

Faculté des sciences

Méthodes de détection d'une valeur aberrante dans une méta-analyse

Auteure : LADEKPO Chikêola

Promotrice : LEGRAND Catherine

Co-promotrice : CHEUVART Brigitte

Lectrice : BUGLI Céline

Année académique 2019-2020

Master en statistique, orientation générale, à finalité spécialisée

Remerciements

A toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce mémoire.

Résumé

La qualité des données étant une garantie du succès d'un essai clinique, la présente étude contribue à développer une méthodologie pour la détection de valeurs aberrantes dans une méta-analyse d'essai clinique multicentriques à exécuter à l'aide du logiciel SAS. Deux méthodes pertinentes au contexte de l'étude ont été identifiées au terme de la revue de littérature : celle des Résidus Externes Studentisés (RES) et celle de Variance Shift Outlier Model (VSOM). Le présent document présente la comparaison de la mise en œuvre et de la performance de ces deux méthodes d'abord dans le cas d'un modèle de régression linéaire classique à un intercept et un modérateur puis dans le cas d'un modèle mixte à un effet fixe et un effet aléatoire avec composante de variance. Des simulations ont été réalisées dans les deux cas et les décisions de détection des deux méthodes reportées ainsi que la distribution des statistiques de détection sur l'ensemble des échantillons.

La méthode des RES apparait plus aisée et plus rapide à mettre en oeuvre que celle du VSOM. Aussi détecte-t-elle les valeurs aberrantes sur plus d'échantillons que la méthode du VSOM. En somme, au vu des résultats des simulations, la méthode des RES telle qu'appliquée dans la présente étude conviendrait mieux dans le contexte de ce travail pour détecter les valeurs aberrantes dans une méta-analyse. Ces deux méthodes appliquée à des données réelles, conduisent à la même décision quant à la déviance d'une nouvelle étude intégrée aux données d'une méta-analyse. Ces deux méthodes peuvent être appliquées de manière complémentaire. Des études ultérieures pourraient permettre de perfectionner leur mise en œuvre afin d'améliorer leur performance quant à la détection des valeurs aberrantes dans une méta-analyse.

Table des matières

Remerciements	i
Résumé	i
Liste des abréviations	vi
Liste des figures	vii
Liste des tableaux	viii
Introduction	1
1 Contexte/Problématique	3
1.1 Généralités	3
1.2 Objectifs de recherche	5
2 Concepts théoriques et état de l’art	6
2.1 Méta-analyses (MA)	6
2.1.1 Méta-analyses basées sur des données agrégées (MA-DA)	6
2.1.2 Méta-analyses basées sur des données individuelles (MA-DI)	7
2.2 Modèles linéaires	7
2.2.1 Modèle linéaire classique (Modèle Linéaire (ML))	8
2.2.1.1 Équation du modèle	8
2.2.1.2 Estimation des paramètres	8
2.2.2 Modèle linéaire mixte (Modèle Linéaire Mixte (MLM))	9
2.2.2.1 Spécification du modèle	9
2.2.2.2 Équation du modèle	10
2.2.2.3 Estimation des paramètres du modèle	11

2.3	Données aberrantes	11
2.3.1	Données aberrantes dans un Modèle Linéaire (ML)	12
2.3.2	Données aberrantes dans un Modèle Linéaire Mixte (MLM)	13
3	Méthodologie	15
3.1	Données	15
3.2	Méthodes de détection de valeurs aberrantes	16
3.2.1	Résidus externes studentisés (RES)	16
3.2.1.1	Cas de la régression linéaire classique	16
3.2.1.2	Cas du modèle mixte	17
3.2.2	Variance Shift Outlier Model (VSOM)	17
3.2.2.1	Cas de la régression linéaire classique	17
3.2.2.2	Cas du modèle mixte	19
3.2.3	Comparaison théorique des deux méthodes	20
3.3	Approche de simulation	20
3.3.1	Cas de la régression linéaire simple avec des observations aberrantes	21
3.3.1.1	Application de la méthode RES dans le cas du Modèle Linéaire (ML)	21
3.3.1.2	Application de la méthode VSOM dans le cas du Modèle Linéaire (ML)	21
3.3.1.3	Comparaison de la performance des deux méthodes dans le cas du Modèle Linéaire (ML)	22
3.3.2	Cas du modèle linéaire mixte avec un cluster aberrant	22
3.3.2.1	Application de la méthode RES dans le cas du Modèle Linéaire Mixte (MLM)	23
3.3.2.2	Application de la méthode VSOM dans le cas du Modèle Linéaire Mixte (MLM)	24
3.3.2.3	Comparaison de la performance des deux méthodes dans le cas du Modèle Linéaire Mixte (MLM)	25
3.4	Application des deux méthodes sur des données réelles	25
3.4.1	Application de la méthode des RES sur les données réelles	26
3.4.2	Application de la méthode du VSOM sur les données réelles	26
4	Résultats	27
4.1	Résultats des simulations	27

4.1.1	Cas de la régression linéaire simple avec des observations aberrantes	27
4.1.1.1	Résultats de l'application de la méthode des RES dans le cas du Modèle Linéaire (ML)	27
4.1.1.2	Résultats de l'application de la méthode du VSOM dans le cas du Modèle Linéaire (ML)	28
4.1.1.3	Résultats de la comparaison de la performance des deux méthodes dans le cas du Modèle Linéaire (ML)	30
4.1.2	Cas du modèle linéaire mixte avec un cluster aberrant	33
4.1.2.1	Résultats de l'application de la méthode des RES dans le cas du Modèle Linéaire Mixte (MLM)	33
4.1.2.2	Résultats de l'application de la méthode du VSOM dans le cas du Modèle Linéaire Mixte (MLM)	33
4.1.2.3	Résultats de la comparaison de la performance des deux méthodes dans le cas du Modèle Linéaire Mixte (MLM)	34
4.2	Résultats de l'application des deux méthodes sur des données réelles	35
4.2.0.1	Résultats de l'application de la méthode des RES sur les données réelles	36
4.2.0.2	Résultats de l'application de la méthode du VSOM sur les données réelles	36
5	Discussion/Conclusion	45
	Références bibliographiques	49
	Annexes	ix
	Annexes confidentielles	x
	Application des deux méthodes aux données réelles	x
	Programme SAS	xiii

Abréviations

BLUE Best Linear Unbiased Estimateur

EMV Estimateur du Maximum de Vraisemblance

IC Intervalle de Confiance

iid indépendant et identiquement distribué

ISBA Institut de Statistique Biostatistique et Sciences Actuarielles

MA Méta-analyses

MA-DA Méta-analyses basées sur des données agrégées

MA-DI Méta-analyses basées sur des données individuelles

ML Modèle Linéaire

MLM Modèle Linéaire Mixte

MV Maximum de Vraisemblance

REMV Estimateur du Maximum de Vraisemblance Restreint

RES Résidus Externes Studentisés

RIS Résidus Internes Studentisés

UCL Université Catholique de Louvain

VSOM Variance Shift Outlier Model

Table des figures

4.1	Histogramme des résidus externes studentisés par observation ($X_1 \sim U[0; 10]$, $\epsilon_1 \sim N(0, 1)$, $\epsilon_{12} = \epsilon_{30} = 5$ et $X_{30} = 20$)	28
4.2	Histogrammes du facteur d'inflation de la variance des effets aléatoires d_1 , d_{12} et d_{30} par rapport à la variance résiduelle et des Résidus Internes Studentisés (RIS) associés	37
4.3	Décision selon les Résidus Internes Studentisés (RIS) t_i^2	38
4.4	Décision selon la statistique du rapport de vraisemblance LRT_i	39
4.5	Pourcentage d'échantillons identifiant chaque observation aberrante en fonction des différentes valeurs de ϵ selon les RES	40
4.6	Pourcentage d'échantillons identifiant chaque observation aberrante en fonction des différentes valeurs de ϵ selon t_i^2	41
4.7	Histogramme des résidus externes studentisés des études A et E ($y_A = y + 10$)	41
4.8	Histogramme des facteurs d'inflation de la variance des effets aléatoires d_A et d_E)	42
4.9	Statistiques de décision de la méthode du VSOM, $y_A = y + 10$	42
4.10	Distribution des RES des études A et E	43
4.11	Distribution des LRT_j des études A et E	44

Liste des tableaux

4.1	Résidus externes studentisés des observations 1, 12 et 30	27
4.2	Résidus internes studentisés des observations 1, 12 et 30	29
4.3	Variance des effets aléatoires	29
4.4	Evolution des statistiques de détection selon la modification de la valeur de X_{30}	31
4.5	Evolution des statistiques de détection selon la modification de la valeur de $\epsilon_{12;30}$	31
4.6	Evolution des statistiques de détections selon le changement de l'amplitude de la déviance des valeurs aberrantes	34
5.1	Statistiques descriptives des modérateurs expliquant la réponse immuni- taire à l'antigène N (n=31.826)	x
5.2	Coefficients estimés du modèle linéaire mixte initial expliquant la concen- tration d'anticorps dans le sang	xi
5.3	Coefficients estimés du modèle linéaire mixte sans la nouvelle étude	xii

Introduction

La médecine factuelle ou "Evidence-Based Medicine" consiste à "intégrer les expertises cliniques individuelles aux meilleures données cliniques externes disponibles provenant de la recherche systématique"[23]. Pour y parvenir, des essais cliniques multicentriques sont de plus en plus réalisés afin de rassembler des données suffisantes et de qualité à même de conduire à des décisions sanitaires bien éclairées. L'essor des essais cliniques multicentriques a promu l'utilisation des modèles mixtes pour la recherche biomédicale. La littérature est largement développée en ce qui concerne le diagnostic d'un Modèle Linéaire (ML) classique. La tâche s'avère plus ardue lorsqu'il s'agit d'un Modèle Linéaire Mixte (MLM) du fait de la dépendance entre les observations induites par les effets aléatoires et/ou la structure de corrélation imposée dans les résidus.

La présente étude porte sur l'identification des valeurs aberrantes dans une méta-analyse à l'aide de SAS. Nous nous limitons au cas du MLM à un effet aléatoire avec une composante de variance. En effet, les méta-analyses visant à synthétiser les résultats issus de divers études réalisées dans des conditions différentes (localisation et période notamment) pour tirer une conclusion globale sur le sujet en cause, leur modélisation inclut de plus en plus des effets aléatoires et/ou une corrélation entre les individus ayant une ou plusieurs caractéristiques communes.

Le présent travail est réalisé dans le cadre de la surveillance de la réponse immunitaire à un vaccin dans le temps et dans l'espace. Chaque année en Belgique et dans d'autres pays européens, une nouvelle formule du vaccin contre la grippe est administrée. L'analyse des prélèvements sanguins de personnes vaccinées permet de mesurer et de comparer la quantité d'anticorps produites dans le sang chaque année et dans chaque pays. Une réponse immunitaire atypique dans un pays à un moment donné nécessiterait une attention particulière, d'où l'intérêt de cette étude.

Une méthode typique de détection des valeurs aberrantes dans un ML classique consiste à examiner les résidus studentisés. Des approches similaires ont été développées pour des

méta-analyses. Cependant les méthodes implémentées en SAS ne prennent pas en compte la variance associée à un effet aléatoire. L'objectif de la recherche est donc de développer une méthodologie en SAS pour détecter un ensemble de données aberrantes qui tienne compte des effets fixes et aléatoires. La méthodologie serait appliquée aux données simulées et réelles. Pour ce faire, deux méthodes de détection de valeurs aberrantes dans une méta-analyse seront comparées du point de vue de leur implémentation et de la performance pour décider si la réponse immunitaire moyenne d'une formulation spécifique d'un vaccin est atypique ou non. Il s'agit de la méthode :

- des Résidus Externes Studentisés (RES) et
- de celle du "Variance Shift Outlier Model" (VSOM) .

Ce document, présente la problématique traitée et les concepts théoriques abordés, la méthodologie utilisée pour répondre à la question de recherche ainsi que les résultats de l'étude puis la discussion et la conclusion des recherches.

Chapitre 1

Contexte/Problématique

1.1 Généralités

Le succès d'un essai clinique est entre autres garanti par la qualité des données générées au cours de sa réalisation. Venet et al. [29] suggèrent d'appliquer une méthodologie statistique à la surveillance de la qualité des données dans le cadre des essais cliniques. Ils proposent une méthodologie qui peut être également appliquée pour la détection de fraudes ou de problèmes de qualité de données dans le domaine de la finance et des assurances [4]. Elle consiste à détecter par des tests statistiques si les données présentant des caractéristiques communes ont une tendance significativement différente de l'ensemble des données contenues dans la base. Les données apparaissant atypiques sont alors examinées de plus près pour mieux comprendre la source de leur déviance afin d'y remédier au besoin. La présente étude s'inscrit dans le cadre d'une initiative globale d'une entreprise pharmaceutique de modéliser et de surveiller la réponse immunitaire à plusieurs antigènes dans le temps et dans l'espace. Ceci permet de s'assurer de maintenir le niveau d'efficacité du vaccin malgré les changements ultérieurs de procédure de fabrication. Contribuant donc à développer une méthodologie dans cette perspective, les statistiques utilisées dans le cadre de ce travail pourraient être intégrées à la batterie de statistiques considérées par l'approche de "Central Statistical Monitoring" proposée par Venet et al [29] pour garantir la qualité des données.

Le caractère multicentrique d'un essai clinique en améliore également la qualité. Il permet de réunir un plus grand nombre de patients présentant des caractéristiques variées générant ainsi des données plus précises et conduisant à des résultats plus généralisables [21]. Cependant, avec ces avantages apparait la complexité liée à la modélisation d'un tel

type d'étude. En effet, pour mieux appréhender la variabilité entre les individus enrôlés dans les différents centres participant à l'étude, il est nécessaire de rajouter au modèle de régression linéaire un effet aléatoire centre et/ou une structure de corrélation entre les individus provenant d'un même centre. Le ML devient de ce fait un MLM. Le diagnostic du modèle faisant partie intégrante de la construction d'un modèle de régression, le sujet est largement traité dans la littérature en ce qui concerne les ML classiques. Dans le cas des MLM par contre, le sujet est actuellement en pleine expansion d'autant plus que les MLM sont de plus en plus utilisés dans le domaine de la recherche médicale. La présente étude est donc réalisée pour combler le déficit de connaissance dans le domaine.

Par ailleurs, l'utilisation de SAS est essentielle à la gestion des données cliniques [9]. Plusieurs auteurs proposent dans la littérature des méthodes de détection de valeurs aberrantes à l'aide de R [30]. On ne retrouve cependant pas des propositions récentes de méthodes implémentées avec SAS. Dans le cas du modèle mixte, les RES sont calculés au niveau individuel avec la procédure "mixed" de SAS mais pas au niveau d'un groupe d'observation. De plus, pour le diagnostic au niveau d'un groupe d'individus la procédure "mixed" de SAS ne propose que des statistiques d'influence et non des statistiques de détection de valeurs aberrantes [24]. Or, la présente étude s'intéresse spécifiquement aux statistiques de détection de valeurs aberrantes et non aux statistiques d'influence puisque l'intérêt du diagnostic réside dans l'identification de groupes d'observations dont la tendance diffère de la tendance globale de l'ensemble des données et pas particulièrement l'influence de ces valeurs sur les résultats de l'analyse puisqu'il pourrait n'y en avoir aucune.

A terme, les résultats de cette étude sont destinés à être exploités dans une entreprise pharmaceutique. Historiquement au sein de cette entreprise, les changements de processus de fabrication (y compris les changements de sites de fabrication) n'avaient pas été systématiquement soutenus par une étude clinique mettant en relation les anciens et les nouveaux processus. Ces modifications étaient généralement validées sur base de données de transition techniques. Cependant, dans les années à venir, il est prévu d'améliorer les processus de fabrication de certains antigènes. Il serait alors judicieux de surveiller plus étroitement les réponses cliniques au vaccin au fur et à mesure de la mise en œuvre de ces changements. Un premier modèle de méta-analyse a été réalisé pour établir une base de référence à partir de laquelle les futures données d'immunogénicité générées avec les antigènes concernés seront comparés. Les études ultérieures sont incluses à la méta-analyse initiale afin d'apprécier si la réponse immunitaire à ces antigènes est affectée par les modifications implémentées. Les études se déroulant dans plusieurs pays, un modèle mixte

à un effet aléatoire pays avec composante de variance est ajusté pour mesurer l'effet de divers facteurs sur la réponse immunitaire du vaccin. Le modèle final est appliqué après avoir identifié et supprimé les valeurs aberrantes confirmées d'étude-pays.

Avant la réalisation du présent travail, les valeurs aberrantes potentielles étude-pays en considérant certains facteurs pronostics inclus dans le modèle étaient détectées à l'aide de la somme des carrés des résidus de l'erreur prédite (PRESS) du modèle. La statistique PRESS pour un échantillon d'observations se calcule comme la somme des carrés de la différence entre la valeur observée pour cet échantillon et la valeur prédite à partir des effets fixes estimés sans les observations de cet échantillon. Cette statistique divisée par le nombre d'observations de l'échantillon mesure la déviation moyenne de l'échantillon par rapport aux autres échantillons. Les valeurs ainsi obtenues sont représentées dans le temps. La statistique T calculée comme le rapport entre la statistique PRESS et l'erreur standard estimée est supposée suivre une distribution de Chi-carré à n degrés de liberté, où n est la taille de l'échantillon. La statistique T est représentée sur un graphique en entonnoir avec les bornes 95% et 99% attendus. Les études-pays en dehors de ces bornes sont investigués comme valeurs aberrantes. Il se pose la question de l'adéquation de cette distribution de Chi-carré à n degrés de liberté et même de l'adéquation de la statistique utilisée pour la détection des valeurs aberrantes. Le présent travail vise à améliorer la méthode de détection des études-pays aberrantes en proposant des techniques plus adaptées au contexte.

