ANNEE 2017 THESE : 2017 – TOU 3 – 4066

CARACTERISATION DES FACTEURS DE RISQUE A PARTIR DE DONNEES ISSUES D'UNE SURVEILLANCE IMPARFAITE : COMPARAISON DES MODELES DE REGRESSION LOGISTIQUE ET DE POISSON ENFLES EN ZERO

THESE
pour obtenir le grade de
DOCTEUR VETERINAIRE

DIPLOME D'ETAT

présentée et soutenue publiquement
devant l'Université Paul-Sabatier de Toulouse

par

COMBELLES, Lisa
Née, le 22/01/1992 à ALBI (81)

Directeur de thèse : M. Fabien CORBIERE

JURY

PRESIDENT :	
M. Alain GRAND	Professeur à l'Université Paul-Sabatier de TOULOUSE
ASSESSEURS :	
M. Fabien CORBIERE	Maître de Conférences à l'Ecole Nationale Vétérinaire de TOULOUSE
Mme Agnès WARET-SZKUTA	Maître de Conférences à l'Ecole Nationale Vétérinaire de TOULOUSE
MEMBRE INVITE :	
M. Timothée VERGNE	Post-doctorant à l'Institut de Recherche pour le Développement

ECOLE NATIONALE VETERINAIRE TOULOUSE

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE L'AGRICULTURE ET DE L'ALIMENTATION

Ministère de l'Agriculture de l'Alimentation
ECOLE NATIONALE VETERINAIRE DE TOULOUSE

Directrice : **Madame Isabelle CHMITELIN**

PROFESSEURS CLASSE EXCEPTIONNELLE

M. **AUTEFAGE André**, *Pathologie chirurgicale*
Mme **CLAUW Martine**, *Pharmacie-Toxicologie*
M. **CONCORDET Didier**, *Mathématiques, Statistiques, Modélisation*
M **DELVERDIER Maxence,** *Anatomie Pathologique*
M. **ENJALBERT Francis**, *Alimentation*
M. **FRANC Michel**, *Parasitologie et Maladies parasitaires*
M. **MILON Alain**, *Microbiologie moléculaire*
M. **PETIT Claude**, *Pharmacie et Toxicologie*
M. **SCHELCHER François,** *Pathologie médicale du Bétail et des Animaux de Basse-cour*

PROFESSEURS 1° CLASSE

M. **BERTAGNOLI Stéphane**, *Pathologie infectieuse*
M. **BERTHELOT Xavier**, *Pathologie de la Reproduction*
M. **BOUSQUET-MELOU Alain**, *Physiologie et Thérapeutique*
M. **BRUGERE Hubert**, *Hygiène et Industrie des aliments d'Origine animale*
Mme **CHASTANT-MAILLARD Sylvie**, *Pathologie de la Reproduction*
M. **DUCOS Alain**, *Zootechnie*
M. **FOUCRAS Gilles**, *Pathologie des ruminants*
Mme **GAYRARD-TROY Véronique**, *Physiologie de la Reproduction, Endocrinologie*
Mme **HAGEN-PICARD Nicole**, *Pathologie de la reproduction*
M. **JACQUIET Philippe**, *Parasitologie et Maladies Parasitaires*
M. **LEFEBVRE Hervé**, *Physiologie et Thérapeutique*
M. **LIGNEREUX Yves**, *Anatomie*
M. **MEYER Gilles,** *Pathologie des ruminants*
M. **PICAVET Dominique**, *Pathologie infectieuse*
M. **SANS Pierre**, *Productions animales*
Mme **TRUMEL Catherine**, *Biologie Médicale Animale et Comparée*

PROFESSEURS 2° CLASSE

M. **BAILLY Jean-Denis**, *Hygiène et Industrie des aliments*
Mme **BOURGES-ABELLA Nathalie**, *Histologie, Anatomie pathologique*
Mme **CADIERGUES Marie-Christine**, *Dermatologie Vétérinaire*
M. **GUERRE Philippe**, *Pharmacie et Toxicologie*
M **GUERIN Jean-Luc**, *Aviculture et pathologie aviaire*
Mme **LACROUX Caroline**, *Anatomie Pathologique, animaux d'élevage*
Mme **LETRON-RAYMOND Isabelle**, *Anatomie pathologique*
M. **MAILLARD Renaud,** *Pathologie des Ruminants*

Université Fédérale Toulouse Midi-Pyrénées

Mise à jour au 06/09/2017

PROFESSEURS CERTIFIES DE L'ENSEIGNEMENT AGRICOLE

Mme **MICHAUD Françoise**, *Professeur d'Anglais*
M **SEVERAC Benoît**, *Professeur d'Anglais*

MAITRES DE CONFERENCES HORS CLASSE

M. **BERGONIER Dominique**, *Pathologie de la Reproduction*
Mme **BOULLIER Séverine**, *Immunologie générale et médicale*
Mme **DIQUELOU Armelle**, *Pathologie médicale des Equidés et des Carnivores*
M. **DOSSIN Olivier**, *Pathologie médicale des Equidés et des Carnivores*
M. **JOUGLAR Jean-Yves**, *Pathologie médicale du Bétail et des Animaux de Basse-cour*
M. **LYAZRHI Faouzi**, *Statistiques biologiques et Mathématiques*
M. **MATHON Didier**, *Pathologie chirurgicale*
Mme **MEYNADIER Annabelle,** *Alimentation*
M. **MOGICATO Giovanni**, *Anatomie, Imagerie médicale*
Mme **PRIYMENKO Nathalie**, *Alimentation*
M. **VERWAERDE Patrick**, *Anesthésie, Réanimation*

MAITRES DE CONFERENCES (classe normale)

M. **ASIMUS Erik**, *Pathologie chirurgicale*
Mme **BENNIS-BRET Lydie**, *Physique et Chimie biologiques et médicales*
Mme **BIBBAL Delphine**, *Hygiène et Industrie des Denrées alimentaires d'Origine animale*
Mme **BOUCLAINVILLE-CAMUS Christelle**, *Biologie cellulaire et moléculaire*
Mme **BOUHSIRA Emilie**, *Parasitologie, maladies parasitaires*
M. **CONCHOU Fabrice**, *Imagerie médicale*
M. **CORBIERE Fabien**, *Pathologie des ruminants*
M. **CUEVAS RAMOS Gabriel**, *Chirurgie Equine*
Mme **DANIELS Hélène**, *Microbiologie-Pathologie infectieuse*
Mme **DEVIERS Alexandra**, *Anatomie-Imagerie*
M. **DOUET Jean-Yves**, *Ophtalmologie vétérinaire et comparée*
Mme **FERRAN Aude**, *Physiologie*
M. **JAEG Jean-Philippe**, *Pharmacie et Toxicologie*
Mme **LAVOUE Rachel**, *Médecine Interne*
M. **LE LOC'H Guillaume**, *Médecine zoologique et santé de la faune sauvage*
M. **LIENARD Emmanuel**, *Parasitologie et maladies parasitaires*
Mme **MEYNAUD-COLLARD Patricia**, *Pathologie Chirurgicale*
Mme **MILA Hanna**, *Elevage des carnivores domestiques*
M. **NOUVEL Laurent,** *Pathologie de la reproduction (en disponibilité)*
Mme **PALIERNE Sophie**, *Chirurgie des animaux de compagnie*
Mme **PAUL Mathilde**, *Epidémiologie, gestion de la santé des élevages avicoles et porcins*
Mme **PRADIER Sophie**, *Médecine interne des équidés*
M. **RABOISSON Didier**, *Productions animales (ruminants)*
M. **VOLMER Romain**, *Microbiologie et Infectiologie*
Mme **WARET-SZKUTA Agnès**, *Production et pathologie porcine*

ASSISTANTS D'ENSEIGNEMENT ET DE RECHERCHE CONTRACTUELS

Mme **COSTES Laura**, *Hygiène et industrie des aliments*
M. **GAIDE Nicolas**, *Histologie, Anatomie Pathologique*
Mme **LALLEMAND Elodie**, *Chirurgie des Equidés*
Mme **SABY-CHABAN Claire**, *Gestion de la santé des troupeaux bovins*

REMERCIEMENTS

A Monsieur le Professeur Alain Grand
Professeur des Universités
Praticien hospitalier
Santé Publique
Qui nous a fait l'honneur d'accepter la présidence de notre jury de thèse.
Hommages respectueux.

A Monsieur le Docteur Fabien Corbière
Maître de Conférences de l'Ecole Nationale Vétérinaire de Toulouse
Pathologie des Ruminants
Qui nous a fait l'honneur de diriger cette thèse.
Très sincères remerciements.

A Madame le Docteur Agnès Waret-Szkuta
Maître de Conférences de l'Ecole Nationale Vétérinaire de Toulouse
Production et pathologie porcines
Qui nous a fait l'honneur de participer à notre jury de thèse.
Très sincères remerciements.

A Monsieur le Docteur Timothée Vergne
Post-doctorant à l'Institut de Recherche pour le Développement
Epidémiologie
Qui nous a fait l'honneur de proposer et d'encadrer cette thèse.
Très sincères remerciements.

A Monsieur le Docteur Didier Calavas
Chef de l'Unité d'Epidémiologie du laboratoire de Lyon de l'ANSES

A Madame le Docteur Viviane Hénaux
Membre de l'Unité d'Epidémiologie du laboratoire de Lyon de l'ANSES

A Madame le Docteur Anne Bronner
Cheffe du projet « Amélioration de la surveillance en santé animale et santé végétale »
Coordinatrice adjointe de la Plateforme ESA
Rédactrice en chef adjointe du Bulletin épidémiologique ANSES-DGAl
SASPP / DGAl

Qui nous ont fait l'honneur de nous donner accès à la base de données d'avortements bovins et qui ont contribué à la relecture de la dernière partie de ce manuscrit.
Très sincères remerciements.

TABLE DES MATIERES

PARAMETRES UTILISES POUR LES SIMULATIONS ET ANALYSES 17

INTRODUCTION GENERALE 19

I- Facteurs de risque et importance de leur identification 19

II- Méthodes d'identification des facteurs de risque de maladies animales 19

1) Etudes de cohortes et études cas-témoins 19

2) Surveillance des maladies animales 20

III- Limites de la surveillance : sous détection et sous déclaration des maladies 23

1) Limites de la surveillance passive dans la collecte des informations 23

a) Limites économiques 23

b) Limites psychologiques et sociales 23

c) Limites à l'échelle mondiale 25

2) Limites de la surveillance active dans la collecte des informations 25

3) Hétérogénéité dans la collecte des informations 25

IV- Objectif de l'étude 26

Partie 1 : METHODES D'ETUDE DES FACTEURS DE RISQUE : REVUE DE LA LITTERATURE 27

I- Etat des lieux sur les modèles utilisés pour l'identification des facteurs de risque de maladies animales à partir de données de surveillance 27

1) Objectifs 27

2) Matériels et méthodes 27

3) Résultats 28

4) Discussion 31

II- Problématique de la sous détection et de la sous déclaration d'une maladie : modélisation de données incomplètes 32

Partie 2 : IMPACT DE LA DETECTION IMPARFAITE SUR L'ETUDE DES FACTEURS DE RISQUE : APPROCHE PAR SIMULATIONS 35

I- Etude de l'influence d'une sensibilité imparfaite sur l'estimation du risque d'une maladie 35

1) Introduction et contexte 35

2) Matériels et méthodes 35

a) Distribution des variables d'intérêt 35

b) Détection de la maladie dans les unités épidémiologiques 36

c) Plan de simulation 38

d) Analyse des données obtenues 39

e) Simulations et analyses : logiciel utilisé 42

3) Résultats 43

a) Impact sur les modèles logistiques d'une sensibilité de détection imparfaite mais homogène 43

b) Impact sur les modèles logistiques d'une sensibilité de détection imparfaite et hétérogène 46

c) Impact sur les modèles logistiques d'une sensibilité de détection imparfaite lorsque les facteurs de risque et de confusion sont identiques 52

d) Bénéfices potentiels à utiliser un modèle de Poisson enflé en zéro si la détection est imparfaite voire hétérogène 55

4) Discussion 62

II- Etude de l'influence d'une spécificité imparfaite sur l'estimation du risque d'une maladie 66
1) Introduction et contexte 66
2) Matériels et méthodes 66
a) Distribution des variables d'intérêts 66
b) Détection de la maladie dans les unités épidémiologiques 67
c) Plan de simulation 68
d) Analyse des données obtenues 68
e) Simulations et analyses : logiciel utilisé 69
3) Résultats : Impact sur les modèles logistiques d'une spécificité de détection imparfaite mais homogène 70
4) Discussion 72
III- Bilan et perspectives 74
Partie 3 : APPLICATION A DES DONNEES REELLES : IDENTIFICATION DE FACTEURS DE RISQUE DES AVORTEMENTS BOVINS EN FRANCE METROPOLITAINE (2010-2011) 77
I- Objectif 77
II- Introduction et contexte 77
III- Matériels et méthodes 78
1) Source des données et population étudiée 78
2) Modélisation et analyse des données 78
a) Variables des modèles 78
b) Sélection des variables et construction des modèles 79
c) Validation des modèles 79
d) Logiciels utilisés 80
IV- Résultats 80
1) Description des données 80
2) Inférence par le modèle logistique 82
3) Inférence par le modèle de Poisson enflé en zéro 83
4) Validation des modèles 84
V- Discussion 85
VI- Conclusion 86
CONCLUSION GENERALE 89
BIBLIOGRAPHIE 91
ANNEXES 97
Annexe 1 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade, valeurs de prev.X0 et valeurs réelles d'OR(X) 97
Annexe 2 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection, valeurs de prev.X0 et valeurs réelles d'OR(X) 98

Annexe 3 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection, valeurs de prev.X0 et valeurs réelles d'OR(X)..... 99

Annexe 4 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4..... 100

Annexe 5 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade, valeurs de prev.X0 et valeurs réelles d'OR(X)..... 101

Annexe 6 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la spécificité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X)..... 102

TABLE DES ILLUSTRATIONS

Figures

Figure 1 : Canaux potentiels de collecte de données pour la surveillance et la maîtrise des événements sanitaires en santé animale, d'après Doherr et al. (2001) 20

Figure 2 : Surveillance passive - Etapes de détection (et déclaration), d'après Doherr et al. (2001) 22

Figure 3 : Organigramme résumant le processus de sélection des articles de la revue 28

Figure 4 : Données enflées en zéro générées par les systèmes de surveillance, d'après Vergne et al. (2015) 32

Figure 5 : Distribution des facteurs X et Y, et leur influence sur la maladie étudiée 36

Figure 6 : Schéma de simulation des données relatives à la maladie étudiée 36

Figure 7 : Variables créées lors des simulations et variables réponses des deux modèles étudiés 40

Figure 8 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs réelles d'OR(X) lorsque prev.X0=0,6 et M=4 43

Figure 9 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4 44

Figure 10 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade lorsque prev.X0=0,5 et OR(X)=5 45

Figure 11 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4 46

Figure 12 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4 *Les barres à droite des graphes indiquent les valeurs médianes de biais($OR(X)_{logist}$) pour 300 simulations.* 47

Figure 13 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection et valeurs réelles d'OR(X) lorsque prev.X0=0,2 48

Figure 14 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4 49

Figure 15 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4 *Les barres à droite des graphes indiquent la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique (sur 300 simulations).* 49

Figure 16 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection et valeurs réelles d'OR(X) lorsque prev.X0=0,2 .. 50

Figure 17 : Evolution de la probabilité que le facteur X soit identifié comme étant un facteur de risque par un modèle logistique en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4 52

Figure 18 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4 53

Figure 19 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs réelles d'OR(X) lorsque prev.X0=0,6 et M=4 55

Figure 20 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade lorsque prev.X0=0,5 et OR(X)=5 56

Figure 21 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par la partie « comptage » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4 59

Figure 22 : Evolution de la probabilité que le facteur X soit identifié comme étant un facteur de risque par la partie « comptage » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4 60

Figure 23 : Schéma de simulation des données lorsque la spécificité de la détection est imparfaite 67

Figure 24 : Evolution de la probabilité que le facteur X soit identifié comme étant un facteur de risque par un modèle logistique en fonction de la spécificité de détection et pour différentes valeurs réelles d'OR(X) lorsque prev.X0=0,2 70

Figure 25 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la spécificité de détection et pour différentes valeurs réelles d'OR(X) lorsque prev.X0=0,2 71

Figure 26 : Distribution du nombre d'élevages de bovins par département français (dans lesquels au moins un élevage a déclaré au moins un avortement entre le 1er août 2010 et le 31 juillet 2011), d'après les informations disponibles dans le jeu de données (Bronner et al., 2013) 80

Figure 27 : Effet de l'interaction entre le type de production et la taille d'élevage sur la valeur de l'odds ratio estimée par le modèle logistique 82

Figure 28 : Effet de l'interaction entre le type de production et la taille d'élevage sur la valeur de l'odds ratio estimée par la partie « logistique » du modèle de Poisson enflé en zéro 84

Figure 29 : Courbe ROC du modèle logistique final *La ligne en pointillée représente la diagonale (Sensibilité=1-Spécificité).* 84

Figure 30 : Probabilité estimée par le modèle logistique final qu'au moins un avortement soit déclaré dans les élevages, selon que les élevages aient effectivement déclaré ou non au moins un avortement *La courbe représente la courbe de densité de probabilité, le point blanc représente la médiane, et le trait noir épais représente l'intervalle interquartile.* 85

Annexe 1 - Figure 31 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade, valeurs de prev.X0 et valeurs réelles d'OR(X) 97

Annexe 2 - Figure 32 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection, valeurs de prev.X0 et valeurs réelles d'OR(X) 98

Annexe 3 - Figure 33 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection, valeurs de prev.X0 et valeurs réelles d'OR(X) 99

Annexe 4 - Figure 34 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4 100

Annexe 5 - Figure 35 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade, valeurs de prev.X0 et valeurs réelles d'OR(X) 101

Annexe 6 - Figure 36 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la spécificité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) 102

Tableaux

Tableau 1 : Etudes identifiant des facteurs de risque pour différentes maladies animales 29

Tableau 2 : Paramètres fixés dans les différentes séries de simulations 38

Tableau 3 : Valeurs de prev.X0 et d'OR(X), et valeurs des prev.X1 qui correspondent *Le contenu de chaque case du tableau est le résultat du calcul de la valeur de prev.X1 qui correspond à chacun des couples prev.X0 et OR(X) envisagés* 39

Tableau 4 : Paramètres fixés dans les différentes séries de simulations lorsque la spécificité est imparfaite 68

Tableau 5 : Interprétation de l'aire sous la courbe d'une courbe ROC, d'après Rakotomalala (2011) 80

Tableau 6 : Taille et type de production des 99 996 élevages du jeu de données 81

Tableau 7 : Répartition du nombre d'avortements déclarés par élevage 81

Tableau 8 : Répartition par catégorie des 99 996 élevages selon qu'ils aient déclaré ou non un avortement 81

Tableau 9 : Modèle logistique final $(OR_{interaction(A\&B)}=OR_A.OR_B.OR_{A:B})$ 82

Tableau 10 : Modèle de Poisson enflé en zéro final $(OR_{interaction(A\&B)}=OR_A.OR_B.OR_{A:B})$ 83

Equations

Équation 1 : Probabilité de présence de la maladie dans une unité épidémiologique 37

Équation 2 : Loi de Poisson tronquée en zéro de paramètre « M » 37

Équation 3 : Loi binomiale de paramètres « m » et « $sensib_i$ » 38

Équation 4 : Equation définissant le modèle logistique 40

Équation 5 : Calcul du biais relatif de l'odds ratio associé au facteur X (régression logistique) 40

Équation 6 : Equation définissant le modèle de Poisson enflé en zéro 41

Équation 7 : Calcul du biais relatif de l'odds ratio associé au facteur X (modèle enflé en zéro) 42

Équation 8 : Loi binomiale de paramètres « U » et « 1-spécificité » 68

PARAMETRES UTILISES POUR LES SIMULATIONS ET ANALYSES

N : Nombre d'unités épidémiologiques

U : Nombre d'unités élémentaires dans les unités épidémiologiques *(utilisé pour l'étude du défaut de spécificité)*

pX : Probabilité de présence du facteur de risque X dans une unité épidémiologique

pY : Probabilité de présence du facteur de confusion Y dans une unité épidémiologique

M : Nombre moyen d'unités élémentaires malades dans une unité épidémiologique malade (équivalent à la prévalence intra-unité épidémiologique)

prev : Probabilité de présence de la maladie dans une unité épidémiologique

prev.X0 : Probabilité de présence de la maladie dans une unité épidémiologique pour laquelle le facteur de risque X est absent

prev.X1 : Probabilité de présence de la maladie dans une unité épidémiologique pour laquelle le facteur de risque X est présent

OR(X) : Odds ratio réel de la présence de la maladie associé au facteur de risque X

$OR(X)_{logist}$: Odds ratio de la présence de la maladie associé au facteur de risque X et estimé par le modèle logistique

biais($OR(X)_{logist}$) : Biais relatif entre OR(X) et $OR(X)_{logist}$ (Équation 5)

$OR(X)_{logit}$: Odds ratio de la présence de la maladie associé au facteur de risque X et estimé par la partie « logistique » du modèle de Poisson enflé en zéro

biais($OR(X)_{logit}$) : Biais relatif entre OR(X) et $OR(X)_{logit}$ (Équation 7)

sensib : Probabilité de détecter chacune des unités élémentaires malades dans une unité épidémiologique malade (sensibilité de détection)

sensib.Y0 : Sensibilité de détection d'une unité élémentaire malade dans une unité épidémiologique malade pour laquelle le facteur de confusion Y est absent

sensib.Y1 : Sensibilité de détection d'une unité élémentaire malade dans une unité épidémiologique malade pour laquelle le facteur de confusion Y est présent

sensib.X0 : Sensibilité de détection d'une unité élémentaire malade dans une unité épidémiologique malade pour laquelle le facteur X est absent (lorsque X est à la fois facteur de risque de présence de la maladie et facteur de confusion de détection de la maladie)

sensib.X1 : Sensibilité de détection d'une unité élémentaire malade dans une unité épidémiologique malade pour laquelle le facteur X est présent (lorsque X est à la fois facteur de risque de présence de la maladie et facteur de confusion de détection de la maladie)

P_i : Variable représentant la présence/absence de la maladie dans une unité épidémiologique i

Mal : Variable représentant le nombre d'unités élémentaires réellement malades dans une unité épidémiologique malade

$Détect_i$: Variable représentant le nombre d'unités élémentaires détectées dans une unité épidémiologique i

INTRODUCTION GENERALE

I- Facteurs de risque et importance de leur identification

Un facteur de risque d'une maladie donnée est un facteur qui augmente la fréquence de cette maladie (Dohoo et al., 2009). Par exemple, une forte densité de canards domestiques ou la proximité d'un point d'eau sont des facteurs de risque de l'influenza aviaire hautement pathogène H5N1 (Gilbert et Pfeiffer, 2012).

Les intérêts de l'identification des facteurs de risque sont multiples. Cela peut permettre de générer des hypothèses sur la cause des maladies, comme cela a été le cas avec une étude qui a démontré l'implication des farines de viande et d'os dans l'apparition de l'encéphalopathie spongiforme bovine (Wilesmith et al., 1992). La connaissance des facteurs de risque peut aussi permettre de comprendre la distribution et la dynamique de la maladie étudiée. La surveillance pourra être ensuite accrue dans les zones où ces facteurs sont présents, afin de pouvoir intervenir de manière plus efficace (détection précoce, protection des populations à risque). La quantification de l'association entre les facteurs de risque et la maladie étudiée permet en outre d'orienter les choix stratégiques d'un point de vue économique (Dohoo et al., 2009).

Il y a cependant des limites à prendre en compte dans les études visant à identifier des facteurs de risque. Tous les facteurs ne sont pas pris en compte dans les modèles statistiques utilisés pour évaluer les facteurs de risque potentiels, à la fois pour des raisons pratiques et afin de respecter le principe de parcimonie (Dohoo et al., 2009). De plus, les données relatives à certains facteurs ne sont pas toujours disponibles, difficulté qui peut parfois être contournée en étudiant un autre facteur approchant (dit aussi « proxy ») : Vergne et al. (2016) ont par exemple utilisé la densité de forêt afin d'approcher la densité de sangliers dans leur étude sur la peste porcine africaine en Russie.

Les limites dans l'identification des facteurs de risque peuvent aussi provenir des données utilisées et des méthodes statistiques employées pour leur analyse.

II- Méthodes d'identification des facteurs de risque de maladies animales

1) Etudes de cohortes et études cas-témoins

L'identification de facteurs de risque repose classiquement sur deux schémas d'étude : les études de cohorte et les études cas-témoins. Les études de cohortes consistent à évaluer la fréquence de la maladie étudiée dans deux groupes d'animaux ou deux groupes d'élevages identiques à l'exception de l'exposition au facteur étudié : un groupe est exposé au facteur tandis que l'autre groupe ne l'est pas (Dohoo et al., 2009). Les études cas-témoins quant à elles consistent à évaluer la fréquence d'exposition au facteur étudié d'une part dans un groupe d'animaux (ou d'élevages) ayant développé la maladie étudiée et d'autre part dans un groupe d'animaux (ou d'élevages) n'ayant pas développé cette même maladie (Dohoo et al., 2009).

Ces études sont généralement contrôlées (contrôle de la taille des groupes étudiés, contrôle de l'exposition au facteur étudié dans les études de cohortes). Elles permettent de mesurer l'effet d'un facteur sur l'apparition d'une maladie, via le calcul du risque relatif ou de l'odds ratio, et ainsi déterminer s'il s'agit d'un facteur de risque (aggravant) ou d'un facteur protecteur. Des biais peuvent intervenir dans la conception de ces études, lors du choix de la taille des groupes étudiés ou de la durée des études par exemple.

D'autres types d'études visant à identifier des facteurs de risque se basent sur des données existantes de surveillance des maladies animales.

2) Surveillance des maladies animales

La surveillance des maladies animales consiste en la collection et l'interprétation d'informations concernant la population animale et la maladie étudiées, ainsi qu'en la diffusion de ces informations aux responsables des mesures de maîtrise et de prévention (Doherr et Audigé, 2001). L'objectif de la surveillance est ainsi de fournir les informations nécessaires à la mise en place des moyens de lutte, que sont la police sanitaire et la vaccination (Vergne, 2012). La surveillance épidémiologique peut en outre être divisée en deux approches : l'épidémiovigilance et l'épidémiosurveillance (Vergne, 2012). L'épidémio-vigilance vise à détecter précocement les maladies émergentes (maladies habituellement absentes du territoire considéré ou apparition d'un nouvel agent pathogène), tandis que l'épidémiosurveillance vise à suivre les maladies déjà connues sur un territoire donné.

Les informations collectées grâce à la surveillance peuvent ensuite être utilisées pour l'identification de facteurs de risque. Ces informations concernant les maladies animales sont obtenus par divers canaux de collecte de données (Figure 1).

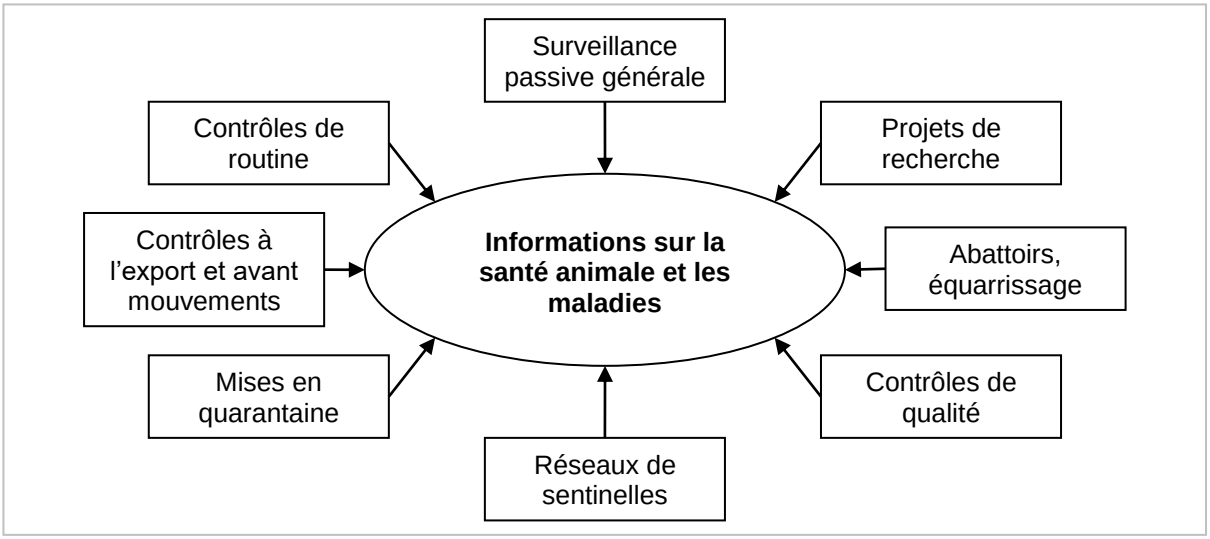


Figure 1 : Canaux potentiels de collecte de données pour la surveillance et la maîtrise des événements sanitaires en santé animale, d'après Doherr et al. (2001)

Doherr et al. (2001) classent ces canaux de collecte de données en différentes catégories. Ils considèrent en premier lieu les sources primaires, que sont les propriétaires d'animaux (les éleveurs, en première ligne pour détecter une anomalie sur leur troupeau), les vétérinaires

praticiens, les laboratoires d'analyses vétérinaires, les différents secteurs de l'élevage (jusqu'à l'abattage des animaux de rente), les instituts de recherche, les exportateurs et les importateurs de bétail. Les données secondaires sont les données disponibles sur Internet et mises à disposition par les différents systèmes de surveillance, par exemple le système de l'OIE (organisation mondiale de la santé animale) (OIE - World Organisation for Animal Health, 2012) ou celui de la FAO (organisation des Nations unies pour l'alimentation et l'agriculture) (FAO - Food and Agriculture Organization of the United Nations, 2014).

Par ailleurs, les données concernant les maladies sont collectées *via* deux modes principaux : la surveillance passive et la surveillance active (Doherr et Audigé, 2001).

La **surveillance passive** repose sur la déclaration spontanée de cas suspects aux autorités compétentes. Dans le cadre des dangers sanitaires de première catégorie et des dangers sanitaires de deuxième catégorie faisant l'objet d'une réglementation, toute suspicion clinique doit faire l'objet d'une déclaration par l'éleveur au vétérinaire sanitaire et à la Direction Départementale de la Protection des Populations (Ganière, 2017), puis les mesures applicables à chaque maladie sont mises en œuvre afin d'une part de limiter le risque de diffusion de la maladie et d'autre part de confirmer la suspicion clinique. Certaines maladies ne font pas l'objet d'une déclaration obligatoire dans le cadre des dangers sanitaires, mais doivent être notifiées à l'OIE (OIE - World Organisation for Animal Health, 2017). C'est par exemple le cas de la babésiose bovine ou de la paratuberculose. Ces maladies sont alors déclarées après que le vétérinaire a réalisé la démarche lui permettant d'établir son diagnostic.

La **surveillance active** est une procédure formelle et programmée, le choix des animaux à tester se fait par les autorités compétentes. Elle repose par exemple sur des tests sérologiques, comme c'est le cas en France pour la surveillance de la brucellose bovine (Ganière et Laaberki, 2017), sur la recherche systématique de lésions en abattoir, comme pour la surveillance de la tuberculose bovine (Bénet et Praud, 2016), ou sur le prélèvement systématique d'organes en abattoir pour des analyses de laboratoire, comme pour la surveillance de l'encéphalopathie spongiforme bovine (Peroz et Ganière, 2017). Dans le cadre de ces maladies, qui sont des dangers sanitaires et donc soumis à une réglementation, les animaux positifs aux tests de routine sont ensuite confirmés par différentes méthodes afin et d'être le plus spécifique possible et d'exclure les faux positifs. En revanche, si le test est réalisé dans une surveillance de routine (par exemple, prophylaxie annuelle pour la brucellose bovine en troupeau allaitant), les animaux négatifs sont directement considérés comme tels. Dans la surveillance active, la sensibilité de la détection dépend donc essentiellement de la sensibilité des tests utilisés, qui n'est jamais parfaite à 100%.

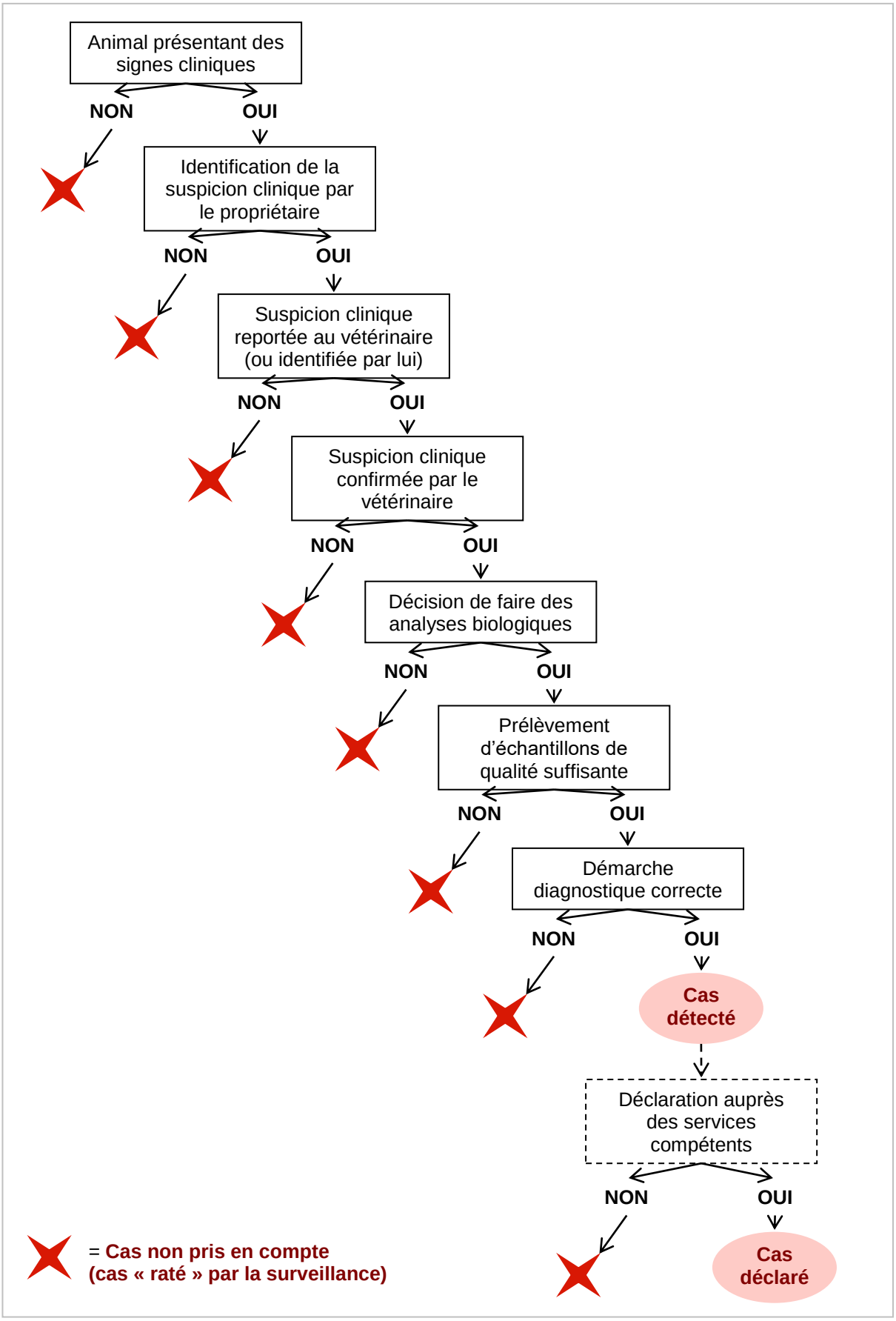


Figure 2 : Surveillance passive - Etapes de détection (et déclaration), d'après Doherr et al. (2001)

III- Limites de la surveillance : sous détection et sous déclaration des maladies

1) Limites de la surveillance passive dans la collecte des informations

La surveillance passive fait appel à diverses étapes, schématisées dans la Figure 2. Chacune de ces étapes peut conduire à une absence de détection, puis à la non déclaration, d'un animal malade. La base du processus est la détection des signes cliniques par l'éleveur, ce qui sous-entend que celui-ci doit passer suffisamment de temps avec ses animaux et être en mesure de détecter une anomalie. L'éleveur doit ensuite faire appel à son vétérinaire qui devra par exemple prendre la décision soit de déclarer une suspicion de danger sanitaire à déclaration obligatoire, soit de confirmer son diagnostic avant notification à l'OIE d'un trouble sanitaire non soumis à la réglementation des dangers sanitaires. La détection dans le cadre de la surveillance passive faisant intervenir de nombreuses étapes est donc nécessairement imparfaite.

