi:10.1016/j.cct.2004.07.002.
5. Bates GE, Neyman J. Contributions to the theory of accident proneness. II. True or false contagion. *Univ. California Publ. Statist.* 1952; 1:255–275.
6. Anisimov VV, Fedorov VV. Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in Medicine* 2007; 26(27):4958–4975, doi:10.1002/sim.2956.
7. Anisimov VV. Recruitment modeling and predicting in clinical trials. *Pharmaceutical Outsourcing* 2009; 10(1):44–48.
8. Anisimov VV. Using mixed poisson models in patient recruitment in multicentre clinical trials. *Proceegings of the World Congress on Ingineering*, vol. II, London, United Kingdom, 2008; 1046–1049.
9. Anisimov VV. Predictive modelling of recruitment and drug supply in multicenter clinical trials. *Proceedings of the Joint Statistical Meeting, ASA*, Washington, USA, 2009; 1248–1259.
10. Mijoule G, Savy N, Savy S. Models for patients recruitment in clinical trials and sensitivity analysis. *Statistics in Medicine* 2012; 31(16):1655–1674.

11. Anisimov VV. Statistical modeling of clinical trials (recruitment and randomization). *Comm. Statist. Theory Methods* 2011; 40(19-20):3684–3699, doi:10.1080/03610926.2011.581189. URL http://dx.doi.org/10.1080/03610926.2011.581189.
12. Anisimov VV. Effects of unstratified and centre-stratified randomization in multi-centre clinical trials. *Pharmaceutical Statistics* 2011; 10(1):50–59, doi:10.1002/pst.412. URL http://onlinelibrary.wiley.com/doi/10.1002/pst.412/abstract.
13. Anisimov VV. Predictive event modelling in multicentre clinical trials with waiting time to response. *Pharmaceutical Statistics* 2011; 10(6):517–522, doi:10.1002/pst.525.
14. Johnson NL, Kotz S, Kemp AW. *Univariate discrete distributions*. Second edn., Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc.: New York, 1992. A Wiley-Interscience Publication.
15. Bernardo JM, Smith AFM. *Bayesian Theory*. John Wiley & Sons: Hoboken, NJ, USA, 2004, doi:10.1002/bimj.4710300220.

Table 1. Description of a set of parameter coherent with the chosen data.

recruitment	instantaneous dropout	dropout during screening
$\alpha = 1.2$	$\psi_1 = 4$	$\alpha_2 = 1$
$\mu = 3.5$	$\psi_2 = 1$	$\mu_2 = 2$

Table 2. Estimated parameters $\widehat{\alpha}$ and $\widehat{\mu}$.

t_1	0.5	1	2
$\widehat{\alpha}$	2.18	1.83	1.37
$\widehat{\mu}$	3.76	3.47	3.37

Table 3. Estimated parameters used in models 1 and 2.

t_1	0.5	1	2
$\widehat{p}$	0.54	0.59	0.63
$\widehat{\psi_1}$ (M2)	1.80	1.88	2.17
$\widehat{\psi_2}$ (M2)	1.46	1.34	1.19
$\widehat{\psi_1}/(\widehat{\psi_1}+\widehat{\psi_2})$ (M2)	0.55	0.58	0.65

Table 4. Estimated parameters used in models 3,4 and 5.

t_1	0.5	1	2
$\widehat{r}$	0.70	0.74	0.77
$\widehat{\theta}$	1.94	2.34	1.96
$\widehat{\alpha}_2$	0.70	0.69	0.78
$\widehat{\mu}_2$	2.26	2.94	2.40
$\widehat{\psi_1}$	1.48	2.10	2.85
$\widehat{\psi_2}$	0.60	0.66	0.72
$\widehat{\psi_1}/(\widehat{\psi_1}+\widehat{\psi_2})$	0.71	0.76	0.80

Table 5. Estimated mean recruitment time, at interim times $t_1 = 0.5, 1$ and 2. Actual recruitment time is 3.20 years.

t_1	0.5	1	2
M1 - $\mathbb{E}[\widehat{T}]$	3.31	3.29	3.13
M2 - $\mathbb{E}[\widehat{T}]$	3.58	3.46	3.16
M3 - $\mathbb{E}[\widehat{T}]$	3.24	3.38	3.22
M4 - $\mathbb{E}[\widehat{T}]$	3.11	3.28	3.12
M5 - $\mathbb{E}[\widehat{T}]$	3.09	3.27	3.12

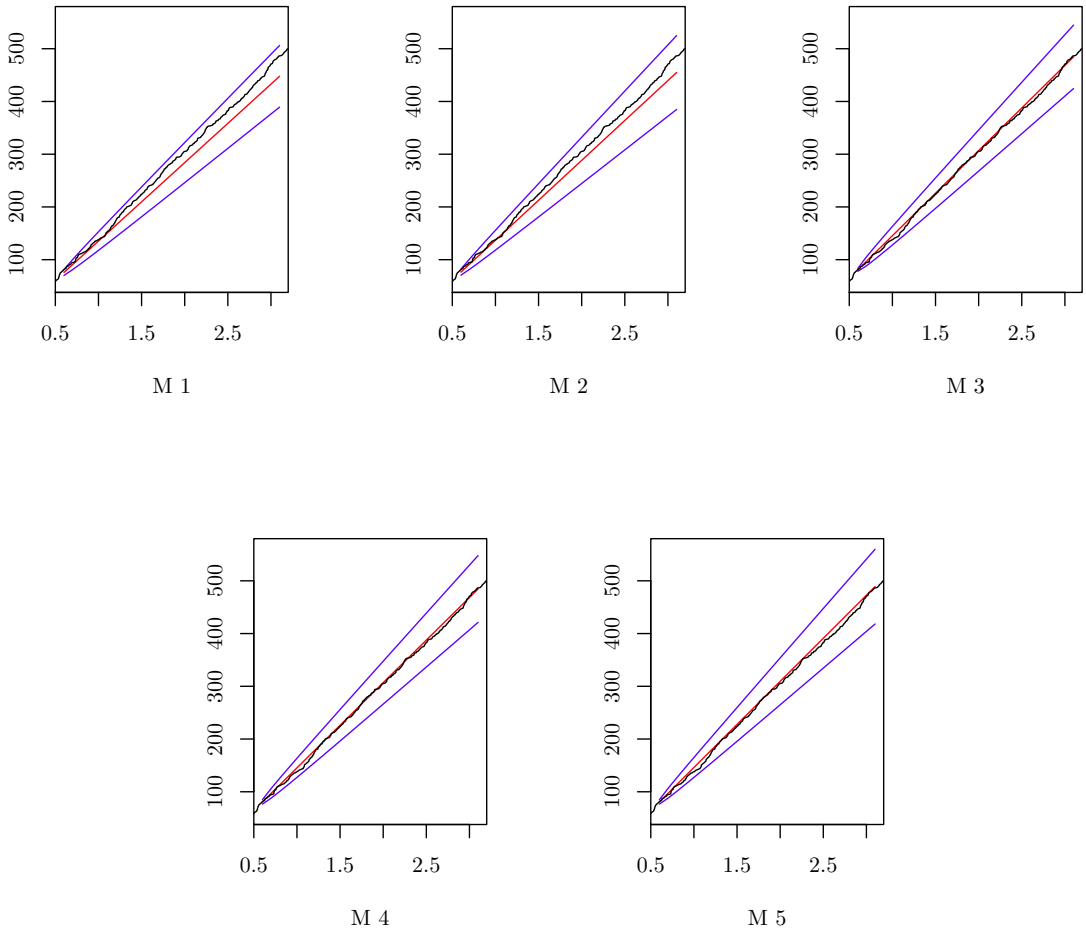

Figure 1. Mean number of randomized patients (red), predictive interval (blue), real data (black) for interim time $t_1 = 0.5$. Top : models 1, 2, 3. Bottom : models 4, 5. Time is in years.

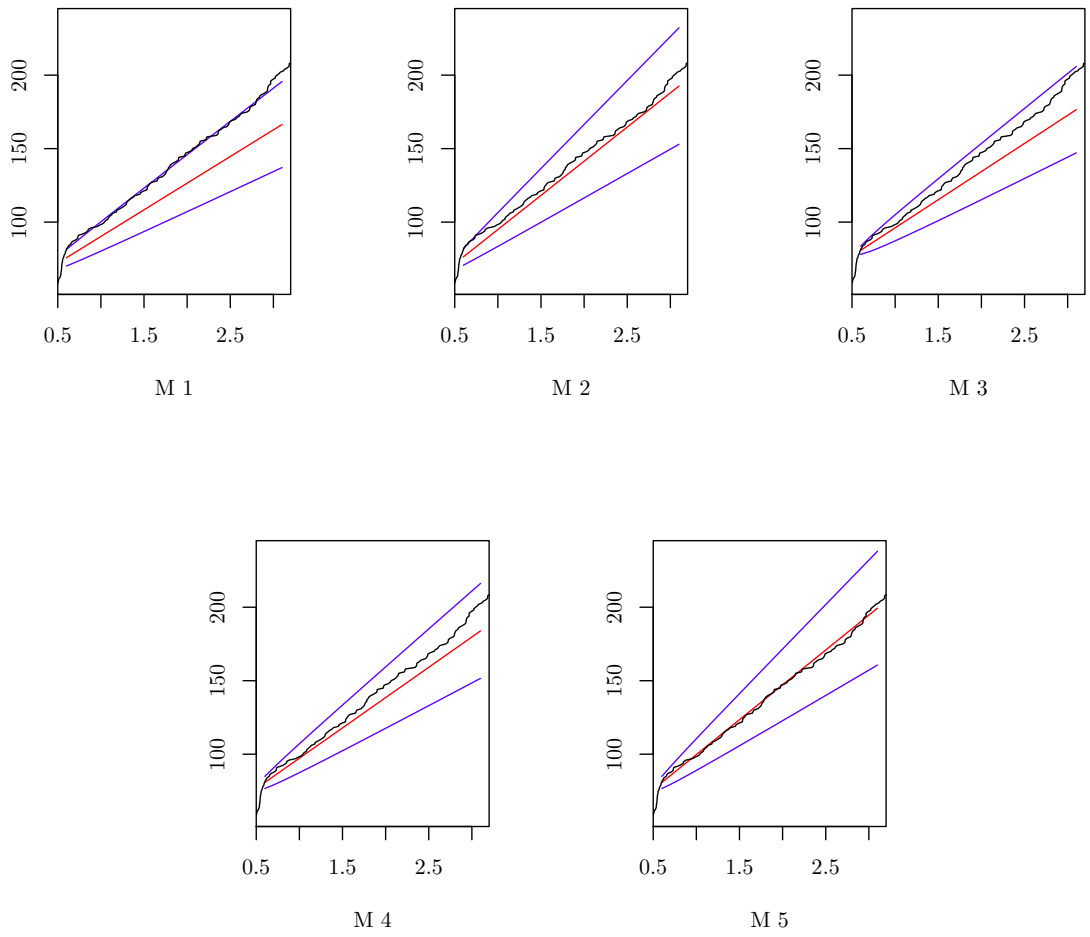

Figure 2. Mean number of randomized patients (red), predictive interval (blue), real data (black) for interim time $t_1 = 0.5$, having closed centres which randomized less than 20% of recruited patients. Top : models 1, 2, 3. Bottom : models 4, 5. Time is in years.