En somme, en répondant à la question de savoir comment détecter des valeurs aberrantes dans une méta-analyse, la présente étude permettra d'améliorer les résultats de méta-analyse, leur interprétation et leur utilisation notamment dans le cadre de la surveillance de la qualité des données cliniques dans le domaine pharmaceutique.

1.2 Objectifs de recherche

Le présent travail vise à étudier des méthodes pour détecter les études ayant des valeurs aberrantes dans une méta-analyse à l'aide du logiciel SAS. Il s'agit de manière spécifique d'évaluer dans un premier temps la performance de la méthode des RES pour détecter les valeurs aberrantes d'une méta-analyse. Ensuite, la performance de la méthode de VSOM pour détecter les valeurs aberrantes d'une méta-analyse est discutée. Enfin les deux méthodes sont comparées du point de vue de leur mise en œuvre et de leur performance à détecter des valeurs aberrantes dans une méta-analyse.

Chapitre 2

Concepts théoriques et état de l'art

2.1 Méta-analyses (MA)

Une méta-analyse (MA) consiste à utiliser des techniques statistiques pour combiner et synthétiser les résultats de plusieurs études scientifiques distinctes afin d'obtenir des estimations d'effets plus précises que celles obtenues en analysant les différentes études qui la composent de manière isolée [19]. La combinaison des échantillons des différentes études aboutit à un échantillon de plus grande taille augmentant ainsi la puissance des tests statistiques réalisés. Aussi la construction d'une base de données élargie permet-elle d'approfondir les analyses d'exploration dans le but de détecter des facteurs susceptibles de modifier l'effet principal étudié. De plus, conduire une MA permet de confronter les "résultats apparemment discordants entre études (...) et/ou d'expliquer la variabilité des résultats". En effet, les résultats d'une MA peuvent permettre de formuler des hypothèses pour des essais ultérieurs dans le but d'améliorer les conclusions de ces derniers [13]. On distingue d'une part les méta-analyses basées sur des données agrégées et d'autre part des méta-analyses basées sur des données individuelles.

2.1.1 Méta-analyses basées sur des données agrégées (MA-DA)

Les méta-analyses basées sur des données agrégées (MA-DA) exploitent les statistiques (moyennes, fréquences, risques relatifs, odds ratio, ...) rapportées par les résultats de toutes les études accessibles (publiées ou non) concernant le sujet en question. Elles sont les plus courantes [17] parce qu'elles sont plus aisées et moins coûteuses à mettre en œuvre. Cependant, l'exactitude des informations collectées lors de la réalisation d'une

MA-DA n'est pas toujours garantie [16]. Selon Stewart et al., les résultats d'une MA-DA à eux seuls peuvent prêter à confusion. Il serait intéressant dans la mesure du possible, de conduire une méta-analyse basée sur les données individuelles afin de réduire les biais et d'obtenir des réponses plus fiables aux questions qui n'ont pas été résolues de manière satisfaisante par les études individuelles [26].

2.1.2 Méta-analyses basées sur des données individuelles (MA-DI)

La méta-analyse basée sur des données individuelles (MA-DI) exploite directement les bases de données contenant les informations sur l'ensemble des individus recrutés, obtenues auprès de chaque investigateur de toutes les études considérées. Elle représente la méthode de référence des MA puisqu'elle permet de vérifier et d'analyser de manière approfondie l'ensemble des données qui la constituent [13]. Malgré la difficulté inhérente à l'obtention de toutes les bases de données des études pour réaliser une MA-DI, elle garantit des analyses statistiques finales plus souples et plus robustes [17]. Elle assure une plus grande fiabilité et permet d'obtenir des résultats plus détaillées même si sa mise en œuvre nécessite plus de temps et est plus onéreuse qu'une MA-DA [25]. D'après Pignon et al. [20], la MA-DI devrait être systématiquement réalisée pour des surveillances d'essais à long terme. La présente étude s'inscrivant dans ce cadre, elle se base sur les données individuelles des patients. De plus, tous les essais étant réalisés par la même organisation conduisant l'analyse de surveillance, ces données sont disponibles et accessibles.

En analysant les données d'une MA-DI, il convient de tenir compte des potentielles variations des effets entre essais. Pour cela, il est d'usage d'inclure à la modélisation des effets aléatoires qui permettent d'examiner et de considérer l'hétérogénéité entre les estimations d'effet d'une étude à l'autre [8]. Michiels et al. ont démontré que les modèles à effets aléatoires permettent d'explorer et de mieux comprendre l'hétérogénéité lors d'une MA-DI dans le cadre d'une analyse de survie [18]. Ils suggèrent que ce type de modélisation pourrait être appliqué dans divers autres contextes.

2.2 Modèles linéaires

Cette partie se base principalement sur les diapositives du cours de "Modèles Linéaires" donné par Christian Hafner à l'Institut de Statistique Biostatistique et Sciences Actua-

rielles (ISBA) de l'Université Catholique de Louvain (UCL) pour le compte de l'année académique 2018/19 [7] ainsi que ceux du cours de "Modèles Linéaires Avancés" donné par Catherine Legrand pour le compte de l'année académique 2019/20 [15].

2.2.1 Modèle linéaire classique (ML)

Le modèle linéaire permet d'une part d'expliquer et de prédire une variable réponse continue par une ou plusieurs variables explicatives continues ou catégorielles et d'autres part d'estimer l'effet marginal d'une variable ajustée sur les autres variables contenues dans le modèle.

2.2.1.1 Équation du modèle

L'équation du ML classique s'écrit :

$$\mathbf{Y} = \mathbf{X}\alpha + \epsilon, \quad (2.1)$$

avec :

- $Y \in R^n$, le vecteur des valeurs observées de la variable dépendante continue où n est la taille de l'échantillon et $Y | X$ est supposé indépendant et identiquement distribué (iid) Normal, d'une observation à l'autre.
- $X \in R_{p+1}^n$, la matrice des effets fixes de rang maximal et $\alpha \in R^{p+1}$, le vecteur de coefficients associés où p est le nombre de modérateurs inclus dans le modèle ;
- $\epsilon \sim N(0, R)$, le vecteur de dimension n des erreurs où $R = \sigma^2 I_n$.

2.2.1.2 Estimation des paramètres

Les coefficients α et les erreurs ϵ n'étant pas observés contrairement à X et Y , la modélisation vise à estimer ces paramètres pour établir le lien entre les valeurs observées. L'estimation des paramètres dans le cas du modèle linéaire classique se fait soit par la méthode des Moindres Carrés Ordinaires (MCO) ou soit par la méthode de Maximum de Vraisemblance (MV).

Cette dernière méthode consiste en une estimation sur la base d'un modèle paramétrique entièrement spécifiée. Elle peut être appliquée dans différents contextes et les estimateurs obtenus possèdent d'excellentes propriétés asymptotiques. L'hypothèse distributionnelle sur ϵ et par conséquent sur $Y | X \sim N(X\alpha, \sigma^2 I_n)$ permet d'obtenir la

vraisemblance du modèle qui après maximisation donne l' "Estimateur du Maximum de Vraisemblance (EMV)".

Soit $\hat{\alpha}_{EMV}$ et $\hat{\sigma}_{EMV}$, les EMV respectifs de α et σ .

$\hat{\alpha}_{EMV} = (X'X)^{-1}X'Y$ est le "Best Linear Unbiased Estimateur (BLUE)" de α et $\hat{\sigma}_{EMV}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - X_i \hat{\alpha})^2$, l'EMV de σ^2 est asymptotiquement non biaisé.

2.2.2 Modèle linéaire mixte (MLM)

2.2.2.1 Spécification du modèle

Le MLM est une généralisation du ML résultant d'une dépendance entre les observations de l'échantillon. Cette dépendance peut être issue de la répétition des mesures de la variable réponse chez les mêmes individus à des moments différents ou du caractère hiérarchique (multiniveaux) des données. La corrélation entre les observations du point de vue de la variable réponse se traduit dans la modélisation de deux manières : soit par des effets/coefficients aléatoires ou soit par une structure de corrélation directement spécifiée dans la matrice des résidus R . Il est également possible au besoin de spécifier à la fois dans un même modèle aussi bien des effets/coefficients aléatoires qu'une structure de corrélation dans la matrice des erreurs. Cette pratique est cependant moins courante car étant assez complexe.

Dans le premier cas, la structure de corrélation est imposée entre les termes d'erreurs relatifs aux observations présentant les mêmes caractéristiques. On parle alors de MLM à structure de covariance. "Une variable bloc (...) est choisie et dans le même bloc les observations peuvent être corrélées. La matrice R prend alors une forme bloc-diagonale". Selon la structure de dépendance spécifiée pour les observations dans chaque bloc, on distingue les types de modèle :

- autorégressif d'ordre 1 (AR1) ou à corrélation décroissante en fonction de la séparation des observations dans le temps ,
- à composante symétrique ou corrélation constante et ceux
- "non spécifié" ou général permettant une valeur de covariance différente pour chaque paire.

Dans le deuxième cas, on ajoute au modèle un ou plusieurs effets/coefficients aléatoires représentant les caractéristiques communes aux observations qui sont corrélées. L'effet aléatoire implique une dépendance entre les observations provenant d'un même niveau de

cet effet. On parle de coefficient aléatoire lorsqu'en plus d'un effet aléatoire, on ajoute au modèle de régression un terme d'interaction aléatoire entre l'effet aléatoire et un effet fixe. "Chaque effet aléatoire introduit dans le modèle une composante de variance" qui quantifie la variation aléatoire relative à cet effet et chaque coefficient aléatoire induit des termes de covariance du fait de la dépendance entre des termes d'intercept et des pentes du modèle. "Les effets/coefficients aléatoires sont supposés être des réalisations d'une distribution aléatoire", typiquement une distribution normale multivariée comme les résidus du modèle. Il est possible de spécifier différentes structures de covariance pour la matrice de variance-covariance associée aux effets aléatoires (composante de variance, composante symétrique ou non-structurée).

La présente étude porte sur un modèle mixte à un effet aléatoire avec une composante de variance.

2.2.2.2 Équation du modèle

Comme dans le cas du ML, l'équation du modèle s'écrit :

$$\mathbf{Y} = \mathbf{X}\alpha + \mathbf{Z}\beta + \epsilon, \quad (2.2)$$

avec :

- X , Y et ϵ définis comme précédemment pour l'équation 2.1 à la seule différence que les $Y | X$ ne sont plus indépendants bien que suivant une distribution Normal d'une observation à l'autre ;
- $Z \in R_q^n$, la matrice des effets aléatoires et $\beta \sim N(0, G)$, le vecteur de dimension q des paramètres associés où $G = \sigma_c^2 I_q$ représente la matrice de variance-covariance des effets aléatoires.

La matrice de variance-covariance V de Y s'écrit :

$V = \text{var}(Y) = ZGZ' + R = Z\sigma_c^2 Z' + \sigma^2 I_n$, la covariance entre β et ϵ étant nulle puisque les effets aléatoires et les résidus sont indépendants.

En posant $\sigma_c^2 = \sigma^2 \gamma$, on obtient :

$$V = \sigma^2(\gamma Z Z' + I_n) = \sigma^2 M.$$

2.2.2.3 Estimation des paramètres du modèle

Dans le cas du MLM, on estime non seulement les paramètres des effets fixes comme dans le cas du ML classique mais aussi les paramètres des distributions associées aux effets aléatoires notamment les paramètres de variances.

La fonction de vraisemblance du modèle basée sur la distribution de $Y \sim N(X\alpha, V)$ donne après maximisation dans le cas du MLM :

$$\hat{\alpha}_{MLM} = (X'V^{-1}X)^{-1}X'V^{-1}Y \text{ et } var(\hat{\alpha}) = (X'V^{-1}X)^{-1}.$$

La matrice de variance-covariance V de Y étant généralement inconnu, l'expression ci-dessus entraîne un biais de sous-estimation de la variance de l'estimateur. Pour y remédier, on a plutôt recours à la méthode de maximum de vraisemblance restreinte qui produit l'Estimateur du Maximum de Vraisemblance Restreint (REMV), un estimateur non biaisé pour les paramètres de variance. La fonction de vraisemblance restreinte est basée sur les résidus marginaux : $y - X\hat{\alpha}$ qui peuvent être considérés comme les termes d'erreurs incluant toutes les sources de variabilité aléatoire par opposition aux résidus conditionnels $\tilde{e} = y - X\hat{\alpha} - Z\hat{\beta}$.

En maximisant la fonction de vraisemblance restreinte et en considérant la reparamétrisation de V à la fin de la section 2.2.2.2 on trouve :

$$\hat{\alpha} = (X'M^{-1}X)^{-1}X'M^{-1}y, \text{ le BLUE de } \alpha \text{ et}$$

$$y'Py = (y - X\hat{\alpha})'M^{-1}(y - X\hat{\alpha}), \text{ la somme du carré des résidus avec } P = M^{-1} - M^{-1}X(X'M^{-1}X)^{-1}X'M^{-1}.$$

Les dérivées de la fonction de log-vraisemblance restreinte dans le processus de maximisation n'étant pas linéaire par rapport aux paramètres de variance, il n'est pas possible de spécifier une solution pour ces derniers par une équation unique comme dans le cas du ML classique. On a alors recours à des processus itératifs d'optimisation numérique tel que l'algorithme de Newton-Raphson pour obtenir une estimation des paramètres de variance.[2]

2.3 Données aberrantes

Les valeurs aberrantes sont les valeurs extrêmes ou atypiques contenues dans la base de données. Elles peuvent être valides ou invalides. Les valeurs aberrantes invalides doivent être corrigées dans la mesure du possible. Quant aux valeurs aberrantes valides, il convient,

de comprendre la source de leur déviation de la tendance globale des données et de mesurer leur impact sur les résultats des analyses.

Avant toute modélisation, il est possible de détecter les valeurs aberrantes d'une base de données par diverses méthodes : la méthode des z-score ou celle du box-plot par exemple. Ces méthodes étant influencées par les valeurs aberrantes et étant plus adaptées à des données qui suivent une distribution symétrique, un groupe de chercheurs de la "Katholieke Universiteit Leuven" (KUL) développe les statistiques robustes de détection et de prise en compte des valeurs aberrantes[22]. Ces statistiques ont pour but de produire des résultats fiables lorsque les données contiennent des valeurs aberrantes. Elles constituent des outils automatiques de détection de valeurs aberrantes. Des estimateurs robustes ont été développés pour différentes techniques statistiques dont la régression.

2.3.1 Données aberrantes dans un ML

Outre la régression robuste, plusieurs autres méthodes sont utilisées pour détecter les valeurs aberrantes dans un modèle de régression linéaire. Kutner et al. [14] présentent dans leur livre "Applied Linear Statistical Models", différentes méthodes de détection des valeurs aberrantes après avoir ajusté un modèle de régression linéaire.

Dans le cas d'une régression ne comportant qu'une ou deux variables explicatives, il est relativement aisé de visualiser les observations aberrantes par rapport à chacun des axes en examinant le graphique des résidus. La tâche devient plus ardue dans le cas de régression impliquant plus de deux dimensions. Dans ce cas, on distingue les observations aberrantes vis-à-vis de la matrice des variables indépendantes X et celles aberrantes vis-à-vis du vecteur réponse Y .

Pour identifier les observations aberrantes par rapport à X , on se sert de la matrice de projection : $H = X(X'X)^{-1}X'$ avec le levier $h_{ii} = (1, X_{i1}, \dots, X_{ip-1})(X'X)^{-1}(1, X_{i1}, \dots, X_{ip-1})'$. X_i est aberrantes si $h_{ii} > \frac{2p}{n}$. Ce cas n'entre pas dans le cadre du présent travail.

Pour identifier les observations aberrantes par rapport à Y , on étudie les résidus du modèle : $e_i = Y_i - \hat{Y}_i$. L'observation i est aberrante si les résidus standardisés encore appelés Résidus Internes Studentisés (RIS) t_i associés sont grands avec :

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}, \quad (2.3)$$

et $\hat{\sigma}$ un estimateur approprié de σ notamment l'EMV ou l'REMV.

Cependant étant donné que la valeur investiguée, Y_i est incluse dans le modèle de régression, elle pourrait influencer les résultats d'estimation. Pour éviter cela, on préfère les résidus standardisés effacés ou résidus externes studentisés (RES) pour identifier les valeurs aberrantes. Les RES sont basés sur l'estimation des coefficients de la régression ajustée en supprimant l'observation i des données.

En posant $s_i = Y_i - \hat{Y}_{i(-i)}$, le RES pour l'observation i est donné par :

$$r_i = \frac{s_i}{std(s_i)}, \text{ avec } std(s_i) \text{ l'estimation de l'écart-type de } s_i.$$

En supposant $r_i \sim t_{n-p}$, Y_i serait donc aberrante si $|r_i| > t_{n-p;0,975}$ pour un seuil de significativité de 5%. [7]

2.3.2 Données aberrantes dans un MLM

Les méthodes de détection de valeurs aberrantes des ML ne peuvent pas s'appliquer telles quelles lors de l'ajustement de MLM car dans un premier temps l'hypothèse d'indépendance des observations sur laquelle se basent les ML est assouplie. Ensuite dans le cas des MLM on est plutôt intéressé à détecter une étude ou un groupe d'observations aberrantes et pas une seule observations atypique. Une valeur aberrante dans une méta-analyse serait une étude dont l'estimation de l'effet (corrélation ou autre valeur statistique) ne semblerait pas cohérente avec celle des autres études soit à cause d'erreurs de collecte, de saisie de données, de calcul ou soit à cause d'une caractéristique inhabituelle au niveau de la conception de l'étude ou des participants inclus. L'effet de ces valeurs aberrantes se traduit généralement par une augmentation de la variabilité résiduelle et/ou un éventuel changement de la mesure de l'effet [11].