Ainsi, dans le cadre de la surveillance passive, la détection puis la déclaration reposent sur la coopération et la confiance entre tous les acteurs (Doherr et Audigé, 2001). La vigilance des éleveurs ou des vétérinaires dépend de leur connaissance de la maladie ainsi que de leur motivation personnelle à vouloir déclarer (Hopp et al., 2007). Cette motivation à déclarer dépend de la motivation financière, mais pas seulement.

a) <u>Limites économiques</u>

Un des premiers facteurs motivant la déclaration est économique. Wineland et al. (1998) ont analysé la déclaration des cas de tremblante ovine aux Etats-Unis entre 1947 et 1992. Lorsque la déclaration de tremblante conduisait à un abattage total du troupeau, les éleveurs étaient réticents à déclarer un cas isolé dans leur troupeau. En revanche, la compensation financière semble inciter les déclarations, puisque chaque augmentation des indemnités s'est traduite par une augmentation du nombre de déclarations de tremblante ovine. Cette tendance est reprise par Kuchler et al. (2000) qui signalent que l'augmentation de l'indemnisation par animal malade encourage les éleveurs à déclarer les cas de tremblante. Ils précisent que ces programmes de compensation doivent inciter les éleveurs à déclarer la maladie plutôt qu'à envoyer leurs animaux à l'abattoir sans déclarer les anomalies qu'ils auraient pu détecter. Barnes et al. (2015) soulignent qu'il faut que la compensation soit suffisante pour favoriser une déclaration précoce, mais pas trop élevée pour que les éleveurs aient un bénéfice à chercher à prévenir la maladie (vaccination, mesures de biosécurité). De plus, il semble que les menaces d'amende en cas de non déclaration soient plus efficaces que les indemnisations en cas de déclaration (Barnes et al., 2015).

b) <u>Limites psychologiques et sociales</u>

Il existe aussi des facteurs psychosociaux qui influencent la détection et la déclaration. Elbers et al. ont investigué ces facteurs en prenant l'exemple de la détection précoce des cas d'influenza aviaire (Elbers, Gorgievski, et al., 2010) et de peste porcine classique (Elbers, Gorgievski-Duijvesteijn, et al., 2010). Il se dégage de ces deux études six axes principaux qui

vont influencer la détection ou la déclaration des cas. Le premier axe concerne la connaissance de la maladie et l'attribution des signes cliniques à une maladie à déclaration obligatoire. La difficulté est que les signes observés ne sont généralement pas spécifiques et que les éleveurs vont avoir tendance à chercher une cause autre qu'une maladie réglementée. La détection de l'influenza aviaire repose par exemple sur une surveillance syndromique, l'éleveur commençant à s'alerter seulement à partir d'un certain seuil de mortalité. Par ailleurs, lorsqu'il s'agit de maladies émergentes, le vétérinaire ne placera peut-être pas ces maladies dans son diagnostic différentiel. Le deuxième axe concerne les sentiments des éleveurs vis-à-vis de l'apparition de la maladie dans leur élevage et de ce que les autres éleveurs peuvent penser, la culpabilité et la honte étant souvent évoquées. Les troisième et quatrième axes concernent les sentiments vis-à-vis des mesures de contrôle et des procédures post-déclaration. Les éleveurs ont souvent une opinion négative sur la prise en charge des maladies réglementées, qui implique le plus souvent un isolement de l'élevage, avec un arrêt complet des échanges qui s'applique aussi aux élevages voisins, voire un abattage des troupeaux. Les cinquième et sixième axes concernent la confiance (souvent mauvaise) dans les autorités sanitaires et la problématique du manque de transparence dans les procédures de déclaration. Des solutions sont proposées à ces problématiques, par les éleveurs mais aussi par les vétérinaires praticiens et les autorités sanitaires. Ces solutions sont par exemple de raccourcir les périodes d'isolement des élevages, afin de réduire les conséquences sociales, d'améliorer les tests diagnostiques, ou encore d'améliorer la communication entre les différents acteurs. La confiance est au cœur de la procédure de déclaration, notamment la confiance au sein du couple éleveur/vétérinaire (le vétérinaire étant celui qui va faire le choix final d'alerter ou non les autorités sanitaires). En outre, une solution financière proposée est d'avoir une indemnité plus importante pour les animaux malades que pour les animaux morts, afin d'inciter à une déclaration la plus précoce possible. Plus récemment, Delgado et al. (2014) ont investigué ces facteurs psychosociaux à propos de la déclaration de fièvre aphteuse dans les élevages de bovins allaitants. Ils sont arrivés à des conclusions similaires (problématique des délais d'action et du manque de transparence des procédures notamment), en soulignant le rôle clé du vétérinaire dans l'accompagnement de l'éleveur. Ils mettent aussi en avant les bénéfices personnels qu'a un éleveur à déclarer un cas de fièvre aphteuse, puisque cela lui permet de savoir ce qu'il se passe dans son élevage et peut lui donner la sensation d'être un « bon éleveur ». Vergne et al. (2016) quant à eux mettent en évidence la problématique du manque de connaissance des procédures de déclaration à travers l'étude de la déclaration des cas de peste porcine africaine, à la fois par les éleveurs de porcs mais aussi par les chasseurs de sangliers.

D'autres facteurs influençant la détection et la déclaration sont évoqués dans des études analysant la volonté des éleveurs à contrôler le statut sanitaire de leur élevage, que ce soit en élevage bovin (Ellis-Iversen et al., 2010) ou porcin (Alarcon et al., 2014). L'aspect financier et le bien-être animal sont les deux principales motivations à contrôler correctement les troubles sanitaires, ainsi que l'image et la réputation de l'élevage. Par ailleurs, l'accent est de nouveau mis sur l'importance de la connaissance des maladies afin de mieux les détecter, ainsi que sur le rôle du vétérinaire et la confiance que lui accordent les éleveurs.

c) Limites à l'échelle mondiale

Jusque-là n'ont été évoquées que les problématiques de détection et déclaration à l'échelle d'un élevage, mais ces problématiques peuvent aussi concerner un pays entier. Perez et al. (2011) se sont intéressés à la surveillance des maladies à l'échelle mondiale. Tous les pays n'ont pas les moyens financiers pour mettre en œuvre une surveillance correcte des maladies animales (formation des vétérinaires, qualité des laboratoires), ce qui va entraîner une sous détection des cas. Par ailleurs, certains pays ne font pas remonter leurs données aux organisations internationales par manque d'intérêt (territoires en guerre) ou par crainte des sanctions commerciales (fermeture des frontières ou limitation des mouvements d'animaux dans le cas de la déclaration de certaines maladies, telle que la fièvre aphteuse (Toma et al., 2017)).

2) Limites de la surveillance active dans la collecte des informations

Dans le cadre d'une surveillance active, la qualité de la détection repose surtout sur la qualité des tests mis en œuvre (sensibilité notamment). Elle peut en outre dépendre de la qualité de réalisation des tests (prélèvement inadapté, mauvaise conservation,…) et de la surveillance exhaustive de tous les animaux à surveiller (difficulté de mise en œuvre de la prophylaxie obligatoire dans certains troupeaux par exemple).

3) Hétérogénéité dans la collecte des informations

Les limites de la surveillance présentées conduisent à l'obtention de données incomplètes, en raison d'une sous détection voire d'une sous déclaration des cas réels. En plus de cela, une hétérogénéité dans la détection ou la déclaration peut s'ajouter, c'est-à-dire que la détection ne se fera pas avec la même sensibilité selon les animaux ou les élevages considérés, notamment si les tests diagnostiques ou de dépistage utilisés sont différents. C'est par exemple le cas dans la surveillance de la brucellose bovine en France, qui ne se fait pas de la même manière pour les troupeaux allaitants et les troupeaux laitiers (Ganière et Laaberki, 2017).

IV- Objectif de l'étude

Les éléments présentés précédemment (sensibilité intrinsèque des tests, sous détection et sous déclaration liées à des facteurs économiques, sociaux ou psychologiques, hétérogénéité dans la détection des cas) sont à l'origine de bases de données imparfaites, et donc incomplètes, pour les études de surveillance des maladies animales. Ces données incomplètes vont potentiellement biaiser les études d'identification de facteurs de risque de maladies animales.

L'objectif de cette étude est d'**évaluer l'importance de l'impact que peut avoir une détection imparfaite, voire hétérogène, d'une maladie sur l'identification de ses facteurs de risque**.

En première partie, nous présenterons un état des lieux des modèles communément utilisés dans l'identification de facteurs de risque de maladies infectieuses animales. En deuxième partie, nous présenterons une étude par simulation pour évaluer l'impact théorique d'une détection imparfaite, homogène ou hétérogène, sur l'identification des facteurs de risque en utilisant un modèle logistique et un modèle de Poisson enflé en zéro. Enfin, la troisième partie illustre l'approche développée en deuxième partie en analysant un jeu de données réelles de surveillance des avortements bovins en France métropolitaine entre le 1er août 2010 et le 31 juillet 2011.

Partie 1 :
METHODES D'ETUDE DES FACTEURS DE RISQUE : REVUE DE LA LITTERATURE

I- Etat des lieux sur les modèles utilisés pour l'identification des facteurs de risque de maladies animales à partir de données de surveillance

1) Objectifs

L'identification des facteurs de risque peut se faire grâce à des approches contrôlées, de type études de cohortes ou études cas-témoins, ou grâce à des approches non contrôlées avec l'utilisation de données de surveillance. Ces données de surveillance sont la plupart du temps incomplètes et imparfaites en raison du manque de sensibilité ou de spécificité de la détection des cas, voire de la déclaration dans le cadre de la surveillance passive.

Les objectifs de cette partie sont de voir quels sont les modèles utilisés pour identifier des facteurs de risque de maladies animales à partir de données de surveillance, et de savoir si l'imperfection, voire l'hétérogénéité, de la détection ayant permis l'obtention de ces données est prise en compte.

2) Matériels et méthodes

Afin de réaliser une revue de la littérature concernant l'identification de facteurs de risque de maladies animales à partir de données de surveillance, une recherche en anglais a été effectuée sur la base de données numérique PubMed (PubMed - NCBI, 2017) le 16 février 2017 avec les mots clés suivants (contenus dans le titre ou dans l'abstract) : « surveillance » *and* « livestock *or* cattle *or* cow *or* cows *or* pig *or* pigs *or* swine *or* sheep *or* goat* *or* ruminant* *or* poultry *or* avian » *and* « disease* *or* health *or* infection* *or* outbreak* » *and* « model* » *and* « risk ». L'astérisque (*) permet d'inclure dans la recherche tous les mots contenant le mot qui la précède. L'objectif étant de fournir une illustration des différentes méthodes et approches pour l'identification des facteurs de risque, nous n'avons pas cherché à être exhaustif, ni à conduire la revue de manière systématique. Une première sélection des articles a ensuite été réalisée sur la base des titres et de la lecture des abstracts. Une seconde sélection a porté sur la lecture intégrale des articles. A cette sélection d'articles, certains articles connus pour leur pertinence au thème d'étude mais n'apparaissant pas dans la recherche sur PubMed ont été ajoutés à la sélection finale.

Suite à la lecture de ces articles, différentes données ont été extraites de chaque article :

- pays d'étude
- maladie étudiée, animaux concernés
- origine des données concernant la maladie (maladie notifiable à l'OIE, programme national de surveillance de la maladie,…)
- unités épidémiologiques considérées (animaux, élevages, villages,…)
- modèle(s) statistique(s) utilisé(s) pour l'identification des facteurs de risque

- variable réponse à expliquer (présence de la maladie, nombre de malades détectés, prévalence)
- liste des variables explicatives : facteurs étudiés
- qualité de la détection des cas
- prise en compte par le modèle ou discussion de la qualité de la détection des cas

3) Résultats

La recherche réalisée sur PubMed a fait ressortir 295 articles, datant de 1998 à 2017. Suite à la lecture des titres et des abstracts, 44 articles ont été retenus, dont finalement 34 sont retenus ici (suite à la lecture intégrale des articles). Trois autres articles connus par ailleurs, et qui ne sont pas ressortis dans la recherche, ont été ajoutés à la liste finale. Le processus de sélection des articles est schématisé dans la Figure 3. Les références des 37 articles ainsi qu'une partie des informations extraites sont présentées dans le Tableau 1.

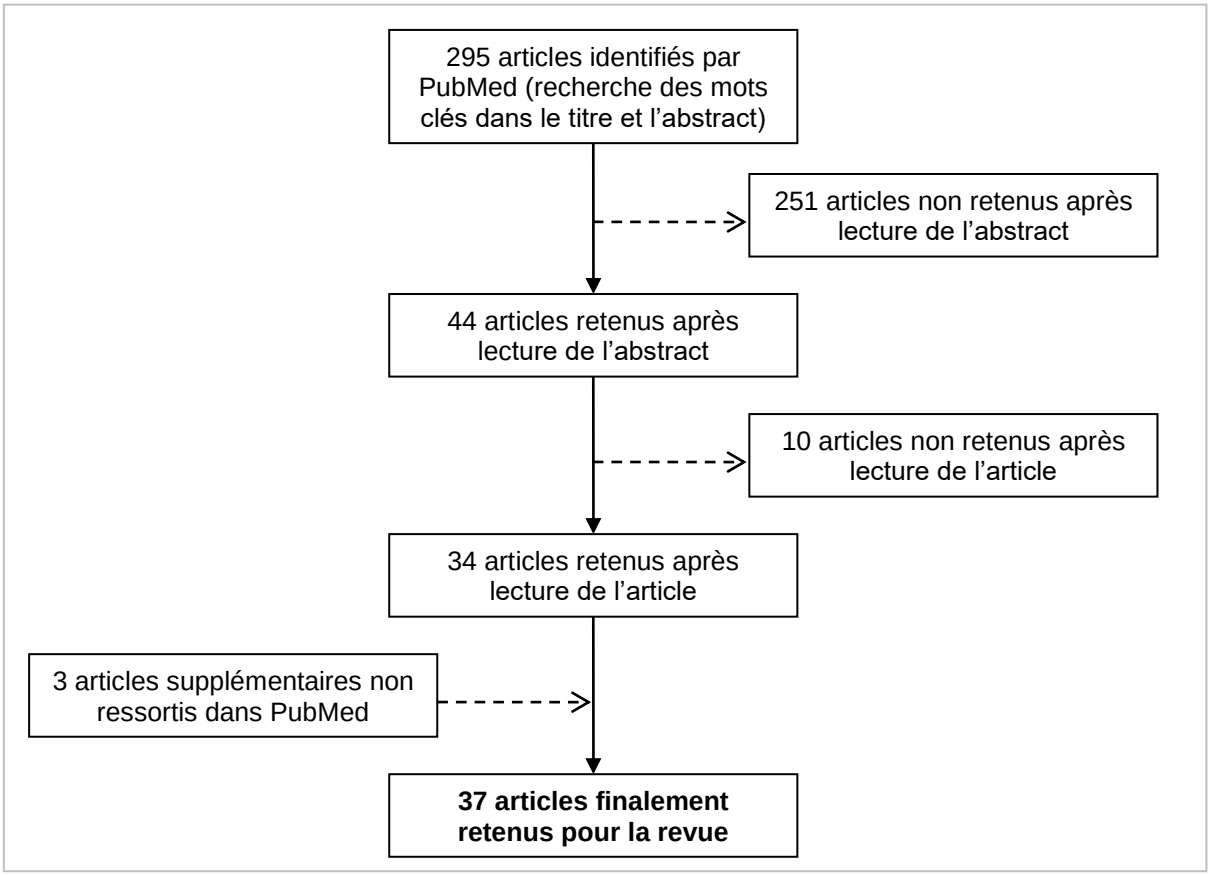


Figure 3 : Organigramme résumant le processus de sélection des articles de la revue

Tableau 1 : Etudes identifiant des facteurs de risque pour différentes maladies animales

Référence	Maladie	Pays	Modèle(s)	Discussion de la détection
Abrial et al. (2005)	ESB [1]	France	Poisson	Non
Pfeiffer et al. (2007)	IAHP [2] (H5N1)	Vietnam	Logistique	Oui : facteurs humains peuvent favoriser détection (urbain *vs* rural)
Porphyre et al. (2008)	Tuberculose bovine	N-Zélande	Poisson	Non
Lee et al. (2009)	Brucellose bovine	Corée du Sud	Logistique	Oui : détection meilleure dans grands troupeaux
Namata et al. (2009)	Salmonellose aviaire	Belgique	Logistique	Non
Baptista et al. (2010)	Salmonellose porcine	Portugal	Logistique	Oui : sensibilité des tests
Benschop et al. (2010)	Salmonellose porcine	Danemark	Binomial enflé en zéro	Oui : taille échantillons
Humblet et al. (2010)	Tuberculose bovine	Belgique	Logistique	Non
Porphyre et al. (2010)	Brucellose	Arménie	Logistique	Oui : sensibilité des tests
Wolfe et al. (2010)	Tuberculose bovine	Irlande	Logistique	Oui : déclaration des lésions variable selon les abattoirs
Del Rio Vilas et al. (2011)	Tremblante ovine	Pays de Galles	Bayésien hiérarchique	Oui : grands troupeaux et compétence du personnel
Gulenkin et al. (2011)	PP [3] africaine	Russie	Linéaire	Non
Loth et al. (2011)	IAHP [2] (H5N1)	Indonésie	Logistique Linéaire (binomial)	Non
Martin et al. (2011)	IAHP [2] (H5N1)	Chine	Logistique Arborescence	Oui : facteurs humains peuvent favoriser détection (urbain *vs* rural)
Rodríguez-Prieto et al. (2012)	Tuberculose bovine	Espagne	Logistique	Non
Trevennec et al. (2012)	Influenza H1N1	Vietnam	Binomial enflé en zéro	Non
Bronner et al. (2013)	Avortements bovins	France	Poisson enflé en zéro Hurdle	Oui : défaut de détection et défaut de déclaration
Shittu et al. (2013)	Tuberculose bovine	G-Bretagne	Logistique	Oui : compétence du personnel et sensibilité de l'inspection (abattoir)
Stevens et al. (2013)	IAHP [2] (H5N1)	Asie	Analyse décisionnelle multicritères	Oui : incertitude sur la sous déclaration
Dhingra et al. (2014)	IAHP [2] (H5N1)	Inde	Logistique Arborescence	Oui : défaut de détection et défaut de déclaration
Korennoy et al. (2014)	PP [3] africaine	Russie	Entropie maximale	Non
Martínez-López et al. (2014)	PP [3] classique	Bulgarie	Logistique	Non
Pascual-Linaza et al. (2014)	FCO [4]	Espagne	Logistique	Oui : sous déclaration minime
Sindato et al. (2014)	Fièvre de la vallée du Rift	Tanzanie	Logistique	Oui : défaut de détection et défaut de déclaration
Thanapongtharm et al. (2014)	SDRP [5]	Vietnam	Logistique Arborescence	Oui : hautement pathogène bien détecté et effet facteurs humains
Vergne et al. (2014)	IAHP [2] (H5N1)	Thaïlande	Poisson enflé en zéro	Oui : défaut de détection et défaut de déclaration
Netrabukkana et al. (2015)	Influenza A	Cambodge	Linéaire	Oui : défaut de déclaration
Saksena et al. (2015)	IAHP [2] (H5N1)	Vietnam	Linéaire Arborescence	Oui : défaut de détection et défaut de déclaration (urbain *vs* rural)
Abdrakhmanov et al. (2016)	Rage	Kazakhstan	Entropie maximale	Oui : facteurs humains peuvent favoriser détection (urbain *vs* rural)
Alkhamis et al. (2016)	IAHP [2] (H5N1)	Moyen-Orient	Entropie maximale	Oui : défaut de déclaration et effet facteurs humains (urbain *vs* rural)
Alkhamis et al. (2016)	Dermatose nodulaire contagieuse bovine	Moyen-Orient	Entropie maximale	Oui : défaut de déclaration et effet facteurs humains (urbain *vs* rural)
Byrne et al. (2016)	Fasciolose bovine	Irlande du Nord	Logistique Linéaire	Oui : compétence du personnel et sensibilité de l'inspection (abattoir)
Cowled et al. (2016)	Paratuberculose ovine	Australie	Logistique	Oui : sensibilité de l'inspection
Hayama et al. (2016)	Fièvre de trois jours	Japon	Logistique	Non
Paul et al. (2016)	IAHP [2] (H5N1)	Thaïlande et Cambodge	Analyse décisionnelle multicritères	Oui : défaut de déclaration (notamment au Cambodge)
Vergne et al. (2016)	PP [3] africaine	Russie	Poisson enflé en zéro	Oui : défaut de détection et défaut de déclaration
Walsh et al. (2016)	IAHP [2] (H5N1)	Afrique, Asie, Europe	Entropie maximale	Oui : défaut de déclaration

1) ESB : encéphalopathie spongiforme bovine ; 2) IAHP : influenza aviaire hautement pathogène ; 3) PP : peste porcine ; 4) FCO : fièvre catarrhale ovine ; 5) SDRP : syndrome dysgénésique et respiratoire porcin

Parmi ces 37 articles, 29 (soit 78%) étudient des maladies notifiables à l'OIE (OIE - World Organisation for Animal Health, 2017) et se basent sur les données de surveillance des pays d'étude. Pour ce qui est des autres articles, 3 articles se basent sur des données de surveillance dans le cadre de la réglementation de l'Union européenne (salmonellose en élevage de volaille (Namata et al., 2009), avortements bovins (Bronner et al., 2013), fasciolose lors de l'abattage des ruminants (Byrne et al., 2016)), 3 se basent sur des programmes nationaux (salmonellose porcine au Danemark (Benschop et al., 2010), influenza A sur les porcs au Cambodge (Netrabukkana et al., 2015), fièvre des trois jours chez les bovins au Japon (Hayama et al., 2016)) et une étude utilise les données d'une étude antérieure (Baptista et al., 2010). Seule l'étude de Trevennec et al. (2012) sur l'influenza H1N1 chez les porcs vietnamiens a été retenue dans la revue bien qu'elle ne fasse pas appel à des données de surveillance existantes, puisque qu'elle s'appuie sur un échantillonnage construit pour l'étude. Elle a cependant été sélectionnée étant donné l'utilisation d'un modèle enflé en zéro pour les analyses.

Le modèle le plus couramment utilisé pour identifier des facteurs de risque est le modèle logistique (19/37 articles, soit 51%). Ce modèle permet d'expliquer la présence ou l'absence de la maladie en question dans une unité épidémiologique (qui peut être un élevage, un village, ou une aire géographique définie), mais ne prend pas en compte de manière explicite un possible défaut de détection de cette maladie. Cependant, les auteurs reviennent en général sur les défauts de sensibilité de la détection dans leur discussion, puisque 13 études sur les 19 utilisant une régression logistique discutent ou évoquent ce potentiel biais. Par exemple, Byrne et al. (2016) ont étudié la prévalence et les facteurs de risque de l'infestation des bovins par la grande douve en Irlande du Nord entre 2011 et 2013. Les données utilisées proviennent de la surveillance systématique des carcasses en abattoir, et ils ont considéré les élevages comme étant les unités épidémiologiques. Ils évoquent en discussion le manque de sensibilité de la détection des lésions en abattoir et la variabilité de cette sensibilité d'un abattoir à l'autre, qui peuvent être dues à la vitesse de la chaîne d'abattage, à la qualité de l'éclairage, ainsi qu'à l'expérience et à la motivation de l'opérateur. Outre la sensibilité de la détection liée à la sensibilité intrinsèque des tests diagnostiques ou aux opérateurs, un autre élément discuté est le rôle des facteurs anthropiques identifiés comme facteurs de risque. C'est par exemple le cas dans l'étude de Pfeiffer et al. (2007), qui vise à étudier les facteurs de risque d'introduction de l'influenza aviaire hautement pathogène dans les populations de volaille au Vietnam entre 2004 et 2006. Les données utilisées proviennent de la surveillance nationale, et les unités épidémiologiques considérées sont les communes. Un des facteurs étudiés est la distance à une zone géographique très peuplée ($\geq$ 50 habitants/km²), qui révèle que le risque d'apparition d'un foyer d'influenza diminue lorsque la distance à une zone très peuplée augmente. Les auteurs discutent du rôle réel de ce facteur : la forte densité humaine pourrait favoriser la dissémination de la maladie (en lien avec des zones d'élevages et d'échanges de volaille), mais pourrait aussi favoriser la détection des foyers existants.

Parmi les 37 articles retenus, 3 articles utilisent le modèle d'entropie maximale en indiquant dans leur discussion que ce modèle permettrait de prendre en compte la sous détection et la sous déclaration des cas (Abdrakhmanov et al., 2016 ; Alkhamis et VanderWaal, 2016 ;

Alkhamis et al., 2016). Ce modèle repose sur la construction de cartes prédictives de distribution, et vise à identifier des variables environnementales comme étant de potentiels facteurs de risque afin de repérer les zones géographiques appropriées pour la maladie.

En outre, 4 articles utilisent un modèle enflé en zéro en prétendant qu'il peut prendre en compte le défaut de détection, voire de déclaration, ainsi qu'une potentielle hétérogénéité dans la détection (Benschop et al., 2010 ; Bronner et al., 2013 ; Vergne, Korennoy, et al., 2016 ; Vergne et al., 2014). En effet, les modèles enflés en zéro permettraient d'identifier en même temps les facteurs influençant la présence de la maladie dans les unités épidémiologiques (comme le ferait une régression logistique) et les facteurs influençant le nombre de cas détectés, ou déclarés, dans une unité épidémiologique affectée.

Par ailleurs, la spécificité est considérée comme parfaite dans l'ensemble des articles. Ceci s'explique par exemple par le fait que les notifications faites à l'OIE sont confirmées par différents tests suite à la suspicion clinique. Les cas conservés sont donc ceux qui ont été confirmés par différents tests, le protocole dépendant de la maladie, ce qui diminue le risque de faux positifs.

4) Discussion

La détection d'une maladie est le plus souvent imparfaite, ne serait-ce qu'en raison de la sensibilité des tests ou des compétences variées des personnes chargées de la détecter (à l'abattoir par exemple). De plus, l'étape de déclaration est aussi importante à prendre en compte lorsqu'on utilise une base de données issue de la surveillance passive, puisqu'elle ajoute un biais dans les informations collectées. En effet, la plupart des maladies étudiées et présentées dans le Tableau 1 sont des maladies à notification obligatoire auprès de l'OIE, qui sont donc enregistrées dans les bases de données suite à une déclaration spontanée de la part des vétérinaires (après appel des éleveurs). Cependant, ces cas déclarés ne reflètent pas forcément de manière exhaustive les cas détectés, qui ne reflètent pas non plus parfaitement les cas réels, comme cela a par exemple été démontré par Bronner et al. (2013) dans leur étude sur les déclarations d'avortements bovins en France entre 2006 et 2011. Selon leur étude, 68% des éleveurs ont détecté au moins un avortement dans leur troupeau, mais seuls 23% des éleveurs ont effectivement déclaré au moins un avortement dans leur troupeau.

Cette différence entre les données disponibles et la réalité de la dynamique des maladies est rarement prise en compte de manière explicite par les modèles couramment utilisés pour l'identification de facteurs de risque, mais est généralement discutée. Le modèle d'entropie maximale permettrait de prendre en compte l'absence de données dans certaines régions en identifiant des variables géographiques associées à la maladie. Quant aux modèles enflés en zéro, ils permettraient la prise en compte du défaut de détection en distinguant les « zéros » issus des unités épidémiologiques non infectées des « zéros » issus des unités épidémiologiques infectées mains non détectées comme telles. Ces modèles enflés en zéro sont expliqués et discutés plus longuement dans la suite du manuscrit.

II- Problématique de la sous détection et de la sous déclaration d'une maladie : modélisation de données incomplètes

La problématique de l'observation imparfaite a notamment été étudiée en écologie. En effet, les études qui visent à détecter une espèce rare dans une aire géographique donnée sont confrontées à la problématique de savoir si l'absence d'observation d'une espèce est due à l'absence réelle de l'espèce ou simplement à un défaut dans sa détection (MacKenzie et al., 2002 ; Royle et al., 2005).

Martin et al. (2005) définissent ainsi deux sources de « zéros » dans les données. Il y a d'une part les « vrais zéros » dus à une faible taille de la population étudiée ou à une espèce qui n'occupe pas l'intégralité d'un territoire qui est pourtant adapté, et d'autre part les « faux zéros » dus à une absence temporaire de l'espèce au moment de l'observation ou à un échec de la détection de l'espèce malgré sa présence (ce qui peut être le cas pour les espèces rares). Afin de prendre en considération ces deux sources de « zéros », ils proposent l'utilisation de modèles enflés en zéro, comme l'avaient déjà fait Welsh et al. (1996). Ces modèles ont été introduits par Lambert (1992) et sont construits en deux étapes : la première étape modélise la présence/absence de l'espèce étudiée avec une loi de Bernoulli, et la seconde étape modélise avec une loi de comptage (loi de Poisson ou loi binomiale) le nombre d'individus de l'espèce étudiée détectés sachant que l'espèce est présente. Ceci permet de modéliser d'une part le processus conduisant aux « vrais zéros » (loi de Bernoulli) et d'autre part le processus conduisant aux « faux zéros » (loi de comptage). Ces modèles sont donc utilisables lorsque la sensibilité dans la détection est imparfaite, et considère en revanche que la spécificité est parfaite.

Cette modélisation de données enflées en zéro peut s'appliquer à l'étude des maladies, comme illustré en Figure 4. En effet, comme déjà vu, la détection est imparfaite quel que soit le mode de surveillance. Les « vrais zéros » sont les animaux non malades, qui sont donc non détectés par la surveillance (qu'elle soit active ou passive), tandis que les « faux zéros » sont les animaux malades qui ne sont pas détectés par le programme de surveillance.

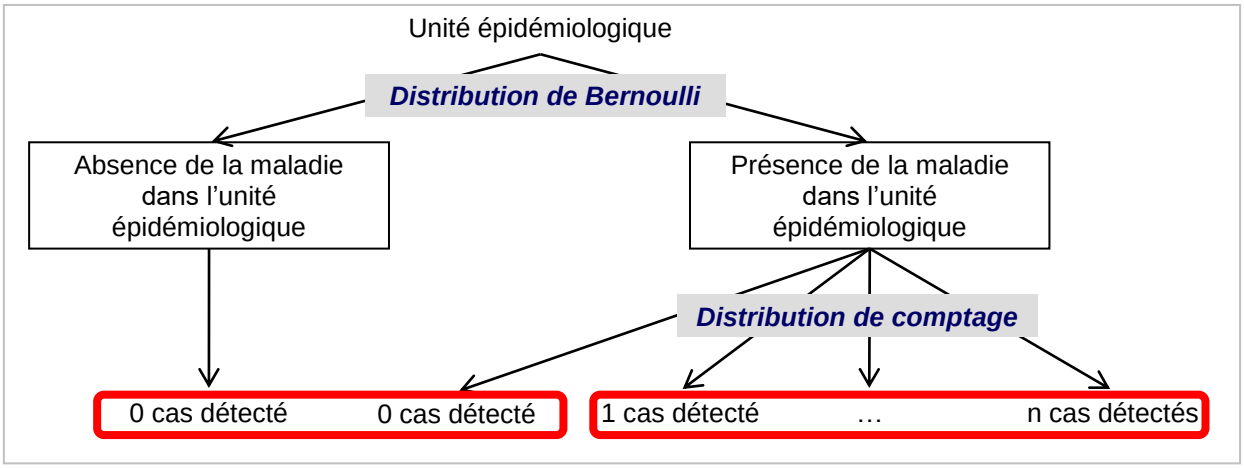


Figure 4 : Données enflées en zéro générées par les systèmes de surveillance, d'après Vergne et al. (2015)

Les modèles enflés en zéro entrent dans la grande catégorie des méthodes de capture-recapture, habituellement utilisés en écologie, et qui pourraient avoir un intérêt en épidémiologie vétérinaire, comme expliqué par Vergne et al. (2012 ; 2015).

L'utilisation particulière du modèle de Poisson enflé en zéro pour l'identification des facteurs de risque est discutée dans la suite du manuscrit. Il est comparé au modèle le plus couramment utilisé, le modèle logistique. Des données simulées où la détection est imparfaite permettent de mettre en évidence les différences entre ces deux modèles.

Partie 2 :
IMPACT DE LA DETECTION IMPARFAITE SUR L'ETUDE DES FACTEURS DE RISQUE : APPROCHE PAR SIMULATIONS

I- Etude de l'influence d'une sensibilité imparfaite sur l'estimation du risque d'une maladie

1) Introduction et contexte

Les études visant à identifier des facteurs de risque se basent sur des données incomplètes, que ce soit par un défaut de détection ou de déclaration. La situation présentée ici se place dans un contexte où la sensibilité de détection est imparfaite, tandis que la spécificité est considérée parfaite. La sensibilité peut par exemple être imparfaite à cause d'une mauvaise sensibilité des tests utilisés lors d'une surveillance active, ou du manque d'observation des signes cliniques lors d'une surveillance passive.

L'étude porte sur N unités épidémiologiques, elles-mêmes composées d'un certain nombre d'unités élémentaires. Ce pourrait par exemple être N élevages, composés d'un certain nombre d'animaux. Ces animaux sont détectés lorsqu'ils sont malades et que le système de surveillance permet de les détecter comme étant effectivement malades.