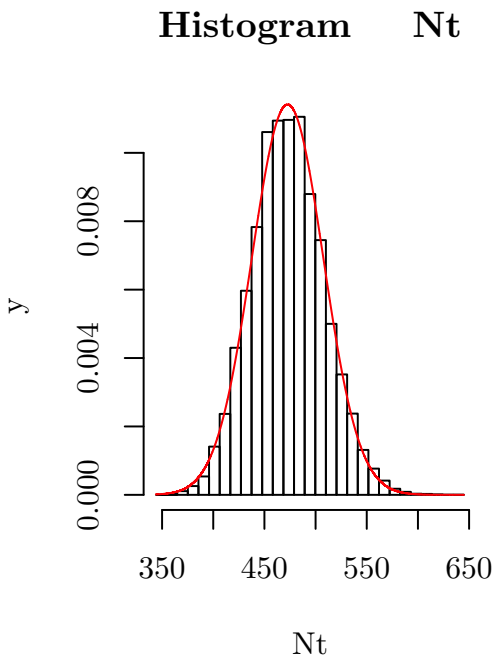

Figure 3. Distribution of the number of randomized patients at time $t = 3$ conditionally on data at time $t_1 = 0.5$, for model 5. Real distribution (histogram), and normal approximation (red line).

Chapitre 4

Un modèle de coût

Ce travail, en collaboration avec Vladimir Anisimov et Nicolas Savy, a été présenté au 7ème International Workshop on Simulation à Rimini en mai 2013 et sera intégré dans les actes de ce workshop [13].

Savy a proposé l'utilisation de processus de Poisson filtrés pour la modélisation du coût ; j'ai utilisé cette idée pour construire la forme générale du coût à partir du modèle 2 du chapitre précédent proposé par Anisimov. J'ai démontré les différents résultats de ce chapitre, et ai élaboré et programmé les simulations.

4.1 Introduction

L'évaluation du coût d'un essai clinique est un problème complexe. Une étude durant plusieurs années et s'appuyant sur plusieurs dizaines de centres implique une logistique très importante, qui engendre des dépenses tout aussi importantes. Néanmoins, il est difficile d'évaluer précisément ce coût, car une multitude de paramètres entrent en jeu : coût de fabrication et d'acheminement des médicaments, coût du personnel, de prise en charge des patients, du management de l'étude...

Nous supposons que l'on peut néanmoins catégoriser ces coûts de la manière suivante :

- **coût fixe par patient screené**,
- **coût fixe par patient randomisé** incluant entre autres le coût du traitement,
- **coût par patient randomisé dépendant du temps passé dans l'essai** : si les patients ne restent pas tous le même temps dans l'essai (à cause par exemple de l'évolution de leur état de santé ; ce sera en particulier le cas dans le cadre des données de survie),
- **coût fixe d'un centre** qui peut inclure le coût de mise en route du centre, de la fourniture en médicaments (en général, la quantité de médicaments fournie à un centre est déterminée au début de l'essai et ne dépend donc pas du processus d'inclusion du centre),
- **coût d'un centre proportionnel à sa durée d'activité** incluant une partie du coût du personnel (par exemple, un médecin est requis en permanence dans un centre ouvert, que le centre recrute ou non), les coûts en énergie, etc.

Notations 4.1.1. – *$N_i^R(t)$ est le nombre de patients randomisés (ou inclus) dans le centre i au temps t.*

– $N_i(t)$ *est le nombre de patients screenés (ou arrivés) dans le centre* i *au temps* t.

Remarque 4.1.2. *Pour tout* $t \geq 0$, $N_i(t) \geq N_i^R(t)$.

Le coût de l'essai pour le centre i à l'instant $t \geq 0$ peut donc s'écrire :

$$C_i(t) = C_1 N_i^R(t) + C_2 N_i(t) + \sum_{0 \leq T_n^i \leq t} g(t, T_n^i) + F_i + G_i t, \tag{4.1}$$

où $(T_n^i)_{n\geq 0}$ sont les instants de randomisation (inclusion) dans le centre i (instants de saut du processus N_i^R), $g(t,s)$ représente le coût à l'instant t d'un patient randomisé à l'instant s, F_i est le coût fixe du centre, G_i le coût par année d'activité du centre, C_1 le coût fixe d'un patient randomisé, C_2 le coût de screening d'un patient. On fait les hypothèses suivantes sur g :

(H1) $g : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ est mesurable,

(H2) g triangulaire au sens où $g(t,s) = 0$ si $t < s$,

(H3) pour tout $t \geq 0$, $g(t,.)$ est continue sur $[0,t]$.

Notons que la somme dans (4.1) peut s'écrire comme une intégrale contre le processus N_i^R :

$$\sum_{0 \leq T_n^i \leq t} g(t, T_n^i) = \int_0^t g(t,s) dN_i^R(s).$$

Si N_i^R est un processus de Poisson d'intensité homogène, le processus $t \mapsto \left(\int_0^t g(t,s) dN_i^R(s)\right)_{t\geq 0}$ est connu sous le nom de processus de Poisson filtré [25, 26, 42, 44].

Le coût global de l'essai à l'instant t est alors défini par

$$C(t) = \sum_{i=1}^{C} C_i(t).$$

L'inclusion de patients s'arrête dès que le processus $N^R = \sum_{i=1}^C N_i^R$ atteint K_f. En posant

$$\tau = \left\{ \inf_{t\geq 0} \ : \ N^R(t) = K_f \right\}$$

le coût total de l'essai est alors $C(\tau)$. On en déduit alors le coût moyen de l'essai, noté $\mathcal{C}$:

Définition 4.1.3 (coût moyen de l'essai). *On appelle coût moyen de l'essai la quantité*

$$\begin{aligned} \mathcal{C} = \mathbb{E}\left[C(\tau)\right] &= C_1 \mathbb{E}\left[N^R(\tau)\right] + C_2 \mathbb{E}\left[N(\tau)\right] \\ &+ \mathbb{E}\left[\int_0^\tau g(\tau, s) dN^R(s)\right] + \sum_{i=1}^{C} F_i + \sum_{i=1}^{C} G_i \mathbb{E}\left[\tau\right]. \end{aligned} \tag{4.2}$$

L'intégrale précédente s'entend au sens de Stieljès : $\int_0^\tau g(\tau,s) dN^R(s) = \sum_{0\leq T_n^i \leq \tau} g(\tau, T_n^i)$, $(T_n^i)_{n\geq 0}$ étant les instants de randomisation dans le centre i. Notons que le processus $t \mapsto \int_0^t g(t,s) dN^R(s)$ n'étant pas une semi-martingale [26], nous ne pouvons pas calculer l'espérance par des techniques de martingale.

On se place dans le cadre du modèle 2 du chapitre 3. On rappelle que le processus N_i est un processus de Poisson d'intensité λ_i, et N_i^R est un processus de Poisson de paramètre $p_i \lambda_i$, où λ_i suit une loi Gamma de paramètres (α, β) et p_i suit une loi Beta de paramètres (ψ_1, ψ_2).

4.2 Préliminaires

Reprenons la formule (4.2). Supposons d'abord $(\lambda_i, p_i)_{1\leq i\leq C}$ connus. Sachant $(p_i, \lambda_i)_{1\leq i\leq C}$, N^R est un processus de Poisson d'intensité $\Lambda = \sum_{i=1}^{C} p_i\lambda_i$. Le processus N est alors représenté par $N \equiv N^R + N^L$ où N^L est un processus de Poisson indépendant de N^R, d'intensité $\Lambda_2 = \sum_{i=1}^{C}(1-p_i)\lambda_i$. On se ramène alors au problème suivant :

Soient N^R et N^L deux processus de Poisson indépendants d'intensités respectives Λ et Λ_2. Posons

$$C(t) = C_1' N^R(t) + C_2 N^L(t) + \int_0^t g(t,s)dN^R(s) + Gt + F, \quad t \geq 0$$

où g vérifie les hypothèses **(H1)**, **(H2)** et **(H3)**. Dans le cadre des essais cliniques, on aura $F = \sum_{i=1}^{C} F_i$, $G = \sum_{i=1}^{C} G_i$, et $C_1' = C_1 + C_2$.

Soit $\tau = \inf\{t \geq 0 \ : \ N^R(t) \geq K_f\}$. Le coût total de l'essai est donc

$$C(\tau) = C_1' K_f + C_2 N^L(\tau) + \int_0^\tau g(\tau,s)dN^R(s) + G\tau + F. \tag{4.3}$$

Sachant Λ, τ suit une loi $\Gamma(K_f, \Lambda)$. Notons p_τ sa densité, et rappelons que

$$p_\tau(dt) = \mathbf{1}_{t>0}\frac{\Lambda^{K_f}}{(K_f-1)!}e^{-\Lambda t}t^{K_f-1}dt. \tag{4.4}$$

Théorème 4.2.1. *Si N^R et N^L sont deux processus de Poisson indépendants d'intensités respectives Λ et Λ_2, si $\tau = \inf\{t \geq 0 \ : \ N^R(t) \geq K_f\}$, et si $C(\tau)$ est définie par* (4.3)*, alors*

$$\mathbb{E}\left[C(\tau)\right] = C_1' K_f + C_2 K_f \frac{\Lambda_2}{\Lambda} + \int_0^{+\infty} g(t,t)p_\tau(dt) + (K_f - 1)\int_0^{+\infty}\int_0^t g(t,s)\,ds\,\frac{p_\tau(dt)}{t} + G\frac{K_f}{\Lambda} + F.$$

La démonstration repose sur les lemmes suivants.