Il est possible de détecter une étude aberrante dans une méta-analyse à l'aide de représentation graphique telle que le graphique en forêt. Le graphique représente les intervalles de confiance des mesures d'effet de chaque étude avec une ligne de référence montrant la valeur de l'effet estimé par le modèle. Les études dont l'intervalle de confiance apparaît en dehors de la tendance générale sont identifiées comme aberrantes [5]. Cependant les méthodes de détection graphique ne tiennent pas nécessairement compte de la différence des tailles d'échantillon entre les études et des modèles incluant des modérateurs à l'analyse des effets[30].

Huffcutt et Arthur ont développé en 1995, la méthode de "Sample-Adjusted Meta-analytic Deviancy" (SAMD) [10] pour la détection de valeurs aberrantes dans une méta-analyse. Cette méthode compare la mesure de l'effet dans chaque étude à la valeur

moyenne de l'effet calculé sur l'ensemble des données en omettant l'étude investiguée, ajustée à la taille de cette dernière. La distribution de l'ensemble des valeurs de SAMD obtenue pour chaque étude est approximé à une distribution de Student. Les études présentant des valeurs extrêmes de SAMD sont désignées comme potentiellement aberrantes [11]. Il s'agit d'une technique essentiellement univariée qui ne correspond pas au contexte de la présente étude qui prend en compte des covariables.

Koller propose une méthode d'estimation robuste de détection des valeurs aberrantes dans un MLM basée sur les modèles de contamination à travers un package R "robustlmm" [12]. Cette méthode s'apparente à celle proposée par Desmet al. [3] : un modèle de "location shift" dans le cadre de la surveillance d'essais cliniques multicentriques. Elle consiste à comparer les paramètres de tendance centrale de variables continues de chaque centre à ceux de l'ensemble des données contenues dans la base. Une p-valeur appréciant la déviance de la moyenne de chaque centre par rapport à la moyenne globale est calculée. En dérivant la distribution sous l'hypothèse nulle de la moyenne de la variable investiguée sur l'ensemble des centres à partir d'un modèle mixte à un effet aléatoire centre, les centres présentant une p-valeur trop petite sont identifiés aberrantes. Une fois de plus, il s'agit essentiellement d'approches univariées qui ne prennent pas en considération l'implication de covariables dans la détection des valeurs atypiques.

Dans le cadre de ce travail, deux autres méthodes jugées plus adaptées au contexte sont étudiées : celle de RES de Viechtbauer et Cheung [30] et celle de VSOM de Gumedze et al.[6]. La première est une extension de l'examen des résidus externes studentisés au cas du MLM et la deuxième consiste à vérifier si l'étude investiguée présente une inflation de la variance résiduelle qui lui est associée. Les détails du fonctionnement de ces deux méthodes sont présentés dans le chapitre suivant.

Chapitre 3

Méthodologie

3.1 Données

La méthodologie développée s'appliquera à terme à un ensemble de données qui regroupe diverses études mesurant les réponses immunitaires à différentes formules d'un vaccin administré dans plusieurs pays au cours du temps. En tenant compte des différentes combinaisons de pays, d'études et de covariables, des clusters : étude-pays sont formés. Il s'agira alors de déterminer si une nouvelle étude réalisée apparaît atypique par rapport aux autres études. Les premières études datent de Juillet 1991 et les dernières de Septembre 2017.

L'ensemble des données regroupe actuellement :

- les caractéristiques des études incluses dans la méta-analyse par groupe d'étude. On dénombre 642 groupes d'études provenant de 71 pays ou combinaisons de pays différents ;
- les caractéristiques des individus de chaque étude soit 133.009 sujets au total pour le compte de toutes les études et
- les résultats des prélèvements sanguins pour chaque sujet à différente visite au nombre de 1.593.973 observations pour toute la méta-analyse concernant les différents antigènes d'intérêt contenus dans le vaccin.

Un modèle linéaire mixte est ajusté aux données pour apprécier l'effet de différents facteurs sur la réponse immunitaire des sujets vaccinés. La réponse immunitaire est mesurée par la concentration moyenne géométrique d'anticorps dans le sang (GMT) par individu supposée suivre une distribution log-normale. La variable dépendante du MLM, Y est donc le logarithme à base 10 de cette concentration pour respecter l'hypothèse de

normalité des $Y \mid X$ sur laquelle repose ce type de modèle. Le modèle inclut un effet aléatoire pays. Les modérateurs considérés dans le modèle sont :

- le temps : variable ordinale considérée comme continue attribuant un rang à chaque étude compte tenu de la date de la première visite du premier sujet ;
- la périodicité des vaccinations précisant le nombre de doses administrés (2 ou 3 doses selon le vaccin) : variable catégorielle ;
- la composition du vaccin : variable catégorielle ;
- la réalisation ou non de vaccination antérieure : variable catégorielle binaire indiquant si le sujet avait reçu une vaccination antérieure contre le même antigène.

L'objectif de l'analyse dans le cadre de ce travail est de déterminer si la réponse immunitaire au vaccin c'est-à-dire la concentration d'anticorps dans le sang (GMT) d'une nouvelle combinaison d'étude-pays ajoutée à la méta-analyse en tenant compte des modérateurs contenus dans le modèle est atypique par rapport aux autres études de la base.

3.2 Méthodes de détection de valeurs aberrantes

Par souci de simplicité, nous considérons dans un premier temps le cas de la régression linéaire classique.

3.2.1 Résidus externes studentisés (RES)

3.2.1.1 Cas de la régression linéaire classique

Dans le cas de la régression linéaire classique, les résidus externes studentisés (r_j) pour l'étude j se calculent par la différence entre la valeur moyenne observée et la valeur prédite obtenue en utilisant toutes les données, sauf les valeurs évaluées comme potentiellement aberrantes.

$$r_j = \frac{y_j - \hat{\mu}_{j(-j)}}{\sqrt{\text{Var}(y_j - \hat{\mu}_{j(-j)})}},$$

avec :

- y_j l'effet moyen observé pour l'étude j et
- $\hat{\mu}_{j(-j)}$ l'effet moyen prédit de l'étude j basé sur le modèle excluant cette étude.

3.2.1.2 Cas du modèle mixte

Selon Viechtbauer et Cheung [30], les RES dans le cas d'un modèle mixte se calculent comme suit :

$$\mathbf{r}_j = \frac{\mathbf{y}_j - \hat{\mu}_{j(-j)}}{\sqrt{\sigma_j^2 + \hat{\sigma}_{c(-j)}^2 + \mathbf{var}(\mu_{j(-j)}^{\hat{}})}}, \quad (3.1)$$

avec : $\mathbf{var}(\mu_j^{\hat{}}) = l_j(\sigma_j^2 + \hat{\sigma}_c^2)$ où l_j est l'élément diagonal à la position j de la matrice de projection dans le cas du MLM : $L = X(X'\tilde{W}X)^{-1}X'\tilde{W}$ et $\tilde{W} = \text{diag}(1/(\sigma^2 + \hat{\sigma}_c^2))$ de dimension $k \times k$ ($k =$ au nombre d'études). l_j est le levier de l'étude j .

Règle de décision : Si $|r_j| > 1,96$, l'étude j serait atypique.

3.2.2 Variance Shift Outlier Model (VSOM)

3.2.2.1 Cas de la régression linéaire classique

Equation du modèle

Gumedze et al. [6] partent du modèle 2.1 avec $\hat{\alpha}_{EMV}$ le BLUE de α tel que défini précédemment. Ils considèrent plutôt $\hat{\sigma}^2 = (y - X\hat{\alpha})'(y - X\hat{\alpha})/(n - p)$, l'estimateur non biaisé, REMV de σ^2 . Il diffère de l'estimateur $\hat{\sigma}_{EMV}^2$ défini à la section 2.2.1.2 par un facteur d'inflation $n/(n-p)$ [28]. Le VSOM pour détecter si l'observation i est aberrante consiste à ajouter un effet aléatoire qui lui est associé au modèle 2.1.

L'équation du VSOM pour l'observation i s'écrit donc :

$$\mathbf{y} = \mathbf{X}\alpha + \delta_i \mathbf{d}_i + \epsilon, \quad (3.2)$$

avec

- d_i un vecteur de dimension n comprenant 1 à la position i et 0 partout ailleurs ;
- $\delta_i \sim N(0, \tau_i)$, l'effet aléatoire associé pour $\tau_i \geq 0$.

La matrice de variance-covariance de y s'écrit alors :

$$\mathbf{var}(y) = \tau_i d_i d_i' + \sigma^2 I_n.$$

En posant $w_i = \tau_i/\sigma^2$, on obtient :

$\mathbf{var}(y) = \sigma^2(w_i d_i d_i' + I_n)$. Cette paramétrisation permet d'avoir une idée de la proportion dans laquelle la variance de l'observation i (qui serait alors $\sigma^2(w_i + 1)$) est accrue par rapport à la variance résiduelle.

Exemple : Supposons un échantillon de $n=4$ individus. On a :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{pmatrix}, \alpha = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}, d_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Pour vérifier si l'observation 4 est aberrante, on ajoute au modèle le vecteur d_4 comme effet aléatoire avec $\delta_4 \sim N(0, \tau_4)$ le coefficient associé. Le VSOM pour l'observation 4 s'écrit donc :

$$y_4 = \alpha_0 + \alpha_1 x_1 + \delta_4 + \epsilon_4,$$

et $\text{var}(y_4) = \sigma^2(w_4 + 1)$ en appliquant la paramétrisation précédente.

Le vecteur α est à estimer ainsi que σ^2 la variance des résidus et w_4 le facteur d'inflation associé à la variance de l'effet aléatoire d_4 . Vérifier si i_4 est atypique revient donc à vérifier que $w_4 > 0$, ce qui signifierait que la variance de l'effet aléatoire associé à cette observation est augmentée par rapport à la variance des erreurs du modèle.

Estimation et test d'hypothèse sur w_i

En utilisant la fonction de log-vraisemblance restreinte du VSOM pour l'observation i , l'estimation $\hat{w}_i$ de w_i a été dérivée par Thompson [28] :

$\hat{w}_i = \frac{(n-p)(t_i^2 - 1)}{(n-p-1)(1 - v_{ii})}$ si $t_i^2 > 1$ et 0 sinon, où $v_{ii} = d_i'X(X'X)^{-1}X'd_i$ est l'élément diagonal i de la matrice de projection sous ce modèle. Etant donné que le modèle ne permet que l'inflation de la variance, w_i serait supérieur à 1 si et seulement si $t_i^2 > 1$.

Sous cette forme, $\hat{w}_i$ s'avère être une fonction du carré du RIS de l'observation i sous le modèle de régression linéaire classique t_i (équation 2.3 défini à la section 2.3.1) noté t_i^2 . Puisque t_i suit une distribution de Student à $(n-p)$ degrés de liberté, $t_i^2/(n-p)$ suit une distribution Beta de paramètres $(1/2, (n-p-1)/2)$.

Après calcul de la fonction de log-vraisemblance restreinte du VSOM pour chaque observation, les observations ayant les plus grandes valeurs de log-vraisemblance restreintes sont investiguées comme valeurs aberrantes. La statistique de rapport de vraisemblance servant à comparer le VSOM pour l'observation i sous l'hypothèse nulle est elle aussi une fonction simple de t_i^2 :

$$LRT_i = (n-p-1) \log \frac{(n-p-1)}{(n-p-t_i^2)} - \log t_i^2, \quad (3.3)$$

si $t_i^2 > 1$ et 0 sinon.

Tester qu'une observation est atypique se résumerait donc à tester $H_0 : w_i = 0$ contre $H_a : w_i > 0$.

Dans le cas où, seulement quelques observations doivent être investiguées comme valeurs aberrantes ce qui correspond au contexte de la présente étude, on pourrait utiliser la distribution connue de $t_i^2/(n-p) \sim \text{Beta}(1/2, (n-p-1)/2)$ pour effectuer le test. Dans le cas contraire, où toutes les observations de l'échantillon doivent être investiguées, on devra réaliser un test multiple. Or t_i^2 n'est pas indépendant d'une observation à l'autre. On se servira donc de la méthode du bootstrap paramétrique pour obtenir une approximation de la distribution jointe t_i^2 de toutes les observations sous l'hypothèse nulle. Ce dernier cas n'est pas abordé dans le présent document.

3.2.2.2 Cas du modèle mixte

D'après Gumedze et al.[6], le VSOM dans le cas d'un modèle mixte à effet aléatoire avec une composante de variance pour l'observation i s'écrit :

$$\mathbf{y} = \mathbf{X}\alpha + \mathbf{Z}\beta + \delta_i \mathbf{d}_i + \epsilon, \quad (3.4)$$

Par analogie au cas précédent, on déduit :

$$\text{var}(y) = \sigma^2(ZGZ' + w_i d_i d_i' + I)$$

Par conséquent, la variance de l'observation i serait accrue de $\sigma^2 w_i$. La covariance avec les autres observations ne change cependant pas.

t_i^2 devient :

$$\mathbf{t}_i^2 = \frac{\tilde{\mathbf{e}}_i^2}{\hat{\sigma}^2 \mathbf{c}_{ii}}$$

avec :

- $\tilde{e} = d_i' P y$, le "Best Linear Unbiased Predictor" de l'erreur e_i ou résidus conditionnels définis de manière générale à la section 2.2.2.3 , et
- $c_{ii} = d_i' P d_i$.

Dans ce cas, un estimateur de w_i est donné par :

$$\hat{w}_i = \frac{(n-p)(t_i^2 - 1)}{c_{ii}(n-p-t_i^2)} \text{ si } t_i^2 > 1 \text{ et } 0 \text{ sinon.}$$

La distribution exact de t_i^2 étant inconnue dans ce cas, Gumedze et al. suggèrent d'utiliser le test de rapport de vraisemblance ou celui de Score pour évaluer l'hypothèse

de nullité de w_i . En appliquant une re-estimation partielle des paramètres de variance, l'expression analytique de la statistique du test de rapport de vraisemblance LRT est la même que celle obtenue dans le cas du ML (équation 3.3). Une mixture 68 :32 de Chi-carré à 0 et 1 degré de liberté pourrait être considérée comme distribution asymptotique pour tester l'hypothèse nulle $H_0 : w_i = 0$ contre l'hypothèse alternative $H_a : w_i > 0$.

La présente étude ayant pour but de détecter une étude aberrante et non une observation aberrante, le VSOM est construit dans ce cadre avec le vecteur d_j constitué de la valeur 1 pour toutes les observations appartenant à l'étude j investiguée et 0 partout ailleurs comme effet aléatoire.

3.2.3 Comparaison théorique des deux méthodes

Pour la méthode de RES, le test statistique servant à l'identification des valeurs aberrantes de Viechtbauer et Cheung porte directement sur les résidus externes studentisés. Gumedze et al. quant à eux suggèrent pour l'application de la méthode de VSOM de commencer par l'investigation du carré des RIS t_i^2 . Ensuite un VSOM devrait être ajusté pour les observations ou groupes d'observations présentant les plus grandes valeurs de t_i^2 afin de confirmer que le facteur d'inflation de la variance w_i est effectivement supérieur à 0. Notons que l'expression de w_i ainsi que celle de la statistique de test pour cette méthode sont des fonctions de t_i^2 . Les deux méthodes reviendraient donc essentiellement à investiguer les résidus du modèle par rapport aux études apparaissant potentiellement aberrantes.

3.3 Approche de simulation

Les simulations sont réalisées à l'aide du logiciel SAS University Edition 2.8.1 9.4 M6. Elle permettent de mettre en œuvre les deux méthodes initialement réalisées sur R par les auteurs à l'aide du logiciel SAS plutôt. A travers les simulations, la difficulté de mise en œuvre ainsi que la performance des deux méthodes sont évaluées afin de s'assurer qu'elles fonctionnent. Elles ont pour but d'une part de vérifier que pour l'application de la méthode de VSOM, le modèle est bien identifiable et d'autre part d'identifier à partir de quelle amplitude d'aberrance chacune des deux méthodes détecte les valeurs atypiques.

Pour cela, dans un premier temps les données sont simulées au départ d'un modèle de régression linéaire simple avec deux (2) observations aberrantes. Ensuite, les données sont

simulées au départ d'un modèle linéaire mixte avec un ensemble d'observations (cluster) aberrant. Les deux méthodes sont appliquées dans chacun de ces cas afin d'apprécier la capacité de détection de chacune d'elle après l'ajustement d'un ML et d'un MLM respectivement. L'approche de simulation dans les deux cas est présentée ci-dessous.

3.3.1 Cas de la régression linéaire simple avec des observations aberrantes

Gumedze et al. [6] ont présenté dans l'article publié, un exemple de simulation illustrant l'utilisation du VSOM dans le cas d'un modèle de régression linéaire. Cette même simulation est textuellement répliquée afin de comparer les résultats obtenus pour garantir l'application correcte de la méthode. Les données sont générées à partir du modèle :

$$y_i = \alpha X_i + b + \epsilon_i, \quad (3.5)$$

avec $i = 1, \dots, 30$; $\alpha = 0,25$; $X \sim U[0; 10]$; $b = 26,25$ et $\epsilon_i \sim N(0, 1)$. 1000 échantillons provenant de ce modèle ont été générés.

Deux observations aberrantes ont été introduites dans chaque échantillon simulé. Il s'agit des observations 12 et 30 : $\epsilon_{12} = \epsilon_{30} = 5$ et $X_{30} = 20$. Tout au long de la simulation, les observations 12 et 30 atypiques sont comparées à l'observation 1 qui ne l'est pas.

3.3.1.1 Application de la méthode RES dans le cas du ML

Les RES sont obtenus par la statistique "RSTUDENT" de la procédure "reg" de SAS qui a permis d'ajuster le modèle de régression linéaire simple. Les résultats de la simulation présentent la moyenne $\pm$ l'écart-type des RES ainsi que l'Intervalle de Confiance (IC) à 95% de la moyenne sur l'ensemble des 1000 échantillons, l'histogramme de la distribution des RES pour chacune des observations investiguées obtenu par la procédure "univariate" de SAS ainsi que le pourcentage d'échantillons identifiant chaque observation aberrante.