Les modèles étudiés sont le modèle logistique, modèle couramment utilisé pour identifier des facteurs de risque, et le modèle de Poisson enflé en zéro. Les objectifs sont de mettre en évidence les capacités de ces deux modèles à correctement identifier des facteurs de risque et à correctement estimer les odds ratios qui correspondent.

2) Matériels et méthodes

a) <u>Distribution des variables d'intérêt</u>

Le **facteur X est défini comme le facteur de risque de présence de la maladie** dans une unité épidémiologique, c'est-à-dire que le risque de présence de la maladie dans les unités épidémiologiques où X est présent est plus élevé que celui dans les unités épidémiologiques où X est absent. La probabilité que X soit présent dans une unité épidémiologique est notée « pX ». Lorsque le facteur X est absent (respectivement présent) d'une unité épidémiologique, la probabilité que la maladie soit présente dans cette unité est notée « prev.X0 » (respectivement « prev.X1 »). L'odds ratio de la présence de la maladie associé à ce facteur est noté « OR(X) ».

Le **facteur Y est défini comme le facteur de confusion de détection de la maladie** dans une unité épidémiologique, c'est-à-dire que la sensibilité de la détection des unités élémentaires malades dans les unités épidémiologiques dépend de la présence de Y. La probabilité que Y soit présent dans une unité épidémiologique est notée « pY ». Lorsque le facteur Y est absent (respectivement présent) d'une unité épidémiologique, la probabilité de détecter chacune des unités élémentaires malades dans cette unité épidémiologique est notée « sensib.Y0 » (respectivement « sensib.Y1 »).

La Figure 5 illustre les relations entre toutes ces variables.

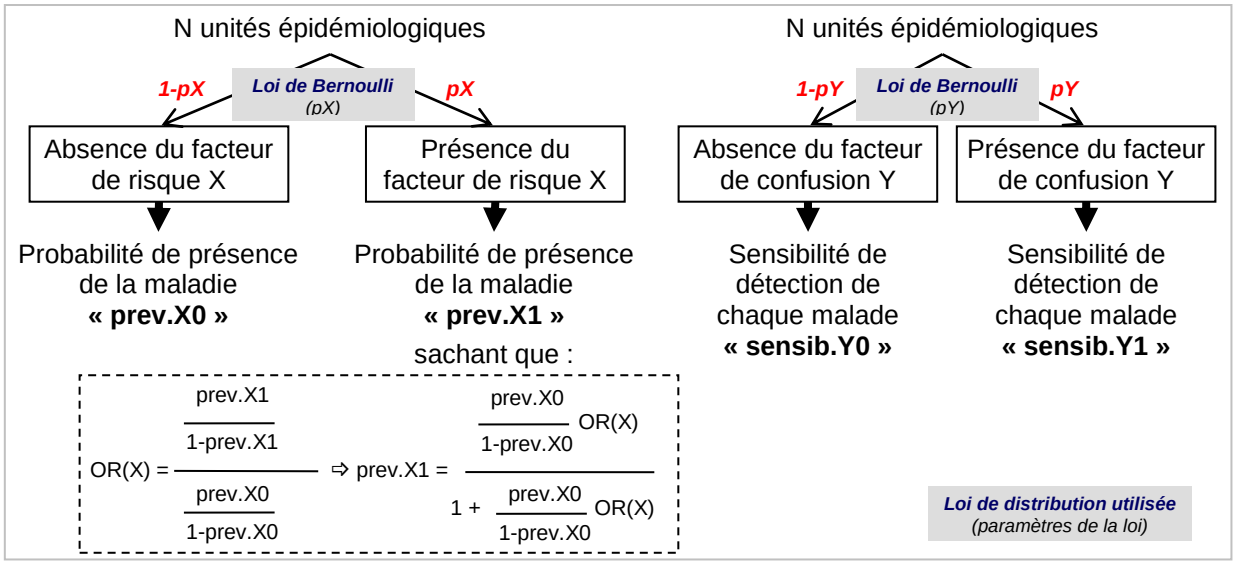


Figure 5 : Distribution des facteurs X et Y, et leur influence sur la maladie étudiée

Pour certaines simulations, seul le facteur X a été défini. Dans ces situations il est à la fois facteur de risque de présence de la maladie et facteur de confusion de détection de la maladie.

b) Détection de la maladie dans les unités épidémiologiques

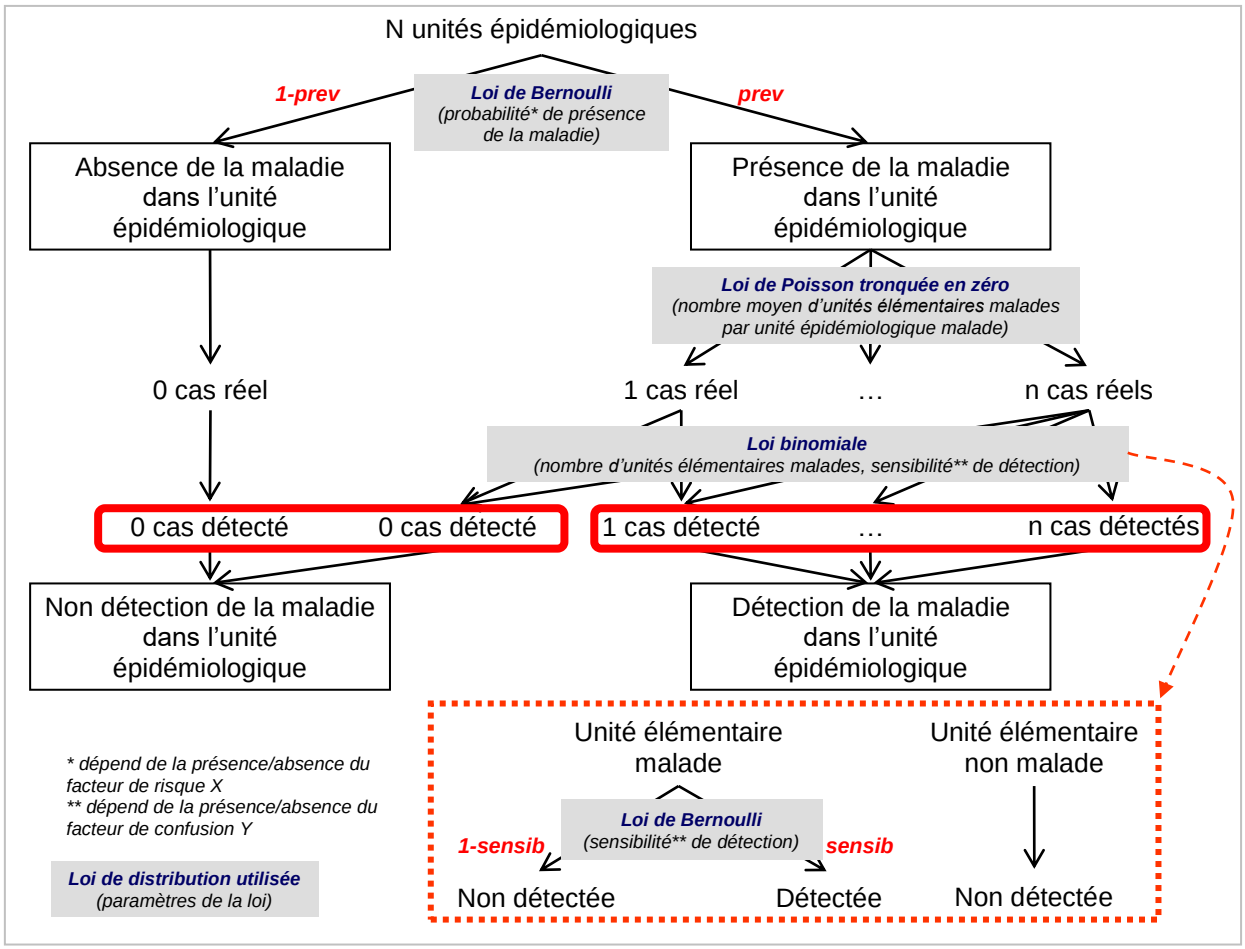


Figure 6 : Schéma de simulation des données relatives à la maladie étudiée

Comme illustré en Figure 6, la détection de la maladie dans les unités épidémiologique a été simulée comme un processus hiérarchique en trois étapes : simulation 1) de la présence de la maladie dans les unités épidémiologiques, 2) du nombre d'unités élémentaires malades dans les unités épidémiologiques malades et 3) du nombre d'unités élémentaires malades et effectivement détectées comme telles dans les unités épidémiologiques malades.

La **présence/absence de la maladie dans une unité épidémiologique i** (notée « P_i ») a été considérée comme une variable aléatoire suivant une loi de Bernoulli de paramètre « $prev_i$ », telle que présentée dans l'Équation 1.

$$P(P_i=1) = prev_i = \begin{cases} prev.X0 & \textit{si X est absent de l'unité épidémiologique i} \\ prev.X1 & \textit{si X est présent dans l'unité épidémiologique i} \end{cases}$$

***P*(P$_i$=1)** : Probabilité que la maladie soit présente dans l'unité épidémiologique i
prev$_i$: Probabilité de présence de la maladie dans l'unité épidémiologique i
prev.X0 : Probabilité de présence de la maladie dans une unité épidémiologique pour laquelle le facteur de risque X est absent
prev.X1 : Probabilité de présence de la maladie dans une unité épidémiologique pour laquelle le facteur de risque X est présent

Équation 1 : Probabilité de présence de la maladie dans une unité épidémiologique

Sachant la maladie présente dans une unité épidémiologique, le **nombre d'unités élémentaires réellement malades** (noté « Mal ») a été considéré comme une variable aléatoire suivant une loi de Poisson tronquée en zéro de paramètre « M » (M étant le nombre moyen d'unités élémentaires malades dans une unité épidémiologique malade, équivalent à la prévalence intra-unité épidémiologique), telle que présentée dans l'Équation 2.

$$P(Mal=m) = \frac{M^m}{m\,!\,(e^M - 1)}$$

***P*(Mal=m)** : Probabilité que le nombre d'unités élémentaires réellement malades dans une unité épidémiologique malade soit égal à m
M : Nombre moyen d'unités élémentaires malades dans une unité épidémiologique malade

Équation 2 : Loi de Poisson tronquée en zéro de paramètre « M »

La sensibilité de détection étant considérée comme imparfaite, toutes les unités élémentaires réellement malades peuvent ne pas être détectées, le nombre d'unités élémentaires détectées est donc inférieur ou égal au nombre d'unités élémentaires malades (la spécificité de détection est considérée parfaite). Par conséquent, le **nombre d'unités élémentaires malades et détectées dans une unité épidémiologique malade i** (noté « $Détect_i$ ») a été considéré comme une variable aléatoire suivant une loi binomiale de paramètres « m » et « $sensib_i$ », telle que présentée dans l'Équation 3.

$$P(\text{Détect}_i=d) = \binom{m}{d} . \text{sensib}_i^{d} . (1 - \text{sensib}_i)^{m-d}$$

$$\textbf{avec } \text{sensib}_i = \begin{cases} \text{sensib.Y0} & \textit{si Y est absent de l'unité épidémiologique i} \\ \text{sensib.Y1} & \textit{si Y est présent dans l'unité épidémiologique i} \end{cases}$$

***P*(Détect$_i$=d)** : Probabilité que le nombre d'unités élémentaires malades et détectées dans l'unité épidémiologique malade i soit égal à d
m : Nombre d'unités élémentaires réellement malades dans l'unité épidémiologique malade i
sensib$_i$: Probabilité de détecter chacune des unités élémentaires malades dans l'unité épidémiologique malade i
sensib.Y0 : Sensibilité de détection d'une unité élémentaire malade dans une unité épidémiologique malade pour laquelle le facteur de confusion Y est absent
sensib.Y1 : Sensibilité de détection d'une unité élémentaire malade dans une unité épidémiologique malade pour laquelle le facteur de confusion Y est présent

Équation 3 : Loi binomiale de paramètres « m » et « sensib$_i$ »

Notons que la maladie peut ne pas être détectée dans les unités épidémiologiques où elle est présente si aucune des unités élémentaires malades n'est détectée. On considère qu'une unité épidémiologique malade est détectée si d > 0. Le nombre d'unités épidémiologiques malades et détectées est donc inférieur ou égal au nombre d'unités épidémiologiques où la maladie est réellement présente.

Dans le cas où le facteur de risque et le facteur de confusion sont identiques, la probabilité de présence de la maladie dans une unité épidémiologique (« prev$_i$ ») définie dans l'Équation 1 et la probabilité de détecter chacune des unités élémentaires malades dans une unité épidémiologique malade (« sensib$_i$ ») présentée dans l'Équation 3 dépendent toutes les deux de la présence du même facteur (défini comme étant le facteur X).

c) <u>Plan de simulation</u>

Tableau 2 : Paramètres fixés dans les différentes séries de simulations

<table>
<tr><th></th><th colspan="2">« Série 1 »</th><th colspan="2">« Série 2 »</th><th colspan="2">« Série 3 » [*]</th><th colspan="2">« Série 4 »</th></tr>
<tr><td>Homogénéité de la détection</td><td colspan="4">Détection homogène</td><td colspan="4">Détection hétérogène</td></tr>
<tr><td>N</td><td colspan="4">10 000</td><td colspan="4">10 000</td></tr>
<tr><td>pX</td><td colspan="4">0,5</td><td colspan="4">0,5</td></tr>
<tr><td>pY</td><td colspan="4">0,4</td><td colspan="4">0,4</td></tr>
<tr><td>M</td><td colspan="2">4</td><td colspan="2">1 à 17
(que les nombres impairs, pour alléger les simulations)</td><td colspan="2">4</td><td colspan="2">1 à 17
(que les nombres impairs, pour alléger les simulations)</td></tr>
<tr><td>prev.X0</td><td colspan="2">0,2 ; 0,4 ; 0,6 ; 0,8</td><td colspan="2">0,1 ; 0,2 ; 0,5</td><td colspan="4">0,1 ; 0,2 ; 0,5</td></tr>
<tr><td>OR(X)</td><td colspan="2">1 à 10</td><td colspan="2">2 ; 5 ; 10</td><td colspan="4">2 ; 5 ; 10</td></tr>
<tr><td>sensib.Y0</td><td colspan="4" rowspan="2">sensib.Y0=sensib.Y1
0,01 à 1 avec un pas de 0,01</td><td colspan="2">0,1 à 1 avec un pas de 0,1</td><td colspan="2">0,3 à 0,9 avec un pas de 0,3</td></tr>
<tr><td>sensib.Y1</td><td colspan="2">0,1 à 1 avec un pas de 0,1</td><td colspan="2">0,3 à 0,9 avec un pas de 0,3</td></tr>
<tr><td>Nombre de simulations (pour N, pX, pY, M, prev.X0 et OR(X) fixés)</td><td colspan="4">1 simulation pour chacune des 100 valeurs de sensibilité testées</td><td colspan="2">300 simulations pour chacun des 100 couples (sensib.Y0,sensib.Y1) testés</td><td colspan="2">300 simulations pour chacun des 9 couples (sensib.Y0,sensib.Y1) testés</td></tr>
</table>

[*]Les paramètres de la « Série 3 » ont aussi été testés en faisant l'hypothèse que le facteur X est identique au facteur Y, avec pX=0,5 (le facteur de risque et le facteur de confusion sont un seul et même facteur, et dans cette situation sensib.Y0 devient sensib.X0 et sensib.Y1 devient sensib.X1)

Le plan de simulation a été défini de manière à évaluer l'influence d'un grand nombre de paramètres d'intérêt sur la caractérisation du risque et fournir une réponse nuancée à la question de recherche.

Le modèle général de simulation comprend 8 paramètres (Tableau 2). Par souci de clarté et de simplicité des interprétations, les valeurs de certains paramètres d'intérêt moindre (N, pX et pY) ont été fixées : le nombre d'unités épidémiologiques étudiées a été fixé à N=10000 (de manière à ce que la taille de l'échantillon ne limite pas la puissance de l'analyse) et les probabilités de présence des facteurs X et Y ont été respectivement fixées à pX=0,5 et pY=0,4. L'influence de tous les autres paramètres (M, prev.X0, OR(X), sensib.Y0, sensib.Y1) a été testée en les faisant varier de manière indépendante ou en combinaison dans des ordres de grandeur réalistes. Les différentes valeurs testées des différents paramètres sont présentées dans le Tableau 2, tandis que le Tableau 3 présente les valeurs de prev.X1 associées à chaque valeur de prev.X0 et d'OR(X).

Pour l'interprétation des résultats, deux niveaux de confusion ont aussi été définis : 1) la sensibilité de détection est imparfaite mais la même pour toutes les unités épidémiologiques (sensibilité de détection homogène entre les unités épidémiologiques : sensib = sensib.Y0 = sensib.Y1) et 2) la sensibilité de détection est imparfaite et différente dans les deux strates de la population (sensibilité de détection hétérogène entre unités épidémiologiques : sensib.Y0 ≠ sensib.Y1).

Tableau 3 : Valeurs de prev.X0 et d'OR(X), et valeurs des prev.X1 qui correspondent
Le contenu de chaque case du tableau est le résultat du calcul de la valeur de prev.X1 qui correspond à chacun des couples prev.X0 et OR(X) envisagés

OR(X) / prev.X0	1	2	3	4	5	6	7	8	9	10
0,1	0,1	0,18	0,25	0,31	0,36	0,4	0,44	0,47	0,5	0,53
0,2	0,2	0,33	0,43	0,5	0,56	0,6	0,64	0,67	0,69	0,71
0,3	0,3	0,46	0,56	0,63	0,68	0,72	0,75	0,77	0,79	0,81
0,4	0,4	0,57	0,67	0,73	0,77	0,8	0,82	0,84	0,86	0,87
0,5	0,5	0,67	0,75	0,8	0,83	0,86	0,88	0,89	0,9	0,91
0,6	0,6	0,75	0,82	0,86	0,88	0,9	0,91	0,92	0,93	0,94
0,7	0,7	0,82	0,88	0,9	0,92	0,93	0,94	0,95	0,95	0,96
0,8	0,8	0,89	0,92	0,94	0,95	0,96	0,97	0,97	0,97	0,98

d) <u>Analyse des données obtenues</u>

Les données obtenues suite aux simulations ont été analysées d'une part avec un modèle logistique et d'autre part avec un modèle de Poisson enflé en zéro (Figure 7).

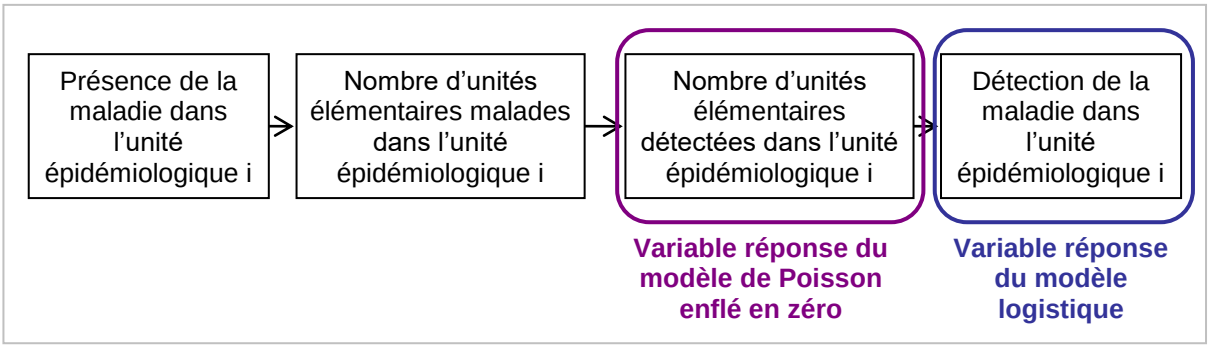


Figure 7 : Variables créées lors des simulations et variables réponses des deux modèles étudiés

- **Analyse par régression logistique**

La **variable à expliquer** du modèle logistique est la **présence/absence observée de la maladie dans les unités épidémiologiques**, c'est-à-dire la détection ou non d'au moins une unité élémentaire malade. Le **facteur de risque X** et le **facteur de confusion Y** sont les **deux variables explicatives** (chacune de ces variables est binaire : présence ou absence dans chacune des unités épidémiologiques étudiées). Le modèle logistique est défini par l'Équation 4.

$$logit(prev) = \alpha_0 + \alpha_x.X + \alpha_y.Y$$

$$\Leftrightarrow$$

$$ln\left(\frac{prev}{1\text{-}prev}\right) = \alpha_0 + ln(OR(X)_{logist}).X + ln(OR(Y)_{logist}).Y$$

prev : Probabilité de présence de la maladie dans une unité épidémiologique
α_0 : Coefficient de l'intercept du modèle logistique
$\alpha_x = ln(OR(X)_{logist})$: Coefficient associé au facteur X, égal au logarithme népérien de l'odds ratio de l'apparition de la maladie associé à X et calculé par le modèle logistique
$\alpha_y = ln(OR(Y)_{logist})$: Coefficient associé au facteur Y, égal au logarithme népérien de l'odds ratio de l'apparition de la maladie associé à Y et calculé par le modèle logistique

Équation 4 : Equation définissant le modèle logistique

Un modèle logistique a été ajusté à chaque jeu de données simulées, et le niveau de significativité de l'association entre chacune des variables explicatives et la variable réponse, ainsi que la valeur de l'$OR(X)_{logist}$ (odds ratio de la présence de la maladie associé au facteur de risque X et estimé par le modèle logistique) ont été enregistrés. Une variable a été considérée comme significativement associée à la variable réponse si le niveau de significativité (**p-value**) était inférieur à 0,05.

L'$OR(X)_{logist}$ a été utilisé pour calculer le **biais relatif** de l'odds ratio associé au facteur X, noté « $biais(OR(X)_{logist})$ » (Équation 5).

$$biais(OR(X)_{logist}) = \frac{(OR(X)_{logist}) - OR(X)}{OR(X)}$$

$biais(OR(X)_{logist})$: Biais relatif entre OR(X) et $OR(X)_{logist}$
$OR(X)_{logist}$: Odds ratio de la présence de la maladie associé au facteur X et estimé par le modèle logistique
OR(X) : Odds ratio réel de la présence de la maladie associé au facteur de risque X

Équation 5 : Calcul du biais relatif de l'odds ratio associé au facteur X (régression logistique)

- **Analyse de Poisson enflée en zéro**

Pour le modèle enflé en zéro, la **variable à expliquer** est le **nombre d'unités élémentaires détectées dans une unité épidémiologique**. Ce nombre d'unités élémentaires détectées est égal à zéro dans les unités épidémiologiques non malades, et supérieur ou égal à zéro dans les unités épidémiologiques malades. L'origine des zéros est donc double : d'une part les « vrais » zéros provenant des unités épidémiologiques non malades, et d'autre part les « faux » zéros provenant des unités épidémiologiques malades mais où aucune des unités élémentaires malades n'a été détectée. Ceci est à l'origine des deux termes de l'équation définissant le modèle de Poisson enflé en zéro (Équation 6), comme introduit par Lambert (1992). On applique d'abord une loi de Bernoulli à chacune des unités épidémiologiques pour déterminer si elle est malade ou non, ce qui est à l'origine de la partie « logistique » du modèle, puis dans les unités épidémiologiques malades on applique une loi de Poisson pour déterminer le nombre d'unités élémentaires malades et détectées, ce qui est à l'origine de la partie « comptage » du modèle. Le **facteur de risque X** et le **facteur de confusion Y** sont de nouveau les **deux variables explicatives** (chacune de ces variables est binaire : présence ou absence dans chacune des unités épidémiologiques étudiées).

$$P(\text{Détect=d}) = \begin{cases} (1 - \text{prev}) + \text{prev}.e^{-\lambda} & \textbf{\textit{si d=0}} \\ \text{prev}.e^{-\lambda}.\dfrac{\lambda^{d}}{d\,!} & \textbf{\textit{si d>0}} \end{cases}$$

avec

$$logit(\text{prev}) = \alpha_0 + \alpha_x.X + \alpha_y.Y \quad \textbf{et} \quad ln(\lambda) = \beta_0 + \beta_x.X + \beta_y.Y$$

***P*(Détect=d)** : Probabilité que le nombre d'unités élémentaires détectées dans une unité épidémiologique égale d
prev : Probabilité de présence de la maladie dans une unité épidémiologique
λ : Nombre moyen d'unités élémentaires malades détectées dans une unité épidémiologique malade
α_0 : Coefficient de l'intercept de la partie « logistique » du modèle de Poisson enflé en zéro
α_x : Coefficient associé au facteur X (égal au logarithme de l'odds ratio de la présence de la maladie associé à X et estimé par la partie « logistique » du modèle enflé en zéro)
α_y : Coefficient associé au facteur Y (égal au logarithme de l'odds ratio de la présence de la maladie associé à Y et estimé par la partie « logistique » du modèle enflé en zéro)
β_0 : Coefficient de l'intercept de la partie « comptage » du modèle de Poisson enflé en zéro
β_x : Coefficient associé au facteur X (égal au logarithme du ratio du taux d'incidence de la maladie associé à X et estimé par la partie « comptage » du modèle enflé en zéro)
β_y : Coefficient associé au facteur Y (égal au logarithme du ratio du taux d'incidence de la maladie associé à Y et estimé par la partie « comptage » du modèle enflé en zéro)

Équation 6 : Equation définissant le modèle de Poisson enflé en zéro

Un modèle de Poisson enflé en zéro a été ajusté à chaque jeu de données simulées, et le niveau de significativité de l'association entre chacune des variables explicatives et la variable réponse, ainsi que la valeur de l'$OR(X)_{logit}$ (odds ratio de la présence de la maladie associé au facteur de risque X et estimé par la partie « logistique » du modèle de Poisson enflé en zéro) ont été enregistrés. Une variable a été considérée comme significativement associée à la variable réponse si le niveau de significativité (**p-value**) était inférieur à 0,05.

L'$OR(X)_{logit}$ a été utilisé pour calculer le **biais relatif** de l'odds ratio associé au facteur X, noté « $biais(OR(X)_{logit})$ » (Équation 7).

$$\text{biais}(OR(X)_{logit}) = \frac{(OR(X)_{logit}) - OR(X)}{OR(X)}$$

biais(OR(X)$_{logit}$) : Biais relatif entre OR(X) et OR(X)$_{logit}$
OR(X)$_{logit}$: Odds ratio de la présence de la maladie associé au facteur X et estimé par la partie « logistique » du modèle de Poisson enflé en zéro
OR(X) : Odds ratio réel de la présence de la maladie associé au facteur de risque X

Équation 7 : Calcul du biais relatif de l'odds ratio associé au facteur X (modèle enflé en zéro)

e) Simulations et analyses : logiciel utilisé

Les simulations et les analyses ont été réalisées avec la version 3.3 du logiciel R (R Core Team, 2017). Le package 'actuar' a été utilisé pour définir la loi de Poisson tronquée en zéro (Goulet et al., 2017) et le package 'pscl' a été utilisé pour définir le modèle enflé en zéro (Jackman et al., 2015). La réalisation de certains graphes a nécessité l'utilisation des packages 'dplyr' (Wickham et Francois, 2016), 'ggplot2' (Wickham et Chang, 2016) et 'RColorBrewer' (Neuwirth, 2014). Certains tracés graphiques ont également nécessité l'utilisation de fonctions disponibles sur Internet (Cookbook for R, 2016 ; Stack Overflow, 2012).

3) Résultats

a) <u>Impact sur les modèles logistiques d'une sensibilité de détection imparfaite mais homogène</u>

Pour une prévalence intra-unité moyenne fixée, lorsque la sensibilité de détection est imparfaite mais est la même pour toutes les unités épidémiologiques, **le modèle logistique identifie correctement et systématiquement le facteur X comme étant un facteur de risque**, et ce quelles que soient les valeurs de la sensibilité de détection dans les unités épidémiologiques (résultats non présentés).

En revanche, **l'odds ratio associé au facteur X pour le modèle logistique est systématiquement sous-estimé**, et **ce biais tend vers 0 lorsque la sensibilité de détection tend vers 1** (Figure 8). En d'autres termes, plus la sensibilité de détection est faible, plus l'odds ratio de la présence de la maladie associé au facteur X et estimé par le modèle logistique sous-estime la valeur réelle de cet odds ratio (Équation 5). Par exemple, comme illustré en Figure 8, pour une sensibilité de détection de 50%, un odds ratio réel de 2 (courbe bleue) aura tendance à être sous-estimé de près de 10% (soit un odds ratio estimé de 1,8).

De plus, **plus la force réelle de l'association entre la présence de la maladie et la présence de la variable X est forte** (c'est-à-dire plus l'odds ratio associé au facteur X est élevé), **plus ce biais va être marqué**, c'est-à-dire plus l'impact d'un défaut de sensibilité va être fort (Figure 8). Par exemple, comme illustré en Figure 8, pour une sensibilité de détection de 50%, un odds ratio réel de 2 (courbe bleue) aura tendance à être sous-estimé de près de 10% (valeur estimée de 1,8), tandis qu'un odds ratio réel de 5 (courbe verte) aura tendance à être sous-estimé de près de 40% (valeur estimée de 3).

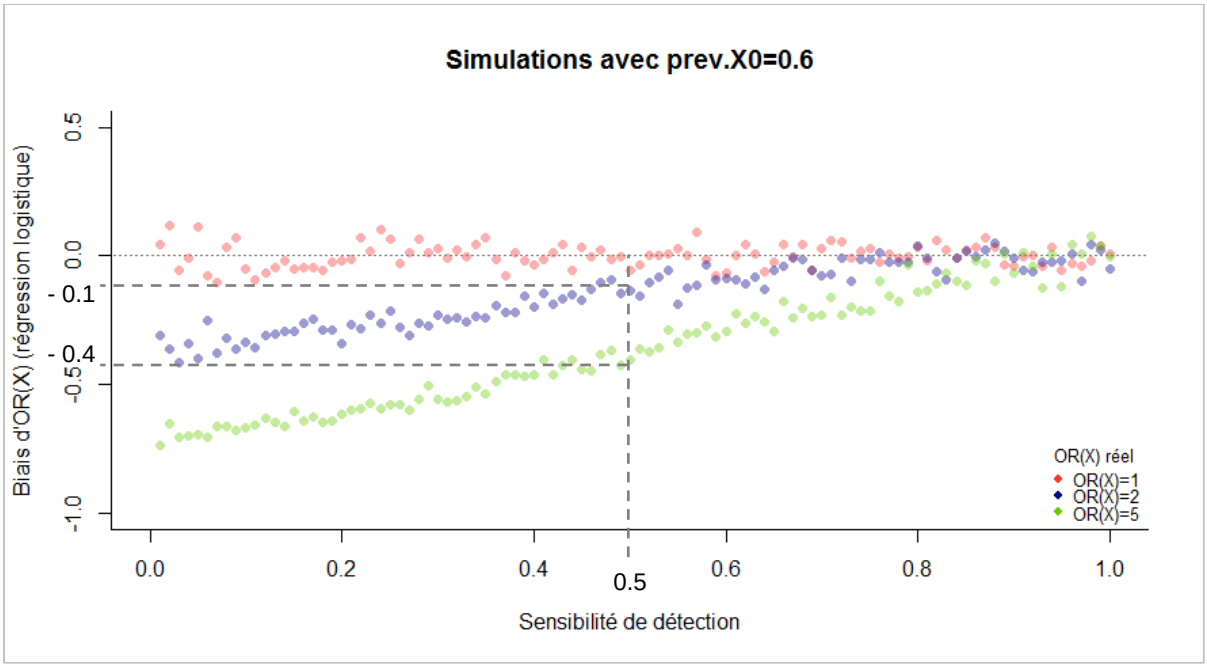


Figure 8 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs réelles d'OR(X) lorsque prev.X0=0,6 et M=4

En outre, **plus la probabilité de présence de la maladie est grande malgré l'absence du facteur X** (c'est-à-dire plus prev.X0 est élevé), **plus le modèle logistique sous-estime la valeur de l'odds ratio** (Figure 9).

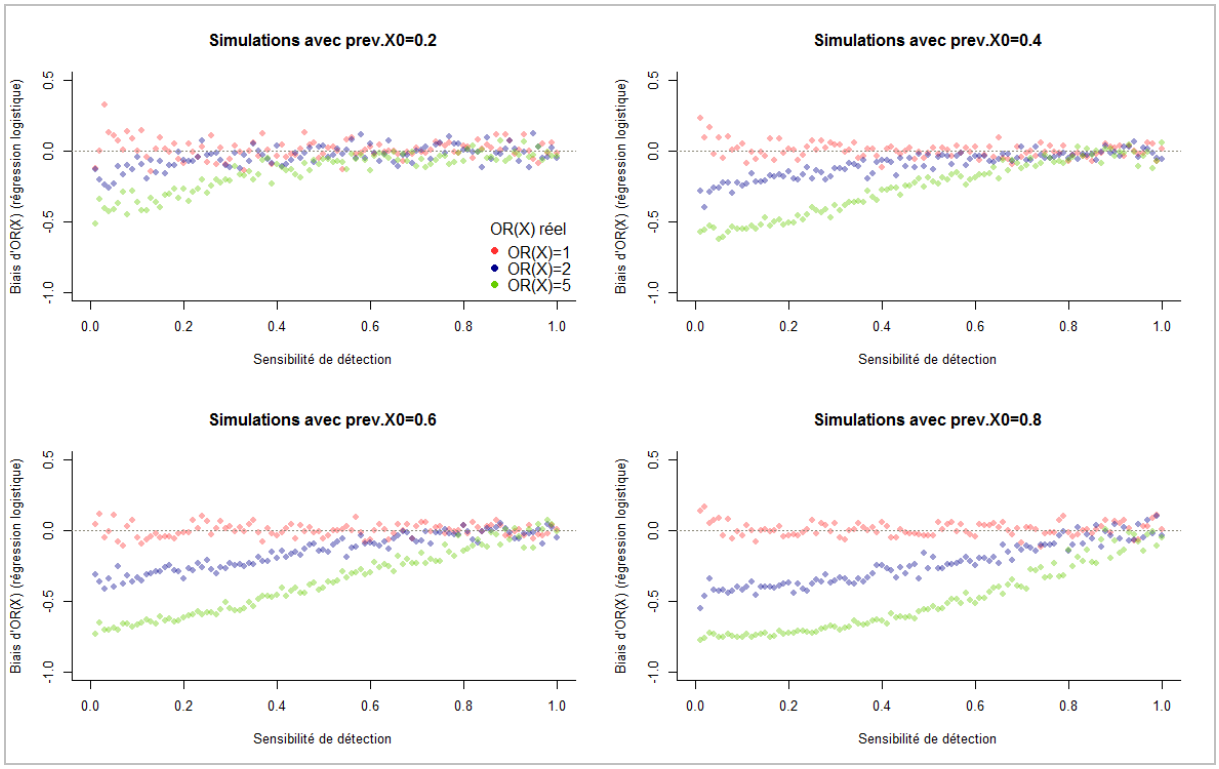


Figure 9 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4

Lorsque la prévalence intra-unité moyenne augmente, c'est-à-dire lorsque le nombre moyen d'unités élémentaires malades par unité épidémiologique malade augmente, **le biais tend rapidement vers 0** même pour de petites sensibilités de détection (Figure 10). Par exemple, comme illustré en Figure 10, pour une sensibilité de détection de 50%, lorsqu'il y a en moyenne 5 unités élémentaires malades par unité épidémiologique malade (courbe bleue) l'odds ratio aura tendance à être sous-estimé de près de 20%, tandis que lorsqu'il y a en moyenne 15 unités élémentaires malades par unité épidémiologique malade (courbe verte) l'odds ratio estimé aura tendance à être proche de la valeur réelle.