Lemme 4.2.2. *Si N^R est un processus de Poisson d'intensité Λ et $\tau = \inf\{t \geq 0 \ : \ N^R(t) \geq K_f\}$, alors*

$$\mathbb{E}\left[\tau\right] = \frac{K_f}{\Lambda}.$$

Démonstration. D'après la proposition 2.2.4, τ suit une loi Gamma de paramètres (K_f, Λ), d'où le résultat. □

Lemme 4.2.3. *Soient N^R et N^L deux processus de Poisson indépendants, d'intensités respectives Λ et Λ_2. Soit $\tau = \inf\{t \geq 0 \ : \ N^R(t) \geq K_f\}$. Alors*

$$\mathbb{E}\left[N^L(\tau)\right] = K_f\frac{\Lambda_2}{\Lambda}. \tag{4.5}$$

Démonstration. Comme la variable aléatoire τ est indépendante de N^L,

$$\mathbb{E}\left[N^L(\tau)\right] = \mathbb{E}\left[\mathbb{E}\left[N^L(\tau)\,|\,\tau\right]\right] = \mathbb{E}\left[\Lambda_2\tau\right] = \Lambda_2\mathbb{E}\left[\tau\right] = K_f\frac{\Lambda_2}{\Lambda}.$$

□

Lemme 4.2.4. *Posons $n = K_f - 1$. Sachant τ, les instants de saut $T_1 \leq T_2 \leq \ldots \leq T_n$ de N^R sur $[0,\tau]$ ont même loi qu'un n-uplet de loi uniforme sur $[0,\tau]^n$ réordonné par ordre croissant.*

Démonstration. En effet, le $(n+1)$-uplet $(T_1, \ldots, T_n, T_{n+1} = \tau)$ admet une densité sur $\mathbb{R}_+^{n+1}$ qui est

$$\begin{aligned} p_{T_1,\ldots,T_{n+1}}(t_1,\ldots,t_{n+1}) &= \mathbf{1}_{0\leq t_1\leq\ldots\leq t_{n+1}} \prod_{i=1}^{n+1} \Lambda e^{-\Lambda(t_i - t_{i-1})} \\ &= \mathbf{1}_{0\leq t_1\leq\ldots\leq t_{n+1}} \Lambda^{n+1} e^{-\Lambda t_{n+1}}, \end{aligned}$$

où par convention $t_0 = 0$. Sachant $T_{n+1} = t$, la densité de $(T_1, \ldots, T_n)$ s'écrit alors

$$\begin{aligned} p_{T_1,\ldots,T_n|T_{n+1}=t}(t_1,\ldots,t_n) &= \frac{\Lambda^{n+1} e^{-\Lambda t}}{p_\tau(t)} \mathbf{1}_{0\leq t_1\leq\ldots\leq t_n\leq t} \\ &= \frac{n!}{t^n} \mathbf{1}_{0\leq t_1\leq\ldots\leq t_n\leq t}. \end{aligned}$$

□

On en déduit le corollaire :

Corollaire 4.2.5. *Sachant τ, pour tout $s < \tau$, $N^R(s)$ suit une loi binomiale $\mathcal{B}\left(K_f - 1, \frac{s}{\tau}\right)$.*

Démonstration. C'est une conséquence directe de la proposition précédente. □

Lemme 4.2.6. *Supposons que g satisfait les hypothèses* **(H1)** *à* **(H3)**. *Alors*

$$\begin{aligned} \mathbb{E}\left[\int_0^\tau g(\tau,s) dN^R(s)\right] &= \int_0^{+\infty} g(t,t) p_\tau(dt) + (K_f - 1) \int_0^{+\infty} \int_0^t g(t,s)\, ds\, \frac{p_\tau(dt)}{t} \qquad (4.6) \\ &= \mathbb{E}\left[g(\tau,\tau)\right] + (K_f - 1)\mathbb{E}\left[G(\tau)\right], \qquad (4.7) \end{aligned}$$

où pour tout $t > 0$, $G(t) = \frac{1}{t}\int_0^t g(t,s)ds$.

Démonstration. On a

$$\mathbb{E}\left[\int_0^\tau g(\tau,s) dN^R(s)\right] = \int_0^{+\infty} \mathbb{E}\left[\int_0^t g(t,s) dN^R(s) \,\middle|\, \tau = t\right] p_\tau(dt),$$

Supposons d'abord que pour tout $t > 0$, la restriction de $s \mapsto g(t,s)$ à $[0,t]$ est dérivable. Notons $\partial_2 g(t,.)$ cette dérivée. Supposons en outre que $\forall t > 0$, $\sup_{0\leq s\leq t} |\partial_2 g(t,s)| < +\infty$. Par une intégration par parties, on a

$$\int_0^t g(t,s) dN^R(s) = \left[g(t,s)N^R(s)\right]_0^t - \int_0^t \partial_2 g(t,s) N^R(s) ds,$$

soit

$$\int_0^t g(t,s) dN^R(s) = g(t,t)N^R(t) - \int_0^t \partial_2 g(t,s) N^R(s) ds. \qquad (4.8)$$

On en déduit que

$$\mathbb{E}\left[\int_0^t g(t,s)dN^R(s) \,\middle|\, \tau = t\right] = g(t,t)K_f - \mathbb{E}\left[\int_0^t \partial_2 g(t,s)N^R(s)ds \,\middle|\, \tau = t\right].$$

Sachant $\{\tau = t\}$, nous pouvons majorer $|\partial_2 g(t,s)N^R(s)| \leq K_f \sup_{0\leq s\leq t} |\partial_2 g(t,s)|$, le théorème de Fubini s'applique et grâce au corollaire 4.2.5,

$$\mathbb{E}\left[\int_0^t \partial_2 g(t,s)N^R(s)ds \,\middle|\, \tau = t\right] = (K_f - 1)\int_0^t \partial_2 g(t,s)\,\frac{s}{t}ds,$$

d'où l'on déduit (4.7) par une intégration par parties.

Supposons à présent que pour tout $t > 0$, $g(t,.)$ est continue sur $[0,t]$. Alors pour tout $\epsilon > 0$, il existe une fonction $\phi^\epsilon(t,.)$ $\mathcal{C}^1$ sur $[0,t]$ telle que $\sup_{0\leq s\leq t} |\phi^\epsilon(t,.) - g(t,.)| \leq \epsilon$, et

$$\begin{aligned}
&\left|\mathbb{E}\left[\int_0^\tau g(\tau,s)dN^R(s)\right] - \int_0^{+\infty} g(t,t)p_\tau(dt) - (K_f-1)\int_0^{+\infty}\int_0^t g(t,s)\,ds\,\frac{p_\tau(dt)}{t}\right| \\
=&\left|\mathbb{E}\left[\int_0^\tau (g-\phi^\epsilon)(\tau,s)dN^R(s)\right] - \int_0^{+\infty}(g-\phi^\epsilon)(t,t)p_\tau(dt)\right. \\
&\qquad\left. - (K_f-1)\int_0^{+\infty}\int_0^t (g-\phi^\epsilon)(t,s)\,ds\,\frac{p_\tau(dt)}{t}\right| \\
\leq& K_f\epsilon + \epsilon + (K_f-1)\int_0^{+\infty} t\epsilon\,\frac{p_\tau(dt)}{t} \\
\leq& (2K_f+1)\epsilon.
\end{aligned}$$

ϵ étant choisi arbitrairement, on en déduit le résultat. □

4.3 Application aux essais cliniques multicentriques

On suppose que la fonction de coût g s'écrit

$$g(t,s) = C_3(t-s)\mathbf{1}_{t\geq s}, \tag{4.9}$$

où $C_3 > 0$. Ceci revient à dire qu'un patient randomisé représente, en plus d'un coût fixe C_1, un coût proportionnel au temps qu'il passe dans l'essai.

4.3.1 Paramètres $(\lambda_i)_{1\leq i\leq C}$ et $(p_i)_{1\leq i\leq C}$ connus

Dans un premier temps, supposons les intensités d'inclusion $(\lambda_i)_{1\leq i\leq C}$ et les probabilités de randomisation $(p_i)_{1\leq i\leq C}$ connues.

L'application du théorème 4.2.1 nous donne le coût moyen :

$$\mathbb{E}\left[C(\tau)\right] = C_1'K_f + C_2K_f\frac{\Lambda_2}{\Lambda} + \frac{1}{2}C_3(K_f-1)\frac{K_f}{\Lambda} + G\frac{K_f}{\Lambda} + F, \tag{4.10}$$

où

$$\Lambda = \sum_{i=1}^{C} p_i \lambda_i,$$
$$\Lambda_2 = \sum_{i=1}^{C} (1-p_i)\lambda_i,$$
$$F = \sum_{i=1}^{C} F_i,$$
$$G = \sum_{i=1}^{C} G_i.$$

A-t-on intérêt à fermer le centre j ? La proposition suivante permet de répondre à cette question.

Proposition 4.3.1. *La fermeture du centre j entraîne une variation du coût moyen de l'essai égale à*

$$\Delta\mathcal{C}_j = K_f \frac{\lambda_j}{\Lambda(\Lambda - p_j\lambda_j)} \left[C_2(p_j\Lambda_2 - (1-p_j)\Lambda) + \frac{1}{2}C_3(K_f - 1)p_j + p_j G - G_j \frac{\Lambda}{\lambda_j} \right] - F_j.$$

Il est rentable de fermer le centre j si cette quantité est négative.

Démonstration. La fermeture du centre j induit une variation de Λ, Λ_2, F et G égales à

$$\Delta\Lambda = -p_j\lambda_j,$$
$$\Delta\Lambda_2 = -(1-p_j)\lambda_j,$$
$$\Delta F = -F_j,$$
$$\Delta G = -G_j.$$

Le coût moyen étant donné par l'équation (4.10), on en déduit le résultat. □

4.3.2 Paramètres $(\lambda_i)_{1\leq i\leq C}$ et $(p_i)_{1\leq i\leq C}$ aléatoires

On se place ici dans le cadre du modèle 2 du chapitre 3 : les intensités d'inclusion des centres $\{\lambda_1, \ldots, \lambda_C\}$ sont indépendantes de loi Gamma de paramètres (α, β), et les probabilités de randomisation $\{p_1, \ldots, p_C\}$ sont indépendantes de loi Beta de paramètres (ψ_1, ψ_2).