3.3.1.2 Application de la méthode VSOM dans le cas du ML

Pour détecter si les observations 1, 12 et 30 sont aberrantes conformément à la méthode proposée par Gumedze et al., les RIS obtenus par la statistique "STUDENT" de la procédure "reg" de SAS sont d'abord examinés. Par la suite, le modèle 3.5 auquel on ajoute trois (3) effets aléatoires d_1 , d_{12} et d_{30} avec une matrice de variance de type

composante de variance est ajusté par la procédure mixed. Les vecteurs d_1 , d_{12} et d_{30} de dimension n égale à la taille de l'échantillon contiennent chacun la valeur 1 respectivement à la position 1, 12 puis 30 et 0 partout ailleurs. Les RIS, les variances des effets aléatoires ainsi que les facteurs d'inflations w_1 , w_{12} et w_{30} associés à chacun d'eux sont présentés dans les résultats.

En considérant $n=30$ et $p=2$, la statistique du test de rapport de vraisemblance LRT_i (équation 3.3) est calculée pour chacune des observations 1, 12 et 30 à partir du carré des RIS t_i^2 .

On sait que $t_i^2/(n-p) \sim \text{Beta}(1/2, (n-p-1)/2)$.

Ce qui revient dans le cadre de la présente simulation à $t_i^2/28 \sim \text{Beta}(1/2; 13, 5)$.

Le quantile 95% de cette distribution $q_{95} = 0,13$.

Le seuil critique pour t_i^2 est donc $q_{95} * (n-p) = 0,13 * 28 = 3,77$.

En appliquant la formule 3.3 au seuil critique de t_i^2 , une observation est détectée aberrante si la statistique LRT_i associée est supérieure à :

$$LRT_i = (n-p-1) \log \frac{(n-p-1)}{(n-p-t_i^2)} - \log t_i^2 = (30-2-1) \log \frac{(30-2-1)}{(30-2-3,77)} - \log 3,77 = 1,60.$$

3.3.1.3 Comparaison de la performance des deux méthodes dans le cas du ML

Pour aller plus loin, les paramètres du modèle sont changés : l'un ou tous les coefficients du modèle, les valeurs ϵ_{12} , ϵ_{30} et X_{30} pour apprécier la réaction des deux méthodes face à ces changements. Les valeurs de l'intercept et de la pente du modèle initial servant de base pour générer les données sont variées l'une après l'autre puis simultanément passant respectivement de 0,25 et 26,25 comme dans l'exemple de Gumedze et al. à 1 et 40. L'amplitude de la déviance des erreurs ϵ_{12} , ϵ_{30} est réduite simultanément et progressivement de 5 à 1. Enfin, la valeur atypique de X_{30} est baissée à 10 puis à 1. Après chacune de ces modifications, les deux méthodes sont à nouveau exécutées et les résultats obtenus comparés.

3.3.2 Cas du modèle linéaire mixte avec un cluster aberrant

Pour détecter les valeurs aberrantes dans le cas du MLM, nous ajustons un modèle mixte comportant un effet fixe traitement à 2 niveaux et un effet aléatoire pays à 10 niveaux. Pour y parvenir, 100 échantillons de taille $n=1000$ sujets sont générés. Les données

de chaque échantillon proviennent de dix (10) pays, recrutant chacun 100 sujets répartis en 2 traitements. Les matrices de design X et Z par échantillon sont extraites à l'aide de la procédure "glimmix". La matrice des effets fixes X comporte 3 colonnes : 1 pour l'intercept et un par niveau de traitement et celle des effets aléatoires Z est composée d'une colonne par niveau de pays soit 10 colonnes. Puisque la procédure "mixed" utilisée pour ajuster le MLM supporte la paramétrisation non-singulière pour les effets catégoriels, les matrices de design X et Z peuvent ne pas être de rang plein.

Le modèle ajusté s'écrit :

$$y_i = b + \alpha X_i + \beta Z_i + \epsilon_i, \quad (3.6)$$

avec :

- $b = 26,25$, $\alpha = 0,25$
- $\beta_i \sim N(0; 0,1)$ et $\epsilon_i \sim N(0; 1)$.

$$\text{— } X(\in R_3^n) = \begin{pmatrix} 1 & 0 & 1 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 1 & 0 \end{pmatrix}, \quad Z(\in R_{10}^n) = \begin{pmatrix} 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

La première colonne de la matrice X est entièrement constituée de la valeur 1 représentant l'intercept du modèle. Les deux dernières colonnes de la matrice X et toute la matrice Z contiennent les valeurs 0 et 1 indiquant respectivement le groupe de traitement et le pays auquel appartient chaque individu de l'échantillon.

Le groupe A du pays 1 auquel est assigné le traitement 1 est atypique. La valeur de y pour ce groupe est majorée de 10 ($y_A = y + 10$). Tout au long de la simulation, ce groupe est comparé au groupe E du pays 3 auquel le même traitement 1 est assigné.

3.3.2.1 Application de la méthode RES dans le cas du MLM

La procédure "mixed" de SAS ne propose le calcul des résidus qu'au niveau des observations et non pour un groupe d'observations. Le calcul des RES n'étant pas directement disponible, ils sont obtenus manuellement. Le processus d'obtention du RES est expliqué ici pour le groupe A. Le même processus est répété pour obtenir le RES du groupe E.

Le modèle mixte 3.6 est ajusté sur les données sans les observations du groupe A. La valeur du RES de l'étude A est calculée en faisant la différence entre la valeur moyenne observée de la variable réponse y_A et la valeur moyenne prédite $\hat{\mu}_A$ divisée par l'écart-type

associé. La valeur moyenne observée y_A est la moyenne des y_i des observations appartenant à ce groupe obtenu du modèle initial de simulation 3.6 au départ des données. La valeur moyenne prédite $\hat{\mu}_A$ est manuellement calculée à partir des coefficients estimés de pente et d'intercept tirés du résultat d'ajustement du même modèle 3.6 sans les observations de l'étude A. La valeur de $\hat{\mu}_A$ pour chaque échantillon résulte de l'addition des valeurs estimées de pente et d'intercept. En ce qui concerne l'écart-type associé, c'est la racine carrée de la somme de la variance des résidus $\hat{\sigma}^2$, celle de l'effet aléatoire pays $\hat{\sigma}_c^2$ et celle de la variance associée à la prédiction de $\hat{\mu}_A$, $\text{var}(\hat{\mu}_A)$. Les deux premiers termes de variance sont directement obtenus de la sortie de la procédure "mixed" de SAS. Quant à la dernière, la variabilité estimée de la prédiction de $\hat{\mu}_A$, elle est déduite de l'addition des erreurs standards associées aux coefficients estimés de l'intercept et de la pente. La distribution des RES des groupes A et E à travers les 100 échantillons est présentée dans le rapport ainsi que le pourcentage d'échantillons identifiant chacun des groupes comme aberrantes. La règle de décision est la même que dans le cas du ML. Une étude est considérée aberrante si la valeur de RES associée est supérieure à 1,96.

Notons que les RES sont calculés à partir des résidus marginaux conformément à l'article de Viechtbauer et Cheung [30], la valeur moyenne des effets aléatoires étant supposée nulle.

3.3.2.2 Application de la méthode VSOM dans le cas du MLM

Pour réaliser la méthode VSOM, un modèle mixte comprenant tous les éléments du modèle 3.6 en plus de deux (2) effets aléatoires : d_A et d_E est ajusté. L'effet aléatoire d_A (respectivement d_E) est un vecteur de dimension n comportant la valeur 1 si l'individu appartient au groupe A (respectivement E) et 0 partout ailleurs. Les RIS des deux groupes (A et E) sont examinés puis comparés ainsi que le paramètre de variance des effets aléatoires et la facteur d'inflation w_A , w_E associé.

Les RIS des individus appartenant aux groupes A et E sont obtenus à partir de la statistique "Student" de l'option "influence" de la procédure "mixed". Cette option calcule les RIS marginaux $(y_i - \hat{\alpha}x_i)$. Les valeurs de RIS par individu obtenues après ajustement du modèle 3.6 tel quel c'est-à-dire sans y rajouter les effets aléatoires d_A et d_E sont élevées au carré. La moyenne des RIS au carré des individus appartenant à l'étude j est considérée comme la valeur t_j^2 de cette étude. Les valeurs de t_j^2 sont alors utilisées pour calculer la statistique du rapport de vraisemblance LRT_j pour les études A et E par la formule 3.3. Sachant que la distribution de la statistique de LRT_j sous l'hypothèse nulle est une

mixture 68 :32 de Chi-carré à 0 et 1 degré de liberté (section 3.2.2.2), le quantile 95% de cette distribution est considéré comme valeur critique. La valeur du quantile est obtenue à l'aide de la fonction "qchibarsq" du package "emdbook" de R. Cette fonction calcule les quantiles de mixture de Chi-carré à df et df-1 degrés de liberté [1]. Elle prend comme arguments : la probabilité d'intérêt 95% dans ce cas-ci, le nombre de degrés de liberté supérieur de la mixture ici df=1 et la fraction de la mixture qui suit une distribution de Chi-carré à df-1 degré de liberté mix=68% dans ce cas. Le seuil critique de la statistique de LRT_j dans ce cadre est 2,01.

3.3.2.3 Comparaison de la performance des deux méthodes dans le cas du MLM

Les coefficients du modèle sont changés comme dans le cas de la régression linéaire classique pour apprécier la réaction des deux méthodes face à ces changements. L'intercept est plutôt fixé à 40 et la pente à 1. Quant à l'amplitude de la déviance des valeurs aberrantes, elle est réduite de 10 à 2 puis à 1. A chaque fois, les RES, les RIS, les valeurs de w_j ainsi que la statistique de test LRT_j sont à nouveau calculés pour les études A et E. La décision de détection des deux méthodes est comparée pour évaluer leur performance en fonction de l'amplitude de la déviance des valeurs aberrantes au niveau des études.

3.4 Application des deux méthodes sur des données réelles

Pour respecter la confidentialité, certaines informations sensibles contenues dans cette partie sont reprises en annexe.

Les deux méthodes sont appliquées sur une partie des données réelles présentées dans la section 3.1 pour un antigène N. Les données d'une nouvelle étude R-90 sont rendues disponibles, formatées et rajoutées à celles de la méta-analyse. L'objectif de l'application est de vérifier si la réponse immunitaire moyenne (GMT) enregistrée pour cette nouvelle étude est aberrante par rapport à celle des études antérieures présentant les mêmes caractéristiques contenues dans la base de données. La méta-analyse regroupe des études administrant le vaccin selon un calendrier vaccinal à 2 ou 3 doses. La nouvelle étude étant une étude à 3 doses, la présente analyse concerne uniquement les études à 3 doses.

Pour l'antigène N en cause, la méta-analyse contient 109 études suivant un calen-

drier vaccinal de 3 doses réalisées dans 35 pays différents. On distingue 179 combinaisons d'étude-pays, covariables différentes. La taille de l'échantillon pour l'ensemble de ces études est de 31.826 sujets-doses. Le tableau 5.1 en annexe présente les statistiques descriptives des modérateurs susceptibles d'expliquer la réponse immunitaire des sujets de l'échantillon à l'antigène N. Les caractéristiques de la nouvelle étude investiguée sont également présentées en annexe ainsi que l'équation générale du modèle mixte ajusté pour expliquer la réponse immunitaire à l'antigène N (équation 5.1) et l'équation spécifique dans le cas de la nouvelle étude (équation 5.2).

3.4.1 Application de la méthode des RES sur les données réelles

Pour calculer le RES de l'étude R-90, il faut la valeur prédite de cette étude obtenue à partir des coefficients du modèle ajusté sur les données excluant les observations de cette étude. Le tableau 5.3 en annexe, présente les coefficients estimés du modèle 5.1 sans la nouvelle étude. La valeur prédite $\hat{\mu}_{R(-R)}$ par ce modèle pour la nouvelle étude est obtenue en remplaçant les coefficients correspondant dans l'équation 5.2. Le carré de la somme des erreurs standards liés à l'approximation de chaque coefficient est considéré comme la variance associée à la prédiction de $\hat{\mu}_{R(-R)}$. La moyenne des valeurs de $\log_{10}(GMT)$ des individus appartenant à l'étude R-90 est considérée comme la valeur moyenne observée y_R de la variable réponse pour cette étude. Avec toutes ces valeurs, le RES de l'étude R-90, RES_R est obtenue par la formule 3.1 et comparé au seuil critique 1,96.

3.4.2 Application de la méthode du VSOM sur les données réelles

Pour l'application de la méthode VSOM, un vecteur d_R est créé dans la base de données. Ce vecteur contient la valeur 1 sur les lignes correspondantes aux observations appartenant à l'étude R-90 et 0 partout ailleurs. Il est ajouté comme effet aléatoire au modèle 5.1. Le RIS de l'étude A est calculé de la même manière que le RES à la seule différence que la valeur prédite considérée est celle obtenue à partir du modèle initial, utilisant les coefficients du tableau 5.2 en annexe. A partir de la valeur de RIS, celle de t_R^2 le carré des RIS et la statistique de rapport de vraisemblance LRT_R sont calculées. La valeur de LRT_R trouvée est comparée au seuil critique 2,01.

Chapitre 4

Résultats

4.1 Résultats des simulations

4.1.1 Cas de la régression linéaire simple avec des observations aberrantes

La valeur moyenne des X_i sur l'ensemble des 1000 échantillons est $5,03 \pm 0,53$ et celle des ϵ_i est de $0,003 \pm 0,18$.

4.1.1.1 Résultats de l'application de la méthode des RES dans le cas du ML

Le tableau 4.1 présente la moyenne des RES des observations 1, 12 et 30 sur l'ensemble des 1000 échantillons. Le RES de l'observation 1 non-atypique est largement inférieur à celui des observations 30 puis 12 respectivement. Selon Viechtbauer et al. , une observation ayant un RES $> 1,96$ serait potentiellement aberrante.

TABLE 4.1 – Résidus externes studentisés des observations 1, 12 et 30

Obs	Moyenne	IC 95%	Ecart-type
1	-0,17	-0,21 ; -0,12	0,72
12	4,10	4,06 ; 4,14	0,65
30	2,57	2,52 ; 2,61	0,78

La figure 4.1 montre la distribution de RES par observation investiguée. Moins de 1% (0,9%) des échantillons identifient à tort la première observation comme aberrante. 100% des simulations identifient correctement l'observation 12 aberrante. 77,5% des simulations

identifient correctement l'observation 30 aberrante.

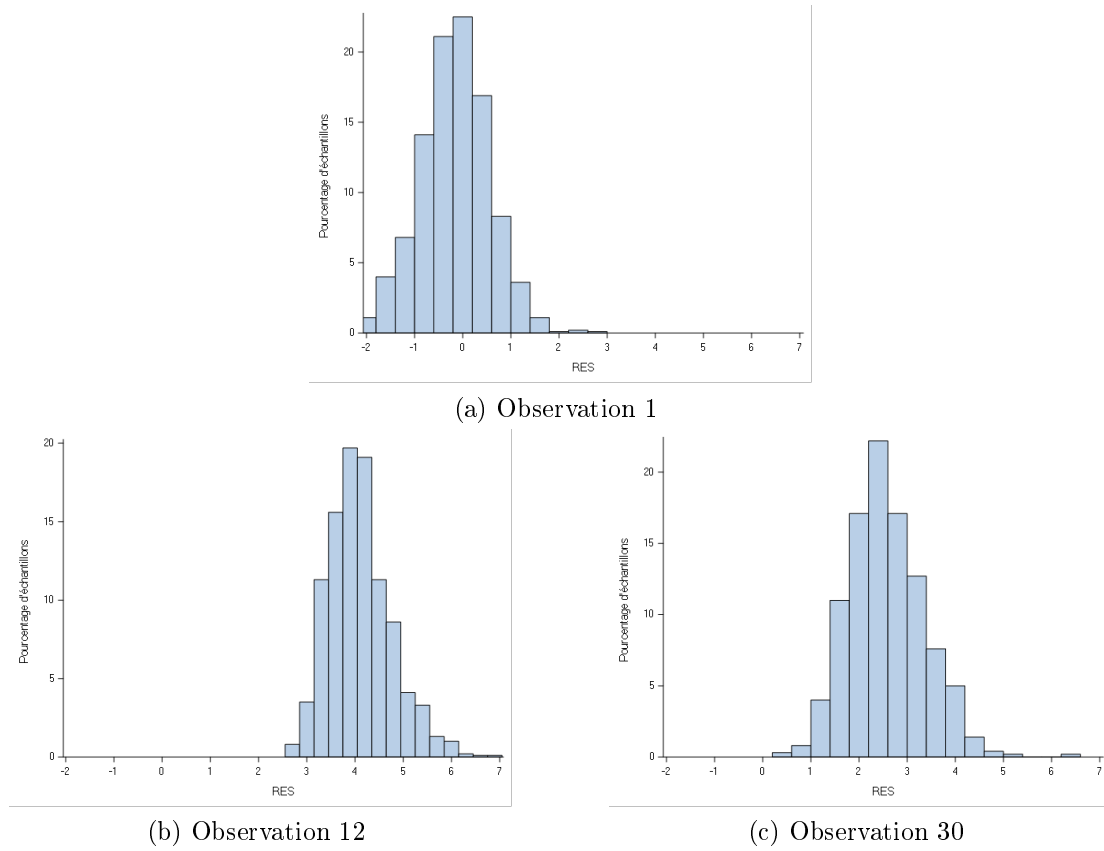


FIGURE 4.1 – Histogramme des résidus externes studentisés par observation ($X_1 \sim U[0; 10]$, $\epsilon_1 \sim N(0, 1)$, $\epsilon_{12} = \epsilon_{30} = 5$ et $X_{30} = 20$)

4.1.1.2 Résultats de l'application de la méthode du VSOM dans le cas du ML

Le tableau 2, présente les résidus internes studentisés des observations investiguées en ajustant le modèle linéaire simple (équation 3.5) précédemment spécifié. Ceux des observations atypiques 12 et 30 sont largement supérieurs à ceux de l'observation 1 qui ne l'est pas. Gumedze et al. ont obtenu respectivement 14,7 et 6 pour le carré des RIS des observations 12 et 30 soit 3,83 et 2,45 pour leur racine carré. Les observations 12 et 30 devraient donc être investiguées par la méthode des VSOM.