Par ailleurs, **les tendances évoquées précédemment sont d'autant plus marquées que le nombre moyen d'unités élémentaires malades par unité épidémiologique malade est faible**, ou que la prévalence intra-unité moyenne diminue : **plus la force réelle de l'association entre la présence de la maladie et la présence de la variable X est forte**, **plus le biais va être marqué**, et **plus la probabilité de présence de la maladie est grande malgré l'absence du facteur X**, **plus le modèle logistique sous-estime la valeur de l'odds ratio** (Annexe 1).

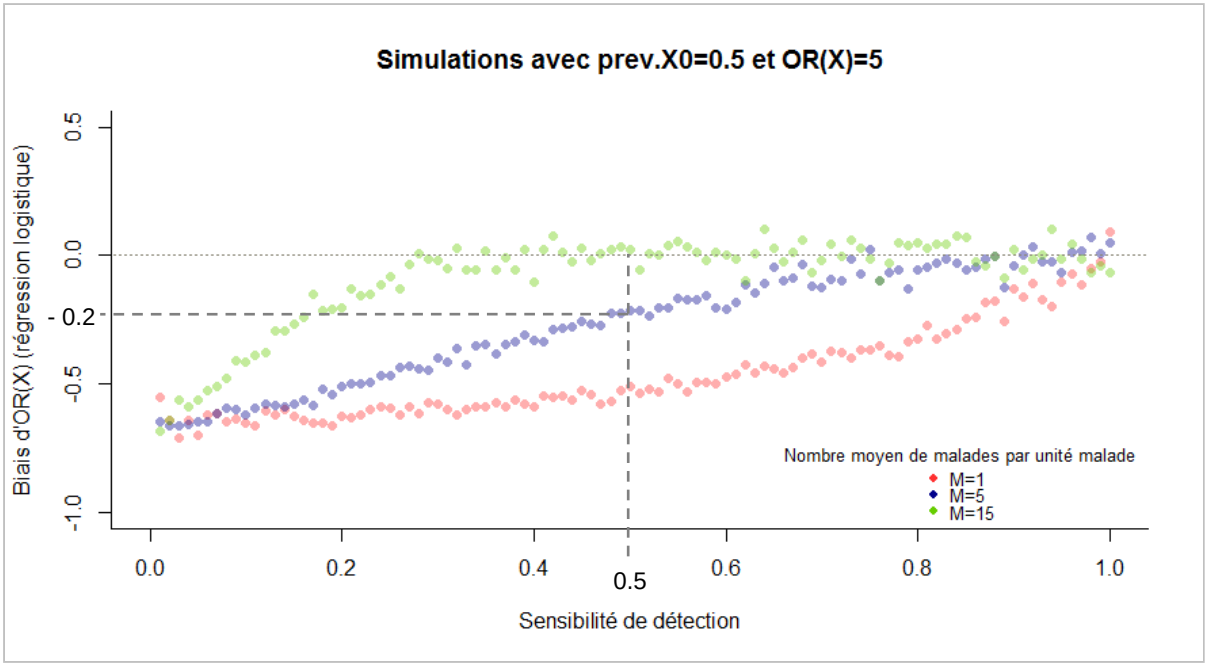


Figure 10 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade lorsque prev.X0=0,5 et OR(X)=5

Sensibilité de détection imparfaite mais homogène : impacts sur les modèles logistiques

Identification du facteur de risque comme étant un facteur de risque ?

Le facteur de risque X est systématiquement identifié comme étant un facteur de risque.

Estimation correcte de l'odds ratio associé au facteur de risque ?

Diminution de la valeur absolue du biais (mais sous-estimation de l'odds ratio) lorsque :

- la sensibilité de détection augmente ;
- l'odds ratio réel diminue ;
- la probabilité de présence de la maladie en l'absence du facteur de risque X diminue ;
- la prévalence intra-unité moyenne (c'est-à-dire le nombre moyen d'unités élémentaires malades dans une unité épidémiologique malade) augmente.

b) <u>Impact sur les modèles logistiques d'une sensibilité de détection imparfaite et hétérogène</u>

Pour toutes les valeurs de prev.X0 et d'OR(X) testées, **le modèle logistique identifie correctement et systématiquement le facteur X comme étant un facteur de risque**, et ce quelles que soient les valeurs de la sensibilité de détection dans les unités épidémiologiques et quelle que soit la prévalence intra-unité moyenne (résultats non présentés). En revanche, en accord avec ce qui a été montré dans le cas d'une sensibilité imparfaite mais homogène, **pour une prévalence intra-unité moyenne donnée**, **l'odds ratio est systématiquement sous-estimé et ce d'autant plus que la sensibilité est mauvaise dans une des sous-populations** (Figure 11). Il est intéressant de noter que la diagonale de la Figure 11 correspond au cas particulier où la détection est imparfaite mais homogène (la sensibilité de la détection dans les unités épidémiologiques avec le facteur Y est la même que celle dans les unités épidémiologiques sans le facteur Y).

De plus, dans les cas où la sensibilité de détection est hétérogène entre les unités épidémiologiques, **le biais relatif se rapproche de 0 dès lors que la sensibilité augmente dans une partie des unités épidémiologiques même si la sensibilité de détection est mauvaise dans l'autre partie des unités épidémiologiques**. Dans les conditions correspondant aux données simulées utilisées pour générer la Figure 11, cela signifie que du moment que la sensibilité de détection est au moins de 75 % dans les unités épidémiologiques avec Y (correspondant à 40% des unités épidémiologiques), le facteur de risque X est correctement identifié comme facteur de risque par le modèle logistique, et son odds ratio n'est pas sous-estimé de plus de 20%, même si la sensibilité de la détection est très mauvaise dans les unités épidémiologiques sans Y (situation correspondant au rectangle blanc en pointillés sur la Figure 11).

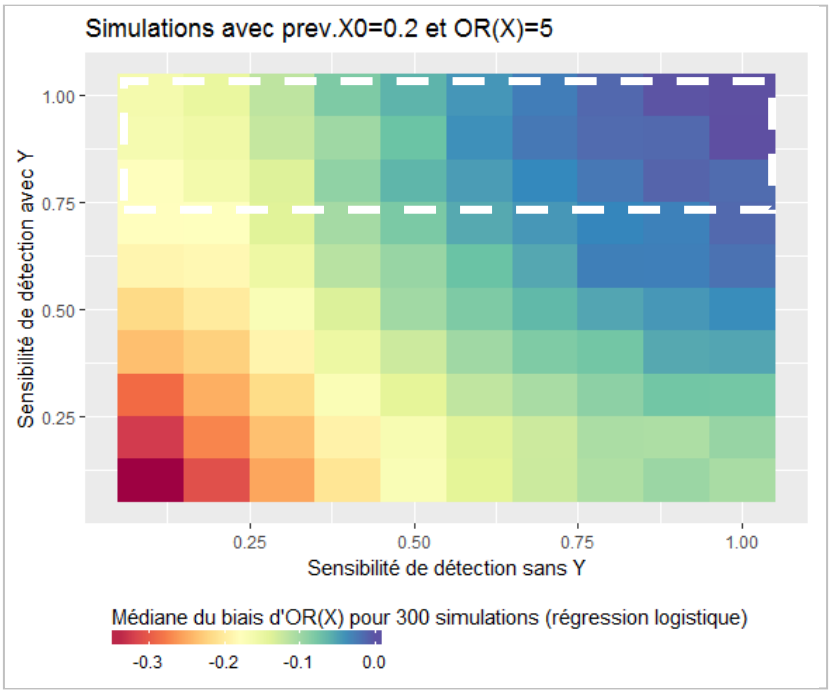


Figure 11 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4

Cependant, dès lors que la **probabilité de présence de la maladie en l'absence du facteur X augmente** et que la **valeur réelle de l'odds ratio associé au facteur X est importante**, il faut une **sensibilité de détection quasi parfaite dans toutes les unités épidémiologiques afin que le biais relatif tende vers 0** et donc que la valeur estimée de l'odds ratio associé au facteur X tende vers la valeur réelle (Figure 12).

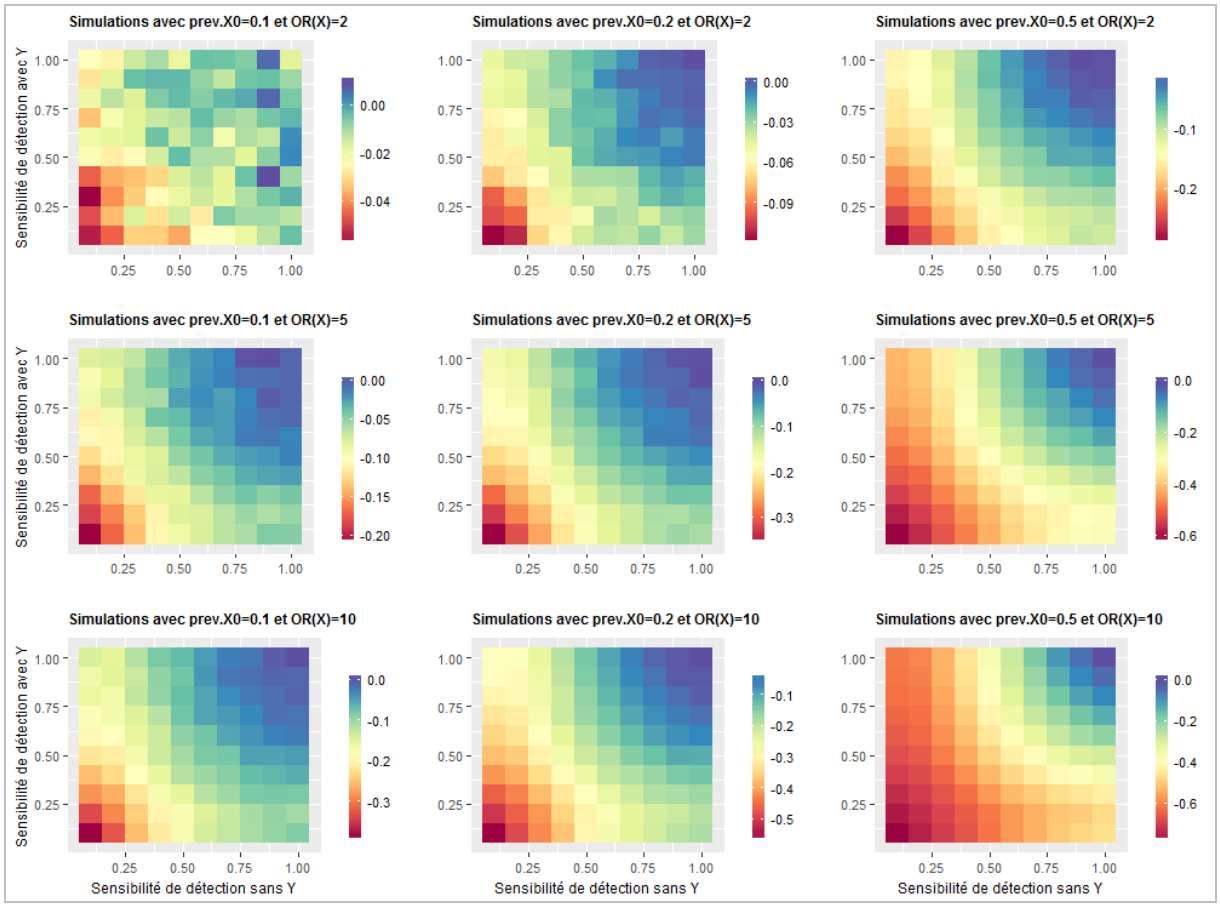


Figure 12 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4
Les barres à droite des graphes indiquent les valeurs médianes de biais($OR(X)_{logist}$) pour 300 simulations.

Lorsque la prévalence intra-unité moyenne augmente, c'est-à-dire lorsque le nombre moyen d'unités élémentaires malades par unité épidémiologique malade augmente, **le biais relatif diminue** (Figure 13). De plus, **lorsque la prévalence intra-unité moyenne diminue**, **plus l'odds ratio réel associé au facteur X est grand**, **plus le biais relatif va être marqué** (Figure 13), et **plus la sensibilité de détection est parfaite**, **plus le biais relatif diminue** (sur la Figure 13, comparaison du couple « sensib.Y0=0,6 et sensib.Y1=0,9 » par rapport au couple « sensib.Y0=0,3 et sensib.Y1=0,9 »).

Par ailleurs, **plus la probabilité de présence de la maladie est grande malgré l'absence du facteur X**, **plus le modèle logistique sous-estime la valeur de l'odds ratio** (Annexe 2).

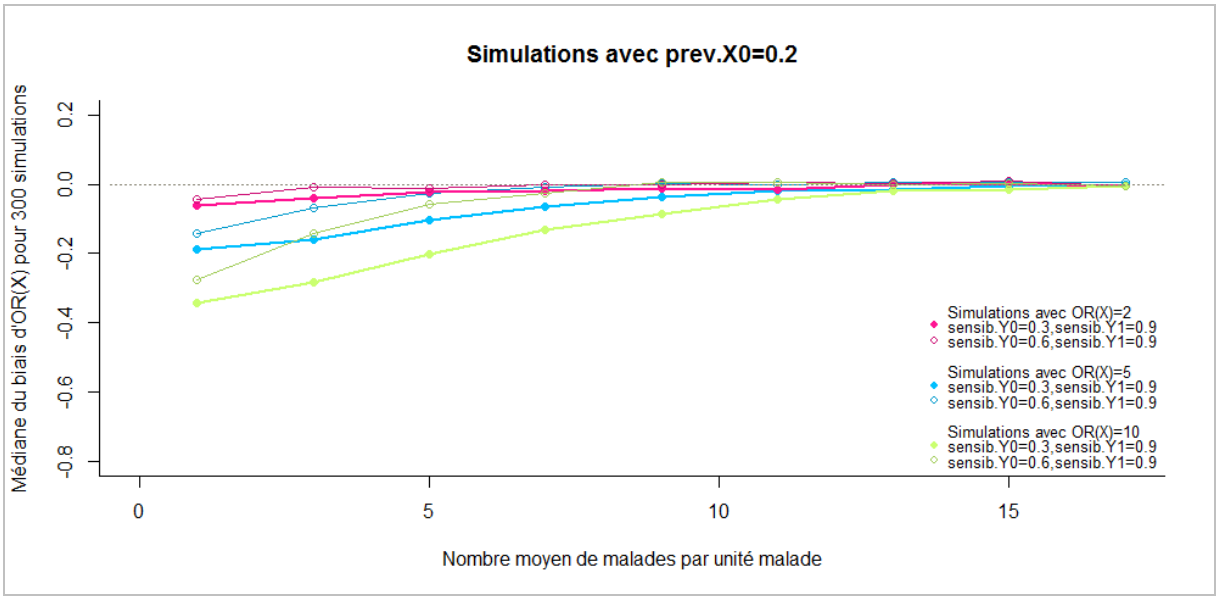


Figure 13 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection et valeurs réelles d'OR(X) lorsque prev.X0=0,2

Le facteur Y, qui fait varier la sensibilité de détection entre les unités épidémiologiques, peut être identifié à tort par le modèle logistique comme étant un facteur de risque, pour une prévalence intra-unité moyenne donnée (Figure 14), alors qu'il s'agit en réalité d'un facteur de confusion. Lorsque la sensibilité de détection est identique entre les unités épidémiologiques, le facteur Y n'a pas d'influence, et il n'est donc pas identifié comme étant un facteur de risque (diagonale rouge sur la Figure 14, correspondant à sensib.Y0=sensib.Y1). **Dès que la sensibilité de détection est hétérogène entre les unités épidémiologiques, donc dépendant de la présence du facteur Y, le modèle logistique identifie Y comme étant un facteur de risque, et ce avec une probabilité d'autant plus grande que la différence de sensibilité entre les unités avec ou sans Y est grande**. Cependant, **pour des sensibilités de détection élevées même si hétérogènes, la probabilité que le modèle logistique identifie à tort Y comme étant un facteur de risque diminue**. Dans les conditions correspondant aux données simulées utilisées pour générer la Figure 14, cela signifie que du moment que la sensibilité de détection est au moins de 75 % dans l'ensemble des unités épidémiologiques, la probabilité que le facteur Y soit identifié à tort comme facteur de risque par le modèle logistique est inférieure à 40% (situation correspondant au rectangle blanc en pointillés sur la Figure 14).

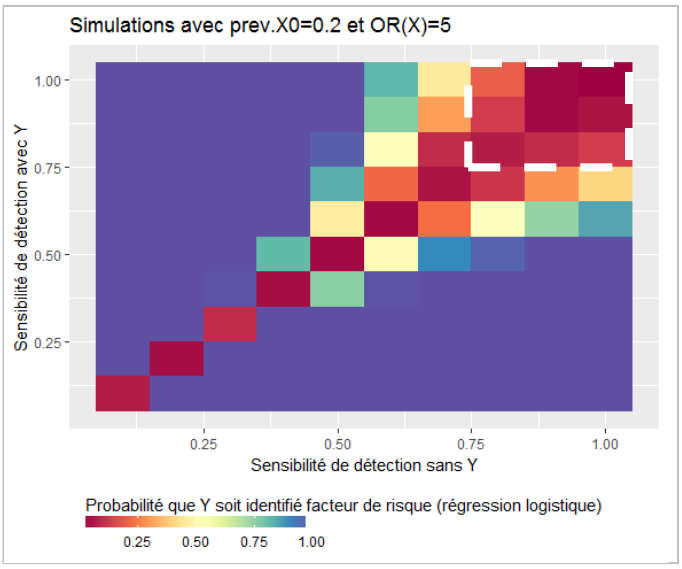


Figure 14 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4

Cependant, dès lors que la **probabilité de présence de la maladie en l'absence du facteur X augmente** et que la **valeur réelle de l'odds ratio associé au facteur X est importante**, il faut une **sensibilité de détection proche de 1 dans toutes les unités épidémiologiques afin que la probabilité que le facteur Y soit identifié comme facteur de risque diminue** (Figure 15).

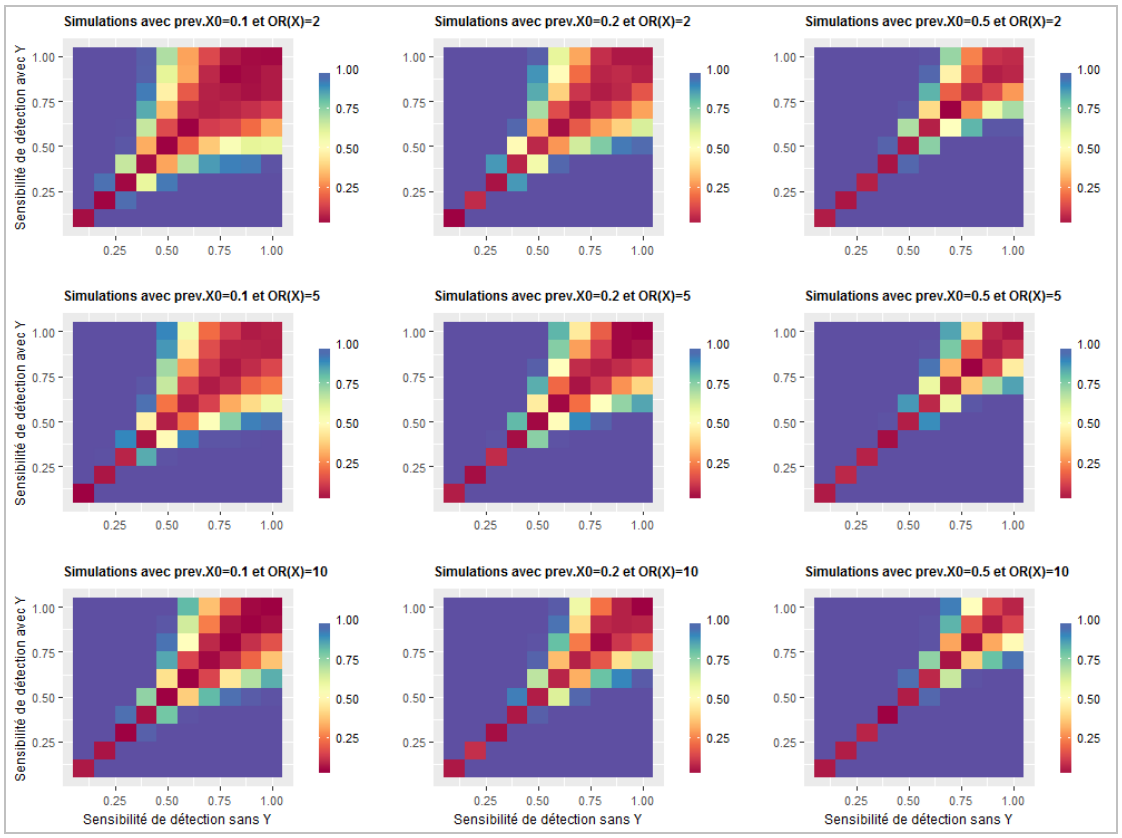


Figure 15 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4 ***Les barres à droite des graphes indiquent la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique (sur 300 simulations).***

Plus la prévalence intra-unité moyenne augmente, soit lorsque le nombre moyen d'unités élémentaires malades par unité épidémiologique malade augmente, **moins il est probable que le modèle logistique identifie le facteur Y comme étant un facteur de risque** (Figure 16). De plus, **plus la sensibilité de détection est parfaite**, **moins il est probable que le modèle logistique identifie le facteur Y comme étant un facteur de risque**, et ce d'autant que la sensibilité de détection est semblable dans l'ensemble des unités épidémiologiques (sur la Figure 16, comparaison du couple « sensib.Y0=0,6 et sensib.Y1=0,9 » par rapport au couple « sensib.Y0=0,3 et sensib.Y1=0,9 »).

En outre, **pour des sensibilités de détection données**, **plus l'odds ratio réel associé au facteur X est grand et plus il est probable que le modèle logistique identifie le facteur Y comme étant un facteur de risque** (Figure 16).

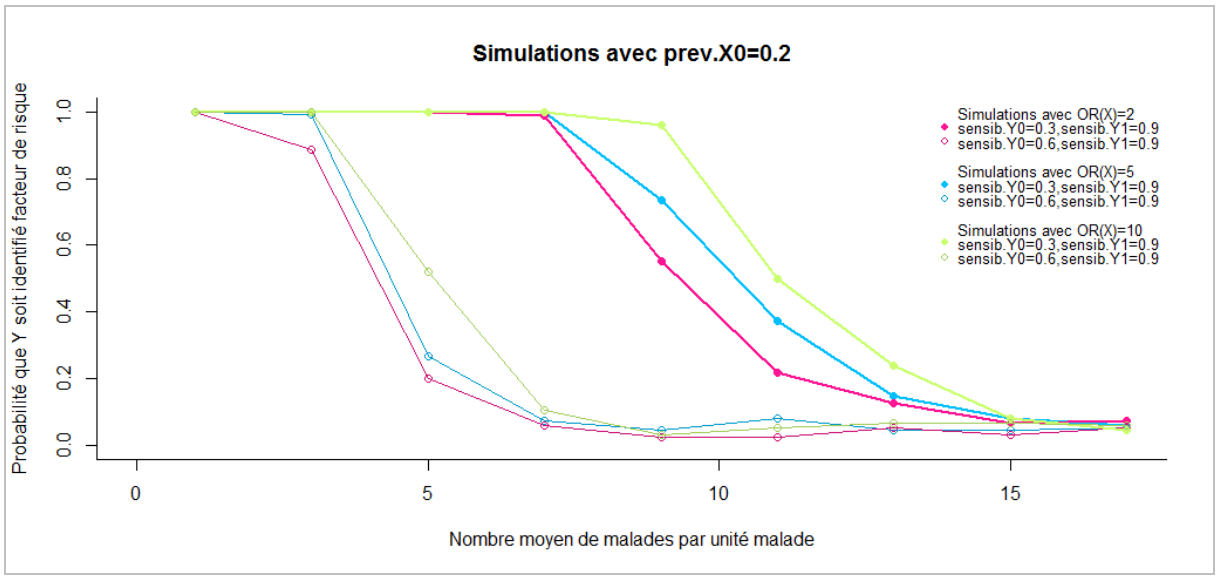


Figure 16 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection et valeurs réelles d'OR(X) lorsque prev.X0=0,2

Par ailleurs, **plus la probabilité de présence de la maladie est grande malgré l'absence du facteur X**, **plus il est probable que le modèle logistique identifie le facteur Y comme étant un facteur de risque** (Annexe 3).

Sensibilité de détection imparfaite et hétérogène : impacts sur les modèles logistiques

Identification du facteur de risque comme étant un facteur de risque ?

Le facteur de risque X est systématiquement identifié comme étant un facteur de risque.

Estimation correcte de l'odds ratio associé au facteur de risque ?

Diminution de la valeur absolue du biais (mais sous-estimation de l'odds ratio) lorsque :

- la sensibilité de détection augmente dans l'ensemble des unités épidémiologiques ;
- la sensibilité de détection est correcte dans une partie des unités épidémiologiques, même si la sensibilité de détection est mauvaise dans l'autre partie des unités épidémiologiques ;
- l'odds ratio réel et la probabilité de présence de la maladie en l'absence du facteur X diminuent ;
- la prévalence intra-unité moyenne (c'est-à-dire le nombre moyen d'unités élémentaires malades dans une unité épidémiologique malade) augmente.

Identification du facteur de confusion comme étant un facteur de risque ?

Le facteur de confusion Y peut être identifié comme étant un facteur de risque dès lors que la sensibilité de détection est différente entre les unités épidémiologiques où le facteur Y est présent et celles où il est absent.

Diminution de la probabilité que le facteur de confusion Y soit identifié à tort comme un facteur de risque lorsque :

- l'hétérogénéité de la sensibilité de détection entre les unités épidémiologiques où Y est absent et celles où il est présent diminue ;
- la sensibilité de détection augmente dans l'ensemble des unités épidémiologiques ;
- l'odds ratio réel associé au facteur X et la probabilité de présence de la maladie en l'absence du facteur X diminuent ;
- la prévalence intra-unité moyenne augmente.

c) Impact sur les modèles logistiques d'une sensibilité de détection imparfaite lorsque les facteurs de risque et de confusion sont identiques

Pour toutes les valeurs de prev.X0 et d'OR(X) testées, **le modèle logistique identifie correctement le facteur X comme étant un facteur de risque**, et ce quelles que soient les valeurs de la sensibilité de détection dans les unités épidémiologiques.

Dans certaines situations, on note cependant que **si la sensibilité de détection est faible dans les unités épidémiologiques où X est présent et correcte dans les unités épidémiologiques où X est absent**, alors **X peut ne pas être identifié comme étant un facteur de risque** (Figure 17). Dans les conditions correspondant aux données simulées utilisées pour générer la Figure 17, on a d'un côté les unités épidémiologiques où X est absent avec une probabilité de présence de la maladie moindre (prev.X0=0,2), et de l'autre les unités épidémiologiques où X est présent avec une probabilité de présence de la maladie plus importante (prev.X1=0,56). Pour sensib.X0=70% (bonne sensibilité de détection de chacune des unités élémentaires malades dans les unités épidémiologiques où X est absent) et sensib.X1=10% (mauvaise sensibilité de détection dans les unités épidémiologiques où X est présent), la probabilité que X soit correctement identifié comme étant un facteur de risque est inférieure à 25% (situation correspondant au rectangle blanc en pointillés sur la Figure 17).

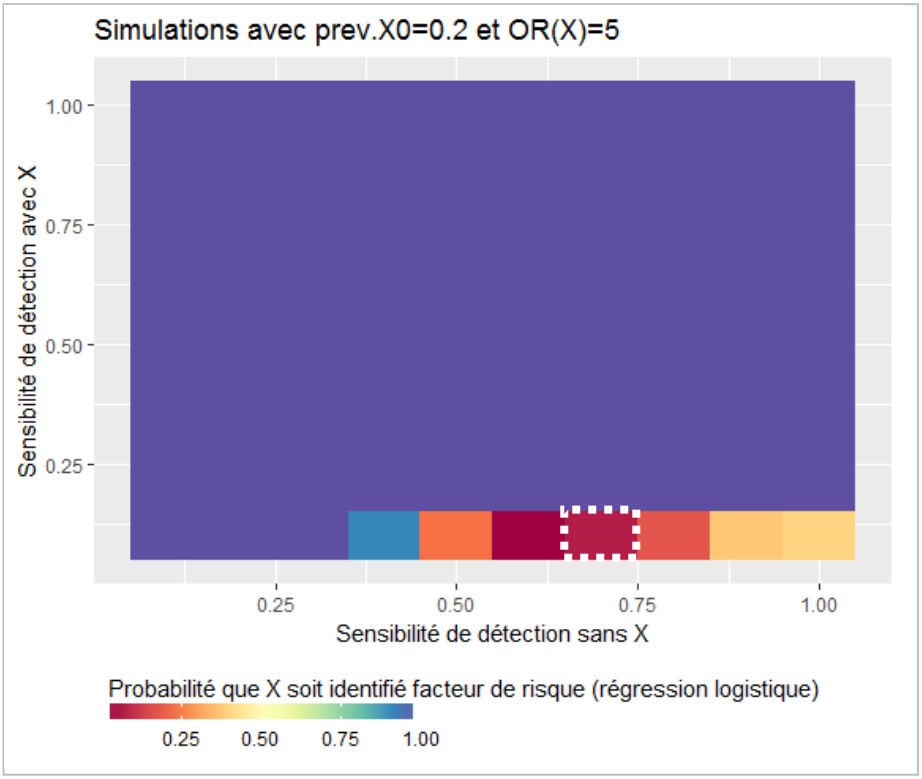


Figure 17 : Evolution de la probabilité que le facteur X soit identifié comme étant un facteur de risque par un modèle logistique en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4

Concernant l'estimation de l'odds ratio associé au facteur X, **pour une prévalence intra-unité moyenne donnée**, **le biais de l'odds ratio peut prendre des valeurs variant de -0,8 à 2,5** (Figure 18), et ceci pour toutes les valeurs de prev.X0 et d'OR(X) testées. Ce biais tend

vers 0 lorsque la sensibilité de détection s'améliore dans l'ensemble des unités épidémiologiques (couleur orangée dans l'angle supérieur droit de la Figure 18).

Contrairement aux situations envisagées précédemment (lorsque le facteur de risque et le facteur de confusion sont deux facteurs distincts) où le biais était négatif et tendait vers 0 lorsque la sensibilité de détection s'améliorait, dans la situation présente **le biais peut être positif, notamment lorsque la sensibilité de détection est faible dans les unités épidémiologiques où le facteur X est absent**. Ceci signifie, étant donnée l'Équation 5 pour le calcul du biais, que si la sensibilité de détection est faible dans les unités épidémiologiques où le facteur X est absent, alors **l'odds ratio calculé par le modèle logistique surestime la valeur réelle**, et ce **d'autant plus que la sensibilité de détection est importante dans les unités épidémiologiques où le facteur X est présent**.

Dans les conditions correspondant aux données simulées utilisées pour générer la Figure 18, on a d'un côté les unités épidémiologiques où X est absent avec une probabilité de présence de la maladie moindre (prev.X0=0,2), et de l'autre les unités épidémiologiques où X est présent avec une probabilité de présence de la maladie plus importante (prev.X1=0,56). Pour sensib.X0<25% et sensib.X1>50%, la détection de chacune des unités élémentaires malades est mauvaise dans les unités épidémiologiques où X est absent et est meilleure dans les unités épidémiologiques où X est présent, et dans ce cas le biais relatif médian de l'odds ratio associé à X est supérieur à 1 (situation correspondant au rectangle blanc en pointillés sur la Figure 18).

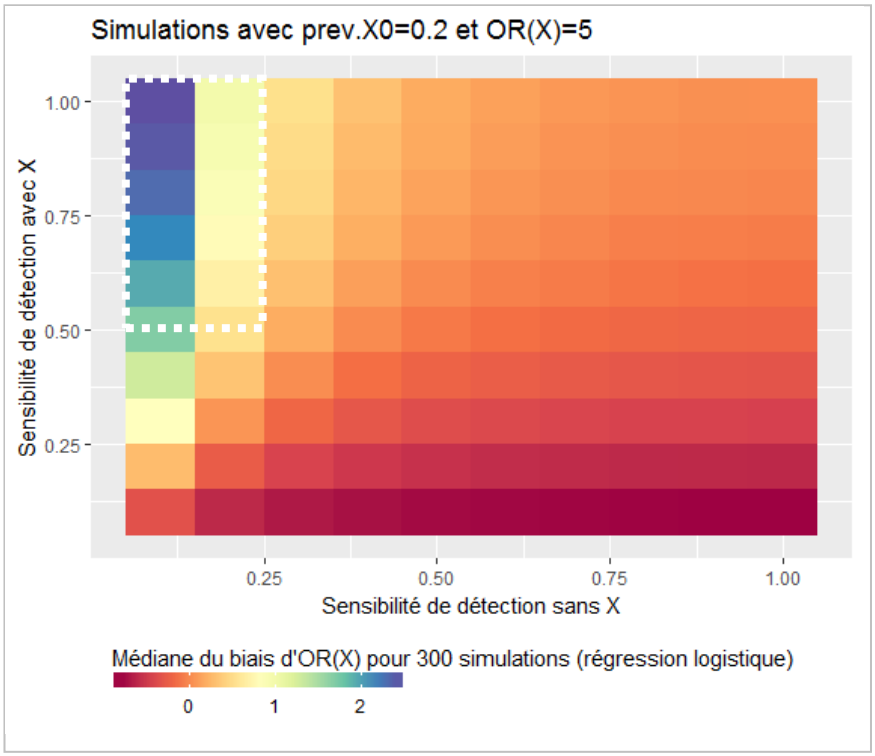


Figure 18 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4

Sensibilité de détection imparfaite lorsque les facteurs de risque et de confusion sont identiques : impacts sur les modèles logistiques

Identification du facteur de risque et de confusion comme étant un facteur de risque ?

Le facteur X est identifié comme étant un facteur de risque. Il peut ne pas l'être dans les situations où la sensibilité de détection est faible dans les unités épidémiologiques où X est présent et correcte dans les unités épidémiologiques où X est absent.

Estimation correcte de l'odds ratio associé au facteur de risque ?

Le biais de l'odds ratio peut prendre des valeurs négatives comme positives, et tend vers 0 lorsque la sensibilité de détection s'améliore dans l'ensemble des unités épidémiologique. L'odds ratio estimé surestime la valeur réelle notamment dans les situations où la sensibilité de détection est correcte dans les unités épidémiologiques où X est présent et faible dans les unités épidémiologiques où X est absent.

d) Bénéfices potentiels à utiliser un modèle de Poisson enflé en zéro si la détection est imparfaite voire hétérogène

Pour mémoire, la partie « logistique » du modèle de Poisson enflé en zéro correspond à l'analyse de l'apparition de la maladie dans les unités épidémiologiques, qui dépend du facteur de risque X, tandis que la partie « comptage » correspond à l'analyse du nombre d'unités élémentaires détectées comme étant malades, qui dépend de la sensibilité de détection et donc du facteur de confusion Y.