On fait une étude à un instant intermédiaire t_1. On considère donc le **coût de l'étude à partir de l'instant** t_1, et non pas le coût global de l'étude. Notons donc K_f le nombre total de patients restant à randomiser après t_1. Sachant l'information à t_1, en notant n_i le nombre de patients recrutés et k_i le nombre de patients randomisés dans le centre i, les lois de λ_i et p_i sont données par (cf chapitre 3) :

$$\lambda_i \overset{\mathcal{L}}{=} \Gamma(\alpha + n_i, \beta + \tau_i),$$
$$p_i \overset{\mathcal{L}}{=} \mathcal{B}(\psi_1 + k_i, \psi_2 + n_i - k_i).$$

Dans ce cadre, la proposition 4.3.1 implique, en moyennant sur les intensités d'inclusion $(\lambda_i)_{1\leq i\leq C}$ et probabilités de randomisation $(p_i)_{1\leq i\leq C}$:

Proposition 4.3.2. *La fermeture du centre j entraîne une variation du coût moyen de l'essai égale à*

$$\Delta\mathcal{C}_j = K_f\mathbb{E}\left[\frac{\lambda_j}{\Lambda(\Lambda - p_j\lambda_j)}\left(C_2(p_j\Lambda_2 - (1-p_j)\Lambda) + \frac{1}{2}C_3(K_f-1)p_j + p_jG - G_j\frac{\Lambda}{\lambda_j}\right)\right] - F_j. \quad (4.11)$$

4.3.3 Applications numériques

Nous testons l'applicabilité du modèle par des simulations numériques. Nous utilisons le modèle 2 du chapitre 3 pour les simulations. Les taux λ_i sont distribués selon une loi Gamma de paramètres (α, β) et les probabilités de screening p_i suivant une loi Beta de paramètres (ψ_1, ψ_2). Les valeurs choisies pour les paramètres sont données dans la table 4.1. Rappelons que $\beta = \alpha/\mu$. On a $\mathbb{E}[\lambda_i] = \mu = 3.5$ (nombre moyen de patients screenés

Intensité d'inclusion	Probabilité de screening
$\alpha = 1.2$	$\psi_1 = 3$
$\mu = 3.5$	$\psi_2 = 1$

TABLE 4.1 – Jeu de paramètres pour la simulation du scénario.

par an et par centre) et $\mathbb{E}[p_i] = \frac{\psi_1}{\psi_1+\psi_2} = 0.75$ (probabilité moyenne qu'un patient screené soit randomisé).

Le nombre de centres est $C = 48$, et nous simulons les données (c'est-à-dire le nombre de patients screenés $(n_i)_{1\leq i\leq C}$ et randomisés $(k_i)_{1\leq i\leq C}$) à l'instant intermédiaire $t_1 = 1$. Tous les centres sont supposés avoir ouvert à $t = 0$. Les données simulées sont écrites dans la table 4.2.

i	1	2	3	4	5	6	7	8	9	10	11	12
k_i	3	3	0	0	6	1	1	6	1	1	0	2
n_i	3	4	0	0	6	2	4	7	1	1	0	2
i	13	14	15	16	17	18	19	20	21	22	23	24
k_i	1	5	3	1	11	2	4	0	4	2	7	1
n_i	4	13	4	1	11	2	6	0	4	2	9	1
i	25	26	27	28	29	30	31	32	33	34	35	36
k_i	3	2	8	1	5	0	4	1	0	0	3	4
n_i	5	5	13	3	6	0	7	1	0	0	4	4
i	37	38	39	40	41	42	43	44	45	46	47	48
k_i	6	0	4	3	1	3	1	0	4	1	3	3
n_i	6	3	6	6	1	3	1	0	4	1	3	8

TABLE 4.2 – Données simulées à t_1. k_i : nombre de patients randomisés. n_i : nombre de patients screenés.

Nous calculons, pour tout j, le coût de la fermeture du centre j, via la formule (4.11). Les paramètres sont estimés grâce à la méthode du modèle 2 du chapitre 3 ; les estimateurs

Intensité d'inclusion	Probabilité de screening
$\widehat{\alpha} = 1.63$	$\widehat{\psi_1} = 3.33$
$\widehat{\mu} = 3.69$	$\widehat{\psi_2} = 1.23$

TABLE 4.3 – Estimateurs des paramètres pour les données simulées à $t_1 = 1$.

apparaissent dans la table 4.3. L'espérance dans (4.11) est calculée par simulations de Monte-Carlo, avec 10^5 simulations.

Rappelons que :

- C_2 est le coût du screening d'un patient,
- C_3 est le coût, par an, d'un patient randomisé,
- F_j est le coût fixe pour garder le centre j ouvert après t_1,
- G_j est le coût de fonctionnement par an du centre j.

Nous étudions différents jeux de valeurs possibles pour ces constantes.

Dans tous les cas, on suppose que l'on veut randomiser 300 patients. A t_1, on a déjà randomisé $K_1 = \sum_{i=1}^{C} k_i = 125$ patients. Il reste donc à randomiser $K_f = 175$ patients.

Cas particulier où $\mathbf{C_3 = 0, F \equiv 0, G \equiv 0}$

Supposons que $C_3 = 0$, $F_j = G_j = 0$ pour tout $1 \leq j \leq C$, et que $C_2 = 2$. Cela signifie que le seul coût est celui du screening d'un patient. On s'attend donc à ce qu'il soit rentable de fermer les centres qui ont le plus de screening failure, c'est-à-dire les centres pour lesquels $(1-p_j)\lambda_j$ est le plus grand. Comme p_j et λ_j ne sont pas atteignables avec les données, nous les remplaçons par leurs estimateurs naturels $\widehat{p}_j = k_j/n_j$ et $\widehat{\lambda}_j = n_j/\tau_j = n_j$; d'où l'on déduit $(1-\widehat{p}_j)\widehat{\lambda}_j = n_j - k_j$.

Dans la table 4.4 est représentée, pour tout j, la variation du coût moyen de l'essai (après t_1) si l'on ferme le centre j. Elle est comparée à $n_j - k_j$. Le centre le plus intéressant à fermer et le centre 14 : c'est celui qui permet en moyenne le maximum d'économies.

j	1	2	3	4	5	6	7	8	9	10	11	12
$\Delta\mathcal{C}_j$	1.35	0.35	0.06	0.07	3.12	-0.45	-2.75	1.94	0.4	0.41	0.07	0.85
$n_j - k_j$	0	1	0	0	0	1	3	1	0	0	0	0
j	13	**14**	15	16	17	18	19	20	21	22	23	24
$\Delta\mathcal{C}_j$	-2.74	**-8.41**	0.35	0.41	6.46	0.85	-0.35	0.07	1.91	0.85	1.24	0.41
$n_j - k_j$	3	**8**	1	0	0	0	2	0	0	0	2	0
j	25	26	27	28	29	30	31	32	33	34	35	36
$\Delta\mathcal{C}_j$	-0.81	-2.43	-2.44	-1.53	1.38	0.07	-1.63	0.4	0.06	0.06	0.35	1.92
$n_j - k_j$	2	3	5	2	1	0	3	0	0	0	1	0
j	37	38	39	40	41	42	43	44	45	46	47	48
$\Delta\mathcal{C}_j$	3.13	-2.97	-0.35	-2.05	0.41	1.36	0.4	0.07	1.91	0.4	1.35	-4.79
$n_j - k_j$	0	3	2	3	0	0	0	0	0	0	0	5

TABLE 4.4 – $\Delta\mathcal{C}_j$: variation du coût moyen de l'essai après t_1 si l'on ferme uniquement le centre j. $n_j - k_j$: estimateur du nombre de patients échouant au screening par an.

La figure 4.1 représente $\Delta\mathcal{C}_j$ en fonction de $n_j - k_j$. La dépendance entre les deux quantités apparait clairement : plus $n_j - k_j$ est grand, plus il est rentable de fermer le centre j.

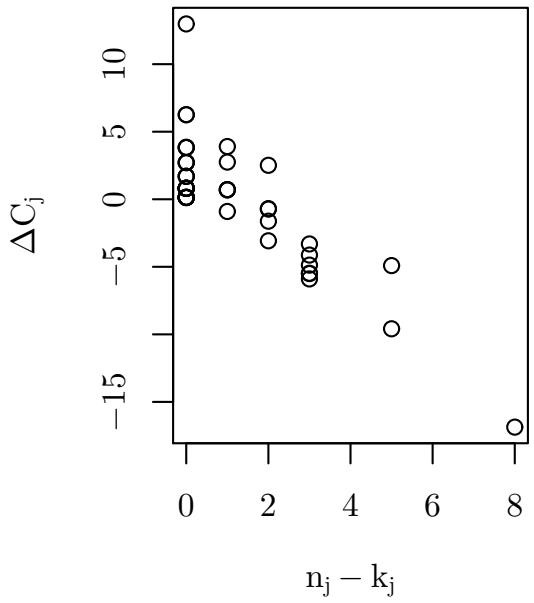

FIGURE 4.1 – $\Delta\mathcal{C}_j$ en fonction de $n_j - k_j$. $C_2 = 2, C_3 = 0, F_j = 0, G_j = 0$.

Cas particulier où $\mathbf{C2 = 0, C_3 = 0}$

Supposons que $C_2 = C_3 = 0$, et que pour tout $1 \leq j \leq C$, $G_j = 10$ et $F_j = 1$. Cela signifie que le coût ne dépend pas du nombre de patients recruté : seuls entrent en jeu pour chaque centre un coût fixe et un coût proportionnel au temps d'ouverture. Comme le coût fixe est le même pour chaque centre, on s'attend donc à ce qu'il soit rentable de fermer les centres qui randomisent le moins de patients, c'est-à-dire les centres pour lesquels $p_j\lambda_j$ est le plus petit. De même que précédemment, nous remplaçons p_j et λ_j par leurs estimateurs naturels $\widehat{p}_j = k_j/n_j$ et $\widehat{\lambda}_j = n_j/\tau_j = n_j$, ce qui donne $\widehat{p}_j\widehat{\lambda}_j = k_j$.

Les centres les plus intéressants à fermer sont comme attendu les centres n'ayant pas recruté (centres 3, 4, 11, 20, 30, 33, 34 et 44) : ce sont eux qui permettent en moyenne le maximum d'économies.

Dans la figure 4.1, on représente $\Delta\mathcal{C}_j$ en fonction de k_j. Là encore, la corrélation apparait clairement : plus le nombre de patients randomisés k_j est grand, moins il est rentable de fermer le centre j.

Un cas général

Prenons $C_2 = 2, C_3 = 0.5, G_j = 5, F_j = 5$.