Sur l'ensemble des 1000 échantillons, le modèle mixte ajusté estime comme valeur moyenne de l'intercepte $\hat{b} = 26,19 \pm 0,40$ et celle de la pente $\hat{\alpha} = 0,27 \pm 0,07$. Gumedze et al. quant à eux ont trouvé $26,4 \pm 0,30$ pour l'intercept et $0,25 \pm 0,06$ pour la pente. Les coefficients estimés se rapprochent non seulement entre les deux études mais aussi

TABLE 4.2 – Résidus internes studentisés des observations 1, 12 et 30

Obs	Moyenne	IC 95%	Ecart-type
1	-0,16	-0,21 ; -0,12	0,71
12	3,25	3,23 ; 3,27	0,31
30	2,29	2,26 ; 2,33	0,56

des valeurs simulées ($b=26,25$ et $\alpha = 0,25$). La méthode de VSOM est donc correctement appliquée telle que décrite par l'article.

Les paramètres de variance obtenus pour les effets aléatoires par le modèle mixte ajusté sont présentés dans le tableau 4.3 ci-dessous. La variance moyenne des effets aléatoires des observations aberrantes est supérieure à celle de la première observation qui ne l'est pas.

TABLE 4.3 – Variance des effets aléatoires

Effet aléatoire	Variance	Ecart-type
d1	0,48	1,30
d12	23,85	2,84
d30	23,82	10,68

La variance moyenne estimée des résidus est $\hat{\sigma}^2 = 0,97 \pm 0,27$ contre 0,51 trouvé par Gumedze et al. Ils ont trouvé comme facteur d'inflation de la variance $\hat{w}_{12} = 51,12$ et $\hat{w}_{30} = 51,27$. En appliquant le facteur d'inflation, on obtient : $\hat{w}_{12}\hat{\sigma}^2 = 26,07$ et $\hat{w}_{30}\hat{\sigma}^2 = 26,15$. Les facteurs d'inflation de la présente simulation sont en moyenne sur l'ensemble des échantillons $\hat{w}_{12} = \hat{\sigma}_{d12}^2/\hat{\sigma}^2 = 24,58$, de même $\hat{w}_{30} = 24,56$.

La figure 4.2 montre la distribution du facteur d'inflation w_i de la variance des effets aléatoires d_1 , d_{12} et d_{30} par rapport à la variance résiduelle ainsi que les valeurs de RIS associées à travers l'ensemble des 1000 échantillons. Sur plus de 80% des échantillons w_1 prend une valeur comprise entre 0 et 1. La variance associée à l'effet aléatoire d_1 ne serait donc pas amplifiée par rapport à la variance résiduelle. Par contre w_{12} est d'office supérieur à 7 pour tous les échantillons. L'observation 12 serait donc correctement détectée aberrante pour 100% des simulations. La plupart des échantillons présente une grande valeur de w_{30} . L'observation 30 serait également détectée aberrante pour la majorité des simulations avec justesse.

Les figures 4.3 et 4.4 confirment les observations précédentes. La décision est pratiquement la même selon que l'on considère les RIS au carré ou la statistique du rapport de vraisemblance. La décision est sans ambiguïté pour l'observation 12 selon les deux statistiques. L'observation 1 n'est pas considérée aberrante à raison sur 99,5% des échan-

tillons aussi bien selon les RIS que la statistique du rapport de vraisemblance. De même, l'observation 30 est détectée aberrante de façon adéquate sur 74% des échantillons selon les RIS et 73,9% selon la statistique du rapport de vraisemblance. La décision selon les RIS est considérée pour comparer la méthode des RES et celle du VSOM dans le cas du ML.

4.1.1.3 Résultats de la comparaison de la performance des deux méthodes dans le cas du ML

Les deux méthodes détectent bien les observations atypiques 12 et 30 dans le cas de la régression linéaire simple. Elles réagissent également bien vis-à-vis de la première observation qui n'est pas atypique.

RES et RIS

Il n'y a presque pas de différence entre la valeur du RES de la première observation (tableau 4.1) et celle du RIS (tableau 4.2) de cette même observation. Les RES des observations 12 et 30 sont supérieures à leur RIS. Aussi bien les valeurs des RES que celles des RIS suggèrent que les observations 12 et 30 sont potentiellement aberrantes.

Changement des paramètres du modèle

Le changement des valeurs de l'intercept et de la pente du modèle n'affecte pas l'identification des valeurs aberrantes pour les deux méthodes. En effet, lorsque les coefficients du modèle sont modifiés, seules les valeurs estimées des modèles de régression ajustés diffèrent. Par exemple, si l'intercept b et la pente α du modèle sont respectivement fixés à 40 et 1, on obtient en moyenne comme valeurs estimées pour le modèle mixte sur l'ensemble des 1000 échantillons : $\hat{b} = 39,9 \pm 0,40$ et $\hat{\alpha} = 1,02 \pm 0,07$.

Le tableau 4.4, montre l'évolution des statistiques de détection des deux méthodes selon le changement de la valeur de X_{30} . Considérer une valeur de $X_{30} = 0$ plutôt que $X_{30} = 20$ dans la simulation des données, n'affecte pas vraiment les résultats des deux méthodes quant à la détection des observations 12 et 30 comme étant aberrantes. Bien que les valeurs 0 et 10 affectée à X_{30} ne soient pas aberrantes, les deux méthodes identifient l'observation 30 comme étant atypique à cause de la valeur aberrante de l'erreur qui lui est associée.

Le tableau 4.5 montre l'évolution des statistiques de détection des deux méthodes en fonction du changement de valeurs de $\epsilon_{12;30}$. Les deux méthodes sont plus sensibles au changement de la valeur aberrante de ϵ_i qu'à celui de X_i . Cependant, les paramètres du

TABLE 4.4 – Evolution des statistiques de détection selon la modification de la valeur de X_{30}

		Obs 1	Obs 12	Obs 30	Paramètres du modèle
$X_{30} = 10$	RES	$-0,19 \pm 0,64$	$3,56 \pm 0,40$	$3,39 \pm 0,42$	$\sigma^2 = 0,97 \pm 0,27$
	IC95% RES	$-0,23 ; -0,15$	$3,54 ; 3,59$	$3,36 ; 3,41$	$\hat{b} = 26,24 \pm 0,40$
	σ_d^2	$0,48 \pm 1,29$	$23,84 \pm 2,75$	$23,89 \pm 3,89$	$\hat{\alpha} = 0,26 \pm 0,07$
$X_{30} = 0$	RES	$-0,19 \pm 0,65$	$3,54 \pm 0,39$	$3,38 \pm 0,44$	$\sigma^2 = 0,97 \pm 0,27$
	IC95% RES	$-0,23 ; -0,15$	$3,52 ; 3,57$	$3,35 ; 3,40$	$\hat{b} = 26,29 \pm 0,40$
	σ_d^2	$0,48 \pm 1,29$	$23,83 \pm 2,75$	$23,98 \pm 3,97$	$\hat{\alpha} = 0,25 \pm 0,07$

modèle ne sont pas tellement affectés par le changement de valeurs de $\epsilon_{12;30}$.

Les RES et la variance des effets aléatoires augmentent avec la valeur des ϵ_i aberrantes. Comme on devrait s'y attendre, la variance de l'effet aléatoire associé à la première observation ne change presque pas à travers les différentes simulations. Par contre, comme pour les 2 autres observations, le RES_1 augmente lorsqu'on augmente la valeur aberrante de ϵ .

TABLE 4.5 – Evolution des statistiques de détection selon la modification de la valeur de $\epsilon_{12;30}$

		Obs 1	Obs 12	Obs 30	Paramètres du modèle
$\epsilon = 4$	RES	$-0,15 \pm 0,79$	$3,50 \pm 0,59$	$2,26 \pm 0,79$	$\sigma^2 = 0,97 \pm 0,27$
	IC95% RES	$-0,20 ; -0,10$	$3,46 ; 3,53$	$2,21 ; 2,31$	$\hat{b} = 26,17 \pm 0,42$
	σ_d^2	$0,49 \pm 1,31$	$14,87 \pm 2,33$	$14,87 \pm 8,63$	$\hat{\alpha} = 0,27 \pm 0,07$
$\epsilon = 3$	RES	$-0,12 \pm 0,86$	$2,77 \pm 0,50$	$1,85 \pm 0,78$	$\sigma^2 = 0,97 \pm 0,27$
	IC95% RES	$-0,17 ; -0,07$	$2,74 ; 2,80$	$1,80 ; 1,90$	$\hat{b} = 26,15 \pm 0,42$
	σ_d^2	$0,49 \pm 1,31$	$7,89 \pm 1,79$	$7,96 \pm 6,54$	$\hat{\alpha} = 0,28 \pm 0,08$
$\epsilon = 2$	RES	$-0,08 \pm 0,94$	$1,93 \pm 0,39$	$1,33 \pm 0,77$	$\sigma^2 = 0,96 \pm 0,27$
	IC95% RES	$-0,14 ; -0,03$	$1,90 ; 1,95$	$1,28 ; 1,38$	$\hat{b} = 26,14 \pm 0,40$
	σ_d^2	$0,48 \pm 1,30$	$2,94 \pm 1,17$	$3,31 \pm 4,36$	$\hat{\alpha} = 0,28 \pm 0,07$
$\epsilon = 1$	RES	$-0,04 \pm 1$	$0,99 \pm 0,27$	$0,69 \pm 0,76$	$\sigma^2 = 0,95 \pm 0,26$
	IC95% RES	$-0,09 ; -0,03$	$0,96 ; 1$	$0,65 ; 0,74$	$\hat{b} = 26,19 \pm 0,35$
	σ_d^2	$0,48 \pm 1,29$	$0,22 \pm 0,39$	$0,96 \pm 2,25$	$\hat{\alpha} = 0,27 \pm 0,06$

Lorsque $\epsilon = 4$ est considérée plutôt que $\epsilon = 5$ comme valeur déviante des erreurs du modèle, les deux méthodes présentent de grandes valeurs pour les statistiques des observations 12 et 30, suggérant qu'elles sont aberrantes. Lorsque $\epsilon = 3$, $RES_{30} < 1,96$. L'observation 30 n'est donc pas identifiée comme aberrante dans ce cas en considérant les RES contrairement à l'observation 12. A partir de $\epsilon = 2$, toutes les valeurs de RES sont inférieures à 1,96. Aucune des deux observations n'apparaît donc aberrante selon la méthode des RES. Quant à la méthode des VSOM, la variance des effets aléatoires des deux observations est beaucoup plus grande que celle de l'obs 1 et que la variance

résiduelle. C'est seulement lorsque $\epsilon = 1$ que $\sigma_{d12}^2 < \sigma_{d1}^2 < \sigma^2 < \sigma_{d30}^2$. La variance des effets aléatoires des observations supposées aberrantes est dans les mêmes proportions que celle de la première observation et de la variance résiduelle. La méthode des VSOM, ne les considère donc pas aberrantes à raison. En effet, les erreurs du modèle simulé étant tirées d'une distribution $N(0, 1)$, une valeur d' $\epsilon = 1$ ne serait effectivement pas aberrante. $\epsilon = 2$ pourrait par contre être considérée aberrante. La méthode des VSOM, serait-elle plus efficace pour détecter les petites déviations par rapport à celle des RES dans le cas de la régression linéaire simple ?

Lorsque $\epsilon_{12;30}=1$, par la méthode des RES les 3 observations (1, 12 et 30) sont identifiées non-aberrantes pour plus de 90% des échantillons (respectivement 94,7%, 99,6% et 95,1% voir figure 4.5 (a) à (c)), ces valeurs ne pouvant effectivement pas être considérées comme aberrantes.

Plus la valeur de $\epsilon_{12;30}$ augmente, pour plus d'échantillons la méthode identifie correctement que la première observation n'est pas aberrante. Le pourcentage d'échantillons pour lesquels les observations 12 et 30 sont correctement identifiées aberrantes augmente avec la valeur de $\epsilon_{12;30}$. Lorsque $\epsilon_{12;30}=2$, on constate une certaine ambiguïté de la décision en ce qui concerne l'observation 12 : sur 58,9%, elle est identifiée non-aberrante. Cette ambiguïté s'observe pour l'observation 30 lorsque $\epsilon_{12;30}=3$ où la méthode ne la détecte pas aberrante sur 59,3% des échantillons. A partir de $\epsilon_{12;30}=4$, la décision est unanime sur l'ensemble des échantillons en ce qui concerne l'observation 12 et sur 63,6% des échantillons, l'observation 30 est bien considérée aberrante. La méthode aurait plus de mal à détectée l'observation 30 aberrante du fait de l'influence de X_{30} ayant une valeur plus grande que pour les autres observations.

Les décisions selon les valeurs de t_i^2 présentent les mêmes tendances que celles selon les RES à quelques nuances près. Le carré des RIS désigne correctement les observations non-aberrantes sur légèrement plus d'échantillon que les RES. Les RES quant à eux détectent mieux les observations aberrantes que lorsque la décision est basée sur les valeurs de t_i^2 . Bien que les variances des effets aléatoire d_{12} et d_{30} apparaissent grandes pour $\epsilon = 2$, les RES détectent ces deux observations aberrantes sur plus d'échantillon que les RIS au carré malgré la petite déviance.

En somme dans le cas de la régression linéaire classique, la méthode des RES telle qu'appliquée dans la présente étude détecte mieux les valeurs aberrantes quelle que soit l'amplitude d'aberrance. En ce qui concerne la méthode de VSOM, elle désigne mieux les observations non-aberrantes.

4.1.2 Cas du modèle linéaire mixte avec un cluster aberrant

4.1.2.1 Résultats de l'application de la méthode des RES dans le cas du MLM

Le RES de l'étude A est en moyenne égale à $9,82 \pm 0,24$ sur l'ensemble des échantillons tandis que celle de l'étude E vaut en moyenne $-0,45 \pm 0,01$.

La méthode des RES détecte l'étude A aberrante sur tous les échantillons. Elle ne considère pas l'étude E aberrante sur l'ensemble des simulations. La décision est correcte à 100% pour cette méthode aussi bien pour identifier l'étude aberrante que celle non aberrante.

4.1.2.2 Résultats de l'application de la méthode du VSOM dans le cas du MLM

Lorsque le VSOM est ajusté aux données, on obtient comme valeur moyenne estimée de l'intercept sur l'ensemble des échantillons $26,25 \pm 0,07$ et comme valeur moyenne estimée de la pente $0,27 \pm 0,07$. La variance estimée des résidus est en moyenne $1,02 \pm 0,05$ et celle de l'effet aléatoire pays est en moyenne $1,92 \cdot 10^{-4} \pm 1,92 \cdot 10^{-3}$. Ces valeurs correspondent bien aux coefficients simulés du modèle de départ.

La variance associée à l'effet aléatoire d_A est en moyenne égale à $100,07 \pm 7,83$ sur l'ensemble des simulations. Celle de l'effet aléatoire d_E quant à elle est égale à $0,11 \pm 0,16$ en moyenne. Les facteurs d'inflation de ces variances par rapport à la variance résiduelle sont respectivement $\hat{w}_A = 98,07 \pm 9,35$ et $\hat{w}_E = 0,11 \pm 0,15$ en moyenne. La figure 4.8 montre la distribution des facteurs d'inflation de la variance des effets aléatoires d_A et d_E sur l'ensemble des 100 simulations. Le facteur d'inflation associé à l'effet aléatoire d_A est largement supérieur à celui associé à l'effet aléatoire d_E .

Le carré des RIS de l'étude A est égale $14,66 \pm 0,26$ en moyenne sur l'ensemble des échantillons et celui de l'étude E vaut en moyenne $0,19 \pm 0,11$. Les valeurs de la statistique du test de rapport de vraisemblance calculée à partir des RIS au carré sont en moyenne $11,07 \pm 0,24$ et 0 respectivement pour les études A et E. La figure 4.9 présente les statistiques de décision de la méthode VSOM. La statistique du rapport de vraisemblance détecte l'étude A aberrante pour toutes les simulations. Elle désigne également correctement l'étude E non-aberrante sur l'ensemble des échantillons.

4.1.2.3 Résultats de la comparaison de la performance des deux méthodes dans le cas du MLM

Changement des coefficients du modèle initial

Le changement des coefficients du modèle a le même effet que dans le cas du modèle linéaire classique. Cela n'affecte que les valeurs estimées de ces coefficients. Les statistiques de décision des deux méthodes restent inchangées. Par exemple, une valeur de 40 est considérée pour l'intercept et 1 pour la pente du modèle initial au lieu de 26,25 et 0,25 respectivement. Dans ce cas, les valeurs moyennes estimées sur l'ensemble des 100 échantillons pour ces coefficients lors de l'ajustement du modèle mixte sont respectivement $39,99 \pm 0,07$ et $1,02 \pm 0,07$.

Changement de l'amplitude de la déviance

Le tableau 4.6 montre l'évolution des statistiques de détection des deux méthodes selon le changement de l'amplitude de déviance des valeurs aberrantes. Comme on devrait s'y attendre, les valeurs des statistiques de détection décroissent lorsque la déviance des valeurs aberrantes diminue.