- **Impact sur les modèles de Poisson enflés en zéro d'une sensibilité de détection imparfaite mais homogène**

Pour une prévalence intra-unité moyenne fixée, lorsque la sensibilité de détection est la même pour toutes les unités épidémiologiques, **le biais de l'odds ratio associé au facteur X pour le modèle de Poisson enflé en zéro tend très rapidement vers 0 lorsque la sensibilité de détection augmente**, quelle que soit la valeur simulée de l'odds ratio associé au facteur X (Figure 19).

La tendance de l'évolution pour des sensibilités très faibles est moins précise que pour le modèle logistique, avec un biais qui oscille autour de 0 avec des valeurs négatives comme positives. Contrairement au modèle logistique, le modèle de Poisson enflé en zéro est donc **très peu biaisé par la variation de sensibilité de détection** (excepté pour des sensibilités très faibles, inférieures à 0,2 pour ce qui est des données illustrées en Figure 19), et ce quelle que soit la valeur réelle de l'odds ratio associé au facteur X.

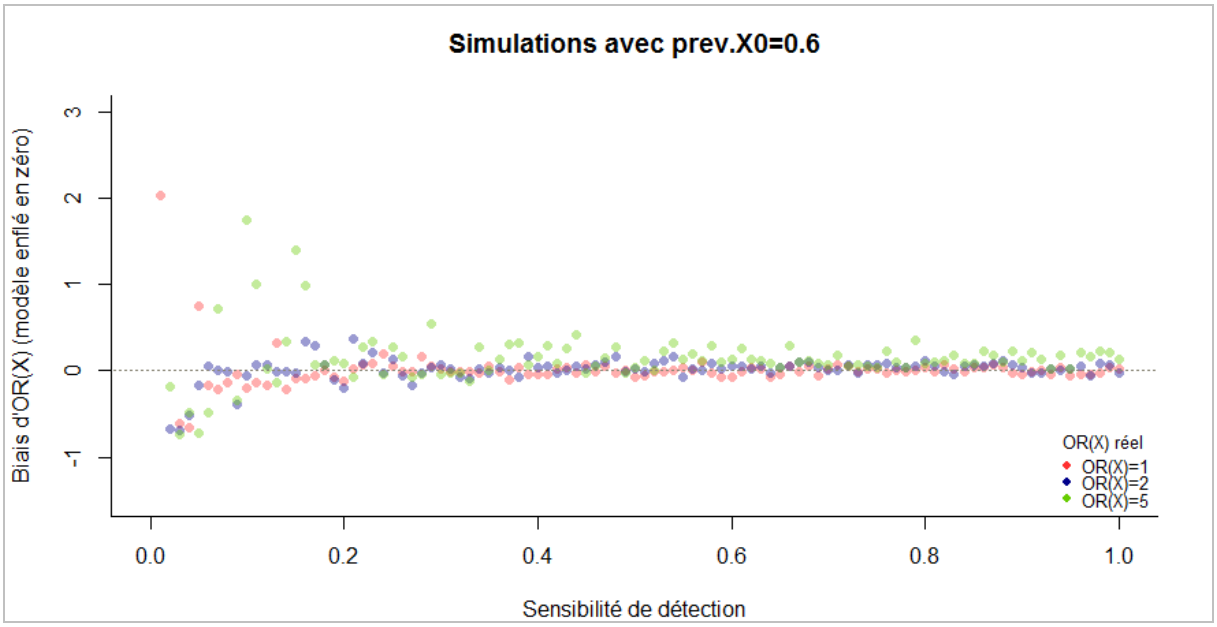


Figure 19 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs réelles d'OR(X) lorsque prev.X0=0,6 et M=4

Cependant, **plus la probabilité de présence de la maladie est grande malgré l'absence du facteur X** (c'est-à-dire plus prev.X0 est élevé), **plus le modèle de Poisson enflé en zéro**

surestime la valeur de l'odds ratio lorsque la valeur réelle de l'odds ratio augmente (Annexe 4).

Pour des valeurs moyennes supérieures à trois unités élémentaires malades par unité épidémiologique malade (M ≥ 3), **le biais de l'odds ratio associé au facteur X pour le modèle de Poisson enflé en zéro ne dépend pas de la prévalence intra-unité moyenne**, c'est-à-dire ne dépend pas du nombre moyen d'unités élémentaires malades par unité épidémiologique malade (Figure 20). En revanche, **lorsqu'il n'y a en moyenne qu'un seul malade par unité épidémiologique malade, le biais augmente de manière importante** (pour la série de données simulées représentée en Figure 20, la valeur moyenne du biais est de 36 840, la médiane est de 8303, et les valeurs extrêmes [-0,89 ; 535 181], toutes sensibilités de détection confondues). De manière moins extrême, lorsqu'il y a en moyenne deux malades par unité épidémiologique malade, et pour la série de données simulées représentée en Figure 20, la valeur moyenne du biais est de 2,5, la médiane est de 0,3, et les valeurs extrêmes [-0,68 ; 189]. A partir d'une moyenne de trois malades par unité épidémiologique malade, la tendance est la même quelle que soit la valeur de cette moyenne de malades, et est celle représentée en Figure 20.

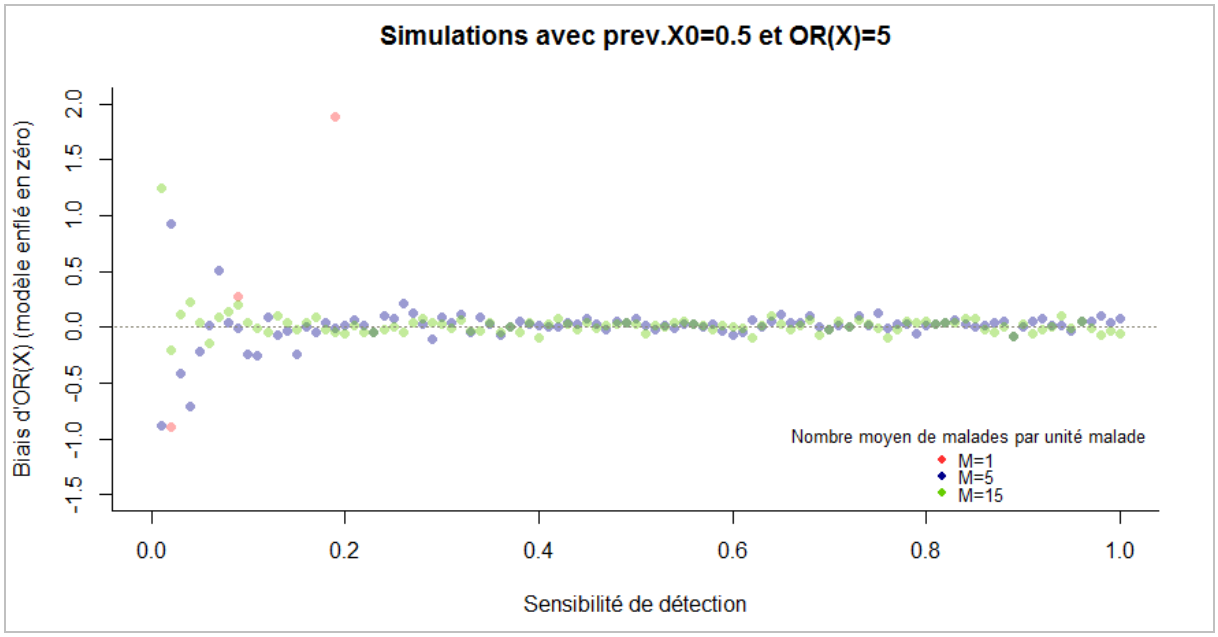


Figure 20 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade lorsque prev.X0=0,5 et OR(X)=5

Ce biais très important lorsqu'il n'y a en moyenne qu'un seul malade par unité épidémiologique malade se retrouve quelle que soit la probabilité de présence de la maladie en l'absence du facteur X, même si les valeurs prises sont moins extrêmes lorsque cette probabilité est faible (Annexe 5).

Sensibilité de détection imparfaite mais homogène :
impacts sur les modèles de Poisson enflés en zéro

Lorsque la ***prévalence intra-unité moyenne est fixée*** (soit lorsque le nombre moyen d'unités élémentaires malades dans une unité épidémiologique malade est fixé) :

- biais de l'odds ratio nul à quasi nul même pour des valeurs de sensibilité de détection médiocre (> 0,3) mais important pour des sensibilités très faibles (< 0,2) ;
- plus l'odds ratio réel est élevé et plus la probabilité de présence de la maladie est grande malgré l'absence du facteur X, plus l'odds ratio est surestimé.

Lorsque la ***prévalence intra-unité moyenne augmente***, cela n'a pas ou peu d'influence sur l'estimation de l'odds ratio. En revanche, les très faibles prévalences intra-unité moyennes (en moyenne une seule ou deux unités élémentaires malades par unité épidémiologique malade), biaisent très fortement l'estimation de l'odds ratio.

- **Impact sur les modèles de Poisson enflés en zéro d'une sensibilité de détection imparfaite et hétérogène**

Pour toutes les valeurs de prev.X0 et d'OR(X) testées, **la partie « logistique » du modèle enflé en zéro identifie correctement le facteur X comme étant un facteur de risque**, et ce quelles que soient les valeurs de la sensibilité de détection dans les unités épidémiologiques et quelle que soit la prévalence intra-unité moyenne, tandis que **la probabilité que la partie « comptage » l'identifie comme facteur de risque est nulle à quasi nulle** (résultats non présentés). Une exception cependant, **lorsqu'il n'y a en moyenne qu'un seul malade par unité épidémiologique malade, on obtient des résultats « aberrants »** : le facteur X peut ne pas être correctement identifié par la partie « logistique » comme étant un facteur de risque, et au contraire être identifié par la partie « comptage » comme étant un facteur de risque.

Concernant **la valeur de l'odds ratio associé à ce facteur et calculée par la partie « logistique » du modèle**, elle est **toujours proche de la valeur réelle, le biais étant nul à quasi nul** quelles que soient les valeurs de la sensibilité de détection dans les unités épidémiologiques et quelle que soit la prévalence intra-unité moyenne (résultats non présentés). Une exception cependant, **lorsqu'il n'y a en moyenne qu'un seul malade par unité épidémiologique malade, on obtient de nouveau des résultats « aberrants »** puisque le biais dans le calcul de l'odds ratio associé à X peut prendre des valeurs de l'ordre du millier.

Pour toutes les valeurs de prev.X0 et d'OR(X) testées, **la probabilité que la partie « logistique » identifie le facteur Y comme facteur de risque est nulle à quasi nulle**, et ce quelles que soient les valeurs de la sensibilité de détection dans les unités épidémiologiques et quelle que soit la prévalence intra-unité moyenne, même pour une moyenne d'un seul malade par unité épidémiologique malade (résultats non présentés).

Quelle que soit la prévalence intra-unité moyenne, même pour une moyenne d'un seul malade par unité épidémiologique malade, **le facteur Y n'est pas identifié par la partie « comptage » du modèle comme étant facteur de risque lorsque la sensibilité de détection est la même dans toutes les unités épidémiologiques** (diagonale rouge sur la Figure 21), et est **correctement identifié comme facteur influençant le nombre d'unités élémentaires détectées dès que la sensibilité de détection est hétérogène entre les unités épidémiologiques** (zones bleu-violet sur la Figure 21). Le même type de graphe que la Figure 21 est obtenu pour toutes les valeurs de prev.X0 et d'OR(X) testées (résultats non présentés).

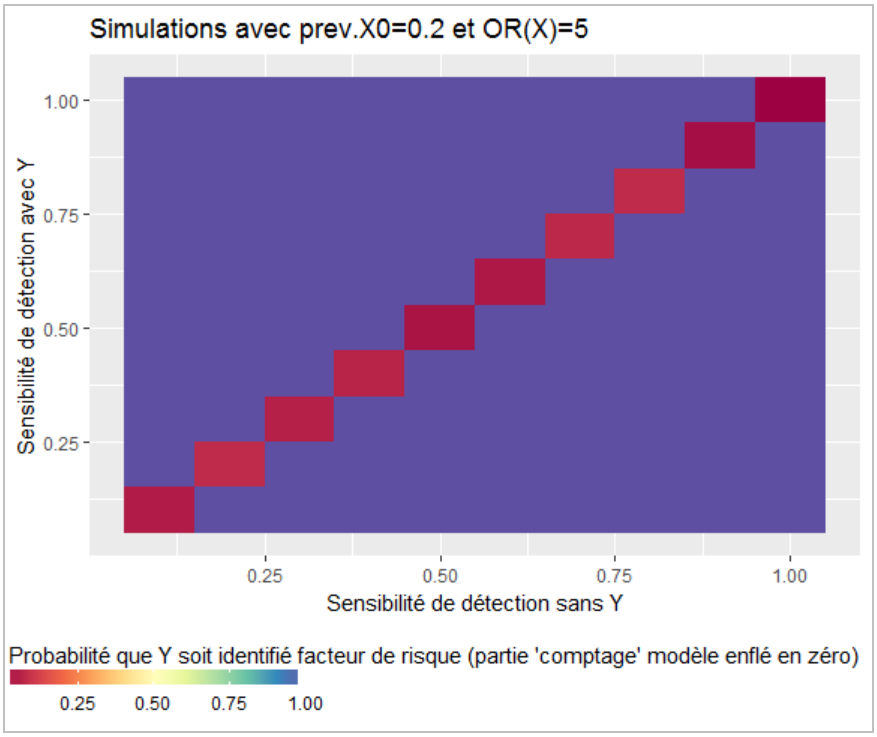


Figure 21 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par la partie « comptage » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4

Sensibilité de détection imparfaite et hétérogène : impacts sur les modèles de Poisson enflés en zéro

Identification du facteur de risque comme étant un facteur de risque ?

Le facteur de risque X est systématiquement identifié comme étant un facteur de risque par la partie « logistique » et pas par la partie « comptage ». Il n'y a pas d'effet de la prévalence intra-unité moyenne, si ce n'est des résultats « aberrants » lorsqu'il n'y a en moyenne qu'une seule unité élémentaire malade par unité épidémiologique malade.

Estimation correcte de l'odds ratio associé au facteur de risque ?

Le biais de l'odds ratio est nul à quasi nul. Il n'y a pas d'effet de la prévalence intra-unité moyenne, si ce n'est des résultats « aberrants » lorsqu'il n'y a en moyenne qu'une seule unité élémentaire malade par unité épidémiologique malade.

Identification du facteur de confusion comme étant un facteur de risque ?

Le facteur de confusion Y n'est pas identifié comme étant un facteur de risque par la partie « logistique », et n'est identifié par la partie « comptage » que dans les situations où la sensibilité de détection est différente entre les unités où Y est présent et celles où Y est absent. Il n'y a pas d'effet de la prévalence intra-unité moyenne, même lorsqu'il n'y a en moyenne qu'une seule unité élémentaire malade par unité épidémiologique malade.

- **Impact sur les modèles de Poisson enflés en zéro d'une sensibilité de détection imparfaite lorsque les facteurs de risque et de confusion sont identiques**

Pour toutes les valeurs de prev.X0 et d'OR(X) testées, **la partie « logistique » du modèle enflé en zéro identifie correctement le facteur X comme étant un facteur de risque**, et ce quelles que soient les valeurs de la sensibilité de détection dans les unités épidémiologiques (résultats non présentés), et **la valeur de l'odds ratio calculée par la partie « logistique » est toujours proche de la valeur réelle**, le biais étant nul à quasi nul quelles que soient les valeurs de la sensibilité de détection dans les unités épidémiologiques (résultats non présentés).

La partie « comptage » du modèle n'identifie pas le facteur X comme étant facteur de risque lorsque la sensibilité de détection est la même dans toutes les unités épidémiologiques (diagonale rouge sur la Figure 22), et **l'identifie correctement comme facteur influençant le nombre d'unités élémentaires détectées dès que la sensibilité de détection est hétérogène entre les unités épidémiologiques** (zones bleu-violet sur la Figure 22). Le même type de graphe que la Figure 22 est obtenu pour toutes les valeurs de prev.X0 et d'OR(X) testées (résultats non présentés).

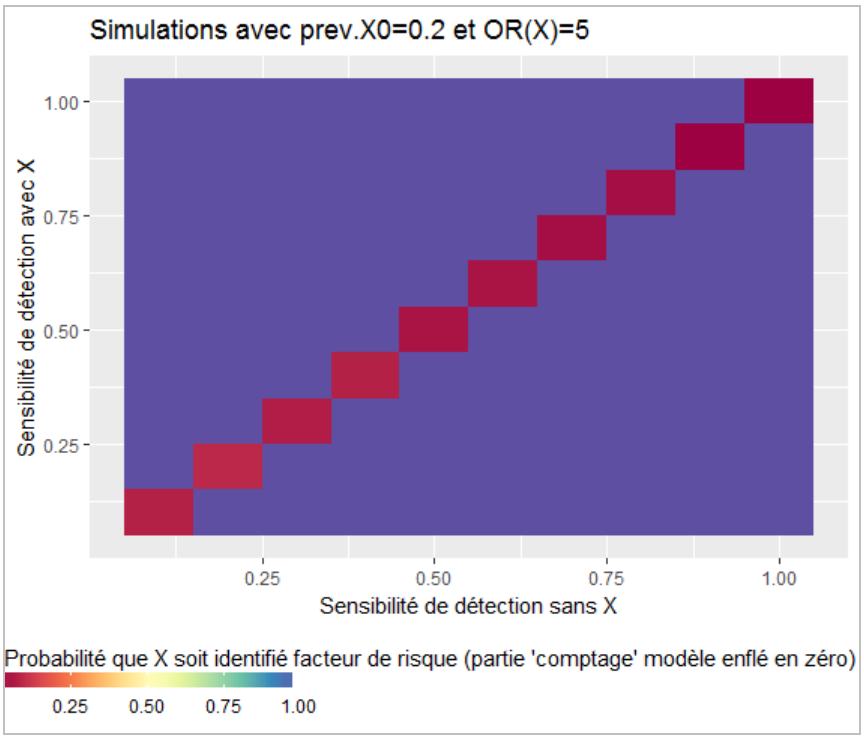


Figure 22 : Evolution de la probabilité que le facteur X soit identifié comme étant un facteur de risque par la partie « comptage » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection lorsque prev.X0=0,2 et OR(X)=5 et M=4

Sensibilité de détection imparfaite lorsque les facteurs de risque et de confusion sont identiques : impacts sur les modèles de Poisson enflés en zéro

Identification du facteur de risque et de confusion comme étant un facteur de risque ?

Le facteur X est identifié comme étant un facteur de risque par la partie « logistique », et n'est identifié par la partie « comptage » que dans les situations où la sensibilité de détection est différente entre les unités où X est présent et celles où X est absent.

Estimation correcte de l'odds ratio associé au facteur de risque ?

Le biais de l'odds ratio est nul à quasi nul.

4) Discussion

Cette section a permis d'étudier les impacts sur les modèles logistiques et les modèles de Poisson enflés en zéro d'une sensibilité de détection imparfaite. La sensibilité a été d'une part envisagée comme étant imparfaite mais homogène (la même pour toutes les unités épidémiologiques étudiées), et d'autre part hétérogène (c'est-à-dire différente entre les unités épidémiologiques, et dépendant d'un facteur de confusion). Le cas où le facteur de risque d'apparition de la maladie au sein d'une unité épidémiologique est identique au facteur de confusion a aussi été envisagé. Cette situation peut se retrouver sur le terrain, par exemple avec une conduite d'élevage qui rend le troupeau plus à risque pour une maladie, et qui dans le même temps permet une détection plus ou moins aisée de cette maladie.

En ce qui concerne le **modèle logistique**, il s'agit d'un **modèle fiable pour identifier un facteur qui est effectivement facteur de risque**, mais qui a **tendance à sous-estimer son odds ratio**. De manière générale, plus la sensibilité de détection est faible, plus le modèle logistique sous-estime la valeur réelle de l'odds ratio, et plus la sensibilité de détection s'améliore, plus l'odds ratio obtenu par régression logistique s'approche de la valeur réelle. En effet, plus la sensibilité de détection est faible, moins il y a d'unités élémentaires réellement malades au sein des unités épidémiologiques qui sont effectivement détectées. Ceci implique qu'il pourra y avoir des unités épidémiologiques malades qui ne seront pas détectées comme telles, si aucune des unités élémentaires malades n'est détectée. Ces faux négatifs vont biaiser le modèle logistique, qui va sous-estimer l'importance réelle du facteur de risque en question dans l'apparition de la maladie, et ce biais est d'autant plus important que l'importance réelle du facteur est grande.

La probabilité de présence de la maladie biaise aussi l'estimation de l'odds ratio. En effet, plus la probabilité de présence de la maladie dans les unités épidémiologiques où il n'y a pas le facteur de risque est grande, moins il y a de différence avec la probabilité de présence de la maladie dans les unités épidémiologiques où ce même facteur est présent, ce qui entraîne un biais dans le modèle logistique.

Dans les cas où la sensibilité de détection est hétérogène entre les unités épidémiologiques, il suffit que la sensibilité soit correcte dans une partie des unités épidémiologiques, pour que le biais tende à être nul même si la sensibilité de détection est mauvaise dans l'autre partie des unités épidémiologiques.

Par ailleurs, plus le nombre moyen d'unités élémentaires malades par unité épidémiologique malade augmente, plus la valeur de l'odds ratio estimée est proche de la valeur réelle. En effet, la probabilité qu'au moins une unité élémentaire malade soit détectée augmente avec le nombre d'unités élémentaires malades dans les unités épidémiologiques malades, et ce malgré une faible sensibilité de détection. Ceci implique que plus il y a de malades dans les unités épidémiologiques malades, plus toutes les unités épidémiologiques malades vont être détectées comme telles, et moins il y aura de faux négatifs. Cette diminution des faux négatifs permet une diminution du biais du modèle logistique.

En revanche, **le modèle logistique a aussi tendance à identifier à tort comme facteur de risque des facteurs qui n'en sont pas**. Lorsque la sensibilité de détection est homogène entre les unités épidémiologiques, et donc que sa valeur ne dépend pas du facteur de confusion, le modèle logistique n'identifie jamais ce facteur comme étant un facteur de risque. Cependant, dès que la sensibilité devient hétérogène entre les unités épidémiologiques, il est probable que le facteur de confusion soit identifié à tort comme étant un facteur de risque. Cette probabilité est d'autant plus faible que la sensibilité de détection est correcte, et d'autant plus qu'elle est semblable dans l'ensemble des unités épidémiologiques. En effet, si la sensibilité de détection est correcte, le nombre de faux négatifs diminue, et l'apparition de la maladie dans une unité épidémiologique est moins bien associée avec le facteur de confusion. Et, plus la sensibilité est la même dans l'ensemble des unités épidémiologiques, moins il y a de différence entre les unités concernant l'influence du facteur de confusion sur la maladie, et donc l'apparition de la maladie dans une unité épidémiologique est d'autant moins bien associée avec le facteur de confusion. De plus, la probabilité de détection du facteur de confusion comme étant à tort facteur de risque est aussi d'autant plus faible que la prévalence intra-unité moyenne augmente. En effet, plus il y a de malades dans une unité épidémiologique, plus il est probable qu'au moins un malade soit détecté même si la sensibilité de détection est mauvaise, et donc la détection de la maladie dans une unité épidémiologique n'est pas influencée par le facteur de confusion.

Par ailleurs, plus l'odds ratio réel associé au véritable facteur de risque (le facteur de risque X dans les simulations) est élevé et plus la probabilité de présence de la maladie est grande malgré l'absence de ce facteur de risque, plus il est probable que le facteur de confusion soit identifié comme étant un facteur de risque. En effet, plus l'odds ratio réel associé au facteur de risque est grand et plus la probabilité de présence de la maladie est grande dans les unités épidémiologiques où ce facteur est absent, plus les probabilités de présence de la maladie vont être semblables dans toutes les unités, que le facteur de risque soit présent ou absent, ce qui pourrait biaiser le modèle logistique dans sa capacité à détecter correctement les facteurs de risque.

Lorsque le facteur de risque et le facteur de confusion sont identiques, il se dégage de nouvelles tendances pour le modèle logistique. Cela reste un modèle fiable pour identifier le facteur comme étant facteur de risque, cependant il peut ne pas arriver à l'identifier si ce facteur est aussi associé à une diminution de la sensibilité de détection. Si la sensibilité de détection est faible en présence du facteur étudié, alors il y a des unités épidémiologiques où la maladie est non détectée, même si le facteur est présent et favorise l'apparition de la maladie, ce qui empêche une identification correcte du facteur de risque par le modèle logistique. L'estimation de l'odds ratio quant à elle surestime ou sous-estime la valeur réelle, et s'approche de la valeur réelle lorsque la sensibilité de détection s'améliore dans l'ensemble des unités épidémiologiques. Le modèle logistique a notamment tendance à surestimer l'odds ratio par rapport à sa valeur réelle lorsque la sensibilité de détection est faible dans les unités épidémiologiques où le facteur étudié est absent. Ceci peut s'expliquer par le fait que la maladie est plus présente et mieux détectée dans les unités épidémiologiques où le facteur est

présent, et moins présente et, de plus, moins bien détectée dans les unités épidémiologiques où il est absent, ce qui exacerbe la différence existant entre les unités épidémiologiques concernant la présence de la maladie.

En ce qui concerne le **modèle de Poisson enflé en zéro**, il s'agit d'un **modèle fiable pour identifier les facteurs de risque et estimer leur odds ratio**, ainsi que pour **identifier les facteurs de confusion**. Quelles que soient les sensibilités de détection dans les unités épidémiologiques, le facteur de risque est correctement identifié comme tel par la partie « logistique » et non par la partie « comptage » du modèle de Poisson enflé en zéro (ce qui est cohérent, puisque le facteur de risque a seulement une influence sur l'apparition de la maladie et non sur sa détection), et le biais dans le calcul de l'odds ratio par la partie « logistique » est nul à quasi nul, sauf pour des sensibilités de détection faible. En revanche, lorsque la valeur réelle de l'odds ratio est élevée et lorsque la probabilité de présence de la maladie est grande malgré l'absence du facteur de risque, l'odds ratio estimé peut être légèrement surestimé. En effet, plus l'odds ratio réel est grand et plus la probabilité de présence de la maladie est grande dans les unités épidémiologiques sans le facteur de risque, plus la probabilité de présence de la maladie dans les unités épidémiologiques avec le facteur de risque est grande. Dans ces situations, il y a de nombreuses unités épidémiologiques malades, diminuant ainsi les unités épidémiologiques non malades et donc le nombre de vrais négatifs pris en compte par le modèle enflé en zéro, ce qui pourrait biaiser le modèle. Par ailleurs, quelles que soient les sensibilités de détection dans les unités épidémiologiques, le facteur de confusion n'est pas identifié comme facteur de risque par la partie « logistique » et est correctement identifié comme facteur influençant le nombre d'unités élémentaires détectées par la partie « comptage » du moment que la sensibilité est hétérogène entre les unités épidémiologiques. En effet, si la sensibilité est identique dans toutes les unités épidémiologiques, le facteur de confusion n'a pas d'influence sur la détection, mais dès lors que la sensibilité est différente entre les unités épidémiologiques alors le facteur de confusion a une influence et est donc un facteur de risque influençant le comptage du nombre d'unités élémentaires détectées.

En revanche, **le modèle de Poisson enflé en zéro semble fortement biaisé s'il y a un nombre trop insuffisant de malades à détecter**. Dans ce cas, il identifie toujours correctement le facteur de confusion comme tel, mais n'identifie pas toujours correctement le facteur de risque, et le biais dans le calcul de son odds ratio est très important. Ces résultats « aberrants » pour une prévalence intra-unité moyenne très faible pourraient s'expliquer par une trop forte proportion de zéros biaisant le modèle (moins il y a d'unités élémentaires réellement malades, moins il y a d'unités élémentaires détectées).

Lorsque le facteur de risque et le facteur de confusion sont identiques, les tendances déjà évoquées se retrouvent pour le modèle de Poisson enflé en zéro. Quelles que soient les sensibilités de détection dans les unités épidémiologiques, l'aspect « facteur de risque » du facteur étudié est correctement identifié comme tel par la partie « logistique » du modèle, et le biais dans le calcul de l'odds ratio est nul à quasi nul, tandis que l'aspect « facteur de confusion » du facteur est correctement identifié comme tel par la partie « comptage » du modèle.

Ces résultats se basent sur des données simulées à partir de modèles statiques utilisant des lois de Bernoulli, des lois de Poisson tronquées en zéro et des lois binomiales. De nombreuses combinaisons de valeurs d'odds ratios réels, de probabilité de présence de la maladie, de prévalence intra-unité et de sensibilités de détection ont été testées, afin d'illustrer des situations diverses. Par exemple, Boussini et al. (2012) ont étudié la prévalence de la tuberculose et de la brucellose dans des élevages bovins laitiers d'Ouagadougou au Burkina Faso. La prévalence de la tuberculose a été estimée à 6,05%, grâce à des tests d'intradermotuberculination simple dont la sensibilité est estimée à 97%, tandis que la prévalence de la brucellose a été estimée à 3,61%, grâce à des épreuves à l'antigène tamponné dont la sensibilité varie entre 91,4% et 100% (Boussini et al., 2012). Un autre exemple, Mercier et al. (2011) rapportent les résultats d'une étude sur la paratuberculose menée sur les élevages caprins français : la prévalence au niveau des troupeaux a été estimée à 63% tandis que la prévalence intra-troupeau a été estimée en moyenne à 11%, à l'aide d'une enquête sérologique réalisée avec des tests ELISA commerciaux dont la sensibilité est estimée à 53%. Ainsi, selon les maladies et les pays considérés, les valeurs de prévalence et de sensibilité de détection sont variables. A la vue des nos résultats, ces différents paramètres sont à prendre en considération lors de l'identification de facteurs de risque et dans l'estimation des odds ratio (risque de sous-estimation ou de surestimation selon les cas).

II- Etude de l'influence d'une spécificité imparfaite sur l'estimation du risque d'une maladie

1) Introduction et contexte

Les situations étudiées précédemment se plaçaient dans un contexte où la sensibilité de détection était imparfaite, tandis que la spécificité était considérée parfaite. La spécificité est souvent considérée comme parfaite car un résultat positif est ensuite confirmé par diverses analyses, comme c'est par exemple le cas pour le diagnostic de la tuberculose bovine (Bénet et Praud, 2016).

La situation présentée ici se place dans le contexte inverse, avec une sensibilité de détection parfaite et une spécificité imparfaite. La spécificité peut être imparfaite à cause d'une mauvaise spécificité des tests utilisés lors d'une surveillance active, ou d'une mauvaise confirmation par des analyses biologiques de signes cliniques observés lors d'une surveillance passive. Le dépistage de la tuberculose bovine par intradermotuberculination peut par exemple entraîner des faux positifs, d'où la nécessité de confirmer les cas positifs avec d'autres analyses (Bénet et Praud, 2016). Ces faux positifs peuvent être dus à une réaction suite à la sensibilisation du bovin par une autre mycobactérie ou à une mauvaise lecture du résultat par le vétérinaire.

L'étude porte sur N unités épidémiologiques, elles-mêmes composées de U unités élémentaires. Ce pourrait par exemple être N élevages, chacun composés de U animaux. Ces animaux sont détectés lorsque le système de surveillance les détecte comme étant malades (ils sont forcément détectés lorsqu'ils sont malades car la sensibilité est parfaite, mais peuvent aussi l'être lorsqu'ils ne sont pas malades).

L'objectif ici est seulement de voir quelle est l'influence d'un défaut de spécificité homogène sur l'identification de facteurs de risque et l'estimation des odds ratios associés, lorsque les données de surveillance sont modélisées avec un modèle logistique. Le modèle enflé en zéro n'est pas une alternative pertinente à évaluer car la structure d'un tel modèle ne peut pas prendre en compte la présence de faux positifs, donc nous nous limiterons à évaluer le modèle logistique.

2) Matériels et méthodes

a) <u>Distribution des variables d'intérêts</u>

Le **facteur X est défini comme le facteur de risque de présence de la maladie** dans une unité épidémiologique. Comme précédemment, la probabilité que X soit présent dans une unité épidémiologique est notée « pX ». Lorsque le facteur X est absent (respectivement présent) d'une unité épidémiologique, la probabilité que la maladie soit présente dans cette unité est notée « prev.X0 » (respectivement « prev.X1 »). L'odds ratio de la présence de la maladie associé à ce facteur est noté « OR(X) ». Les relations entre ces variables ont déjà été illustrées avec la Figure 5.

b) Détection de la maladie dans les unités épidémiologiques

Comme illustré en Figure 23, la détection de la maladie dans les unités épidémiologique a été simulée comme un processus hiérarchique en deux étapes principales : simulation 1) de la présence de la maladie dans les unités épidémiologiques, 2) du nombre d'unités élémentaires détectées par unité épidémiologique (qu'elles soient malades ou non).

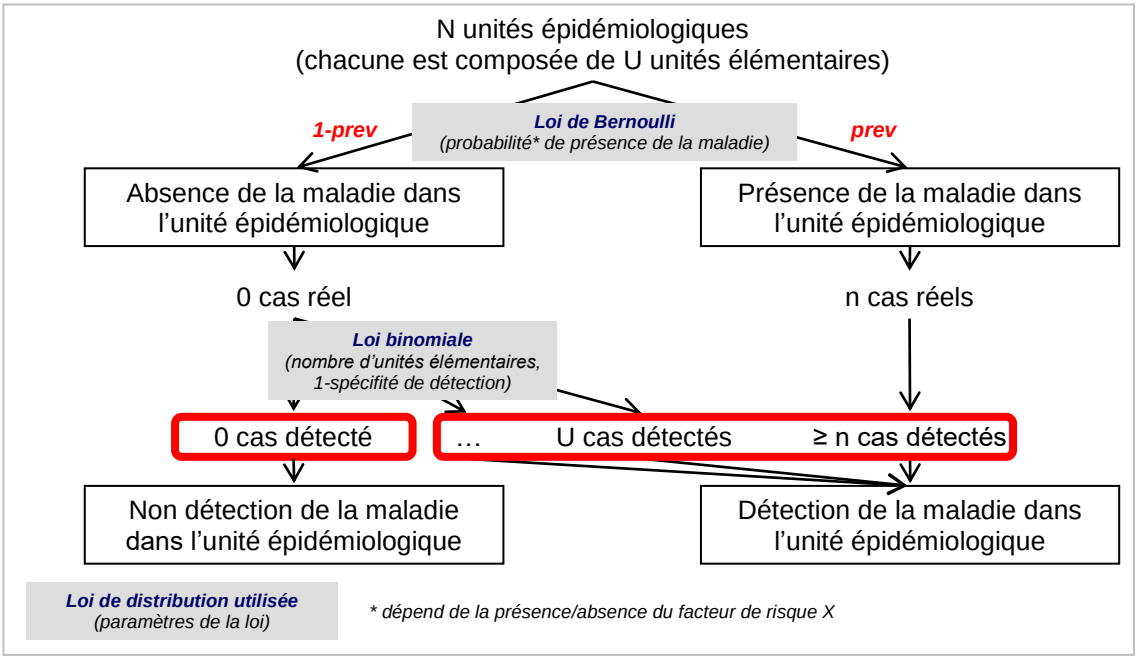


Figure 23 : Schéma de simulation des données lorsque la spécificité de la détection est imparfaite

La **présence/absence de la maladie dans une unité épidémiologique i** (notée « P_i ») a de nouveau été considérée comme une variable aléatoire suivant une loi de Bernoulli de paramètre « $prev_i$ », telle que présentée dans l'Équation 1.