Les résultats montrent que le centre le plus intéressant à fermer est le centre 38.

Néanmoins, et contrairement aux deux exemples précédents, les corrélations entre $\Delta\mathcal{C}_j$ et $n_j - k_j$ d'une part et entre $\Delta\mathcal{C}_j$ et k_j d'autre part sont beaucoup moins prononcées (cf figure 4.3).

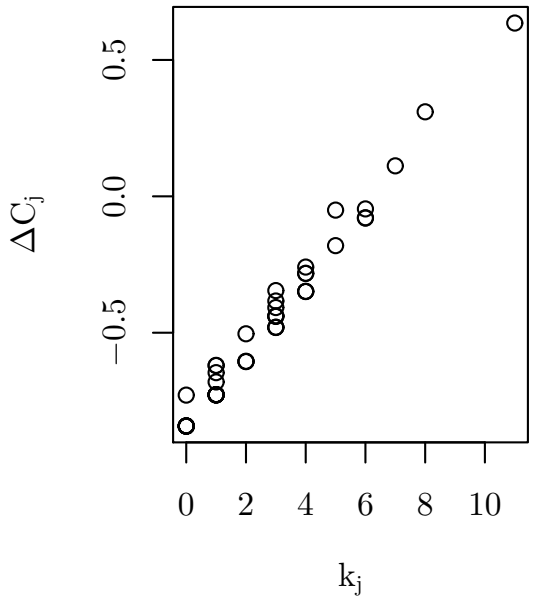

FIGURE 4.2 – $\Delta\mathcal{C}_j$ en fonction de k_j. $C_2 = 0, C_3 = 0, G_j = 10, F_j = 1$.

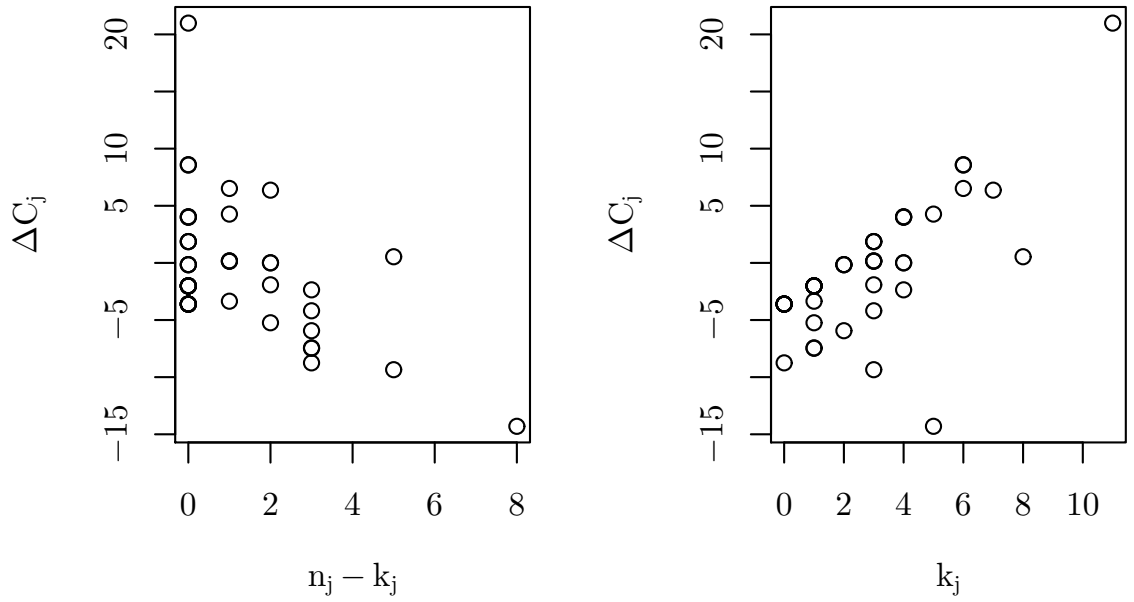

FIGURE 4.3 – $\Delta\mathcal{C}_j$ en fonction de $n_j - k_j$ (gauche) et k_j (droite). $C_2 = 2, C_3 = 0.5$, $G_j = 5, F_j = 5$.

Chapitre 5

Conclusion et perspectives

5.1 Conclusion

Les modèles bayésiens utilisés par Anisimov [11] et leurs extensions et variantes [35] se révèlent pertinents pour la modélisation du recrutement de la phase III d'un essai clinique de par leur bonne adéquation aux données réelles, comme le montrent les tests effectués sur les courbes d'occupation, et la qualité de la prédiction qu'ils permettent. La théorie de l'estimation permet en outre de quantifier l'erreur commise sur les estimateurs, et donc sur la prédiction, à partir de données recueillies à une étude intermédiaire. On a vu que les modèles pouvaient servir de base à des modèles plus complexes prenant en compte la perte de patients en phase de screening, eux-mêmes servant de base pour un modèle de coût. De plus, ils peuvent être simplement mis en oeuvre numériquement et utilisés pour un usage en routine.

5.2 Perspectives

Les perspectives de recherches s'articulent suivant deux axes, un axe théorique dont l'objectif est de mettre en avant les propriétés des modèles introduits dans cette thèse et un axe appliqué dont l'objectif est de développer ces outils et de les faire connaitre au monde médical. Pour ce dernier point nous disposons du soutien de l'IRESP qui a financé, dans le cadre de l'Appel d'Offre "Soutien à la recherche mathématique et statistique appliquée à la cancérologie" le projet "Statistic Modelling for Patients recruitment" porté par l'Institut de Mathématiques de Toulouse et l'Unité INSERM 1027 en partenariat avec Vladimir Anisimov.

5.2.1 Perspectives appliquées

Saturation des centres

Dans certains essais, les centres arrivent au maximum de leur capacité d'accueil de patients et le recrutement ralentit. Il serait possible de modéliser ce phénomène en supposant que l'intensité (stochastique) λ_i du processus N_i dépend de N_i :

$$\forall t \geq 0,\ \lambda_i(t) = \lambda_i^0\, \sigma_i(N_i(t))$$

où $\sigma_i : \mathbb{N} \to \mathbb{R}_+$ est une certaine fonction décroissante et λ_i^0 suit une loi Gamma. Par exemple, si la capacité du centre est de M_i patients, on aura $\sigma_i(n) = 0$ pour tout $n \geq M_i$.

Données de survie

Dans le cas des données de survie, un nouveau problème apparait : ce qu'il est utile de connaitre n'est pas nécessairement le nombre de patients mais le nombre d'événements d'intérêt à la fin de l'essai. Avec l'utilisation d'un modèle Gamma-Poisson pour l'inclusion et un modèle exponentiel pour les données de survie, Anisimov [7] obtient de bonnes prédictions du nombre d'événements.

Notre but est de développer un modèle plus fin que le modèle précédent, dans lequel une chaîne de Markov en temps continu est associée à chaque patient et dont les états représentent les différentes étapes de la progression de la maladie durant l'essai (à risque, perdu de vue, dans l'étude, décédé,...). En estimant les probabilités de transition, il sera possible de prédire plus précisément le nombre et le type d'événements à venir.

Modèle de coût

Une discussion avec les cliniciens est nécessaire pour déterminer l'applicabilité du modèle de coût présenté dans le chapitre 4. Le modèle aura très certainement à être ajusté. La difficulté sera de déterminer avec précision les valeurs des paramètres de coût utilisés (coût fixe d'un patient, d'un centre, etc). Ainsi, il sera utile d'étudier la sensibilité du coût aux différents paramètres, afin de mesurer comment se répercute une erreur d'un paramètre sur le coût total.

Covariables déterminant les taux d'inclusion

Le point clé pour une bonne prédiction du temps de recrutement est une estimation la plus précise possible des intensités de recrutement de chaque centre. Rechercher des covariables pouvant régir la valeur de ces intensités permettrait d'augmenter la précision de la prédiction. Ces covariables pourraient être la taille de la ville où le centre est implanté, la prédominance de la maladie étudiée dans la région, le fait qu'un autre centre soit ouvert à proximité, etc. Par exemple, nous pourrions utiliser un modèle de régression où le taux d'inclusion λ d'un centre vérifie

$$\lambda = \lambda_0 + \gamma^T X + \epsilon$$

où X est le vecteur de covariables et ϵ est un bruit. Un test de la nullité de γ permettrait de tester la pertinence du modèle.

5.2.2 Perspectives théoriques

Les perspectives théoriques concernent essentiellement les processus de Cox.

En premier lieu, nous envisageons d'affiner l'étude de sensibilité en cherchant à estimer la distance de Wasserstein entre deux processus de Cox, vus comme variables aléatoires

dans l'espace des configurations (cf [24]). On s'appuiera sur des travaux analogues de Barbour [14] dans le cas Poisson et Deucreusefond et al [24] dans le cas de processus ponctuels.

D'autre part, le modèle de coût fait apparaitre un processus N de la forme $t \mapsto \int_0^t g(t,s)dK_s$ où K est un processus de Cox. Ce type de processus est encore très peu étudié dans la littérature. Si toute la trajectoire du processus N est disponible, alors, sous de bonnes hypothèses sur g, on peut récupérer les instants de saut du processus sous-jacent K en regardant les variations de la dérivée de N. On se penchera sur la question de retrouver ces instants de saut dans le cas où l'on observe N seulement à des instants discrets. D'autre part, on cherchera à établir un théorème de type Girsanov, dans l'optique de l'estimation de paramètres par la méthode du maximum de vraisemblance.

Bibliographie

[1] I. Abbas, J. Rovira, and J. Casanovas, *Clinical trial optimization : Monte Carlo simulation Markov model for planning clinical trials recruitment*, Contemporary clinical trials **28** (2007), 220–231.

[2] V. V. Anisimov, *Using Mixed Poisson Models in patient recruit in multicentre clinical trials*, Proceegings of the World Congress on Ingineering (London, United Kingdom), vol. II, 2008.

[3] _____, *Using mixed poisson models in patient recruitment in multicentre clinical trials*, Proceegings of the World Congress on Ingineering (London, United Kingdom), vol. II, 2008, pp. 1046–1049.

[4] _____, *Predictive modelling of recruitment and drug supply in multicenter clinical trials*, Proceedings of the Joint Statistical Meeting, ASA (Washington, USA), August 2009, pp. 1248–1259.

[5] _____, *Recruitment modeling and predicting in clinical trials*, Pharmaceutical Outsourcing **10** (2009), no. 1, 44–48.