TABLE 4.6 – Evolution des statistiques de détections selon le changement de l'amplitude de la déviance des valeurs aberrantes

		Etude A	Etude E
y+2	RES	$1,97 \pm 0,05$	$-0,18 \pm 6,23.10^{-3}$
$\hat{\sigma}^2 = 1,02 \pm 0,05$	$\hat{\sigma}_{dij}^2$	$3,98 \pm 1,58$	$0,11 \pm 0,15$
y+1	RES	$0,99 \pm 0,02$	$-0,08 \pm 2,28.10^{-3}$
$\hat{\sigma}^2 = 0,92 \pm 0,20$	$\hat{\sigma}_{dij}^2$	$1,05 \pm 0,78$	$0,11 \pm 0,16$

On observe sur la figure 4.10 la distribution des RES pour l'ensemble des échantillons selon la déviance des valeurs aberrantes de l'étude A. Lorsque la variable réponse y est majorée de 2 plutôt que 10 pour l'étude A, cette étude est détectée aberrante sur 59% des échantillons par les RES. Pour une majoration de 1 de la variable réponse, l'étude A ne peut être considérée aberrante. De ce fait, elle est détectée non-aberrante sur l'ensemble des échantillons par la méthode des RES. Les RES ne considèrent pas à raison l'étude E comme aberrante sur l'ensemble des échantillons quelle que soit la valeur de la déviance de l'étude A.

Lorsque la valeur de la variable réponse de l'étude A est majorée de 2 plutôt que de 10, la moyenne des RIS au carré sur l'ensemble des échantillons est égale à $2,75 \pm 0,89$ pour cette étude. Celle de l'étude E est égale à $0,11 \pm 0,14$.

Lorsque la valeur de la variable réponse est majorée de 1 plutôt que 2, le RIS au carré de

l'étude A vaut en moyenne $0,86 \pm 0,56$ sur l'ensemble des échantillons et la statistique du rapport de vraisemblance associée vaut en moyenne $0,04 \pm 0,09$.

La figure 4.11 montre la distribution des LRT_j des études A et E sur l'ensemble des échantillons selon la valeur de la déviance des valeurs aberrantes. L'étude A n'est détectée aberrante que sur un seul échantillon sur 100. Les LRT_j désignent correctement l'étude E non-aberrante sur l'ensemble des échantillons. Lorsque $y_A = y + 1$, l'étude A ne peut être considérée aberrante. Elle est désignée non-aberrante sur l'ensemble des échantillons par les LRT_j . La valeur de la statistique du rapport de vraisemblance de l'étude E ne change pas au fil des simulations quelle que soit l'amplitude de la déviance des valeurs de l'étude A.

En somme, la méthode des RES détecte l'étude A lorsque la mesure réponse est majorée de 2 sur certains échantillons tandis que celle de VSOM telle qu'appliquée ne la détecte aberrante que sur un seul échantillon. La méthode des RES telle qu'appliquée dans la présente étude apparaît également dans le cas du MLM plus efficace que celle du VSOM pour détecter une petite déviance.

4.2 Résultats de l'application des deux méthodes sur des données réelles

Le tableau 5.2 en annexe présente les coefficients estimés du modèle linéaire mixte initial expliquant la concentration d'anticorps dans le sang ainsi que les p-valeurs associés. Si on considère $\hat{\alpha}_k$ les coefficients obtenus directement des sorties du logiciel SAS pour chaque effet fixe, les coefficients présentés dans ce tableau sont les valeurs transformées : $10^{\hat{\alpha}_k}$. Cette transformation permet d'interpréter l'effet des facteurs directement en terme de changement d'unités de GMT induit par chaque coefficient. La variance estimée des résidus ($\hat{\sigma}^2$) de ce modèle est égale à 0,11 et celle de l'effet aléatoire pays ($\hat{\sigma}_c^2$) vaut 0,03.

En partant des coefficients estimés obtenues directement des sorties du logiciel, la valeur prédite de GMT pour la nouvelle étude R-90 en logarithme à base 10 est calculée. Les détails de ce calcul ainsi que la valeur obtenue sont présentés en annexe. Ce sont ces valeurs prédites du modèle en logarithme à base 10 qui sont considérées pour le calcul des résidus et des statistiques de détection de valeurs aberrantes.

4.2.0.1 Résultats de l'application de la méthode des RES sur les données réelles

La valeur moyenne prédite de la nouvelle étude calculée à partir des coefficient estimés du modèle ajusté sans les observations de cette étude $\hat{\mu}_{R(-R)}$ et la variance associée $\text{var}(\hat{\mu}_{R(-R)})$ sont reportées en annexe. Il en est de même pour la valeur moyenne observée y_R de la variable réponse pour cette étude. La variance des résidus de ce modèle et celle de l'effet aléatoire pays ont la même valeur que celles du modèle initial. La valeur absolue du RES de l'étude R-90 (voir annexe) est inférieure à 1,96. Selon la méthode des RES telle qu'appliquée, la nouvelle étude n'est pas aberrante.

4.2.0.2 Résultats de l'application de la méthode du VSOM sur les données réelles

Lorsque le VSOM est ajusté au modèle 5.1 en y ajoutant l'effet aléatoire d_R , les coefficients estimés du modèle ne changent pas de même que les variances associées aux résidus et à l'effet aléatoire pays. La variance de l'effet aléatoire d_R quant à elle est égale à 0,04. Elle n'est pas majorée par rapport à la variance résiduelle du modèle.

En considérant la valeur prédite pour la nouvelle étude à partir du modèle initial (tableau 5.2 en annexe), $\hat{\mu}_R$ et la variance associée $\text{var}(\hat{\mu}_R)$, le RIS de la nouvelle étude obtenu est reporté en annexe ainsi que le carré du RIS, t_R^2 . La valeur de t_R^2 trouvée étant inférieure à 1, la statistique du rapport de vraisemblance pour l'étude R-90 est donc égale à 0. Elle est inférieure à 2,01. La nouvelle étude n'apparaît donc pas non plus aberrante selon la méthode du VSOM.

Aucune des deux méthodes ne considère la nouvelle étude aberrante contrairement à la décision de la statistique PRESS adaptée (voir section 1.1) utilisée au départ pour détecter les valeurs aberrantes dans cette méta-analyse. Notons que la valeur du RIS et celle du RES de la nouvelle étude sont assez proches. Cette étude ne serait donc effectivement pas aberrante.

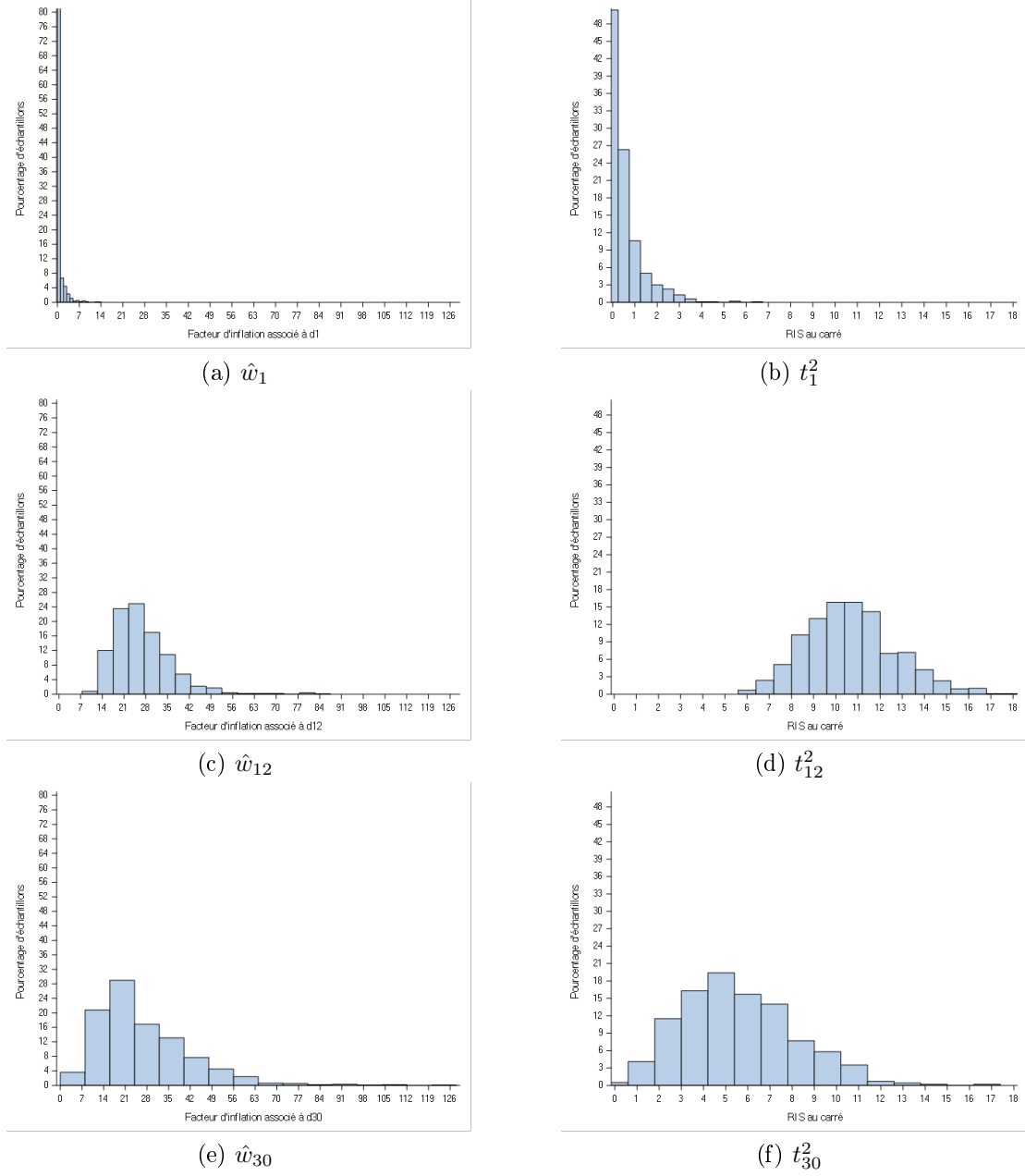
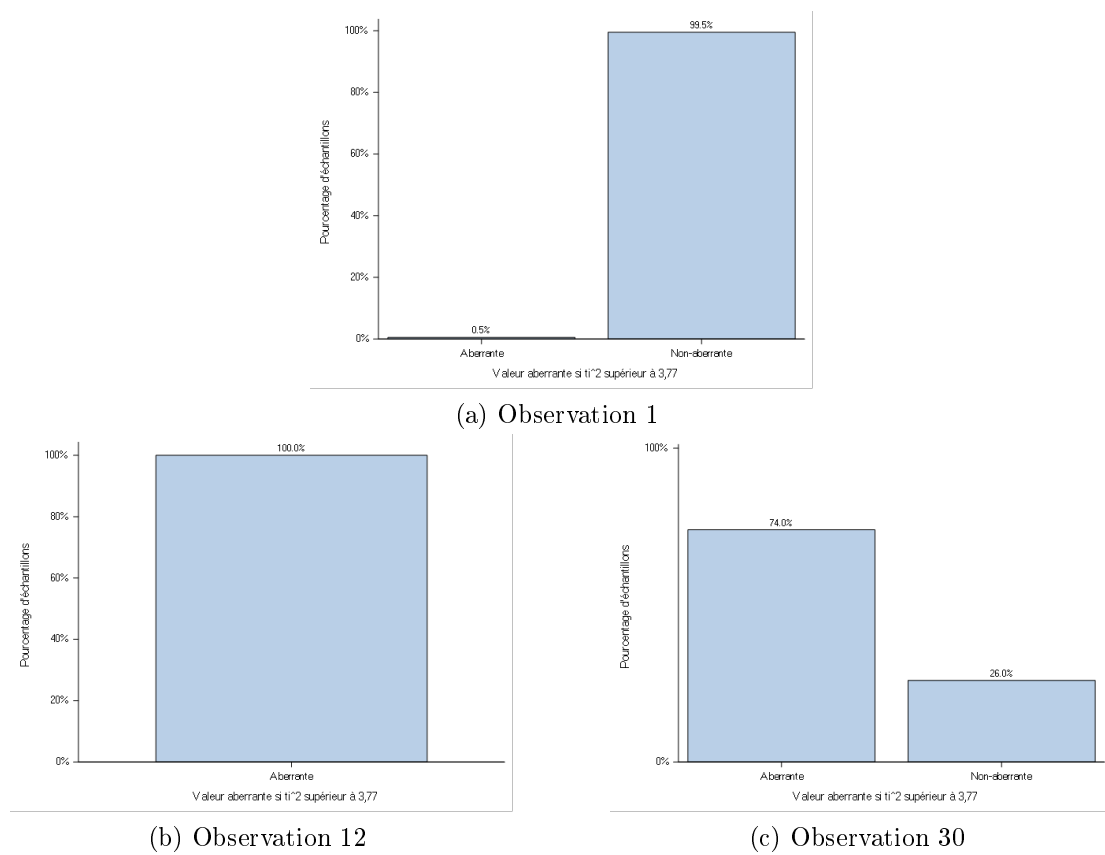
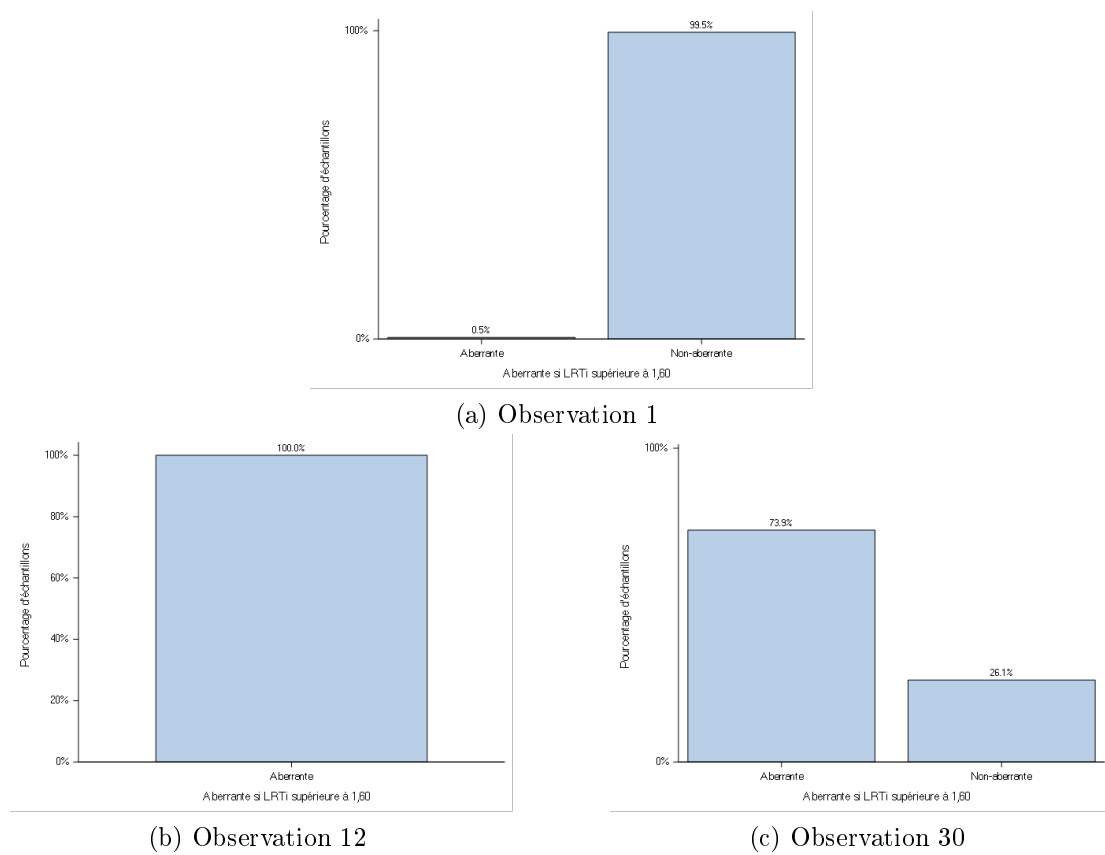


FIGURE 4.2 – Histogrammes du facteur d'inflation de la variance des effets aléatoires d_1 , d_{12} et d_{30} par rapport à la variance résiduelle et des RIS associés

FIGURE 4.3 – Décision selon les RIS t_i^2

FIGURE 4.4 – Décision selon la statistique du rapport de vraisemblance LRT_i

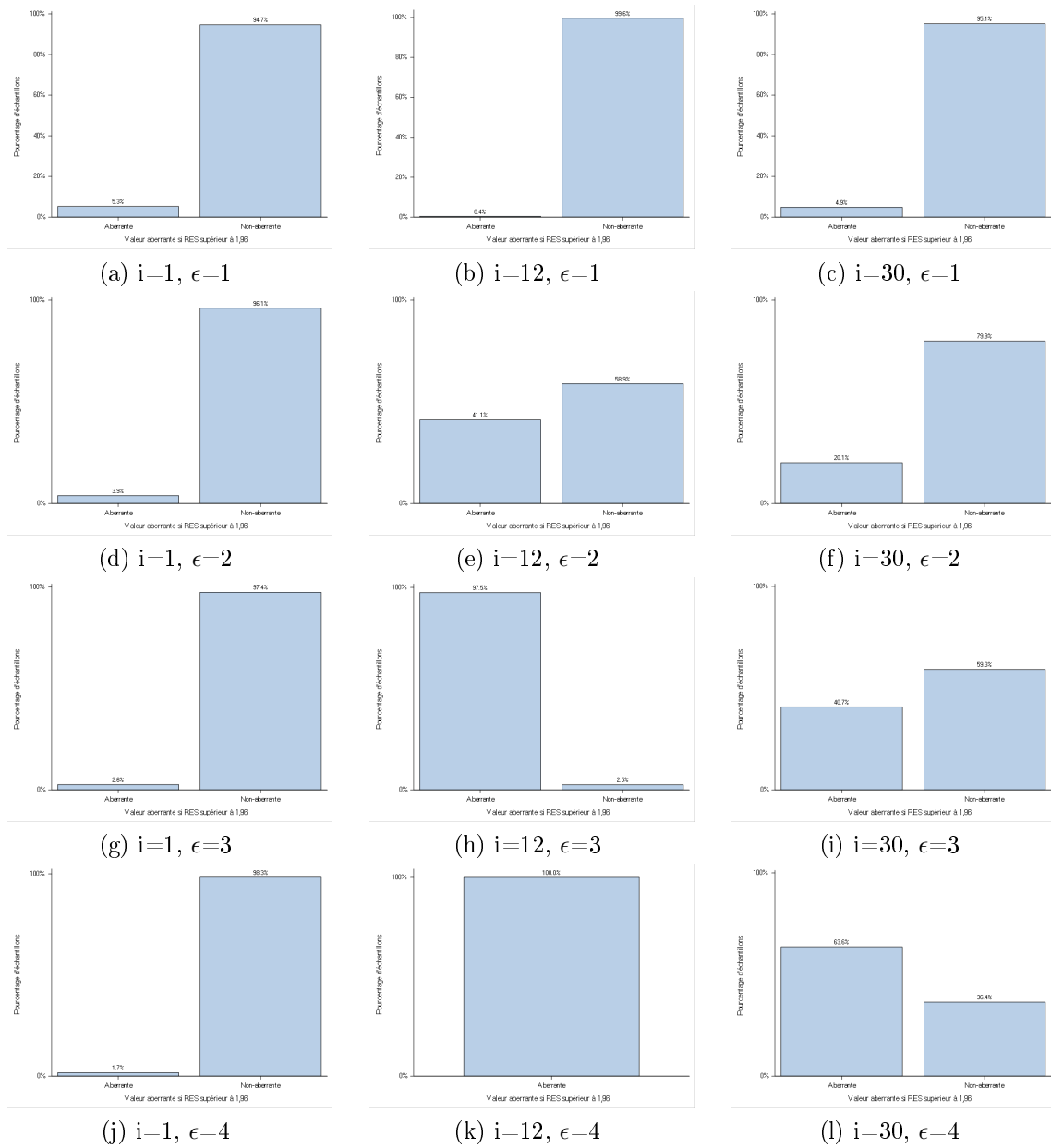


FIGURE 4.5 – Pourcentage d'échantillons identifiant chaque observation aberrante en fonction des différentes valeurs de ϵ selon les RES