La sensibilité de détection étant parfaite, toutes les unités élémentaires malades sont détectées, donc la maladie est détectée dans toutes les unités épidémiologiques où elle est présente. Par conséquent, le nombre de cas réels par unité épidémiologique n'a pas été modélisé explicitement.

La spécificité de détection étant considérée comme imparfaite, des unités élémentaires non malades peuvent être détectées, le nombre d'unités élémentaires détectées est donc nécessairement supérieur ou égal au nombre d'unités élémentaires malades (la sensibilité de détection est considérée parfaite). Le **nombre d'unités élémentaires détectées dans une unité épidémiologique non malade i** (noté « $Détect_i$ ») a été considéré comme une variable aléatoire suivant une loi binomiale de paramètres « U » et « 1-spécificité » (probabilité de détecter une unité élémentaire non malade comme étant malade), telle que présentée dans l'Équation 8.

$$P(\text{Détect}_i=d) = \begin{pmatrix} U \\ d \end{pmatrix} . (1\text{-spécificité})^d . (1 - (1\text{-spécificité}))^{U-d}$$

***P*(Détect$_i$=d)** : Probabilité que le nombre d'unités élémentaires détectées dans l'unité épidémiologique non malade i soit égal à d
U : Nombre d'unités élémentaires dans chaque unité épidémiologique
1-spécificité : Probabilité de détecter une unité élémentaire non malade comme étant malade

Équation 8 : Loi binomiale de paramètres « U » et « 1-spécificité »

La maladie pouvant être détectée dans les unités épidémiologiques où elle est absente, le nombre d'unités épidémiologiques détectées est donc supérieur ou égal au nombre d'unités épidémiologiques où la maladie est réellement présente.

c) Plan de simulation

De la même manière que dans la partie précédente, le plan de simulation a été défini de manière à évaluer l'influence d'un grand nombre de paramètres d'intérêt sur la caractérisation du risque et fournir une réponse nuancée à la question de recherche.

Le modèle général de simulation comprend 6 paramètres (Tableau 4). Par souci de clarté et de simplicité des interprétations, les valeurs de certains paramètres d'intérêt moindre (N, U et pX) ont été fixées : le nombre d'unités épidémiologiques étudiées a été fixé à N=10000 (de manière à ce que la taille de l'échantillon ne limite pas la puissance de l'analyse), le nombre d'unités élémentaires dans les unités épidémiologiques à été fixé à U=50 (valeur réaliste dans un contexte d'élevage de bovins français), et la probabilité de présence du facteur X a été fixée à pX=0,5. L'influence de tous les autres paramètres (prev.X0, OR(X), spécificité) a été testée en les faisant varier de manière indépendante ou en combinaison dans des ordres de grandeur réalistes. Les différentes valeurs testées des différents paramètres sont présentées dans le Tableau 4.

Tableau 4 : Paramètres fixés dans les différentes séries de simulations lorsque la spécificité est imparfaite

	« Série A »	**« Série B »**
N	10 000	
U	50	
pX	0,5	
prev.X0	0,1 ; 0,2 ; 0,5	0,2 ; 0,4 ; 0,6 ; 0,8
OR(X)	2 ; 5 ; 10	1 ; 2 ; 5
spécificité	0,85 à 1 avec un pas de 0,001	0,85 à 1 avec un pas de 0,01
Nombre de simulations (pour N, U, pX, prev.X0 et OR(X) fixés)	**1 simulation** pour chacune des **151 valeurs de spécificité** testées	**300 simulations** pour chacune des **16 valeurs de spécificité** testées

d) Analyse des données obtenues

Les données obtenues suite aux simulations ont été analysées uniquement avec un **modèle logistique**, afin d'étudier le potentiel biais provoqué par un défaut de spécificité dans la détection de facteurs de risque.

Comme précédemment, la **variable à expliquer** du modèle logistique est la **présence/absence observée de la maladie dans les unités épidémiologiques**, c'est-à-dire la détection ou non d'au moins une unité élémentaire considérée comme malade. Le **facteur de risque X** est la seule **variable explicative** (il s'agit d'une variable binaire : présence ou absence dans chacune des unités épidémiologiques étudiées), étant donné qu'il s'agit d'un contexte où la spécificité est imparfaite mais homogène.

Comme précédemment, un modèle logistique a été ajusté à chaque jeu de données simulées, et le niveau de significativité de l'association entre la variable explicative et la variable réponse, ainsi que la valeur de l'$OR(X)_{logist}$ ont été enregistrés. La variable a été considérée comme significativement associée à la variable réponse si le niveau de significativité (**p-value**) était inférieur à 0,05.

L'$OR(X)_{logist}$ a été utilisé pour calculer le **biais relatif** de l'odds ratio associé au facteur X, noté « biais($OR(X)_{logist}$) » comme décrit dans l'Équation 5.

e) <u>Simulations et analyses : logiciel utilisé</u>

Les simulations et les analyses ont été réalisées avec la version 3.3 du logiciel R (R Core Team, 2017).

3) Résultats : Impact sur les modèles logistiques d'une spécificité de détection imparfaite mais homogène

Plus la spécificité est faible, et moins il est probable que le modèle logistique identifie correctement le facteur X comme étant un facteur de risque (Figure 24). En outre, **moins l'influence réelle du facteur X sur l'apparition de la maladie est importante** (OR(X) faible)**, meilleure doit être la spécificité pour que le facteur X soit identifié comme facteur de risque**. Par exemple, comme illustré en Figure 24, lorsque l'odds ratio réel est égal à 2 (courbe bleue), le facteur X est systématiquement correctement identifié comme étant un facteur de risque pour une spécificité allant de 98% à 100%, tandis que lorsque l'odds ratio réel est égal à 5 (courbe verte), le facteur X est systématiquement identifié comme étant un facteur de risque pour une spécificité allant de 94% à 100%. L'évolution de cette probabilité d'identification du facteur X comme étant un facteur de risque est semblable pour les différentes valeurs de prev.X0 testées (résultats non présentés).

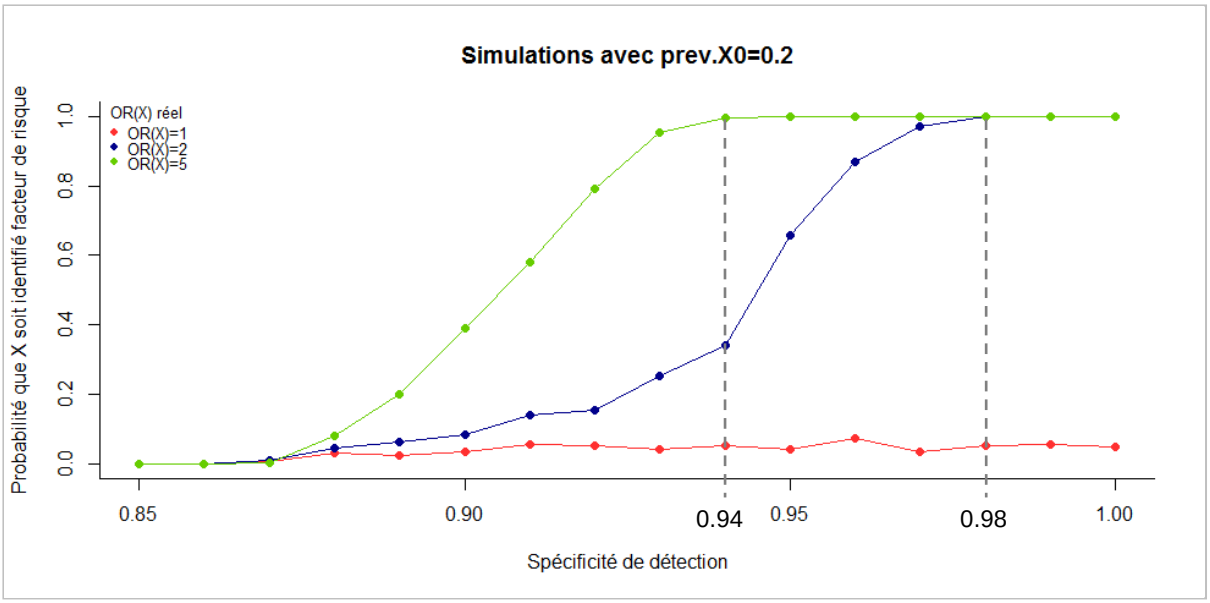


Figure 24 : Evolution de la probabilité que le facteur X soit identifié comme étant un facteur de risque par un modèle logistique en fonction de la spécificité de détection et pour différentes valeurs réelles d'OR(X) lorsque prev.X0=0,2

Dès lors que la spécificité de détection est inférieure à 1, l'odds ration associé au facteur X est systématiquement sous-estimé. Le biais négatif augmente très rapidement pour des valeurs de spécificité proches de 1 pour atteindre un plateau dès lors que la spécificité est inférieure à 98%, dont la valeur semble dépendre de l'odds ratio réel. Par exemple, comme illustré en Figure 25, pour une spécificité de détection de 95%, un odds ratio réel de 2 (courbe bleue) aura tendance à être sous-estimé de près de 30%, tandis qu'un odds ratio réel de 5 (courbe verte) aura tendance à être sous-estimé de près de 60%. Il est à noter que pour une spécificité de détection inférieure à 90%, l'estimation du biais est imprécise (Figure 25). Autrement dit, **plus la spécificité de détection est faible (risque élevé de faux positifs), plus l'odds ratio de l'apparition de la maladie associé au facteur X et calculé par le modèle logistique est inexact et imprécis**. A noter que cette imprécision dans l'estimation de l'odds

ratio correspond à des valeurs de spécificités pour lesquelles il est peu probable que le modèle logistique identifie correctement le facteur X comme étant un facteur de risque, comme illustré en Figure 24.

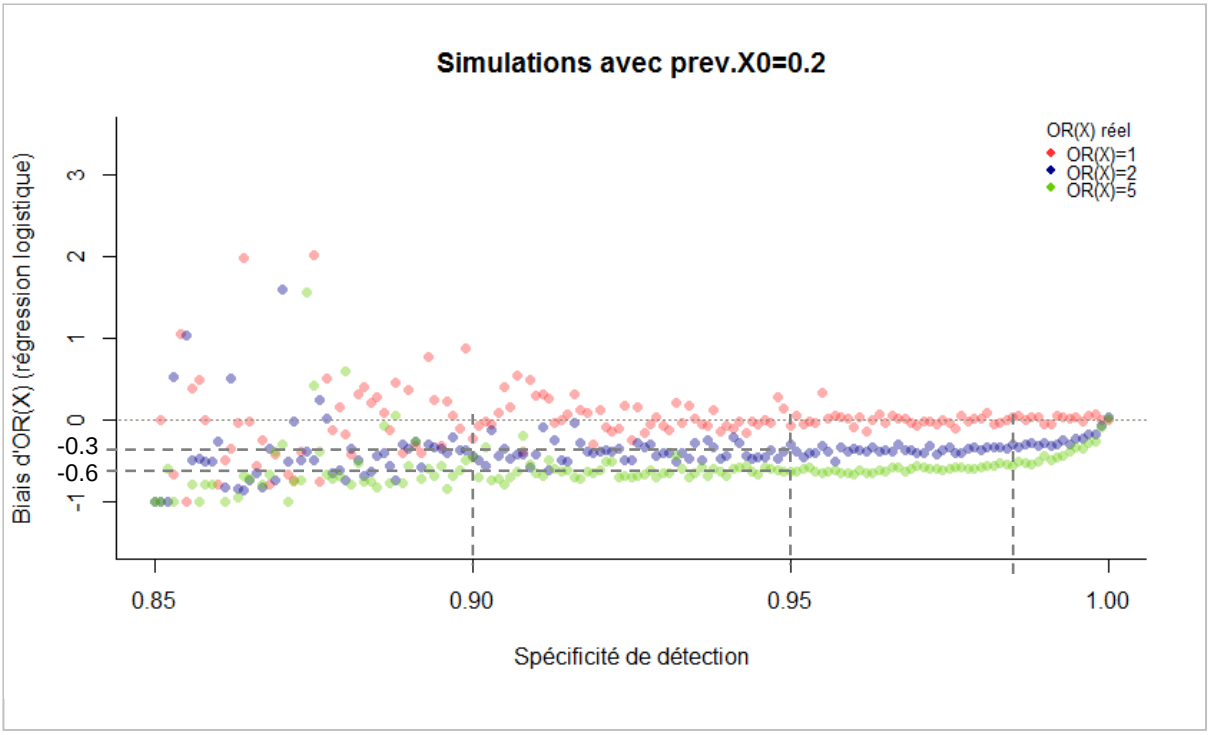


Figure 25 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la spécificité de détection et pour différentes valeurs réelles d'OR(X) lorsque prev.X0=0,2

En outre, **plus la probabilité de présence de la maladie est grande malgré l'absence du facteur X** (c'est-à-dire plus prev.X0 est élevé), **plus la spécificité pour laquelle l'estimation de l'odds ratio se précise augmente**, et **plus le biais dans l'estimation de l'odds ratio tend rapidement vers 0 lorsque ce biais se précise** (Annexe 6).

Spécificité de détection imparfaite mais homogène :
impacts sur les modèles logistiques

Identification du facteur de risque comme étant un facteur de risque ?

Augmentation de la probabilité que le facteur de risque soit correctement identifié lorsque :

- la spécificité de détection augmente ;
- l'odds ratio réel augmente.

Estimation correcte de l'odds ratio associé au facteur de risque ?

L'odds ratio est sous-estimé, et plus la spécificité est faible plus l'estimation est imprécise.

Plus la probabilité de présence de la maladie est grande malgré l'absence du facteur de risque, plus la spécificité à partir de laquelle l'estimation de l'odds ratio se précise est élevée.

4) Discussion

Cette partie a permis d'étudier les impacts sur les modèles logistiques d'une spécificité de détection imparfaite. Contrairement à l'étude sur les impacts d'une sensibilité de détection imparfaite, seule l'influence d'une spécificité imparfaite mais homogène a été étudiée (l'effet d'une spécificité imparfaite et hétérogène n'a pas été envisagé). La sensibilité de détection a été considérée comme parfaite pour pouvoir étudier le seul effet d'un défaut de spécificité, bien que cette situation sur le terrain soit rare.

Plus la spécificité de détection est faible, moins il est probable que le modèle logistique identifie correctement les facteurs de risque. En effet, il y a d'autant plus de faux positifs que la spécificité est faible, et dans ce cas le facteur de risque ne paraît pas influencer l'apparition de la maladie. Si on prend l'exemple de la situation illustrée en Figure 24 pour un facteur de risque ayant une association modérée avec la présence de la maladie (odds ratio réel de 2, illustré par la courbe bleue), si la spécificité de la détection n'est que de 95% alors le facteur de risque n'est pas identifié par le modèle logistique dans 40% des cas (la probabilité d'identification du facteur étant 60%).

De plus, lors d'un défaut de spécificité, **l'odds ratio estimé sous-estime la valeur réelle, et ce d'autant plus que l'odds ratio réel est grand**. Plus la spécificité de détection est faible, plus il y a d'unités élémentaires non malades qui sont détectées comme malades au sein des unités épidémiologiques. Ceci implique qu'il y a plus d'unités épidémiologiques non malades qui sont détectées à tort comme malades, si au moins une unité élémentaire est détectée comme malade. Ces faux positifs vont biaiser le modèle logistique, qui va sous-estimer l'importance réelle qu'a le facteur de risque dans l'apparition de la maladie dans les unités épidémiologiques, biais qui est d'autant plus important que l'importance réelle de ce facteur est grande.

L’imprécision observée pour des spécificités très faibles vient du fait que s’il y a beaucoup de faux positifs, de très nombreuses unités épidémiologiques finissent par être détectées (même si elles ne sont pas malades) menant à un nombre d’unités épidémiologiques non détectées trop faible pour permettre au modèle logistique d’identifier les différences entre les unités épidémiologiques détectées comme malades et les non détectées. Cette imprécision existe même pour des spécificités de détection relativement élevées si la probabilité de présence de la maladie est importante dans les unités épidémiologiques où il n’y a pas le facteur de risque. En effet, dans ces conditions la maladie est présente dans de nombreuses unités épidémiologiques, que le facteur de risque soit présent ou non, et à ces vrais positifs s’ajoutent les faux positifs que sont les unités épidémiologiques non malades qui sont détectées à tort comme malades. Ce trop grand nombre de positifs va biaiser le modèle logistique et conduire à l’imprécision observée.

Ainsi, le défaut de spécificité et le grand nombre de faux positifs qu’il entraîne vont biaiser le modèle logistique et empêcher l’identification correcte des facteurs de risque. Il est donc important d’avoir une spécificité la plus parfaite possible avant d’utiliser un jeu de données pour l’identification de facteurs de risque, et il est donc intéressant de recourir à des tests de confirmation des cas positifs afin d’améliorer au maximum la spécificité.

III- Bilan et perspectives

Les défauts de sensibilité et de spécificité de détection ont un impact sur l'étude des facteurs de risque. La spécificité peut être améliorée en testant avec diverses analyses les cas positifs, afin de confirmer qu'ils sont réellement positifs. La sensibilité pourrait quant à elle être améliorée en testant de nouveau tous les cas négatifs, mais cela paraît peu réalisable en pratique et d'un point de vue économique.

Le défaut de spécificité peut aboutir à une non identification des facteurs de risque par le modèle logistique. A l'inverse, lors d'un défaut de sensibilité, les facteurs de risque vont être correctement identifiés par le modèle logistique, qui va aussi potentiellement identifier à tort d'autres facteurs (tels que les facteurs de confusion). Dans tous les cas, les odds ratios obtenus sont généralement sous-estimés. Les odds ratio sont surestimés par le modèle logistique seulement dans le cas où le facteur de risque est aussi un facteur de confusion influençant la sensibilité de détection de la maladie. La spécificité semble donc le paramètre le plus important à prendre en compte avant d'identifier des facteurs de risque, dans le sens où un défaut de spécificité peut faire manquer un facteur de risque potentiel.

Concernant le modèle de Poisson enflé en zéro, il a seulement été utilisé dans le cas d'un défaut de sensibilité. Les résultats obtenus avec les données simulées indiquent qu'il semble être un modèle permettant de différencier les facteurs de risque influençant l'apparition de la maladie dans une unité épidémiologique des facteurs de confusion influençant la détection du nombre d'unités élémentaires malades. De plus, les odds ratios estimés pour le facteur de risque sont plus proches de la valeur réelle que ceux estimés par le modèle logistique.

Ces résultats se basent sur des données simulées à partir de modèles statiques utilisant des lois de Bernoulli, des lois de Poisson tronquées en zéro et des lois binomiales. De nombreuses combinaisons de valeurs d'odds ratios réels, de probabilité de présence de la maladie, de prévalence intra-unité, de sensibilités et de spécificités de détection ont été testées, afin d'illustrer des situations diverses. Cependant, ces simulations ne modélisent pas de manière parfaite l'évolution d'une maladie. Une étude dynamique serait intéressante, ainsi que la prise en compte d'une auto-corrélation spatiale. En effet, la plupart des maladies évoluent dans le temps, et sont contagieuses. Ainsi, l'apparition de la maladie dans une unité épidémiologique dépend de la présence de cette maladie dans les unités voisines. Ces aspects n'ont pas été pris en compte dans les simulations présentées, et pourraient avoir une influence sur les modèles utilisés pour l'identification des facteurs de risque. De plus, les unités épidémiologiques n'ont pas de taille fixée dans l'étude du défaut de sensibilité, et ont toutes une taille identique dans l'étude du défaut de spécificité. Une étude avec des tailles variables d'unités épidémiologiques afin de refléter la réalité des variations de taille de troupeaux serait aussi peut-être plus pertinente. Enfin, seulement deux facteurs ont été étudiés, avec chacun un effet différent selon qu'il est présent ou absent. Or, la réalité est parfois plus complexe, avec des effets supplémentaires, avec pourquoi pas un effet « combinaison » en présence de plusieurs facteurs (interaction entre facteurs).

Des améliorations ainsi que des complexifications sont encore possibles afin de comparer et comprendre les résultats obtenus lors de la modélisation de jeux de données dans l'objectif d'identifier des facteurs de risque. La présente étude a permis d'envisager certains aspects simples, et de mettre en évidence l'attention qu'il faut porter sur les défauts de collection des données.

Partie 3 :
APPLICATION A DES DONNEES REELLES : IDENTIFICATION DE FACTEURS DE RISQUE DES AVORTEMENTS BOVINS EN FRANCE METROPOLITAINE (2010-2011)

I- Objectif

Dans la Partie 2, nous avons utilisé des données simulées pour mettre en évidence les différences qui existent dans l'identification de facteurs de risque entre le modèle logistique, classiquement utilisé pour l'identification de facteurs de risque, et le modèle de Poisson enflé en zéro. L'objectif de cette troisième partie est de revisiter un jeu de données réelles afin d'interpréter les différences qui existent entre ces deux modèles à la lumière des conclusions obtenues dans la Partie 2. Pour cela, nous avons utilisé les données de déclarations d'avortements bovins en France métropolitaine entre le 1er août 2010 et le 31 juillet 2011 (Bronner et al., 2013), auxquelles nous avons ajusté un modèle logistique et un modèle de Poisson enflé en zéro.

II- Introduction et contexte

Chez les bovins, un avortement est défini comme étant l'expulsion avant terme d'un fœtus ou l'expulsion à terme d'un veau né mort ou qui meurt dans les 48 heures après sa naissance (Ganière et Laaberki, 2017). Tout avortement chez un bovin doit obligatoirement faire l'objet d'une déclaration au vétérinaire sanitaire dans le cadre de la surveillance passive de la brucellose bovine (Ganière et Laaberki, 2017 ; Ministre de l'agriculture et de la pêche, 2008). Cette surveillance systématique de tous les avortements bovins vise à la détection précoce d'une introduction de brucellose en France.

Les avortements bovins ont des causes variées, majoritairement infectieuses. Ces infections peuvent causer des « flambées » d'avortements au sein d'un élevage, sans pour autant concerner les élevages voisins. Cependant, certaines maladies ont entraîné des avortements à l'échelle nationale. En 2006 et 2007, la France a été touchée par une épizootie de fièvre catarrhale ovine, qui a entre autre entraîné des avortements chez les bovins (Bronner et al., 2013 ; Peroz et Ganière, 2017). Fin 2011 et en 2012, c'est une épizootie de maladie de Schmallenberg qui a touché la France (Gache et al., 2015). Ces deux maladies sont dues à des virus transmis par des culicoïdes, ce qui explique leur propagation.

En théorie, les données disponibles sur les avortements bovins en France devraient donc rassembler l'ensemble des avortements bovins. Cependant, les avortements ne sont pas forcément tous détectés par les éleveurs (Forar et al. (1995) estiment que seulement 30% des morts fœtales sont observées visuellement), et les avortements détectés ne sont pas systématiquement déclarés au vétérinaire sanitaire (Bronner et al. (2013) estiment que seuls 23% des éleveurs qui ont détecté un avortement en ont effectivement déclaré au moins un). Les données de surveillance des avortements bovins en France sont donc partielles. Bronner et al. (2013) ont montré que la probabilité de détection d'au moins un avortement dans un

élevage était significativement associée au type d'élevage (laitier, allaitant ou mixte) et que le nombre d'avortements déclarés dans les élevages où au moins un avortement a été détecté était significativement associé au type de production et à la taille de l'élevage. Ce jeu de données entre donc parfaitement dans le cadre de l'étude de comparaison du modèle logistique et du modèle de Poisson enflé en zéro pour l'identification de facteurs de risque lorsque la sensibilité de la surveillance est imparfaite.

III- Matériels et méthodes

1) Source des données et population étudiée

Les données utilisées sont extraites de la base de données française Sigal, le système d'informations de la DGAl (direction générale de l'Alimentation), et de la BDNI (base de données nationale d'identification des bovins).

Les données utilisées ici concernent les déclarations d'avortements bovins en France métropolitaine entre le 1er août 2010 et le 31 juillet 2011. Le nombre d'avortements déclarés, le département, la taille de l'élevage (en nombre de bovins-jours) et le type de production (allaitant, laitier ou mixte) sont connus pour chaque exploitation présente dans la BDNI.

2) Modélisation et analyse des données

Afin d'utiliser les résultats obtenus avec les données simulées de la Partie 2, un modèle logistique et un modèle de Poisson enflé en zéro ont été ajustés aux données d'avortements. Pour chacun des deux modèles, l'unité épidémiologique est l'élevage, et les unités élémentaires sont les bovins de l'élevage. Le jeu de données a été scindé en deux de manière aléatoire, avec d'un côté 90% des élevages utilisés pour estimer les paramètres des modèles, et d'un autre côté 10% des élevages pour une étape de validation des modèles.

a) <u>Variables des modèles</u>

Pour le modèle logistique, la variable à expliquer est l'existence d'au moins une déclaration d'avortement dans une unité épidémiologique, tandis que la variable à expliquer du modèle de Poisson enflé en zéro est le nombre d'avortements déclarés dans une unité épidémiologique. Pour rappel, le modèle de Poisson enflé en zéro décrit cette variable en deux étapes : une partie « logistique » qui correspond à la probabilité qu'un éleveur détecte ou pas au moins un avortement dans son élevage et une partie « comptage » décrivant le nombre d'avortements déclarés (0, 1, 2, etc.) par les éleveurs ayant détecté au moins un avortement. Pour les deux modèles, les variables explicatives testées sont le type de production (allaitant, laitier ou mixte) et la taille d'élevage (divisée en trois catégories, selon les terciles).

La période étudiée, entre le 1er août 2010 et le 31 juillet 2011, a été choisie car elle correspond à une période sans épizootie à l'origine de vagues d'avortements (fièvre catarrhale ovine ou maladie de Schmallenberg). Ainsi, il n'y a pas eu besoin de prendre en compte une possible autocorrélation spatiale des avortements survenant dans des élevages voisins.

b) Sélection des variables et construction des modèles

L'absence de corrélation entre les deux variables explicatives a été vérifiée grâce à un test de Kendall, les deux variables explicatives étant catégorielles (Baudot, 2014 ; McLeod, 2015). Les deux variables ont donc été conservées pour construire les modèles.

Les modèles ont ensuite été construits suivant une procédure de « backward elimination » et de « forward selection » (Dohoo et al., 2009), identique pour le modèle logistique et le modèle de Poisson enflé en zéro (Zeileis et al., 2008). Dans un premier temps, un modèle contenant toutes les variables explicatives a été construit (modèle « complet »), puis chacune des variables a été enlevée du modèle (obtention de modèles « réduits »). La différence d'ajustement aux données entre le modèle « complet » et chacun des modèles « réduits » a été évaluée grâce à un « likelihood ratio test » (Dohoo et al., 2009). Si le test met en évidence une détérioration significative de l'ajustement (p-value < 0,05) entre le modèle « complet » et le modèle « réduit », le modèle « complet » est considéré comme mieux ajusté aux données. Si en revanche il n'y a pas de différence significative entre le modèle « complet » et le modèle « réduit », le modèle « réduit » est sélectionné afin de satisfaire le principe de parcimonie. Suite à cette première étape, l'interaction potentielle entre les variables sélectionnées a été testée. Le principe est le même que précédemment, le modèle « réduit » étant désormais le modèle intermédiaire, et le modèle « complet » étant le modèle avec l'interaction. A la fin de cette étape de « forward selection », le modèle sélectionné est le modèle final.

c) Validation des modèles

La validité du modèle logistique final et du modèle de Poisson enflé en zéro final a été évaluée en utilisant les 10% d'élevages qui n'ont pas été utilisés pour la construction des modèles. Pour ces données, l'adéquation entre les prédictions et les observations a été évaluée à l'aide d'une courbe « receiver operating characteristic », ou courbe ROC (Dohoo et al., 2009).

Une courbe ROC permet d'évaluer les capacités d'un modèle à correctement prédire les données observées. Pour chaque élevage, les probabilités de déclarer au moins un avortement selon chacun des modèles ont été calculées à partir des résultats d'estimation des coefficients de chacun des modèles et à l'aide de l'Équation 4 (pour le modèle logistique) ou de l'Équation 6 (pour le modèle de Poisson enflé en zéro). La présence réelle d'au moins une déclaration d'avortement dans les élevages a ainsi été confrontée aux probabilités prédites par chacun des modèles.

L'interprétation des courbes ROC se base sur le calcul de l'aire sous la courbe (« area under curve » ou AUC), comme indiqué dans le Tableau 5. Si le modèle est complètement non-informatif et prédit les observations positives (c'est-à-dire au moins un avortement) et négatives (c'est-à-dire pas d'avortement) de manière totalement aléatoire, l'AUC sera égale à 0,5 (la distribution de la probabilité d'être un positif est la même chez les positifs que chez les négatifs). A l'inverse, si le modèle prédit parfaitement les observations, l'AUC sera égale à 1.

En général, l'AUC est donc comprise entre 0,5 et 1 avec des valeurs proches de 1 qui témoignent d'un modèle à fort pouvoir prédictif (Tableau 5).

Tableau 5 : Interprétation de l'aire sous la courbe d'une courbe ROC, d'après Rakotomalala (2011)

Aire sous la courbe (AUC)	Interprétation
AUC = 0,5	Pas de distinction entre les positifs et les négatifs
0,7 ≤ AUC < 0,8	Distinction acceptable
0,8 ≤ AUC < 0,9	Distinction excellente
0,9 ≤ AUC	Distinction exceptionnelle

d) Logiciels utilisés

Les analyses ont été réalisées avec la version 3.3 du logiciel R (R Core Team, 2017). Le package 'Kendall' a été utilisé pour le test de Kendall (McLeod, 2011). Le package 'pscl' a été utilisé pour définir le modèle de Poisson enflé en zéro (Jackman et al., 2015). Le package 'lmtest' a été utilisé pour la méthode des likelihood ratio tests (Hothorn et al., 2017), le package 'pROC' a été utilisé pour les courbes ROC (Robin et al., 2017) et le package 'vioplot' a été utilisé pour le tracé des violin plots (Adler, 2005). Certains calculs (odds ratios et ratio du taux d'incidence) ont été réalisés avec le logiciel Excel (Microsoft, 2003).

IV- Résultats

1) Description des données

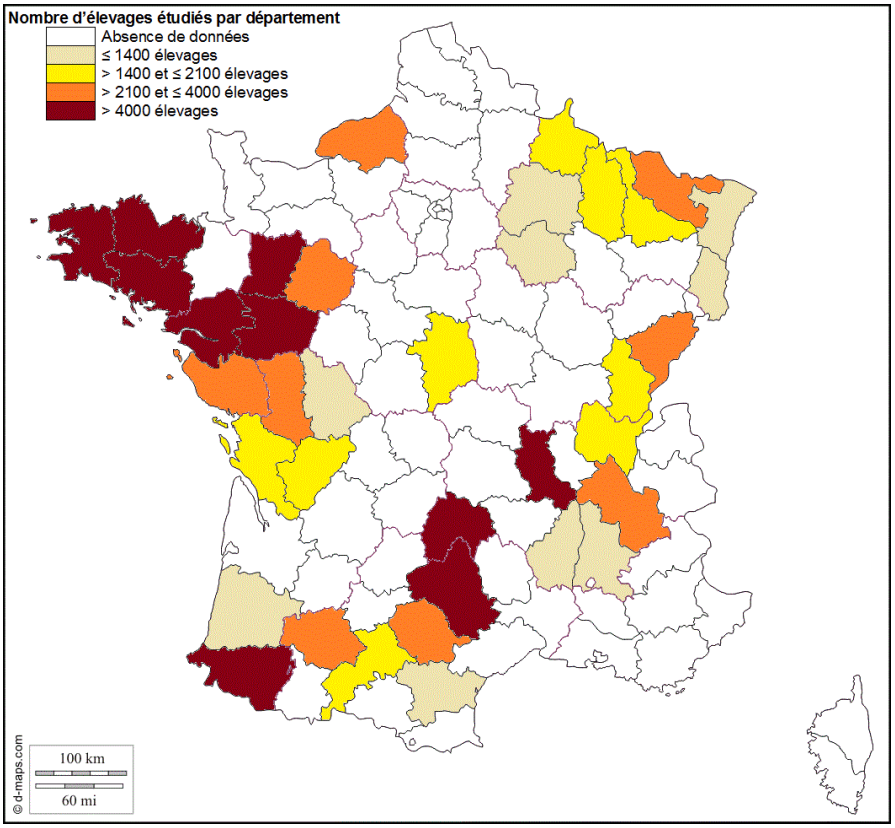


Figure 26 : Distribution du nombre d'élevages de bovins par département français (dans lesquels au moins un élevage a déclaré au moins un avortement entre le 1er août 2010 et le 31 juillet 2011), d'après les informations disponibles dans le jeu de données (Bronner et al., 2013)

La base de données comprenait 99 996 élevages, répartis dans 37 départements de France métropolitaine (Figure 26). Seuls les départements dans lesquels au moins un élevage a déclaré au moins un avortement ont été inclus dans l'étude (Bronner et al., 2013). Nous avons sélectionné aléatoirement 89 996 (soit 90%) de ces élevages pour permettre l'estimation des paramètres des modèles, les 10 000 (soit 10%) restants servant à valider les modèles.

La taille moyenne des 99 996 élevages était de 15 482 bovins-jours, et la taille médiane est de 13 485 (minimale = 1 ; maximale = 276 605). Les catégories de taille utilisées pour la construction des modèles sont présentées dans le Tableau 6. Concernant les types de production, 58 979 élevages (soit 59%) étaient des élevages allaitants, 29 275 élevages (soit 29%) sont des élevages laitiers, et 11 742 élevages (soit 12%) sont des élevages ayant une production mixte. Parmi les 99 996 élevages, 19 200 élevages (19%) ont déclaré au moins un avortement entre le 1er août 2010 et le 31 juillet 2011.

Les répartitions des tailles, des types de production et du nombre d'avortements déclarés par élevages sont présentées dans le Tableau 6, le Tableau 7 et le Tableau 8 pour l'ensemble de la base de données (99 996 élevages).