[6] _____, *Effects of unstratified and centre-stratified randomization in multi-centre clinical trials*, Pharmaceutical Statistics **10** (2011), no. 1, 50–59.

[7] _____, *Predictive event modelling in multicentre clinical trials with waiting time to response*, Pharmaceutical Statistics **10** (2011), no. 6, 517–522.

[8] _____, *Statistical modeling of clinical trials (recruitment and randomization)*, Comm. Statist. Theory Methods **40** (2011), no. 19-20, 3684–3699. MR 2860767 (2012i :62311)

[9] V. V. Anisimov, D. Downing, and V. V. Fedorov, *Recruitment in multicentre trials : prediction and adjustment*, 8th International Workshop in model-oriented design and analysis, Physica-Verlag/Springer, Heidelberg, June 2007, pp. 1–8. MR MR2409023

[10] V. V. Anisimov and V. V. Fedorov, *Modeling of enrolment and estimation of parameters in multicentre trials*, Tech. report, GSK BDS Technical Report, 2005.

[11] _____, *Modelling, prediction and adaptive adjustment of recruitment in multicentre trials*, Statistics in Medicine **26** (2007), no. 27, 4958–4975. MR MR2405491

[12] V. V. Anisimov, G. Mijoule, and N. Savy, *Statistical modelling of recruitment in multicentre clinical trials with patients' dropout*, Submitted to Statistics in Medicine, 2013.

[13] _____, *A toy model for the modelling of the cost of a clinical trial*, Proceedings of the 7th International Workshop on Simulation (Rimini, Italy), 2013.

[14] A. D. Barbour, T. C. Brown, and A. Xia, *Point processes in time and Stein's method*, Stochastics Stochastics Rep. **65** (1998), no. 1-2, 127–151. MR 1708412 (2000d :60085)

[15] A. D. Barbour, L. Holst, and S. Janson, *Poisson approximation*, Oxford Studies in Probability, vol. 2, The Clarendon Press Oxford University Press, New York, 1992, Oxford Science Publications. MR 1163825 (93g :60043)

[16] K. D. Barnard, L. Dent, and A. Cook, *A systematic review of models to predict recruitment to multicentre trials*, BMC Medical Research Methodology **63** (2010), no. 10.

[17] G. E. Bates and J. Neyman, *Contributions to the theory of accident proneness. II. True or false contagion*, Univ. California Publ. Statist. **1** (1952), 255–275. MR 0050837 (14,390a)

[18] J. M. Bernardo and A. F. M. Smith, *Bayesian theory*, John Wiley & Sons, Hoboken, NJ, USA, 2004.

[19] P. Brémaud, *Point processes and queues*, Sprfinger-Verlag, New York, 1981, Martingale dynamics, Springer Series in Statistics. MR 636252 (82m :60058)

[20] R. E. Carter, *Application of stochastic processes to participant recruitment in clinical trials*, Controlled Clinical Trials **25** (2004), no. 5, 429–436.

[21] R. E. Carter, S. C. Sonne, and K. T. Brady, *Practical considerations for estimating clinical trial accrual periods : application to a multi-center effectiveness study*, BMC Medical Research Methodology **5** (2005), no. 11, 1–5.

[22] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes. Vol. I*, second ed., Probability and its Applications (New York), Springer-Verlag, New York, 2003, Elementary theory and methods. MR 1950431 (2004c :60001)

[23] ———, *An introduction to the theory of point processes. Vol. II*, second ed., Probability and its Applications (New York), Springer, New York, 2008, General theory and structure. MR 2371524 (2009b :60150)

[24] L. Decreusefond, A. Joulin, and N. Savy, *Upper bounds on Rubinstein distances on configuration spaces and applications*, Commun. Stoch. Anal. **4** (2010), no. 3, 377–399. MR 2677197 (2011d :60148)

[25] L. Decreusefond and N. Savy, *Filtered Brownian motions as weak limit of filtered Poisson processes*, Bernoulli **11** (2005), no. 2, 283–292. MR 2132727 (2005j :60154)

[26] ———, *Anticipative calculus with respect to filtered Poisson processes*, Ann. Inst. H. Poincaré Probab. Statist. **42** (2006), no. 3, 343–372. MR 2219714 (2007b :60132)

[27] B. J. Gajewski, S. D. Simon, and S. E. Carlson, *Predicting accrual in clinical trials with Bayesian posterior predictive distributions*, Statistics in Medicine **27** (2008), no. 13, 2328–2340. MR 2432492

[28] P. E. Greenwood and M. S. Nikulin, *A guide to chi-squared testing*, Wiley Series in Probability and Statistics : Applied Probability and Statistics, John Wiley & Sons Inc., New York, 1996, A Wiley-Interscience Publication. MR 1379800 (97e :62002)

[29] E. G. Haug, *The complete guide to option pricing formulas*, McGraw-Hill, September 1997.

[30] N. L. Johnson, S. Kotz, and A. W. Kemp, *Univariate discrete distributions*, second ed., Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics, John Wiley & Sons Inc., New York, 1992, A Wiley-Interscience Publication. MR 1224449 (95d :62018)

[31] D. Lai, *Estimating the Hurst effect and its application in monitoring clinical trials*, Computational Statistics and Data Analysis **45** (2004), no. 3, 549–562. MR MR2050255 (2005a :62060)

[32] D. Lai, R. D. Barry, and J. H. Robert, *Fractional Brownian motion and clinical trials*, Journal of Applied Statistics **27** (2000), no. 1, 103–108.

[33] D. Lai, L. A. Moyé, B. R. Davis, L. E. Brown, and F. M. Sacks, *Brownian motion and long-term clinical trial recruitment*, Journal of Statistical Planning and Inference **93** (2001), no. 1-2, 239–246. MR MR1822399

[34] Y. J. Lee, *Interim recruitment goals in clinical trials*, Journal of Chronic Diseases **36** (1983), no. 5, 379–389.

[35] G. Mijoule, N. Savy, and S. Savy, *Models for patients recruitment in clinical trials and sensitivity analysis*, Statistics in Medicine **31(16)** (2012), 1655–1674.

[36] D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gotzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman, *Consort 2010 explanation and elaboration : updated guidelines for reporting parallel group randomised trials*, Journal of Clinical Epidemiology **In Press, Corrected Proof** (2010), –.

[37] T. M. Morgan, *Nonparametric estimation of duration of accrual and total study length for clinical trials*, Biometrics **43** (1987), no. 4, 903–912. MR MR920474

[38] International Conference on Harmonisation, *Statistical principles for clinical trials (ICH E9)*, Statistics in Medicine **18** (1999), 1905–1942.

[39] S. Piantadosi and B. Patterson, *A method for predicting accrual, cost, and paper flow in clinical trials*, Controlled Clinical Trials **8** (1987), no. 3, 202–215.

[40] M. A. Rojavin, *Recruitment index as a measure of patient recruitment activity in clinical trials*, Contemporary Clinical Trials **26** (2005), no. 5, 552–556.

[41] _____, *Patient recruitment and retention : From art to science*, Contemporary Clinical Trials **30** (2009), no. 5, 387–387.

[42] G. Samorodnitsky, *A class of shot noise models for financial applications*, Athens Conference on Applied Probability and Time Series Analysis, Vol. I (1995), Lecture Notes in Statist., vol. 114, Springer, New York, 1996, pp. 332–353. MR 1466727 (98f :60215)

[43] N. Savy, G. Mijoule, and S. Savy, *Approche méthodologique du recrutement de patients*, Bulletin du Cancer **97 (4)** (2010), S80.

[44] N. Savy and J. Vives, *Anticipative integrals with respect to a filtered Lévy process and Lévy-Itô decomposition*, Submitted to Bernoulli, 2013.

[45] K. F. Schulz, D. G. Altman, and D. Moher, *Consort 2010 statement : Updated guidelines for reporting parallel group randomised trials*, Journal of Clinical Epidemiology **In Press, Corrected Proof** (2010), –.

[46] S. Senn, *Statistical issues in drug development*, John Wiley & Sons, Chichester, 1997.

[47] _____, *Some controversies in planning and analysing multi-centre trials*, Statistics in Medicine **17** (1998), 1753–1765.

[48] A. W. van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 3, Cambridge University Press, Cambridge, 1998. MR 1652247 (2000c :62003)

[49] A-P. Vlasto, *Brevets et médicament en france. pourquoi l'application du droit des brevets au médicament est-elle autant critiquée ?*, Médecine & Droit **2007** (2007), no. 82, 25–32.

[50] W. O. Williford, S. F. Bingham, D. G. Weiss, J. F. Collins, K. T. Rains, and W. F. Krol, *The "constant intake rate" assumption in interim recruitment goal methodology for multicenter clinical trials.*, Journal of chronic diseases **40** (1987), no. 4, 297–307.

[51] A. Yagouti, I. Abi-Zeid, Taha B. M. J. Ouarta, and B. Bobée, *Revue de processus ponctuels et synthèse de tests statistiques pour le choix d'un type de processus (review and classification of statistical tests applied to point processes)*, Revue des sciences de l'eau **14** (2001), no. 3, 323–361.

Chapitre 6

Annexes

6.1 Données

Les quatre jeux de données utilisés dans la partie 1 et tirés de [10] sont les suivants.
Ici, $\nu = (\nu(1), \nu(2), \nu(3), \ldots)$.
Etude A : $C = 91$, ν = (7,11,8,8,9,8,9,7,2,4,1,3,3,4,0,0,2,1,1,2,1).
Etude B : $C = 54$, ν = (1,4,2,3,5,1,4,4,3,2,4,1,5,2,0,2,1,2,1,0,1,3,0,0,1,1,0,0,0,0,0,0,0,0,0,0,1).
Etude C : $C = 65$, ν = (1,1,1,0,3,5,2,1,2,2,5,3,7,3,5,1,1,0,4,2,4,2,2,0,1,2,2,1,1,0,0,0,0,0,0,0,0,0,0,1).
Etude D : $C = 59$, ν = (2,5,6,4,3,7,2,8,3,5,3,1,1,3,2,2,0,0,0,0,0,0,0,0,0,1,0,0,1).