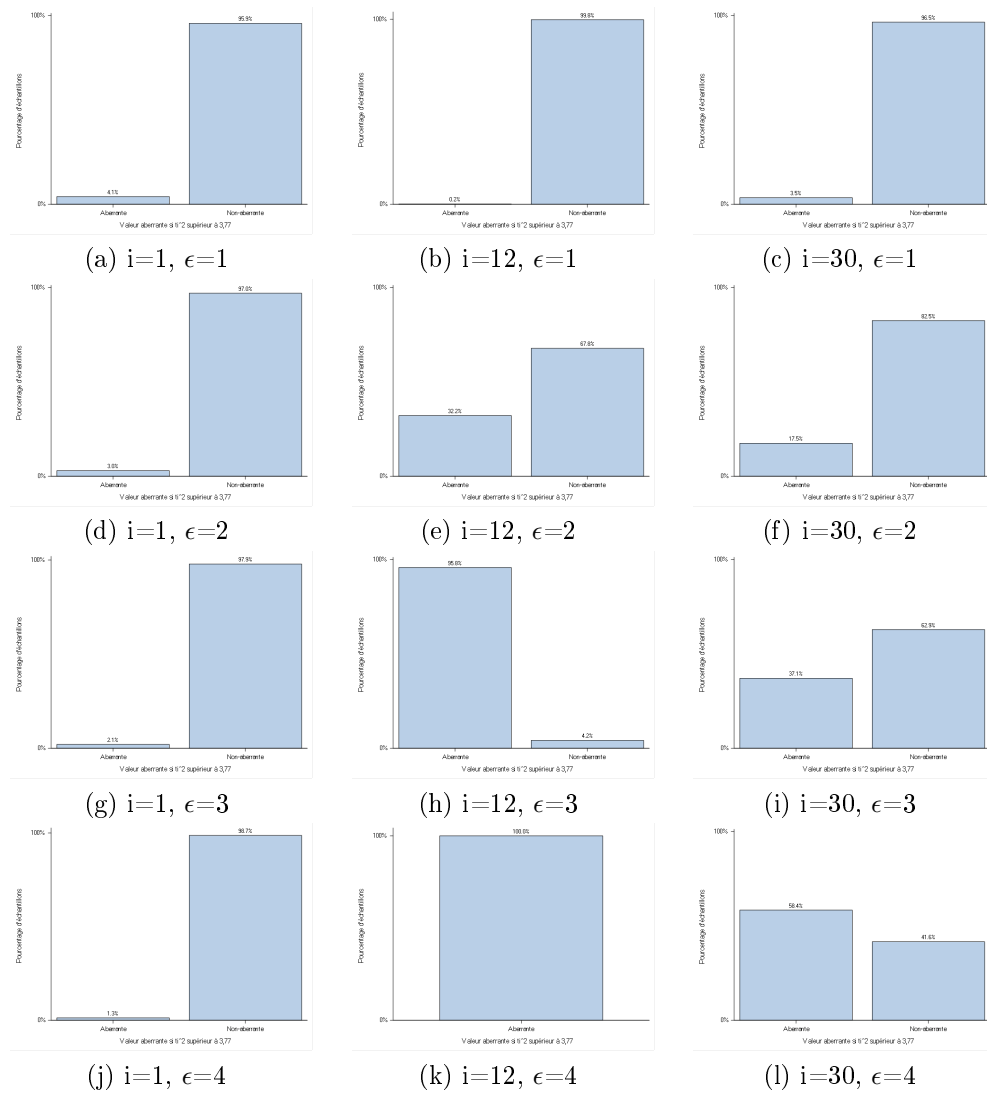


FIGURE 4.6 – Pourcentage d'échantillons identifiant chaque observation aberrante en fonction des différentes valeurs de ϵ selon t_i^2

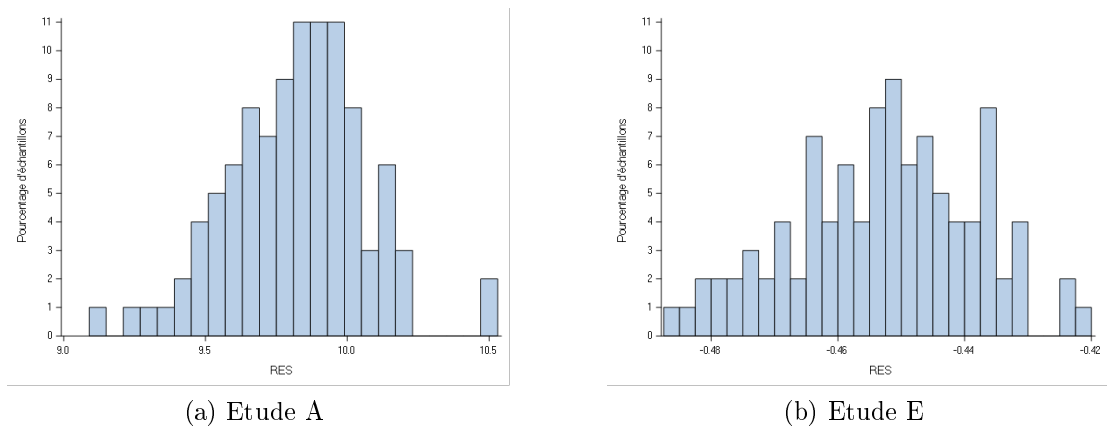


FIGURE 4.7 – Histogramme des résidus externes studentisés des études A et E ($y_A = y+10$)

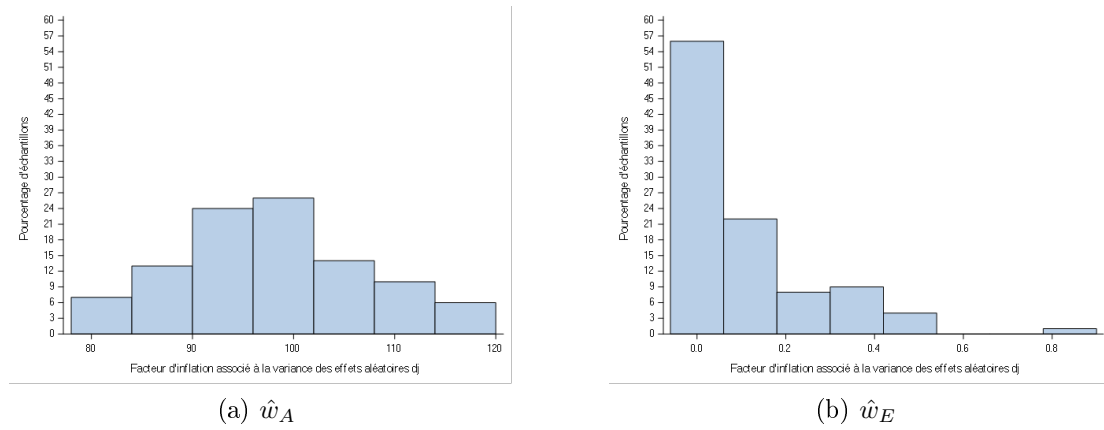


FIGURE 4.8 – Histogramme des facteurs d'inflation de la variance des effets aléatoires d_A et d_E)

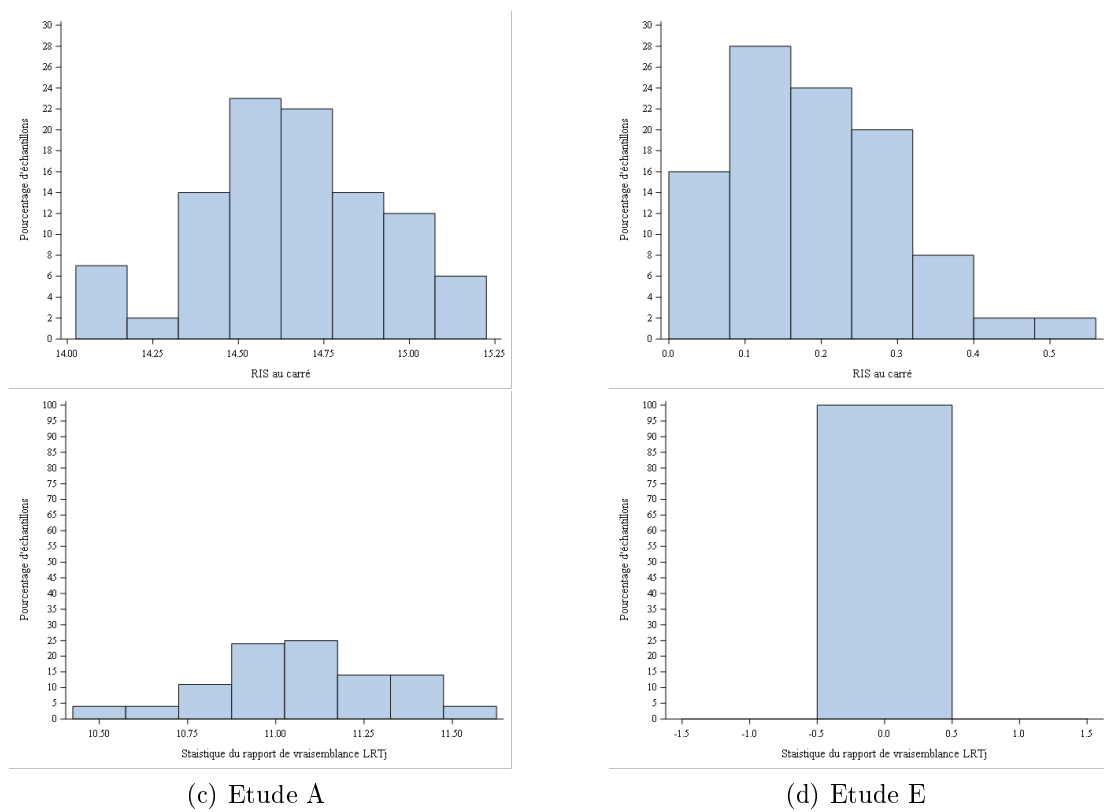


FIGURE 4.9 – Statistiques de décision de la méthode du VSOM, $y_A = y + 10$

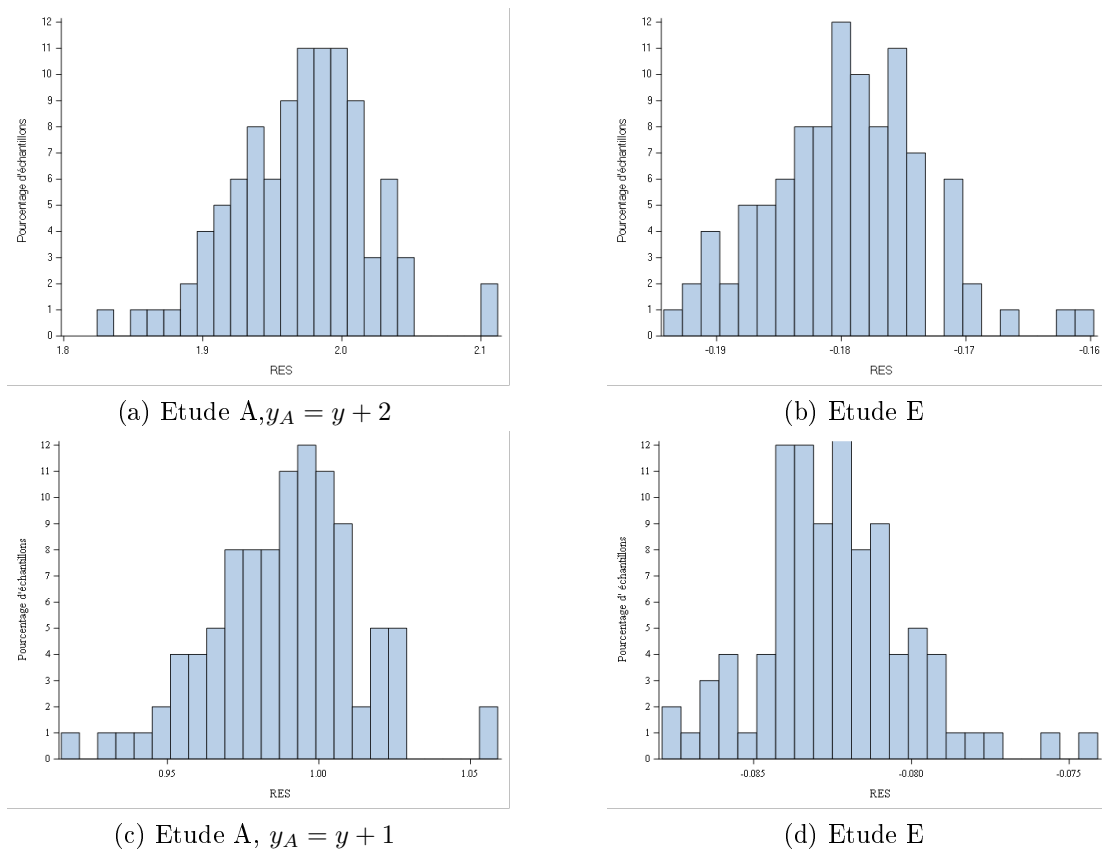
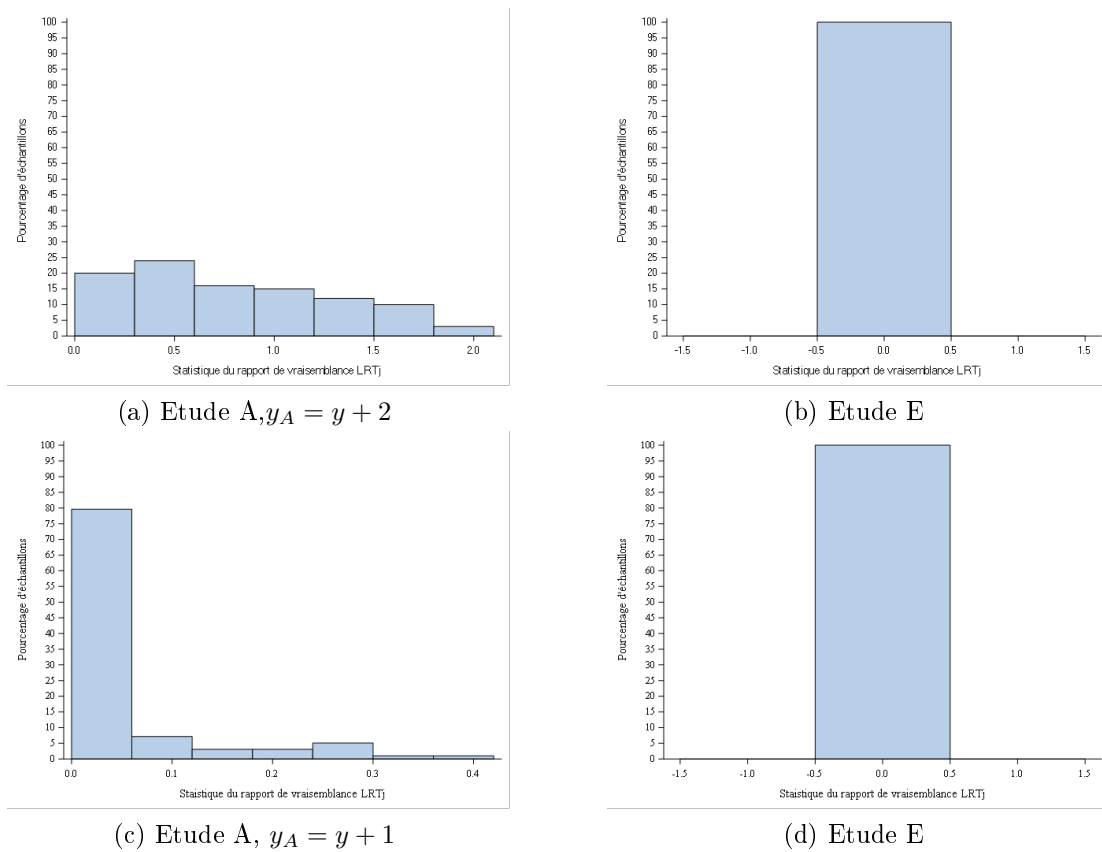


FIGURE 4.10 – Distribution des RES des études A et E

FIGURE 4.11 – Distribution des LRT_j des études A et E

Chapitre 5

Discussion/Conclusion

Pour comparer la mise en œuvre et la performance de la méthode des RES et celle du VSOM dans la détection des valeurs aberrantes dans une méta-analyse, le modèle de régression linéaire classique nous a permis de comprendre le fonctionnement de ces méthodes dans un cas simple afin d'avancer progressivement vers le cas complexe du modèle mixte à effet aléatoire avec une composante de variance. La méthode des RES apparaît plus aisée et plus rapide à mettre en œuvre que celle du VSOM. L'exécution des simulations de modèle mixte nécessite beaucoup plus de temps. Elles sont d'autant plus longues qu'il y a d'effets aléatoires dans le modèle. Or la méthode du VSOM requiert l'ajout d'un effet aléatoire au modèle par observation ou étude investiguée.