Tableau 6 : Taille et type de production des 99 996 élevages du jeu de données

	Catégories (nombre d'élevages par catégorie)		
Taille des élevages (en bovins-jours)	≤ 7 686 (33 058 élevages soit 33%)	> 7 686 et ≤ 18 586 (33 050 élevages soit 33%)	> 18 586 (33 888 élevages soit 34%)
Type de production	Allaitant (58 979 élevages soit 59%)	Laitier (29 275 élevages soit 29%)	Mixte (11 742 élevages soit 12%)

Tableau 7 : Répartition du nombre d'avortements déclarés par élevage

Nombre d'avortements déclarés par élevage	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Nombre d'élevages	80 796 (81%)	11 632	4 081	1 878	824	392	197	87	43	30
Nombre d'avortements déclarés par élevage	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**25**	**96**
Nombre d'élevages	14	9	4	2	1	2	1	1	1	1

Tableau 8 : Répartition par catégorie des 99 996 élevages selon qu'ils aient déclaré ou non un avortement

	Elevages n'ayant déclaré aucun avortement			**Elevages ayant déclaré ≥ 1 avortement**		
Production / Taille	**Allaitant**	**Laitier**	**Mixte**	**Allaitant**	**Laitier**	**Mixte**
≤ 7 686 bovins-jours	28129	2728	1028	876	254	43
> 7 686 et ≤ 18 586 bovins-jours	13680	9753	2236	1678	4868	835
> 18 586 bovins-jours	11895	6605	4742	2721	5067	2858

2) Inférence par le modèle logistique

Le modèle logistique final contient les variables « type de production » et « taille d'élevage » ainsi que l'interaction entre ces deux variables (Tableau 9).

Tableau 9 : Modèle logistique final $(OR_{interaction(A\&B)}=OR_A.OR_B.OR_{A:B})$

Variable		Catégories	Odds ratio	Intervalle de confiance (95%)
Type de production		Allaitant	référence	référence
		Laitier	2,96	2,53-3,44
		Mixte	1,35	0,95-1,85
Taille des élevages (en bovins-jours)		≤ 7 686	référence	référence
		> 7 686 et ≤ 18 586	3,91	3,58-4,28
		> 18 586	7,40	6,81-8,04
Interaction	***Production***	***Taille***		
Type de production & Taille des élevages (en bovins-jours)	Laitier	> 7 686 et ≤ 18 586	16,22	10,77-24,45
	Mixte	> 7 686 et ≤ 18 586	11,93	5,56-25,54
	Laitier	> 18 586	24,91	16,66-37,25
	Mixte	> 18 586	19,28	9,13-40,75

Le modèle logistique indique que le type de production et la taille des élevages bovins sont des facteurs associés aux les avortements : pour une taille d'élevage donnée, les avortements sont significativement plus fréquents dans les élevages laitiers (OR=2,96 [2,53-3,44]) que dans les élevages allaitants ; de même pour un type de production donné, la fréquence d'observation d'au moins un avortement est significativement augmentée dans les élevages de taille moyenne (OR=3,9 [3,58-4,28]) ou grande (OR=7,4 [6,81-8,04]), par rapport aux élevages de petite taille.

La prise en compte dans le modèle logistique final de l'interaction entre le type de production et la taille de l'élevage (Tableau 9) permet de mettre en évidence que l'influence du type d'élevage varie selon sa taille (Figure 27).

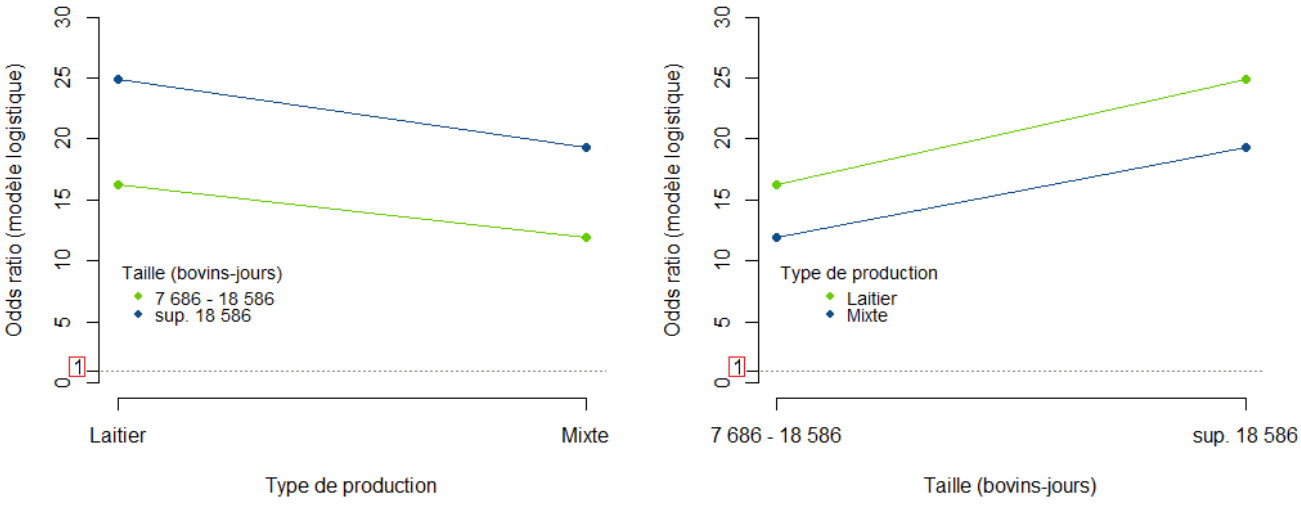


Figure 27 : Effet de l'interaction entre le type de production et la taille d'élevage sur la valeur de l'odds ratio estimée par le modèle logistique

3) Inférence par le modèle de Poisson enflé en zéro

Le modèle de Poisson enflé en zéro final contient les variables « type de production » et « taille d'élevage » à la fois pour la partie « logistique » et pour la partie « comptage », ainsi que l'interaction entre ces deux variables dans la partie « logistique » uniquement (Tableau 10).

Tableau 10 : Modèle de Poisson enflé en zéro final $(OR_{interaction(A\&B)}=OR_A.OR_B.OR_{A:B})$

Partie « logistique »				
Variable		**Catégories**	**Odds ratio**	**Intervalle de confiance (95%)**
Type de production		Allaitant	référence	référence
		Laitier	1,97	1,65-2,37
		Mixte	0,90	0,63-1,29
Taille des élevages (en bovins-jours)		≤ 7 686	référence	référence
		> 7 686 et ≤ 18 586	2,35	2,00-2,78
		> 18 586	3,30	2,81-3,87
Interaction	***Production***	***Taille***		
Type de production & Taille des élevages (en bovins-jours)	Laitier	> 7 686 et ≤ 18 586	7,86	4,58-13,48
	Mixte	> 7 686 et ≤ 18 586	5,78	2,36-14,16
	Laitier	> 18 586	8,79	5,17-14,93
	Mixte	> 18 586	6,94	2,89-16,69
Partie « comptage »				
Variable		**Catégories**	**Ratio du taux d'incidence**	**Intervalle de confiance (95%)**
Type de production		Allaitant	référence	référence
		Laitier	1,75	1,67-1,83
		Mixte	1,60	1,52-1,69
Taille des élevages (en bovins-jours)		≤ 7 686	référence	référence
		> 7 686 et ≤ 18 586	2,05	1,78-2,35
		> 18 586	3,21	2,80-3,69

La partie « logistique » du modèle de Poisson enflé en zéro indique que le type de production et la taille des élevages bovins sont des facteurs de risque pour la détection des avortements : pour une taille d'élevage donnée, la fréquence d'observation d'au moins un avortement est significativement plus élevée dans les élevages laitiers (OR=1,97 [1,65-2,37]) que dans les élevages allaitants ; de même pour un type de production donné, des élevages de taille moyenne ou grande sont associées à des fréquences de détection d'au moins un avortement significativement augmentées (respectivement OR=2,35 [2,00-2,78] et OR=3,3 [2,81-3,87]) par rapport aux élevages de petite taille.

La prise en compte dans la partie « logistique » du modèle de Poisson enflé en zéro final de l'interaction entre le type de production et la taille de l'élevage (Tableau 10) permet de mettre en évidence que l'influence du type d'élevage est fonction de sa taille (Figure 28).

La partie « comptage » du modèle de Poisson enflé en zéro indique par ailleurs que le type de production et la taille des élevages bovins français sont aussi des facteurs de confusion influençant le nombre d'avortements déclarés puisqu'un éleveur ayant détecté au moins un avortement peut n'en déclarer aucun. En effet, par rapport aux élevages allaitants, le nombre de déclarations d'avortements est significativement plus élevé dans les élevages ayant une activité mixte (1,60 [1,52-1,69]) et dans les élevages laitiers (1,75 [1,67-1,83]). De plus, quel

que soit le type de production, plus l'élevage est grand, plus le nombre d'avortements déclarés est important, puisque par rapport aux élevages de petite taille, le nombre d'avortements déclarés est significativement plus élevé dans les élevages de taille moyenne (2,05 [1,78-2,35]) et dans les élevages de grande taille (3,21 [2,80-3,69]).

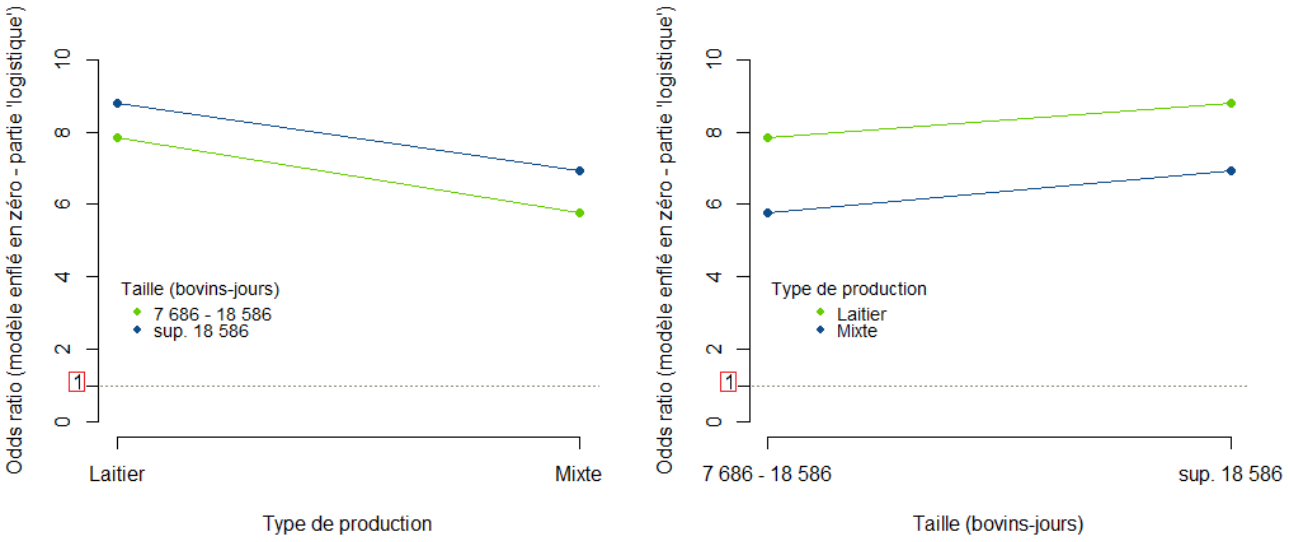


Figure 28 : Effet de l'interaction entre le type de production et la taille d'élevage sur la valeur de l'odds ratio estimée par la partie « logistique » du modèle de Poisson enflé en zéro

4) Validation des modèles

La courbe ROC pour le modèle logistique final est représentée sur la Figure 29. L'aire sous la courbe est égale à 0,7601 donc ce modèle réalise une distinction acceptable entre les élevages qui n'ont pas déclaré d'avortement et ceux qui en ont déclaré au moins un. La Figure 30 représente la probabilité estimée par le modèle logistique final qu'au moins un avortement soit déclaré dans les élevages, selon que les élevages aient effectivement déclaré ou non au moins un avortement.

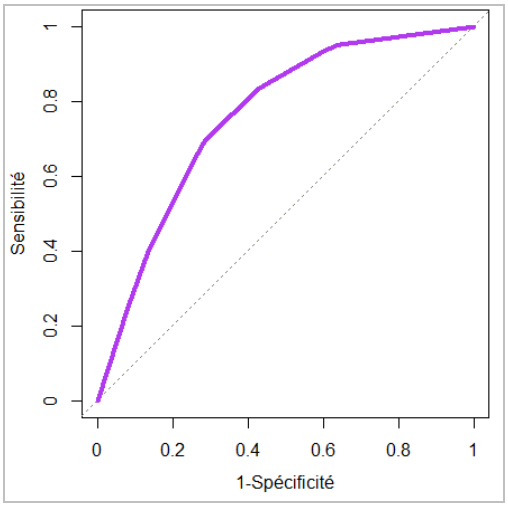


Figure 29 : Courbe ROC du modèle logistique final
La ligne en pointillée représente la diagonale (Sensibilité=1-Spécificité).

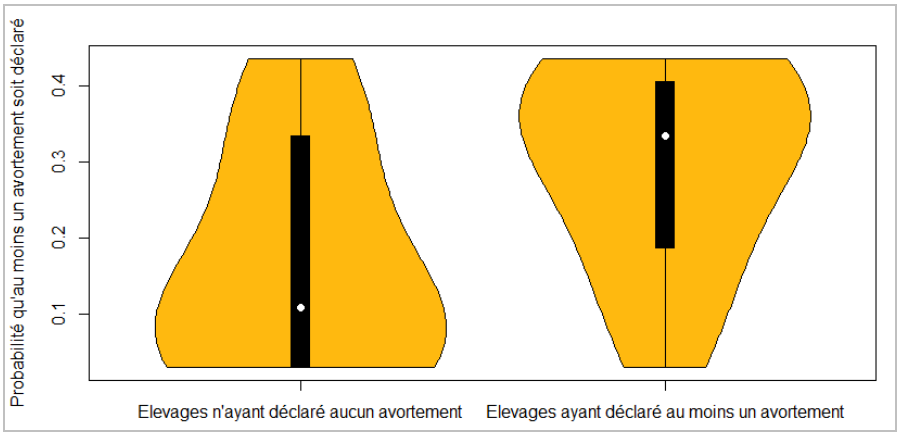


Figure 30 : Probabilité estimée par le modèle logistique final qu'au moins un avortement soit déclaré dans les élevages, selon que les élevages aient effectivement déclaré ou non au moins un avortement
La courbe représente la courbe de densité de probabilité, le point blanc représente la médiane, et le trait noir épais représente l'intervalle interquartile.

La courbe ROC pour le modèle de Poisson enflé en zéro final n'est pas présentée, elle est semblable à celle obtenue pour le modèle logistique final (Figure 29). L'aire sous la courbe est égale à 0,7601 donc ce modèle réalise aussi une distinction acceptable entre les élevages qui n'ont pas déclaré d'avortement et ceux qui en ont déclaré au moins un.

V- Discussion

Dans cette troisième partie, l'étude réalisée sur les facteurs de risque d'avortements bovins illustre très clairement les résultats théoriques obtenus suite à l'étude par simulations (Partie 2), notamment ceux obtenus lorsque le facteur de confusion est identique au facteur de risque.

Même si les facteurs « type de production », « taille des élevages » et leur interaction sont bien identifiés comme significativement associés à la survenue des avortements, les odds ratio estimés diffèrent très largement entre les deux approches statistiques, avec des valeurs très supérieures dans le modèle logistique. La taille et le type d'élevage sont des variables connues pour influencer l'efficacité de la surveillance des avortements bovins (Bronner et al., 2013). Il était donc attendu que les résultats du modèle logistique diffèrent de ceux du modèle de Poisson enflé en zéro : d'après la Partie 2, les odds ratios estimés par le modèle logistique sont supposés surestimer les odds ratio réels si la sensibilité de détection est plus faible dans les unités épidémiologiques où la variable de confusion est absente. Dans le cas des avortements bovins, d'après les résultats de la partie « comptage » du modèle de Poisson enflé en zéro, les deux facteurs étudiés améliorent la déclaration des avortements. La sensibilité de déclaration est donc plus faible dans les élevages allaitants que dans les élevages laitiers ou mixtes, et plus faible dans les élevages de moins de 7 686 bovins-jours que dans les élevages de plus de 7 686 bovins-jours (Tableau 10). Ainsi, il était attendu que le modèle logistique sélectionné pour modéliser les avortements bovins surestime les odds ratios associés à chacun des facteurs (car il ne tient pas compte de l'absence de déclaration dans certains élevages où au moins un avortement a été détecté). D'après les résultats des simulations, cette surestimation peut être de 200% (Figure 18) voire 300% dans certaines situations (résultats non présentés). D'après les résultats de l'analyse des données d'avortements présentés dans

les Tableau 9 et Tableau 10, les odds ratio estimés par le modèle logistique sont 1,5 fois plus grands que ceux estimés par le modèle de Poisson enflé en zéro (jusqu'à plus de 2 fois plus grand en ce qui concerne l'odds ratio associé aux élevages de grande taille). Ces différences entre les deux modèles sont celles mises en évidence lors des simulations lorsque les facteurs de risque et de confusion sont identiques, en considérant que les estimations obtenues avec le modèle de Poisson enflé en zéro sont les plus proches de la réalité.

Avec les données mises à notre disposition, les modèles sélectionnés permettent d'identifier que le type de production et la taille des élevages bovins sont des facteurs associés à la déclaration d'au moins un avortement dans les élevages bovins en France métropolitaine. Le modèle de Poisson enflé en zéro permet d'identifier que ces deux facteurs sont aussi des facteurs de confusion influençant la déclaration des avortements et d'estimer la force de l'association entre ces variables et la survenue d'au moins une déclaration d'avortement tout en prenant en compte l'hétérogénéité de la détection des avortements. Statistiquement, plus un élevage est grand, plus il est probable qu'au moins un avortement survienne (et soit détecté) sur une année (pris en compte par la partie « logistique » du modèle de Poisson enflé en zéro), et plus le nombre attendu d'avortements déclarés est élevé (pris en compte par la partie « comptage »). De plus, les grands élevages étant généralement associés à des meilleures pratiques de gestion de la reproduction, il semble logique que la probabilité de détecter puis déclarer au moins un avortement sur une année soit plus élevée dans les grands élevages, si au moins un survient.

Concernant l'influence du type de production, les éleveurs de troupeaux allaitants passent en général moins de temps avec leurs animaux que les éleveurs de troupeaux laitiers (qui observent les animaux deux fois par jour à la traite), notamment lors des périodes de pâture, ce qui peut expliquer qu'il y ait une meilleure détection des avortements en élevage laitier et donc un plus grand nombre moyen d'avortements déclarés dans les élevages avec au moins un avortement détecté.

Le fait que la partie « comptage » du modèle de Poisson enflé en zéro identifie le type de production et la taille des élevages bovins comme étant des facteurs de confusion nous renseigne sur l'impact qu'ont ces facteurs dans le processus de déclaration des avortements. Cependant, ce modèle seul ne permet pas de conclure si ces facteurs ont un impact sur la détection des avortements en augmentant le nombre d'avortements par élevage ou en favorisant le processus de détection en lui-même (Vergne, Korennoy, et al., 2016 ; Vergne et al., 2014)

VI- Conclusion

La détection et la déclaration des avortements bovins en France métropolitaine étant imparfaite (Bronner et al., 2013), les données disponibles sont incomplètes. L'étude menée dans cette partie consistait à identifier les facteurs de risque de ces avortements à l'aide d'un modèle logistique d'une part et d'un modèle de Poisson enflé en zéro d'autre part.

La modélisation des données avec chacun de ces modèles conduit à des résultats différents. Les deux facteurs étudiés, la taille d'élevage et le type de production, sont tous les deux identifiés comme étant des facteurs de risque. Cependant, le modèle de Poisson enflé en zéro apporte une information supplémentaire : en plus d'être associés à la détection d'avortements dans un élevage, ces deux facteurs favorisent la déclaration ou pas des avortements dans les élevages. De plus, à la lumière de ce qui a été vu dans la Partie 2, le modèle logistique surestime ici les odds ratios associés à chacun des facteurs.

L'identification des facteurs de risque est correcte avec les deux modèles, cependant, dans un contexte où les données disponibles sont incomplètes, le modèle de Poisson enflé en zéro permet de mettre en évidence que certains facteurs de risque sont aussi associés à une meilleure déclaration de la maladie étudiée, ici les avortements, et les estimations des coefficients sont plus fiables qu'avec un modèle logistique.

CONCLUSION GENERALE

Les données de surveillance passive et de surveillance active concernant les maladies animales sont le plus souvent imparfaites, en raison d'un manque de sensibilité et/ou de spécificité des méthodes de surveillance dû à l'utilisation de tests diagnostiques imparfaits, aux comportements humains face aux épidémies, etc... Ceci conduit à des informations incomplètes et potentiellement hétérogènes sur la distribution des cas de maladies, qui peuvent biaiser l'identification des facteurs de risque.

Dans cette thèse, nous avons montré grâce à une étude par simulation que l'utilisation de modèles logistiques pour décrire la distribution d'une maladie lorsque la détection de celle-ci est imparfaite conduit très rapidement à des biais sur l'estimation des forces d'association des facteurs de risque réels mais aussi à l'identification en tant que facteur de risque des facteurs de confusion associés à l'hétérogénéité de la détection. Nous avons aussi montré dans cette partie qu'en présence d'un défaut de sensibilité de la détection, l'utilisation d'un modèle de Poisson enflé en zéro serait une alternative intéressante au modèle logistique car, en permettant l'identification simultanée des facteurs de risque de présence de la maladie et des facteurs influençant la détection de la maladie (confusion), il permet d'estimer les forces d'association des facteurs de risque de présence ajustées aux facteurs de confusion.

Nous avons ensuite appliqué cette approche aux données réelles de surveillance des avortements bovins français et montré que les modèles logistiques et de Poisson enflés en zéro produisaient des résultats différents et que ces différences s'expliquaient très simplement grâce à l'étude par simulation. Lors de l'utilisation de modèles logistiques pour modéliser des données de surveillance potentiellement imparfaites, il est donc indispensable de bien identifier les potentielles sources d'imperfection de la surveillance et de discuter les effets qu'elles pourraient avoir sur les résultats obtenus. Si les données le permettent, et si une hétérogénéité de la détection est suspectée, nous recommandons même de privilégier l'utilisation de modèles de comptage enflés en zéro pour identifier les facteurs de risque d'une maladie.

ECOLE NATIONALE VETERINAIRE TOULOUSE

UNIVERSITÉ TOULOUSE III PAUL SABATIER

AGREMENT SCIENTIFIQUE

En vue de l'obtention du permis d'imprimer de la thèse de doctorat vétérinaire

Je soussigné, Fabien CORBIERE, Enseignant-chercheur, de l'Ecole Nationale Vétérinaire de Toulouse, directeur de thèse, certifie avoir examiné la thèse de **Lisa COMBELLES** intitulée **« Caractérisation des facteurs de risque à partir de données issues d'une surveillance imparfaite : comparaison des modèles de régression logistique et de Poisson enflés en zéro»** et que cette dernière peut être imprimée en vue de sa soutenance.

Fait à Toulouse, le 26 septembre 2017
Docteur Fabien CORBIERE
Maître de Conférences
de l'Ecole Nationale Vétérinaire de Toulouse

Dr. F. CORBIERE

Vu :
La Directrice de l'Ecole Nationale Vétérinaire de Toulouse
Isabelle CHMITELIN

ECOLE NATIONALE VÉTÉRINAIRE DE TOULOUSE

Vu :
Le Président du jury :
Professeur Alain GRAND

Pr A. GRAND

Vu et autorisation de l'impression :
Président de l'Université
Paul Sabatier
Monsieur Jean-Pierre VINEL

Le Président de l'Université Paul Sabatier
par délégation,
La Vice-Présidente de la CFVU

Régine ANDRE-OBRECHT

Mlle Lisa COMBELLES
a été admis(e) sur concours en : 2012
a obtenu son diplôme d'études fondamentales vétérinaires le : 23/6/2016
a validé son année d'approfondissement le : 22/06/2017
n'a plus aucun stage, ni enseignement optionnel à valider.

Université de Toulouse

BIBLIOGRAPHIE

ABDRAKHMANOV SK, SULTANOV AA, BEISEMBAYEV KK, KORENNOY FI, KUSHUBAEV DB, KADYROV AS (2016). Zoning the territory of the Republic of Kazakhstan as to the risk of rabies among various categories of animals. *Geospatial Health*, 11, 174-181.

ABRIAL D, CALAVAS D, JARRIGE N, DUCROT C (2005). Poultry, pig and the risk of BSE following the feed ban in France--a spatial analysis. *Veterinary Research*, 36, 615-628.

ADLER D (2005). *vioplot: Violin plot* [En ligne]. Disponible sur : https://cran.r-project.org/web/packages/vioplot/index.html (consulté le 8 juillet 2017)

ALARCON P, WIELAND B, MATEUS ALP, DEWBERRY C (2014). Pig farmers' perceptions, attitudes, influences and management of information in the decision-making process for disease control. *Preventive Veterinary Medicine*, 116, 223-242.

ALKHAMIS MA, VANDERWAAL K (2016). Spatial and Temporal Epidemiology of Lumpy Skin Disease in the Middle East, 2012-2015. *Frontiers in Veterinary Science*, 3, Article 19.

ALKHAMIS M, HIJMANS RJ, AL-ENEZI A, MARTÍNEZ-LÓPEZ B, PEREA AM (2016). The Use of Spatial and Spatiotemporal Modeling for Surveillance of H5N1 Highly Pathogenic Avian Influenza in Poultry in the Middle East. *Avian Diseases*, 60, 146-155.

BAPTISTA FM, ALBAN L, NIELSEN LR, DOMINGOS I, POMBA C, ALMEIDA V (2010). Use of herd information for predicting Salmonella status in pig herds. *Zoonoses and Public Health*, 57, 49-59.

BARNES AP, MOXEY AP, VOSOUGH AHMADI B, BORTHWICK FA (2015). The effect of animal health compensation on 'positive' behaviours towards exotic disease reporting and implementing biosecurity: A review, a synthesis and a research agenda. *Preventive Veterinary Medicine*, 122, 42-52.

BAUDOT J-Y (2014). Corrélation de Kendall. *Techniques et concepts de l'entreprise, de la finance et de l'économie (et fondements mathématiques)* [En ligne]. Disponible sur : http://www.jybaudot.fr/Correl_regress/kendall.html (consulté le 14 juin 2017)

BÉNET J, PRAUD A (2016). La tuberculose animale. *Polycopié des Unités de maladies contagieuses des Ecoles Nationales Vétérinaires françaises*, Mérial (Lyon), 102 pages.

BENSCHOP J, SPENCER S, ALBAN L, STEVENSON M, FRENCH N (2010). Bayesian zero-inflated predictive modelling of herd-level Salmonella prevalence for risk-based surveillance. *Zoonoses and Public Health*, 57, 60-70.

BOUSSINI H, TRAORÉ T, TAMBOURA H, BESSIN R, BOLY H, OUÉDRAOGO A (2012). Prévalence de la tuberculose et de la brucellose dans les élevages bovins laitiers intra-urbains et périurbains de la ville d'Ouagadougou au Burkina Faso. *Revue scientifique et technique de l'Office international des épizooties*, 31, n°3, 943-951.

BRONNER A, HÉNAUX V, VERGNE T, VINARD J-L, MORIGNAT E, HENDRIKX P, CALAVAS D, GAY E (2013). Assessing the Mandatory Bovine Abortion Notification System in France Using Unilist Capture-Recapture Approach. *PLOS ONE*, 8, e63246.

BYRNE AW, MCBRIDE S, LAHUERTA-MARIN A, GUELBENZU M, MCNAIR J, SKUCE RA, MCDOWELL SWJ (2016). Liver fluke (Fasciola hepatica) infection in cattle in Northern Ireland: a large-scale epidemiological investigation utilising surveillance data. *Parasites & Vectors*, 9, Article 209.

COOKBOOK FOR R (2016). Multiple graphs on one page (ggplot2). *Cookbook for R* [En ligne]. Disponible sur : http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/ (consulté le 10 mai 2017)

COWLED BD, STEVENSON MA, MADIN B (2016). An assessment of the association between soil pH and ovine Johne's disease using Australian abattoir surveillance data. *Preventive Veterinary Medicine*, 126, 208-219.

DEL RIO VILAS VJ, ANCELET S, ABELLAN JJ, BIRCH CPD, RICHARDSON S (2011). A Bayesian hierarchical analysis to compare classical and atypical scrapie surveillance data; Wales 2002-2006. *Preventive Veterinary Medicine*, 98, 29-38.

DELGADO AH, NORBY B, SCOTT HM, DEAN W, MCINTOSH WA, BUSH E (2014). Distribution of cow–calf producers' beliefs about reporting cattle with clinical signs of foot-and-mouth disease to a veterinarian before or during a hypothetical outbreak. *Preventive Veterinary Medicine*, 117, 505-517.

DHINGRA MS, DISSANAYAKE R, NEGI AB, OBEROI M, CASTELLAN D, THRUSFIELD M, LINARD C, GILBERT M (2014). Spatio-temporal epidemiology of highly pathogenic avian influenza (subtype H5N1) in poultry in eastern India. *Spatial and Spatio-Temporal Epidemiology*, 11, 45-57.

DOHERR MG, AUDIGÉ L (2001). Monitoring and surveillance for rare health-related events: a review from the veterinary perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356, 1097-1106.

DOHOO I, MARTIN W, STRYHN H (2009). *Veterinary Epidemiologic Research*. 2nd edition. Charlottetown, CA : VER Inc. 865 p. ISBN : 978-0-919013-60-5.

ELBERS ARW, GORGIEVSKI MJ, ZARAFSHANI K, KOCH G (2010). To report or not to report: a psychosocial investigation aimed at improving early detection of avian influenza outbreaks. *Revue scientifique et technique / Office International des Epizooties*, 29, 435-449.

ELBERS ARW, GORGIEVSKI-DUIJVESTEIJN MJ, VAN DER VELDEN PG, LOEFFEN WLA, ZARAFSHANI K (2010). A socio-psychological investigation into limitations and incentives concerning reporting a clinically suspect situation aimed at improving early detection of classical swine fever outbreaks. *Veterinary Microbiology*, 142, 108-118.

ELLIS-IVERSEN J, COOK AJC, WATSON E, NIELEN M, LARKIN L, WOOLDRIDGE M, HOGEVEEN H (2010). Perceptions, circumstances and motivators that influence implementation of zoonotic control programs on cattle farms. *Preventive Veterinary Medicine*, 93, 276-285.

FAO - FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS (2014). EMPRES-i - Global Animal Disease Information System. Disponible sur : http://empres-i.fao.org/eipws3g/ (consulté le 20 juillet 2017)

FORAR AL, GAY JM, HANCOCK DD (1995). The frequency of endemic fetal loss in dairy cattle: A review. *Theriogenology*, 43, n°6, 989-1000. Disponible sur : < http://dx.doi.org/10.1016/0093-691X(95)00063-E >

GACHE K, HOSTEING S, PERRIN J-B, ZIENTARA S, TOURATIER A (2015). Surveillance de l'infection congénitale par le virus Schmallenberg en France : une circulation moins intense en 2013. *Bulletin épidémiologique, santé animale et alimentation*, n°67, 15-18.

GANIÈRE J-P (2017). Cours de réglementation sanitaire générale. *Polycopié des Unités de maladies contagieuses des Ecoles Nationales Vétérinaires françaises*, Mérial (Lyon), 124 pages.

GANIÈRE J-P, LAABERKI M-H (2017). La brucellose animale. *Polycopié des Unités de maladies contagieuses des Ecoles Nationales Vétérinaires françaises*, Mérial (Lyon), 58 pages.

GILBERT M, PFEIFFER DU (2012). Risk factor modelling of the spatio-temporal patterns of highly pathogenic avian influenza (HPAIV) H5N1: A review. *Spatial and Spatio-temporal Epidemiology*, 3, 173-183.

GOULET V, AUCLAIR S, DUTANG C, MILHAUD X, OUELLET T, POULIOT L-P, PIGEON M (2017). *actuar: Actuarial Functions and Heavy Tailed Distributions* [En ligne]. Disponible sur : https://cran.r-project.org/web/packages/actuar/index.html (consulté le 29 mars 2017)

GULENKIN VM, KORENNOY FI, KARAULOV AK, DUDNIKOV SA (2011). Cartographical analysis of African swine fever outbreaks in the territory of the Russian Federation and computer modeling of the basic reproduction ratio. *Preventive Veterinary Medicine*, 102, 167-174.

HAYAMA Y, MORIGUCHI S, YANASE T, SUZUKI M, NIWA T, IKEMIYAGI K, NITTA Y, YAMAMOTO T, KOBAYASHI S, MURAI K, TSUTSUI T (2016). Epidemiological analysis of bovine ephemeral fever in 2012-2013 in the subtropical islands of Japan. *BMC veterinary research*, 12, Article 47.

HOPP P, VATN S, JARP J (2007). Norwegian farmers' vigilance in reporting sheep showing scrapie-associated signs. *BMC Veterinary Research*, 3, Article 34.

HOTHORN T, ZEILEIS A, (PAN.F) RWF, (PAN.F) CC, MILLO G, MITCHELL D (2017). *lmtest: Testing Linear Regression Models* [En ligne]. Disponible sur : https://cran.r-project.org/web/packages/lmtest/index.html (consulté le 5 juillet 2017)

HUMBLET M-F, GILBERT M, GOVAERTS M, FAUVILLE-DUFAUX M, WALRAVENS K, SAEGERMAN C (2010). New assessment of bovine tuberculosis risk factors in Belgium based on nationwide molecular epidemiology. *Journal of Clinical Microbiology*, 48, 2802-2808.

JACKMAN S, TAHK WITH CONTRIBUTIONS FROM A, ZEILEIS A, FEARON CM AND J (2015). *pscl: Political Science Computational Laboratory, Stanford University* [En ligne]. Disponible sur : https://cran.r-project.org/web/packages/pscl/index.html (consulté le 29 mars 2017)

KORENNOY FI, GULENKIN VM, MALONE JB, MORES CN, DUDNIKOV SA, STEVENSON MA (2014). Spatio-temporal modeling of the African swine fever epidemic in the Russian Federation, 2007-2012. *Spatial and Spatio-Temporal Epidemiology*, 11, 135-141.

KUCHLER F, HAMM S (2000). Animal disease incidence and indemnity eradication programs. *Agricultural Economics*, 22, 299-308.

LAMBERT D (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34, 1-14.

LEE B-Y, HIGGINS IM, MOON O-K, CLEGG TA, MCGRATH G, COLLINS DM, PARK J-Y, YOON H-C, LEE S-J, MORE SJ (2009). Surveillance and control of bovine brucellosis in the Republic of Korea during 2000-2006. *Preventive Veterinary Medicine*, 90, 66-79.

LOTH L, GILBERT M, WU J, CZARNECKI C, HIDAYAT M, XIAO X (2011). Identifying risk factors of highly pathogenic avian influenza (H5N1 subtype) in Indonesia. *Preventive Veterinary Medicine*, 102, 50-58.

MACKENZIE DI, NICHOLS JD, LACHMAN GB, DROEGE S, ANDREW ROYLE J, LANGTIMM CA (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83, 2248-2255.

MARTIN TG, WINTLE BA, RHODES JR, KUHNERT PM, FIELD SA, LOW-CHOY SJ, TYRE AJ, POSSINGHAM HP (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8, 1235-1246.

MARTIN V, PFEIFFER DU, ZHOU X, XIAO X, PROSSER DJ, GUO F, GILBERT M (2011). Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. *PLoS pathogens*, 7, e1001308.