6.2 Codes R

```
# %%%% loi de k_i %%%%
# calcule pour chaque modele Proba(k_i=n) Ã  t1

loi_gammapoisson = function(alpha, mu, tau, n) {
beta = alpha/mu
res = exp(lgamma(alpha + n) - lgamma(n + 1) - lgamma(alpha)) * betaˆalpha * tauˆn/(beta +
tau)ˆ(alpha + n)
res
}

loi_paretopoisson = function(gamma, delta, tau, n) {
res = deltaˆgamma * gamma * tauˆgamma * incgamma(n - gamma, delta * tau)/gamma(n +
1)
res
}

loi_gammapoissonuni = function(alpha, mu, u, t1, n) {
if (u > t1) {
res=0
} else {
beta = alpha/mu
res = exp(lgamma(alpha + n) - lgamma(n + 1) - lgamma(alpha)) * betaˆalpha/u *
B(alpha, beta, n, t1 - u, t1)
}
res
}
```

```
# %%%% log-vraisemblances %%%%

loglike_gamma = function(k, tau, alpha, mu) {
C = length(k)
y = sum(lgamma(alpha + k) - lgamma(alpha) - (alpha + k) * log(alpha/mu + tau) +
alpha * log(alpha/mu))
y/C
}

loglike_pareto = function(k, tau, gam, delta) {
C = length(k)
y = 0
for (i in 1:C) {
if (tau[i] > 0) {
y = y + gam * log(delta) + log(gam) + log(incgamma(k[i] - gam,
delta * tau[i])) + gam * log(tau[i])
}
}
y/C
}

loglike_gammauni = function(k, u, t1, alpha, mu) {
#u : vecteurs premiers instants d'inclusion de chaque centre
C = length(k)
y = sum(lgamma(alpha + k) - lgamma(alpha) + alpha * log(alpha/mu))
for (i in 1:C) {
y = y + log(B(alpha, alpha/mu, k[i], t1 - u[i], t1))
}
y/C
}

loglike_model2 = function(k, n, psi1, psi2) {
# deuxieme terme de la log-vrais. dans le modele 2 de dropout
sum(lbeta(psi1 + k, psi2 + n - k) - lbeta(psi1, psi2))
}

# %%%% matrices de variance-covariance globales %%%%
# M = 1/C sum_i=1^C J(theta, tau_i)

covar_gamma = function(alpha, mu, tau, nSimu = 1000) {
res = array(0, c(2, 2))
C = length(tau)
for (i in 1:C) {
res[1, 1] = res[1, 1] - mu * tau[i]/alpha/(alpha + mu * tau[i])
res[2, 2] = res[2, 2] + alpha * tau[i]/alpha/(alpha + mu * tau[i])
}
res[1, 1] = res[1, 1] + sum(esppsi(alpha, mu, tau, nSimu))
-res/C
}

covar_pareto = function(gamma, delta, tau, dg = 0.001, dd = 0.001, nSimu = 1000) {
temp_uni = runif(nSimu, min = 0, max = 1)
C = length(tau)
lambda = delta*(1-temp_uni)^(-1/gamma)
```

```
# lambdapg = delta*(1-temp_uni)^(-1/(gamma+dg))
# lambdamg = delta*(1-temp_uni)^(-1/(gamma-dg))
# lambdapd = (delta+dd)*(1-temp_uni)^(-1/gamma)
# lambdamd = (delta-dd)*(1-temp_uni)^(-1/gamma)
res = array(0, c(2, 2))
for (i in 1:C) {
k = rpois(nSimu, lambda * tau[i])
f = loglike_pareto(k, tau[i], gamma, delta)
fpgpd = loglike_pareto(k, tau[i], gamma*(1+dg), delta*(1+dd))
fpg = loglike_pareto(k, tau[i], gamma*(1+dg), delta)
fmg = loglike_pareto(k, tau[i], gamma*(1-dg), delta)
fpd = loglike_pareto(k, tau[i], gamma, delta*(1+dd))
fmd = loglike_pareto(k, tau[i], gamma, delta*(1-dd))
res[1, 1] = res[1, 1] + (fpg + fmg - 2 * f)/dg^2/gamma^2
res[2, 2] = res[2, 2] + (fpd + fmd - 2 * f)/dd^2/delta^2
res[1, 2] = res[1, 2] + (fpgpd - fpg - fpd + f)/dg/dd/gamma/delta
}
res[2, 1] = res[1, 2]
-res/C
}


covar_gammauni = function(alpha, mu, u, t1, da = 0.001, dmu = 0.001, nSimu = 1000) {
temp_uni = runif(nSimu, min = 0, max = 1)
C = length(u)
lambda = rgamma(nSimu, alpha, alpha/mu)
arrayone = array(1, nSimu)
res = array(0, c(2, 2))
for (i in 1:C) {
k = rpois(nSimu, lambda * (t1 - u[i]*temp_uni))
utemp = u[i] * arrayone
f = loglike_gammauni(k, utemp, t1, alpha, mu)
fpapm = loglike_gammauni(k, utemp, t1, alpha*(1 + da), mu*(1 + dmu))
fpa = loglike_gammauni(k, utemp, t1, alpha*(1 + da), mu)
fma = loglike_gammauni(k, utemp, t1, alpha*(1 - da), mu)
fpm = loglike_gammauni(k, utemp, t1, alpha, mu*(1 + dmu))
fmm = loglike_gammauni(k, utemp, t1, alpha, mu*(1 - dmu))
res[1, 1] = res[1, 1] + (fpa + fma - 2 * f)/da^2/alpha^2
res[2, 2] = res[2, 2] + (fpm + fmm - 2 * f)/dmu^2/mu^2
res[1, 2] = res[1, 2] + (fpapm - fpa - fpm + f)/da/dmu/alpha/mu
}
res[2, 1] = res[1, 2]
-res/C
}

# %%%%% fonctions annexes %%%%

incgamma = function(x, y) {
# fonction gamma incomplete : int_y^+infty exp(-t) t^(x-1)dt
if (x > 0) {
res = gamma(x) * (1 - pgamma(y, x))
} else {
a = -floor(x)
atilde = a + x
if (atilde == 0) {
```

```
res = expint_E1(y)
} else {
res = gamma(atilde) * (1 - pgamma(y, atilde))
}
for (j in 1:a) {
res = 1/(atilde - 1) * (res - exp(-y) * (y)^(atilde - 1))
atilde = atilde - 1
}
}
res
}

incgammainv = function(y,delta,gamma){

}

esppsi = function(alpha, mu, tau, nSimu) {
# Calcule E[psi(alpha)-psi(alpha+k_i)] par simulations de MC dans le modele gamma-poisson
C = length(tau)
temp = 0
res = array(0, c(C))
beta = alpha/mu
for (i in 1:nSimu) {
lambda = rgamma(1, alpha, beta)
n = rpois(C, lambda * tau)
res = res + (trigamma(alpha) - trigamma(alpha + n))/nSimu
}
res
}

B = function(a, b, k, Sp, S) {
# fonction utilisÃ©e dans la log-vrais. du modele gamma uniforme
res = 0
if (a>1){
res=b^(1-a)*(pbeta(b/(Sp+b),a-1,k+1)-pbeta(b/(S+b),a-1,k+1) )*beta(a-1,k+1)
}
else if ((a<1) && k>0){
res = 1/(a-1)*(Sp^k/(Sp+b)^(a+k-1)-S^k/(S+b)^(a+k-1)+b^(1-a)*k*beta(a,k)*(pbeta(b/(Sp+b),a,k)-pbeta(
}
else if (a<1){
res = 1/(1-a) * ((S + b)^( - a + 1) - (Sp + b)^( - a + 1))
}
else if (a == 1) {
f = function (x){ x^(-k-1)*(x-b)^k}
res=integrate(f,Sp+b,S+b)
}
res
}

# %%%% Calcul de l'EMV %%%%%
# k et tau supposes donnes

calculate_emv = function(k, tau) {
# calcule l'EMV dans les modeles gamma-poisson et pareto-poisson
# renvoie une liste avec param. pour gamma-poisson (alpha,mu) et pour pareto-poisson (gamma,delta)
if (length(tau) == 1) {
```

```
tau = array(tau,c(length(k)))
}
tomin1 = function(x) {
# x[1]=alpha, x[2]=mu
-loglike_gamma(k, tau, x[1], x[2])
}
tomin2 = function(x) {
# x[1]=gamma, x[2]=delta
-loglike_pareto(k, tau, x[1], x[2])
}
res = list(alpha = 0, mu = 0, gamma = 0, delta = 0)
a = optim(c(2, 2), tomin1, gr = NULL, method = "L-BFGS-B", lower = c(0.001, 0.001),
upper = c(10, 1000))
param = a$par
res$alpha = param[1]
res$mu = param[2]
#a = optim(c(2, 2), tomin2, gr = NULL, method = "L-BFGS-B", lower = c(0.001, 0.001),
# upper = c(100, 100))
param = a$par
res$gamma = param[1]
res$delta = param[2]
res
}

calculate_emv_gammauni = function(k, u, t1) {
# calcule l'EMV dans le modÃ¨le Gamma-Poisson uniforme
# k : vecteurs de patients inclus
# u : dates de premiere inclusion de chaque centre
C = length(k)
if (length(u) == 1) {
u = array(u, c(C))
}
u2=u
for (i in 1:C) {
u2[i]=min(u2[i],t1)
}
tomin = function(x) {
alpha = x[1]
mu = x[2]
-loglike_gammauni(k, u2, t1, alpha, mu)
}
res = list(alpha = 0, mu = 0)
a = optim(c(2, 2), tomin, gr = NULL, method = "L-BFGS-B", lower = c(0.001, 0.001),
upper = c(100, 100))
param = a$par
res$alpha = param[1]
res$mu = param[2]
res
}

calculate_emv_dropout_2 = function(k, n, tau) {
# calcul par emv de alpha,mu,psi1,psi2 dans le modele 2 de dropout
# k : patients randomises.
# n : patients screenes.
if (length(tau) == 1) {
```

```
tau = array(tau,c(length(k)))
}
tomin1 = function(x) {
#alpha = x[1], mu = x[2]
-loglike_gamma(n, tau, x[1], x[2])
}
tomin2 = function(x) {
#psi1 = x[1], psi2 = x[2]
-loglike_model2(k, n, x[1], x[2])
}
res = list(alpha = 0, mu = 0, psi1 = 0, psi2 = 0)
a = optim(c(2, 2), tomin1, gr = NULL, method = "L-BFGS-B", lower = c(0.001, 0.001),
upper = c(100, 100))
param = a$par
res$alpha = param[1]
res$mu = param[2]
a = optim(c(2, 2), tomin2, gr = NULL, method = "L-BFGS-B", lower = c(0.001, 0.001),