Dans les cas de la régression linéaire classique les deux méthodes étudiées : celle des RES et celle du VSOM dont la décision est établie en considérant les RIS s'équivalent pratiquement. Elles consistent toutes les deux essentiellement à étudier les résidus du modèle ajusté. La première se base sur les résidus calculés en excluant l'observation investiguée tandis que les résidus sont calculés pour la méthode du VSOM à partir de toutes les observations. L'une ou l'autre pourrait être utilisée selon l'intérêt de l'analyse. La méthode des RES détecte plus les observations aberrantes alors que celle du VSOM désigne plus aisément les observations non-aberrantes. Dans le cas du MLM également, la méthode des RES a permis de détecter l'étude aberrante même avec une petite déviance. Etant donné le contexte de la présente étude où l'analyse est réalisée dans le but de garantir la qualité des données d'essais cliniques, il serait plus approprié de se servir des RES pour détecter les valeurs aberrantes car il vaudrait mieux détecter à tort une étude aberrante que de manquer de l'identifier.

D'ailleurs Viechtbauer et Cheung [30], encouragent l'utilisation des RES plutôt que

celle des RIS pour détecter les valeurs atypiques car les dernières incluant l'étude aberrante dans leur calcul sont plus susceptibles d'être influencées par cette valeur. En effet si l'étude j est aberrante, alors $\hat{\mu}_j$, la valeur prédite pour cette étude sera plus rapprochée de la valeur observée aberrante y_j , la différence entre les deux sera donc plus petite ce qui rendra difficile l'identification de la valeur aberrante. De plus, la présence d'une valeur aberrante augmenterait la valeur estimée de la variance de l'effet aléatoire σ_c^2 . La variance résiduelle serait alors surestimée et les résidus studentisés résultant trop petits pourraient ne pas apparaître aberrants.

Le même principe s'applique dans le cas de la régression linéaire classique. Les deux méthodes détectent plus aisément l'observation dont seule la valeur de l'erreur est déviante que l'observation ayant non seulement une valeur d'erreur déviante mais aussi une valeur de X majorée par rapport aux autres observations. Aussi bien les RES que les RIS, sont constitués au dénominateur de la variance de $\hat{\epsilon}_i$. Dans le cas d'une matrice de design X à deux colonnes pour une régression linéaire classique telle que considérée dans la simulation de la présente étude, $var(\hat{\epsilon}_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \right)$ [27].

La valeur de l'observation ayant une valeur de X majorée étant éloignée de la moyenne des observations, la variance des erreurs pour cette observation influence la valeur obtenue des résidus. Cette observation en plus d'être aberrante est un point ayant un grand levier.

En outre, Viechtbauer et al. propose de calculer les RES avec la vraie valeur de la variance de l'erreur σ^2 et non une estimation. Ayant calculé les RES dans la présente étude en utilisant une estimation de la variance $\hat{\sigma}^2$, les résultats obtenus seraient moins précis. Des études ultérieures pourraient étudier l'impact de l'utilisation d'une estimation de la variance plutôt que la vraie valeur.

Aussi l'auteur suggère-t-il que sous l'hypothèse nulle, les RES suivent approximativement une distribution normale standard. Le seuil critique considéré pour les RES est donc fixé à 1,96 tandis que les RIS lors de l'application de la méthode du VSOM sont comparés au quantile 95% d'une distribution Beta $(1/2, (n - p - 1)/2)$ dans le cas de la régression linéaire classique. Cette dernière valeur critique tient compte de la taille de l'échantillon alors que la première en est indépendante. La taille d'échantillon $n=30$ utilisée dans le cadre du présent travail pour le cas du ML serait limite pour supposer une normalité asymptotique. De prochaines simulations pourraient exploiter des échantillons de taille plus grande ou considérer pour les RES un seuil critique tenant compte de la taille de l'échantillon tel qu'un quantile d'une distribution de Student. Notons également que dans le cas du MLM les RES et les RIS ne sont pas calculés de la même manière. Il serait plus

approprié de les calculer à l'aide de la même méthode à la seule différence de l'extraction des observations investiguées pour les RES afin que la comparaison de la décision parte d'une même base.

Par ailleurs, le percentile 95% de la distribution Beta $(1/2, (n - p - 1)/2)$ utilisé comme seuil critique de décision pour $t_i^2/(n - p)$ dans le cas du ML ne serait probablement pas le plus adapté au contexte. Afin de tenir compte de la dépendance entre les RIS, Gumedze et al. [6] ont plutôt fait recours à une méthode de bootstrap paramétrique pour approximer la distribution jointe des t_i^2 sous l'hypothèse nulle. Les observations présentant les plus grandes valeurs de t_i^2 sont alors comparées au percentile $100(1 - \alpha)$ de la distribution obtenue par la méthode du bootstrap, α représentant le taux d'erreur de type I souhaité. Il en est de même pour le cas du MLM. Au lieu de la mixture 68 :32 de Chi-carré à 0 et 1 degré de liberté Gumedze et al. proposent d'utiliser la même méthode de bootstrap paramétrique pour identifier la distribution asymptotique de la statistique du rapport de vraisemblance sous l'hypothèse nulle. Il serait intéressant à l'avenir d'apprécier les conséquences de cette amélioration sur les résultats de détection par la méthode du VSOM.

Les RES et les RIS sont calculés ici à partir des résidus marginaux $y - \hat{\alpha}X$ considérés comme les termes d'erreurs incluant toutes les sources de variabilité aléatoire (l'effet aléatoire et les résidus du modèle)[15]. Afin de considérer l'impact de l'effet aléatoire pays inclus dans le modèle, il serait plus adéquat de calculer les valeurs des RES et les RIS au départ des résidus conditionnels $y - X\hat{\alpha} - Z\hat{\beta}$ qui ne déduisent pas les valeurs de $Z\hat{\beta}$.

Pour le modèle mixte initial servant à générer les données au départ des simulations, la variance de l'effet aléatoire est assez petite. Elle vaut 0,1. La variance des erreurs quant à elle est fixée à 1. L'effet de l'augmentation de la variance de l'effet aléatoire ou les cas dans lesquels la variance résiduelle pourrait être modélisée (la valeur de la variance augmentant linéairement avec l'âge du sujet par exemple) pourrait également faire l'objet de recherches ultérieures.

Les deux méthodes appliquées sur des données réelles ne considèrent pas aberrante une étude qui avait été détectée aberrante par une version adaptée de la statistique PRESS. Notons que, l'étude ne faisait pas partie des premières études les plus aberrantes selon la statistique antérieurement utilisée. La statistique utilisée pour la détection des études atypiques pour l'analyse préalable à la présente étude se base sur la déviation vis-à-vis de la valeur prédite. Elle est influencée par la moyenne, la variabilité observée et la taille de l'échantillon. La désignation de cette étude comme atypique serait due non seulement à la grande taille de l'échantillon (981 observations) pour cet antigène par rapport aux autres mais aussi à la variabilité introduite dans le modèle par d'autres procédures intrinsèques

à l'analyse telle qu'une imputation réalisée pour certains antigènes par exemple.

Somme toute, ce travail a permis d'étudier deux méthodes de détection de valeurs aberrantes dans une méta-analyse implémentée à l'aide du logiciel SAS contrairement à pratiquement toutes les méthodes actuelles retrouvées dans la littérature qui proposent une implémentation sur R. À l'aide de données simulées, on conclut que la méthode des RES et celle de VSOM permettent de détecter des valeurs aberrantes dans une méta-analyse sur SAS. Ces méthodes appliquées sur des données réelles ont permis d'apprécier la déviance d'une nouvelle étude par rapport à la tendance générale des données. Les deux méthodes apparaissent utiles aussi bien l'une que l'autre pour détecter les valeurs aberrantes dans la méta-analyse ayant suscité la présente étude. Elles peuvent être appliquées de manière complémentaire. La méthode des RES pourrait contribuer à effectuer une première investigation des données. Celle du VSOM peut ensuite être réalisée pour confirmer la décision des RES et attribuer un poids aux études apparaissant aberrantes. Ainsi, l'impact de ces valeurs sur les résultats de la modélisation sera réduit. Il ne sera donc pas nécessaire de les retirer de l'analyse. L'information qu'elles apportent pourra alors être maintenue sans que leur déviance n'affecte considérablement les résultats de l'analyse. De futurs perfectionnements permettront d'améliorer leur mise en œuvre et leur performance afin de favoriser leur exploitation dans l'industrie.

Bibliographie

- [1] Ben BOLKER et al. *Emdbook : Support Functions and Data for "Ecological Models and Data"*. Version 1.3.12. 19 fév. 2020. URL : <https://CRAN.R-project.org/package=emdbook> (visité le 21/08/2020).
- [2] Helen BROWN et Robin PRESCOTT. *Applied Mixed Models in Medicine*. John Wiley & Sons, 1^{er} mai 2006. 479 p. ISBN : 978-0-470-02357-0. Google Books : 7BdjFUUwTzgC.
- [3] L. DESMET et al. « Linear Mixed-Effects Models for Central Statistical Monitoring of Multicenter Clinical Trials ». In : *Statistics in Medicine* 33.30 (2014), p. 5265-5279. ISSN : 1097-0258. DOI : 10.1002/sim.6294. URL : <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6294> (visité le 11/06/2020).
- [4] Lieven DESMET. « Efficient Statistical Detection of Fraud in Multicentre Experiments. May 2020 » (ISBA, UCL). 5 août 2020.
- [5] Freedom N. GUMEDZE et Dan JACKSON. « A Random Effects Variance Shift Model for Detecting and Accommodating Outliers in Meta-Analysis ». In : *BMC Medical Research Methodology* 11.1 (16 fév. 2011), p. 19. ISSN : 1471-2288. DOI : 10.1186/1471-2288-11-19. URL : <https://doi.org/10.1186/1471-2288-11-19> (visité le 06/08/2020).
- [6] Freedom N. GUMEDZE et al. « A Variance Shift Model for Detection of Outliers in the Linear Mixed Model ». In : *Computational Statistics & Data Analysis* 54.9 (1^{er} sept. 2010), p. 2128-2144. ISSN : 0167-9473. DOI : 10.1016/j.csda.2010.03.019. URL : <http://www.sciencedirect.com/science/article/pii/S0167947310001210> (visité le 09/07/2020).
- [7] Christian HAFNER. « LSTAT2120 Linear Models » (ISBA, UCL). 2018-19.
- [8] A B HAIDICH. « Meta-Analysis in Medical Research ». In : *Hippokratia* 14 (Suppl 1 déc. 2010), p. 29-37. ISSN : 1108-4189. pmid : 21487488. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049418/> (visité le 04/08/2020).

- [9] Center for Devices and Radiological HEALTH. « Clinical Data for Premarket Submissions ». In : *FDA* (Fri, 02/08/2019 - 23 :58). URL : <https://www.fda.gov/medical-devices/premarket-submissions/clinical-data-premarket-submissions> (visité le 24/07/2020).
- [10] Allen HUFFCUTT et Jr ARTHUR Winfred. « Development of a New Outlier Statistic for Meta-Analytic Data ». In : *Journal of Applied Psychology* 80 (1^{er} avr. 1995), p. 327-334. DOI : 10.1037/0021-9010.80.2.327.
- [11] Winfred Arthur JR, Winston BENNETT et Allen I. HUFFCUTT. *Conducting Meta-Analysis Using SAS*. Psychology Press, 1^{er} juin 2001. 232 p. ISBN : 978-1-135-64345-4. Google Books : cVN4AgAAQBAJ.
- [12] Manuel KOLLER. « **Robustlmm** : An *R* Package for Robust Estimation of Linear Mixed-Effects Models ». In : *Journal of Statistical Software* 75.6 (2016). ISSN : 1548-7660. DOI : 10.18637/jss.v075.i06. URL : <http://www.jstatsoft.org/v75/i06/> (visité le 04/08/2020).
- [13] Andrew KRAMAR et Simone MATHOULIN-PÉLISSIER. *Méthodes biostatistiques appliquées à la recherche en cancérologie*. Montrouge : J. Libbey Eurotext, 2011. ISBN : 978-2-7420-0774-5.
- [14] Michael H. KUTNER. *Applied Linear Statistical Models*. McGraw-Hill Irwin, 2005. 1396 p. ISBN : 978-0-07-112221-4. Google Books : 0xqCAAAACAAJ.
- [15] Catherine LEGRAND. « LSTAT2210 : Modèles Linéaires Avancés » (ISBA, UCL). 2019-20.
- [16] Gary H LYMAN et Nicole M KUDERER. « The Strengths and Limitations of Meta-Analyses Based on Aggregate Data ». In : *BMC Medical Research Methodology* 5 (25 avr. 2005), p. 14. ISSN : 1471-2288. DOI : 10.1186/1471-2288-5-14. pmid : 15850485. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1097735/> (visité le 04/08/2020).
- [17] Patrick MAISON. « La méta-analyse sur données résumées ». In : *Recherche en soins infirmiers* N° 101.2 (2010), p. 18-24. ISSN : 0297-2964. URL : <https://www.cairn.info/revue-recherche-en-soins-infirmiers-2010-2-page-18.htm> (visité le 03/08/2020).
- [18] S. MICHIELS et al. « Random Effects Survival Models Gave a Better Understanding of Heterogeneity in Individual Patient Data Meta-Analyses ». In : *Journal of Clinical Epidemiology* 58.3 (mar. 2005), p. 238-245. ISSN : 0895-4356. DOI : 10.1016/j.jclinepi.2004.08.013. pmid : 15718112.

- [19] David MOHER et al. « Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement ». In : *Systematic Reviews* 4.1 (1^{er} jan. 2015), p. 1. ISSN : 2046-4053. DOI : 10.1186/2046-4053-4-1. URL : <https://doi.org/10.1186/2046-4053-4-1> (visité le 28/07/2020).
- [20] Jean-Pierre PIGNON et Catherine HILL. « Meta-Analyses of Randomised Clinical Trials in Oncology ». In : *The Lancet Oncology* 2.8 (1^{er} août 2001), p. 475-482. ISSN : 1470-2045, 1474-5488. DOI : 10.1016/S1470-2045(01)00453-3. pmid : 11905723. URL : [https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045\(01\)00453-3/abstract](https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(01)00453-3/abstract) (visité le 04/08/2020).
- [21] Annie ROBERT. *Statistiques Des Essais Cliniques STAT 2330*. 2008-2009.
- [22] ROBUST@Leuven. URL : <https://wis.kuleuven.be/statdatascience/robust/home> (visité le 04/08/2020).
- [23] D. L. SACKETT et al. « Evidence Based Medicine : What It Is and What It Isn't. » In : *BMJ : British Medical Journal* 312.7023 (13 jan. 1996), p. 71-72. ISSN : 0959-8138. pmid : 8555924. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2349778/> (visité le 20/08/2020).
- [24] SAS Help Center : Residuals and Influence Diagnostics. URL : https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_mixed_details24.htm&docsetVersion=15.1&locale=en#statug.mixed.mixeddetailinf (visité le 24/07/2020).
- [25] Catrin Tudur SMITH et al. « Individual Participant Data Meta-analyses Compared with Meta-analyses Based on Aggregate Data ». In : *Cochrane Database of Systematic Reviews* 9 (2016). ISSN : 1465-1858. DOI : 10.1002/14651858.MR000007.pub3. URL : <https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.MR000007.pub3/abstract/fr> (visité le 03/08/2020).
- [26] L. A. STEWART et M. K. PARMAR. « Meta-Analysis of the Literature or of Individual Patient Data : Is There a Difference ? » In : *Lancet (London, England)* 341.8842 (13 fév. 1993), p. 418-422. ISSN : 0140-6736. DOI : 10.1016/0140-6736(93)93004-k. pmid : 8094183.
- [27] Studentized Residual. In : *Wikipedia*. 26 avr. 2020. URL : https://en.wikipedia.org/w/index.php?title=Studentized_residual&oldid=953251003 (visité le 16/08/2020).

- [28] R. THOMPSON. « A Note on Restricted Maximum Likelihood Estimation with an Alternative Outlier Model ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 47.1 (1985), p. 53-55. ISSN : 2517-6161. DOI : 10.1111/j.2517-6161.1985.tb01329.x. URL : <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1985.tb01329.x> (visité le 10/08/2020).
- [29] David VENET et al. « A Statistical Approach to Central Monitoring of Data Quality in Clinical Trials : » in : *Clinical Trials* (8 juin 2012). DOI : 10.1177/1740774512447898. URL : <https://journals.sagepub.com/doi/10.1177/1740774512447898> (visité le 11/06/2020).
- [30] Wolfgang VIECHTBAUER et Mike W.-L. CHEUNG. « Outlier and Influence Diagnostics for Meta-Analysis ». In : *Research Synthesis Methods* 1.2 (2010), p. 112-125. ISSN : 1759-2887. DOI : 10.1002/jrsm.11. URL : <https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.11> (visité le 16/03/2020).