upper = c(100, 100))
param = a$par
res$psi1 = param[1]
res$psi2 = param[2]
res
}

courbe_occup = function(nu, tau, param, writeinfile = FALSE, name = "occup_curve") {
#trace les courbes d'occupation dans les modeles Gamma-Poisson et Pareto-Poisson
C = sum(nu)
if (length(tau) == 1) {
temp = tau
tau = array(temp, c(C))
}
alpha = param$alpha
mu = param$mu
gamma = param$gamma
delta = param$delta
x1 = 1:length(nu)
x2 = seq(0.5, length(nu) + 0.5, by = 0.01)
y = nu
ygamma = array(0, c(length(x2)))
ypareto = array(0, c(length(x2)))
for (i in 1:length(x2)) {
for (j in 1:C) {
ygamma[i] = ygamma[i] + loi_gammapoisson(alpha, mu, tau[j], x2[i])
ypareto[i] = ypareto[i] + loi_paretopoisson(gamma, delta, tau[j], x2[i])
}
}
ymax = 1.1 * max(c(y, ygamma, ypareto))
if (writeinfile) {
postscript(paste(name, ".eps", sep = ""), width = 3, height = 4, horizontal = FALSE,
onefile = FALSE, paper = "special", family = "ComputerModern", encoding = "TeXtext.enc")
}
plot(x1, y, type = "l", col = "green", xlab = "", ylab = "nu", ylim = c(0, ymax))
lines(x2, ygamma, type = "l", col = "red")
lines(x2, ypareto, type = "l", col = "blue")
if (writeinfile) {
dev.off()
```

```
}
}

courbe_occup_gammauni = function(k, u, t1, param, writeinfile = FALSE, name = "occup_curve_uni") {
#trace la courbe d'occupation dans le modele Gamma-Poisson uniforme
C = length(k)
u2 = u
for (i in 1:C) {
u2[i] = min(u2[i], t1)
}
alpha = param$alpha
mu = param$mu
xmax = max(k)
x1 = 1:xmax
x2 = seq(0, xmax + 0.5, by = 0.1)
y = array(0, c(xmax))
for (i in 1:C) {
y[k[i]] = y[k[i]] + 1
}
ygamma = array(0, c(length(x2)))
for (i in 1:length(x2)) {
for (j in 1:C) {
ygamma[i] = ygamma[i] + loi_gammapoissonuni(alpha, mu, u2[j], t1, x2[i])
}
}
ymax = 1.1 * max(c(y, ygamma))
if (writeinfile) {
postscript(paste(name, ".eps", sep = ""), width = 3, height = 4, horizontal = FALSE,
onefile = FALSE, paper = "special", family = "ComputerModern", encoding = "TeXtext.enc")
}
plot(x1, y, type = "l", col = "green", xlab = "i", ylab = expression(nu[i]),
ylim = c(0, ymax))
lines(x2, ygamma, type = "l", col = "red")
if (writeinfile) {
dev.off()
}
}

test_occup = function(nu, tau, param) {
# Effectue le test du Chi 2 sur les courbes d'occupation pour les modeles Gamma-Poisson et Pareto-Po
# tau est le meme pour tous les centres
indices = c()
C = sum(nu)
m = length(nu)
alpha = param$alpha
mu = param$mu
gamma = param$gamma
delta = param$delta
p_gamma = c()
p_pareto = c()
nu_gamma = c()
nu_pareto = c()
i = m
while (i >= 1) {
nu_gamma = c(nu[i], nu_gamma)
```

```
p_gamma = c(loi_gammapoisson(alpha, mu, tau, i), p_gamma)
j = 0
while ((nu_gamma[1] < 5) && (i - j >= 2) && (i - j - 1 > 0.2 * m)) {
j = j + 1
p_gamma[1] = p_gamma[1] + loi_gammapoisson(alpha, mu, tau, i - j)
nu_gamma[1] = nu_gamma[1] + nu[i - j]
}
i = i - j - 1
}
i = m
while (i >= 1) {
nu_pareto = c(nu[i], nu_pareto)
p_pareto = c(loi_paretopoisson(gamma, delta, tau, i), p_pareto)
j = 0
while ((nu_pareto[1] < 5) && (i - j >= 2) && (i - j - 1 > 0.2 * m)) {
j = j + 1
p_pareto[1] = p_pareto[1] + loi_paretopoisson(gamma, delta, tau, i -
j)
nu_pareto[1] = nu_pareto[1] + nu[i - j]
}
i = i - j - 1
}
res = list(pgamma = p_gamma, ppareto = p_pareto, nugamma = nu_gamma, nupareto = nu_pareto)
list(pgamma = p_gamma, nugamma = nu_gamma, statgamma = sum((res$nug - C * res$pgam)^2/C/res$pgam),
deglibgamma = length(nu_gamma) - 1, ppareto = p_pareto, nupareto = nu_pareto,
statpareto = sum((res$nup - C * res$ppar)^2/C/res$ppar), deglibpareto = length(nu_pareto) -
1)
}

### Fonctions msigma2 : calculent le vecteur E[lambda_i | k_i] dans chaque modele

msigma2gamma = function(alpha, mu, k, tau) {
res = list(m = (alpha + k)/(alpha/mu + tau), sigma2 = (alpha + k)/(alpha/mu +
tau)^2)
}

msigma2pareto = function(gamma, delta, k, tau) {
C = length(k)
res = list(m = array(0, C), sigma2 = array(0, C))
if (length(tau) == 1) {
tau = array(tau, C)
}
for (i in 1:C) {
res$m[i] = incgamma(k[i] - gamma + 1, delta * tau[i])/incgamma(k[i] - gamma,
delta * tau[i])/tau[i]
res$sigma2[i] = incgamma(k[i] - gamma + 2, delta * tau[i])/incgamma(k[i] -
gamma, delta * tau[i])/tau[i]^2 - res$m[i]^2
}
}

msigma2gammauni = function(alpha, mu, k, Sp, S) {
res = list(m = (alpha + k)/(S - Sp) * log((alpha/mu + S)/(alpha/mu + Sp)), sigma2 = (alpha +
k)/(alpha/mu + Sp)/(alpha/mu + S))
}

###############
```

```
AB = function(m, sigma2) {
#Calcule paramtres A,B de la loi gamma approchant somme(lambda_i)
c((sum(m))^2/sum(sigma2), sum(m)/sum(sigma2))
}

##############
espproT = function(A, B, K, x, p) {
#calcule Esp[T],quantile d'ordre p de T, et Pro[T<x]
# K : nbre de patients restant Ã  recruter
pro = pbeta(x/(x + B), K, A)
temp1 = 0
temp2 = 10
res1 = pbeta(temp1/(temp1 + B), K, A)
res2 = pbeta(temp2/(temp2 + B), K, A)
while (abs(res1 - res2) > 1e-07) {
temp = 0.5 * (temp1 + temp2)
res = pbeta(temp/(temp + B), K, A)
if (res < p) {
temp1 = temp
res1 = res
} else {
temp2 = temp
res2 = res
}
}
if (A > 1) {
c(K * B/(A - 1), temp, pro)
} else {
c(0, temp, pro)
}
}


gradient_muTqpT = function(alpha, mu, k, u, t1, Kf, del = 1e-03,p=0.95) {
# pour le modele gamma uniforme
# calcule muT=E[T] et qpT tq Pro[T<qpT]=p, ainsi que le gradient de ces quantites par rapport aux pa
K1 = sum(k)
temp = msigma2gammauni(alpha, mu, k, t1 - u, t1)
tempa = msigma2gammauni(alpha*(1 + del), mu, k, t1 - u, t1)
tempa2 = msigma2gammauni(alpha*(1 - del), mu, k, t1 - u, t1)
tempmu = msigma2gammauni(alpha, mu*( 1+ del), k, t1 - u, t1)
tempmu2 = msigma2gammauni(alpha, mu*( 1- del), k, t1 - u, t1)
temp = AB(temp$m, temp$sigma2)
tempa = AB(tempa$m, tempa$sigma2)
tempa2 = AB(tempa2$m, tempa2$sigma2)
tempmu = AB(tempmu$m, tempmu$sigma2)
tempmu2 = AB(tempmu2$m, tempmu2$sigma2)
res = list(muT = 0, qpT = 0, gradmuT = c(0, 0), gradqpT = c(0, 0),pro=0)
res$pro=pbeta((3-t1)/((3-t1) + temp[2]), Kf-K1, temp[1])
temp = espproT(temp[1], temp[2], Kf-K1, 3 - t1, p)
tempa = espproT(tempa[1], tempa[2], Kf-K1, 3 - t1, p)
tempa2 = espproT(tempa2[1], tempa2[2], Kf-K1, 3 - t1, p)
tempmu = espproT(tempmu[1], tempmu[2], Kf-K1, 3 - t1, p)
tempmu2 = espproT(tempmu2[1], tempmu2[2], Kf-K1, 3 - t1, p)
res$muT = t1+temp[1]
```

```
res$qpT = t1+temp[2]
res$gradmuT = c((tempa[1] - tempa2[1])/del/alpha/2, (tempmu[1] - tempmu2[1])/del/mu/2)
res$gradqpT = c((tempa[2] - tempa2[2])/del/alpha/2, (tempmu[2] - tempmu2[2])/del/mu/2)
res
}



cout = function(alpha, mu, psi1, psi2, k, n, tau, Gj, Fj, C2, C3, K, nSimu = 1e4) {
# Calcule le vecteur de differentiel de cout Delta C_j en fermant le centre j
C = length(k)
beta = alpha/mu
if (length(Gj) == 1) {
Gj = array(Gj, c(C))
}
if (length(Fj) == 1) {
Fj = array(Fj, c(C))
}
sumG = sum(Gj)
sumF = sum(Fj)
res = array(0, c(C))
if (length(tau) == 1) {
tau = array(tau, c(C))
}
for (i in 1:nSimu) {
lambda = rgamma(C, alpha + n, beta + tau)
p = rbeta(C, psi1 + k, psi2 + n - k)
Lambda = sum(lambda * p)
Lambda2 = sum(lambda * (1 - p))
res = res + (K * C2* (Lambda2 - lambda * (1 - p))/(Lambda - lambda * p) + 0.5 *
C3 * (K - 1) * K/(Lambda - p * lambda) + K * Gj/(Lambda - p * lambda))/nSimu
res = res - (K * C2* Lambda2/Lambda + 0.5 * C3 * (K - 1) * K/(Lambda) + K *
Gj/(Lambda))/nSimu
}
res = res - Fj
res
}
```