MARTÍNEZ-LÓPEZ B, ALEXANDROV T, MUR L, SÁNCHEZ-VIZCAÍNO F, SÁNCHEZ-VIZCAÍNO JM (2014). Evaluation of the spatial patterns and risk factors, including backyard pigs, for classical swine fever occurrence in Bulgaria using a Bayesian model. *Geospatial Health*, 8, 489-501.

MCLEOD AI (2011). *Kendall: Kendall rank correlation and Mann-Kendall trend test* [En ligne]. Disponible sur : https://cran.r-project.org/web/packages/Kendall/index.html (consulté le 30 juin 2017)

MCLEOD AI (2015). Package 'Kendall'. 12 pages.

MERCIER P, MÉSI F, MÉMETEAU S (2011). Paratuberculose : éléments d'épidémiologie et description du plan de lutte français. *Bulletin épidémiologique, santé animale et alimentation*, n°47, 2-7.

MICROSOFT (2003). Microsoft Excel.

MINISTRE DE L'AGRICULTURE ET DE LA PÊCHE (2008). *Arrêté du 22 avril 2008 fixant les mesures techniques et administratives relatives à la prophylaxie collective et à la police sanitaire de la brucellose des bovinés*.

NAMATA H, WELBY S, AERTS M, FAES C, ABRAHANTES JC, IMBERECHTS H, VERMEERSCH K, HOOYBERGHS J, MÉROC E, MINTIENS K (2009). Identification of risk factors for the prevalence and persistence of Salmonella in Belgian broiler chicken flocks. *Preventive Veterinary Medicine*, 90, 211-222.

NETRABUKKANA P, CAPPELLE J, TREVENNEC C, ROGER F, GOUTARD F, BUCHY P, ROBERTSON ID, FENWICK S (2015). Epidemiological Analysis of Influenza A Infection in Cambodian Pigs and Recommendations for Surveillance Strategies. *Transboundary and Emerging Diseases*, 62, 37-44.

NEUWIRTH E (2014). *RColorBrewer: ColorBrewer Palettes* [En ligne]. Disponible sur : https://cran.r-project.org/web/packages/RColorBrewer/index.html (consulté le 25 avril 2017)

OIE - WORLD ORGANISATION FOR ANIMAL HEALTH (2012). OIE World Animal Health Information System. Disponible sur : http://www.oie.int/wahis_2/public/wahid.php/Wahidhome/Home (consulté le 20 juillet 2017)

OIE - WORLD ORGANISATION FOR ANIMAL HEALTH (2017). Maladies de la Liste de l'OIE 2017. Disponible sur : http://www.oie.int/fr/sante-animale-dans-le-monde/oie-listed-diseases-2017/ (consulté le 27 mars 2017)

PASCUAL-LINAZA AV, MARTÍNEZ-LÓPEZ B, PFEIFFER DU, MORENO JC, SANZ C, SÁNCHEZ-VIZCAÍNO JM (2014). Evaluation of the spatial and temporal distribution of and risk factors for Bluetongue serotype 1 epidemics in sheep Extremadura (Spain), 2007-2011. *Preventive Veterinary Medicine*, 116, 279-295.

PAUL MC, GOUTARD FL, ROULLEAU F, HOLL D, THANAPONGTHARM W, ROGER FL, TRAN A (2016). Quantitative assessment of a spatial multicriteria model for highly pathogenic avian influenza H5N1 in Thailand, and application in Cambodia. *Scientific Reports*, 6, 31096.

PEREZ A, ALKHAMIS M, CARLSSON U, BRITO B, CARRASCO-MEDANIC R, WHEDBEE Z, WILLEBERG P (2011). Global animal disease surveillance. *Spatial and Spatio-temporal Epidemiology*, 2, 135-145.

PEROZ C, GANIÈRE J-P (2017). Dangers sanitaires de 1ère et 2ème catégories chez les ruminants. *Polycopié des Unités de maladies contagieuses des Ecoles Nationales Vétérinaires françaises*, Mérial (Lyon), 132 pages.

PFEIFFER DU, MINH PQ, MARTIN V, EPPRECHT M, OTTE MJ (2007). An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. *The Veterinary Journal*, 174, 302-309.

PORPHYRE T, JACKSON R, SAUTER-LOUIS C, WARD D, BAGHYAN G, STEPANYAN E (2010). Mapping brucellosis risk in communities in the Republic of Armenia. *Geospatial Health*, 5, 103-118.

PORPHYRE T, STEVENSON MA, MCKENZIE J (2008). Risk factors for bovine tuberculosis in New Zealand cattle farms and their relationship with possum control strategies. *Preventive Veterinary Medicine*, 86, 93-106.

PUBMED - NCBI (2017). PubMed - NCBI. *PubMed - NCBI* [En ligne]. Disponible sur : https://www.ncbi.nlm.nih.gov/pubmed/ (consulté le 16 février 2017)

R CORE TEAM (2017). The Comprehensive R Archive Network. Disponible sur : https://cran.r-project.org/ (consulté le 29 mars 2017)

RAKOTOMALALA R (2011). Pratique de la régression logistique. *Régression Logistique Binaire et Polytomique, Université Lumière Lyon*, 2, 258 pages.

ROBIN X, TURCK N, HAINARD A, TIBERTI N, LISACEK F, SANCHEZ J-C, MÜLLER M, CODE) SS (FAST D (2017). *pROC: Display and Analyze ROC Curves* [En ligne]. Disponible sur : https://cran.r-project.org/web/packages/pROC/index.html (consulté le 30 juin 2017)

RODRÍGUEZ-PRIETO V, MARTÍNEZ-LÓPEZ B, BARASONA JA, ACEVEDO P, ROMERO B, RODRIGUEZ-CAMPOS S, GORTÁZAR C, SÁNCHEZ-VIZCAÍNO JM, VICENTE J (2012). A Bayesian approach to study the risk variables for tuberculosis occurrence in domestic and wild ungulates in South Central Spain. *BMC veterinary research*, 8, Article 148.

ROYLE JA, NICHOLS JD, KÉRY M (2005). Modelling occurrence and abundance of species when detection is imperfect. *Oikos*, 110, 353-359.

SAKSENA S, FOX J, EPPRECHT M, TRAN CC, NONG DH, SPENCER JH, NGUYEN L, FINUCANE ML, TRAN VD, WILCOX BA (2015). Evidence for the Convergence Model: The Emergence of Highly Pathogenic Avian Influenza (H5N1) in Viet Nam. *PloS One*, 10, e0138138.

SHITTU A, CLIFTON-HADLEY RS, ELY ER, UPTON PU, DOWNS SH (2013). Factors associated with bovine tuberculosis confirmation rates in suspect lesions found in cattle at routine slaughter in Great Britain, 2003-2008. *Preventive Veterinary Medicine*, 110, 395-404.

SINDATO C, KARIMURIBO ED, PFEIFFER DU, MBOERA LEG, KIVARIA F, DAUTU G, BERNARD B, PAWESKA JT (2014). Spatial and temporal pattern of Rift Valley fever outbreaks in Tanzania; 1930 to 2007. *PloS One*, 9, e88897.

STACK OVERFLOW (2012). Any way to make plot points in scatterplot more transparent in R?. *Stack Overflow* [En ligne]. Disponible sur : http://stackoverflow.com/questions/12995683/any-way-to-make-plot-points-in-scatterplot-more-transparent-in-r (consulté le 31 mars 2017)

STEVENS KB, GILBERT M, PFEIFFER DU (2013). Modeling habitat suitability for occurrence of highly pathogenic avian influenza virus H5N1 in domestic poultry in Asia: a spatial multicriteria decision analysis approach. *Spatial and Spatio-Temporal Epidemiology*, 4, 1-14.

THANAPONGTHARM W, LINARD C, PAMARANON N, KAWKALONG S, NOIMOH T, CHANACHAI K, PARAKGAMAWONGSA T, GILBERT M (2014). Spatial epidemiology of porcine reproductive and respiratory syndrome in Thailand. *BMC veterinary research*, 10, Article 174.

TOMA B, DUFOUR B, RIVIÈRE J, PICAVET D, MOUTOU F, ZIENTARA S (2017). La fièvre aphteuse. *Polycopié des Unités de maladies contagieuses des Ecoles Nationales Vétérinaires françaises*, Mérial (Lyon), 67 pages.

TREVENNEC K, LEGER L, LYAZRHI F, BAUDON E, CHEUNG CY, ROGER F, PEIRIS M, GARCIA J-M (2012). Transmission of pandemic influenza H1N1 (2009) in Vietnamese swine in 2009-2010. *Influenza and Other Respiratory Viruses*, 6, 348-357.

VERGNE T, GUINAT C, PETKOVA P, GOGIN A, KOLBASOV D, BLOME S, MOLIA S, PINTO FERREIRA J, WIELAND B, NATHUES H, PFEIFFER D (2016). Attitudes and Beliefs of Pig Farmers and Wild Boar Hunters Towards Reporting of African Swine Fever in Bulgaria, Germany and the Western Part of the Russian Federation. *Transboundary and Emerging Diseases*, 63, 194-204.

VERGNE T, KORENNOY F, COMBELLES L, GOGIN A, PFEIFFER D (2016). Modelling African swine fever presence and reported abundance in the Russian Federation using national surveillance data from 2007 to 2014. *Spatial and Spatio-Temporal Epidemiology*, 19, 70-77.

VERGNE T (2012). *Les méthodes de capture-recapture pour évaluer les systèmes de surveillance en santé animale*. PhD thesis, Université Paris Sud - Paris XI, 228 p.

VERGNE T, DEL RIO VILAS VJ, CAMERON A, DUFOUR B, GROSBOIS V (2015). Capture–recapture approaches and the surveillance of livestock diseases: A review. *Preventive Veterinary Medicine*, 120, 253-264.

VERGNE T, PAUL MC, CHAENGPRACHAK W, DURAND B, GILBERT M, DUFOUR B, ROGER F, KASEMSUWAN S, GROSBOIS V (2014). Zero-inflated models for identifying disease risk factors when case

detection is imperfect: Application to highly pathogenic avian influenza H5N1 in Thailand. *Preventive Veterinary Medicine*, 114, 28-36.

WALSH MG, AMSTISLAVSKI P, GREENE A, HASEEB MA (2016). The Landscape Epidemiology of Seasonal Clustering of Highly Pathogenic Avian Influenza (H5N1) in Domestic Poultry in Africa, Europe and Asia. *Transboundary and Emerging Diseases*, 14 pages.

WELSH AH, CUNNINGHAM RB, DONNELLY CF, LINDENMAYER DB (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88, 297-308.

WICKHAM H, CHANG W (2016). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* [En ligne]. Disponible sur : https://cran.r-project.org/web/packages/ggplot2/index.html (consulté le 14 avril 2017)

WICKHAM H, FRANCOIS R (2016). *dplyr: A Grammar of Data Manipulation* [En ligne]. Disponible sur : https://cran.r-project.org/web/packages/dplyr/index.html (consulté le 25 avril 2017)

WILESMITH JW, RYAN JBM, HUESTON WD (1992). Bovine spongiform encephalopathy: case-control studies of calf feeding practices and meat and bonemeal inclusion in proprietary concentrates. *Research in Veterinary Science*, 52, n°3, 325-331.

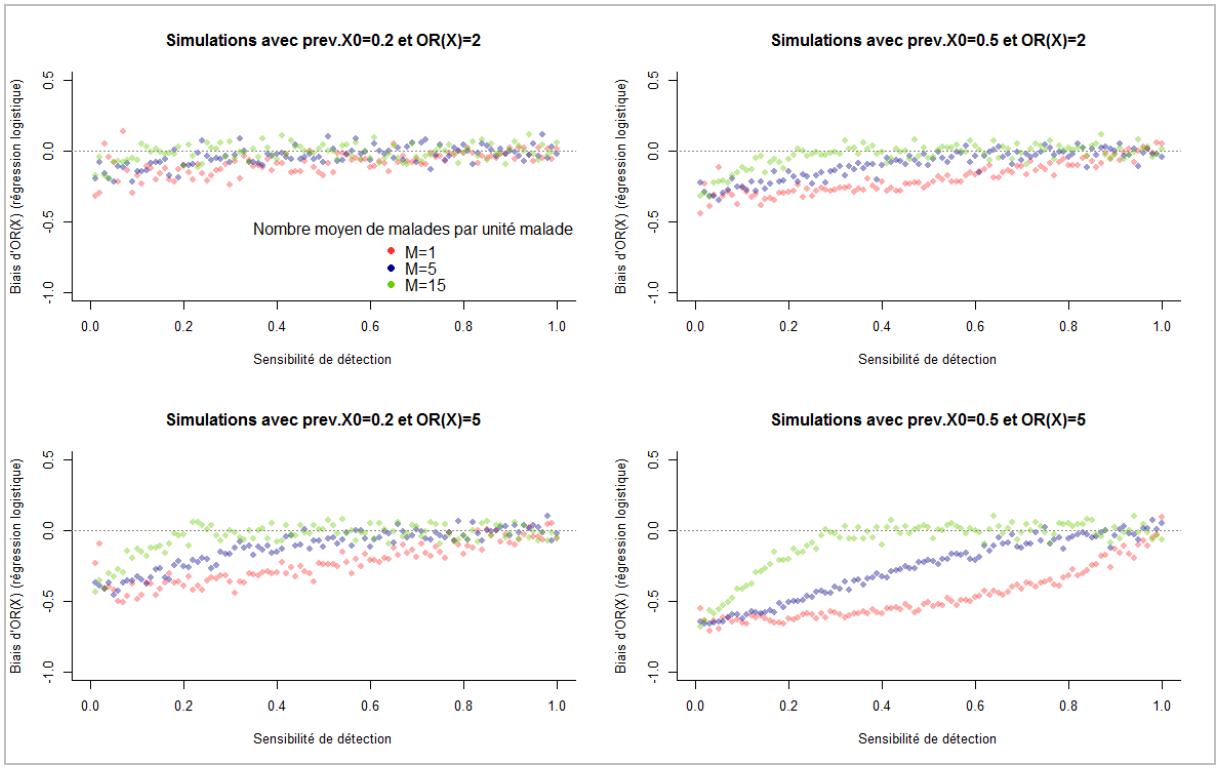
WINELAND NE, DETWILER LA, SALMAN MD (1998). Epidemiologic analysis of reported scrapie in sheep in the United States: 1,117 cases (1947-1992). *Journal of the American Veterinary Medical Association*, 212, 713-718.

WOLFE DM, BERKE O, KELTON DF, WHITE PW, MORE SJ, O'KEEFFE J, MARTIN SW (2010). From explanation to prediction: a model for recurrent bovine tuberculosis in Irish cattle herds. *Preventive Veterinary Medicine*, 94, 170-177.

ZEILEIS A, KLEIBER C, JACKMAN S (2008). Regression models for count data in R. *Journal of statistical software*, 25 pages.

ANNEXES

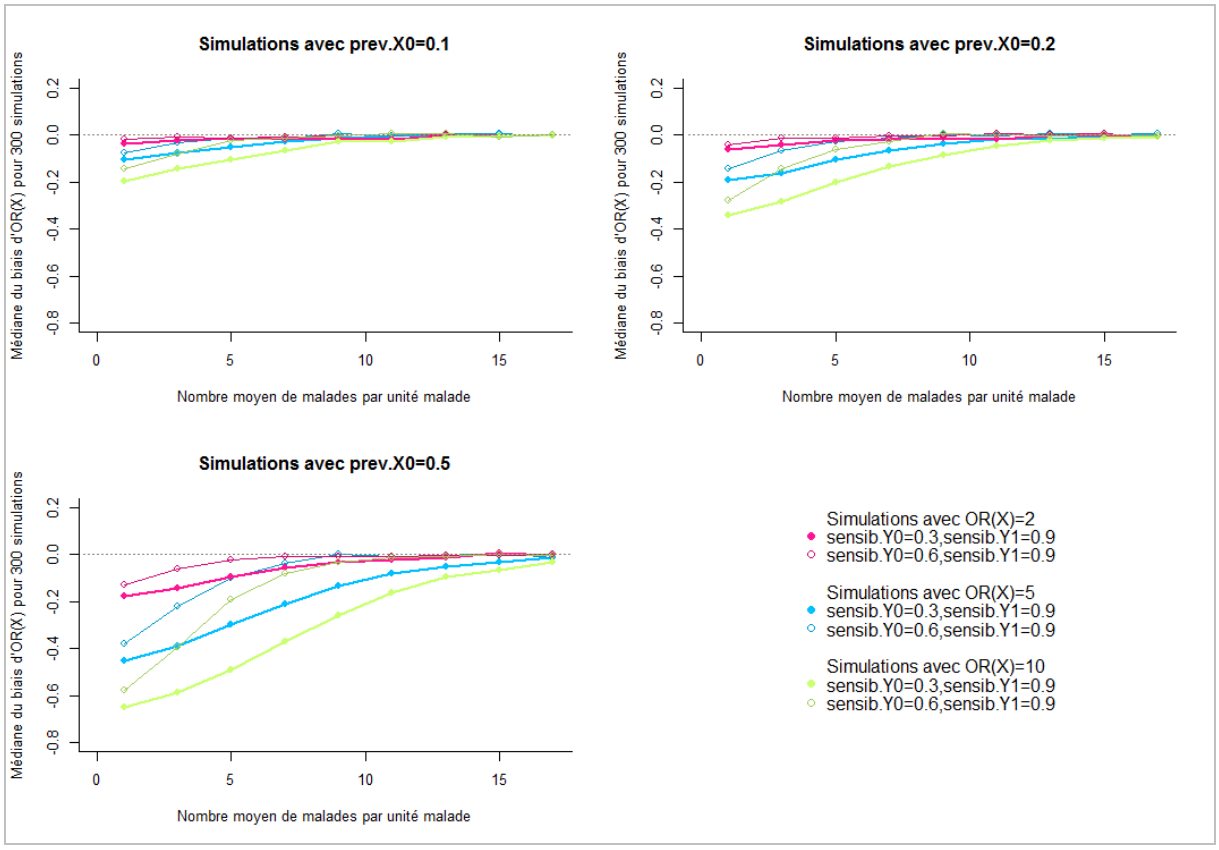
Annexe 1 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade, valeurs de prev.X0 et valeurs réelles d'OR(X)



Annexe 1 - Figure 31 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade, valeurs de prev.X0 et valeurs réelles d'OR(X)

Lorsque la sensibilité de détection est la même pour toutes les unités épidémiologiques, l'odds ratio associé au facteur X pour le modèle logistique est systématiquement sous-estimé, et ce biais tend vers 0 lorsque la sensibilité de détection tend vers 1. En outre, quelles que soient les situations, lorsque la prévalence intra-unité moyenne augmente, le biais tend rapidement vers 0 même pour de petites sensibilités de détection ; ceci est d'autant plus marqué que l'odds ratio réel et la probabilité de présence de la maladie dans les unités épidémiologiques où il n'y a pas le facteur X sont grands. Cette augmentation du biais avec l'augmentation de l'odds ratio réel et de la probabilité de présence de la maladie dans les unités épidémiologiques où il n'y a pas le facteur X a déjà été vue dans la Figure 9, et ce biais est cette fois d'autant plus marqué que le nombre moyen d'unités élémentaires malades par unité épidémiologique malade est faible.

Annexe 2 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection, valeurs de prev.X0 et valeurs réelles d'OR(X)

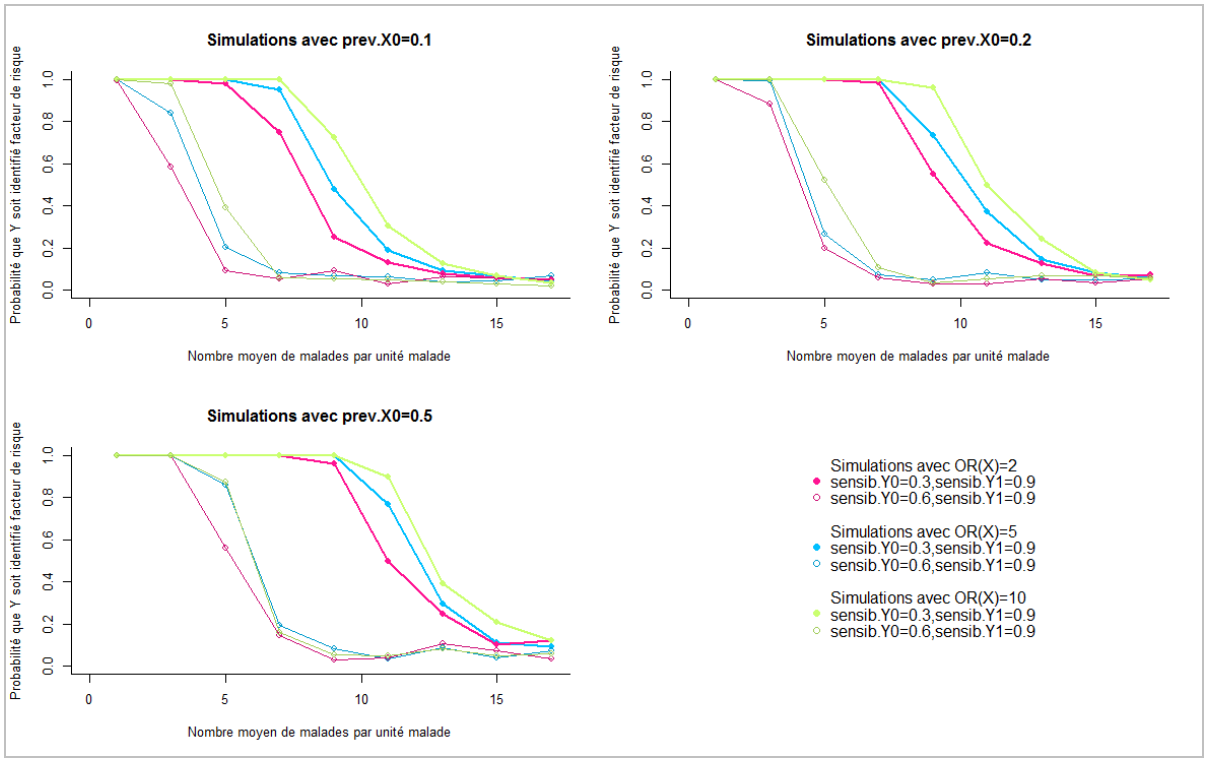


Annexe 2 - Figure 32 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection, valeurs de prev.X0 et valeurs réelles d'OR(X)

L'odds ratio calculé dépend des valeurs de la sensibilité de détection et de la prévalence intra-unité moyenne, ainsi que de la probabilité de présence de la maladie dans les unités épidémiologiques où il n'y a pas le facteur X et de la valeur réelle de l'odds ratio. On retrouve les tendances déjà décrites précédemment :

- plus la prévalence intra-unité moyenne augmente, plus le biais diminue ;
- plus la sensibilité de détection est parfaite, plus le biais diminue (comparaison du couple « sensib.Y0=0,6 et sensib.Y1=0,9 » par rapport au couple « sensib.Y0=0,3 et sensib.Y1=0,9 ») ;
- plus OR(X) augmente, plus le biais augmente ;
- plus prev.X0 augmente, plus le biais augmente.

Annexe 3 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection, valeurs de prev.X0 et valeurs réelles d'OR(X)

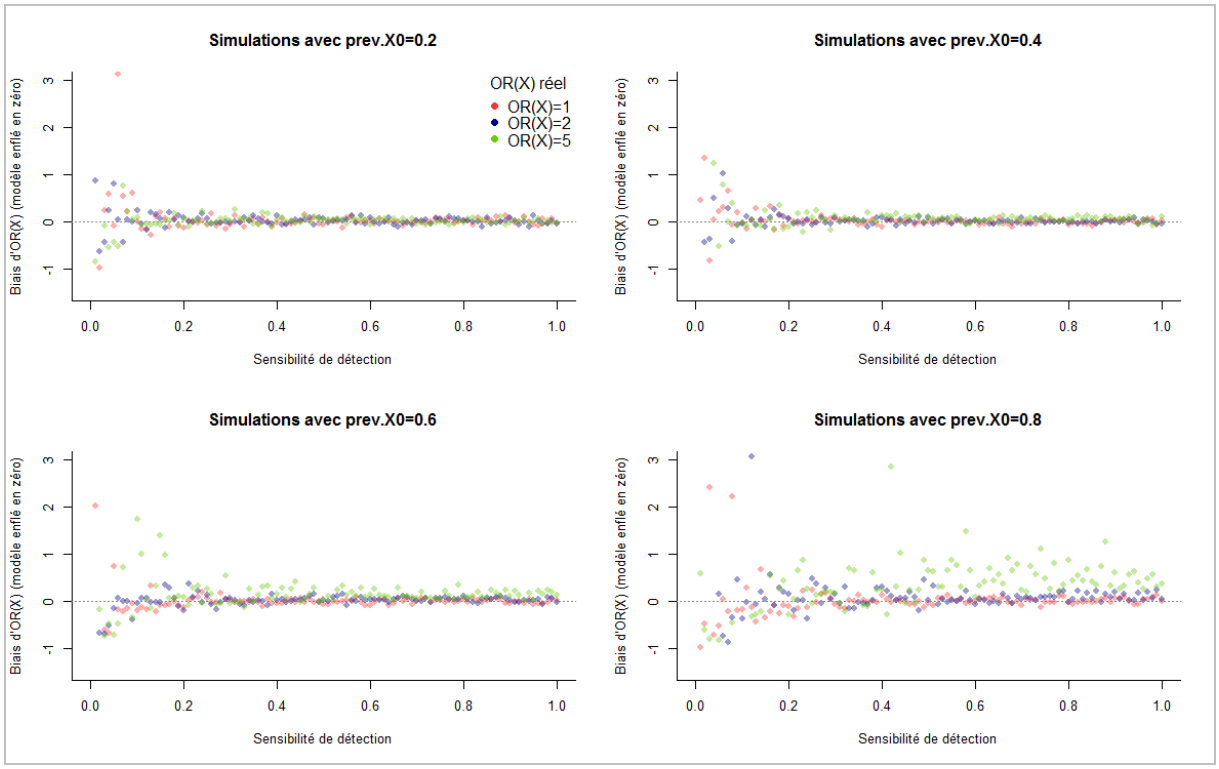


Annexe 3 - Figure 33 : Evolution de la probabilité que le facteur Y soit identifié comme étant un facteur de risque par un modèle logistique en fonction du nombre moyen d'unités élémentaires malades par unité épidémiologique malade et pour différentes valeurs de sensibilité de détection, valeurs de prev.X0 et valeurs réelles d'OR(X)

De manière générale, la probabilité que le facteur Y soit identifié (à tort) par le modèle logistique comme étant un facteur de risque diminue lorsque la prévalence intra-unité moyenne augmente. Par ailleurs, on retrouve les tendances déjà décrites dans la Figure 15 :

- plus la sensibilité de détection est parfaite, moins il est probable que le modèle logistique identifie le facteur Y comme étant un facteur de risque (comparaison du couple « sensib.Y0=0,6 et sensib.Y1=0,9 » par rapport au couple « sensib.Y0=0,3 et sensib.Y1=0,9 ») ;
- plus OR(X) est grand, plus il est probable que le modèle logistique identifie le facteur Y comme étant un facteur de risque ;
- plus prev.X0 est grand, plus Y est identifié comme étant un facteur de risque.

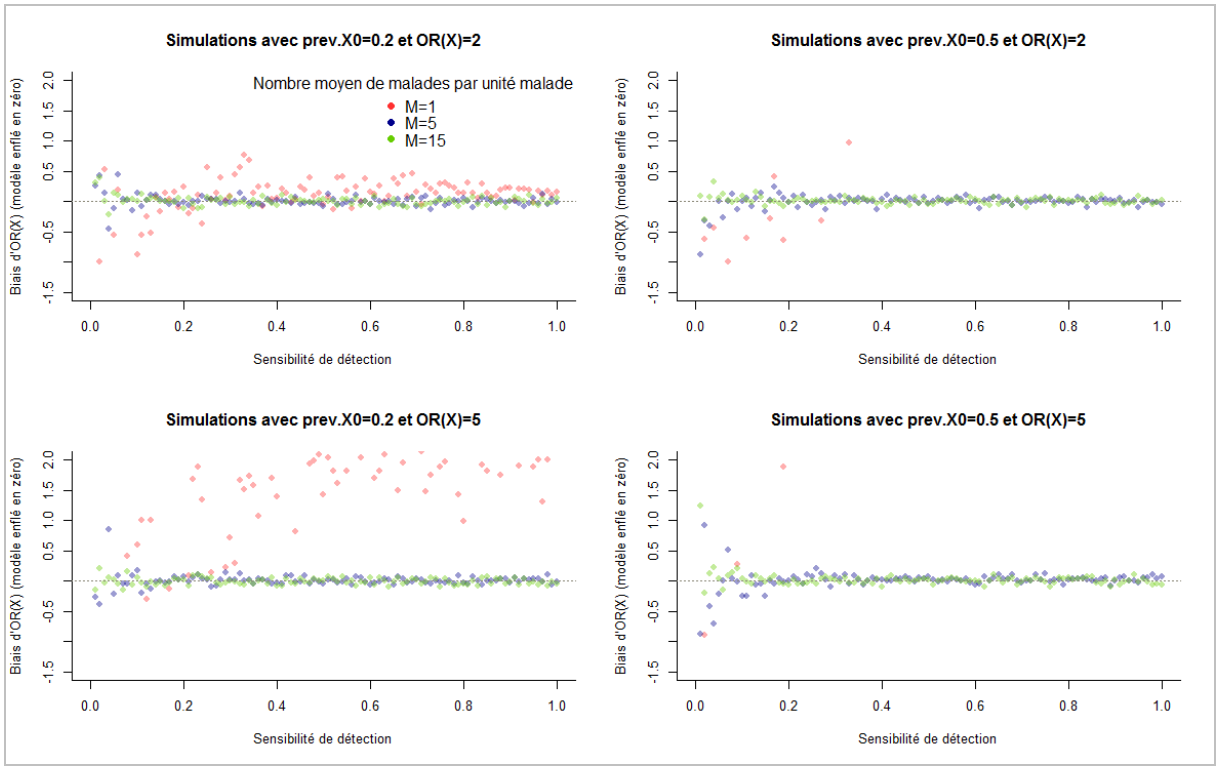
Annexe 4 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4



Annexe 4 - Figure 34 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X) lorsque M=4

Pour une prévalence intra-unité moyenne fixée, lorsque la sensibilité de détection est la même pour toutes les unités épidémiologiques, le biais de l'odds ratio associé au facteur X pour le modèle de Poisson enflé en zéro tend très rapidement vers 0 lorsque la sensibilité de détection augmente, et ce quelle que soit la valeur de la probabilité de présence de la maladie dans les unités épidémiologiques où il n'y a pas le facteur X. Cependant, on peut noter que pour des probabilités de présence de la maladie importantes malgré l'absence du facteur X (prev.X0 élevée), le biais a une légère tendance à augmenter et à rester positif lorsque l'odds ratio réel augmente. Le biais relatif est ainsi important pour des valeurs de prev.X0=0,8 et OR(X)=5 (panneau inférieur droit).

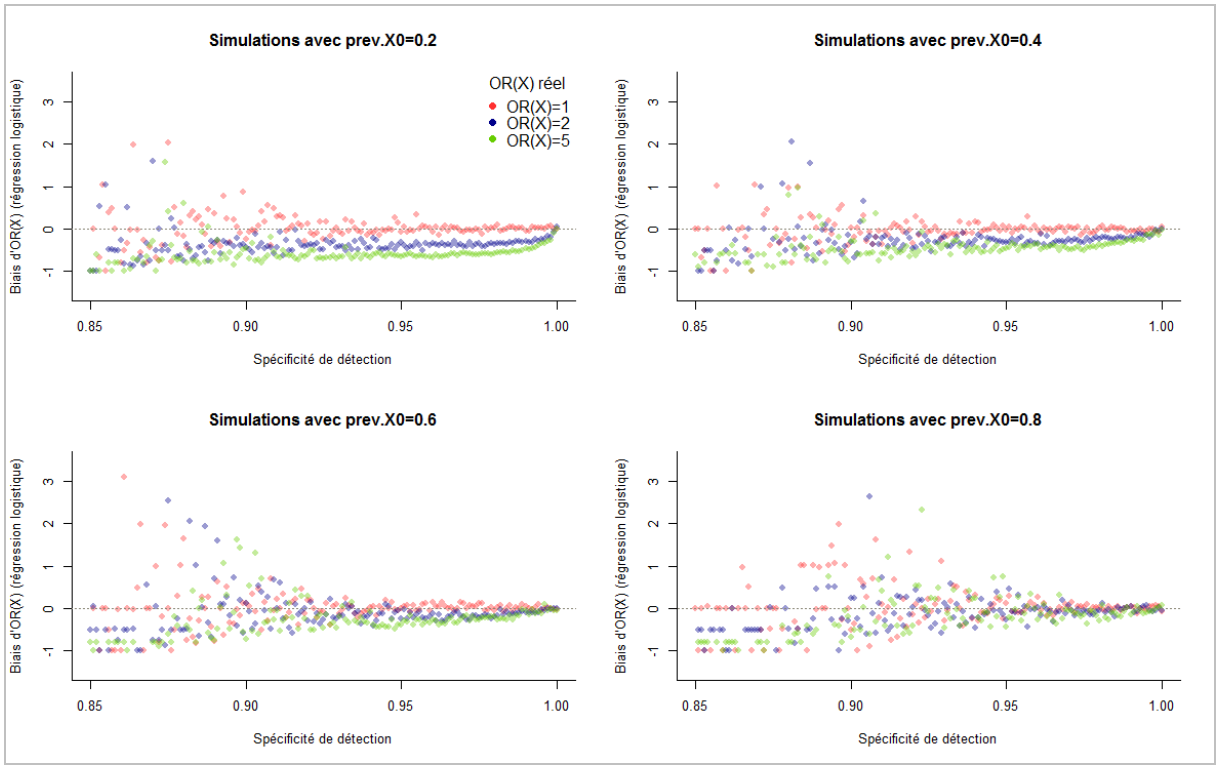
Annexe 5 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade, valeurs de prev.X0 et valeurs réelles d'OR(X)



Annexe 5 - Figure 35 : Evolution du biais relatif de l'odds ratio de X estimé par la partie « logistique » d'un modèle de Poisson enflé en zéro en fonction de la sensibilité de détection et pour différentes valeurs du nombre moyen d'unités élémentaires malades par unité épidémiologique malade, valeurs de prev.X0 et valeurs réelles d'OR(X)

Lorsque la sensibilité de détection est la même pour toutes les unités épidémiologiques, le biais de l'odds ratio associé au facteur X pour le modèle de Poisson enflé en zéro tend très rapidement vers 0 lorsque la sensibilité de détection augmente, et cette tendance ne dépend ni de l'odds ratio réel, ni de la probabilité de présence de la maladie dans les unités épidémiologiques où il n'y a pas le facteur X, ni du nombre moyen d'unités élémentaires malades par unité épidémiologique malade. Une exception cependant, lorsqu'il n'y a en moyenne qu'un seul malade par unité épidémiologique malade, le biais augmente de manière importante, et ce d'autant plus que l'odds ratio réel et la probabilité de présence de la maladie dans les unités épidémiologiques où il n'y a pas le facteur X sont grands.

Annexe 6 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la spécificité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X)



Annexe 6 - Figure 36 : Evolution du biais relatif de l'odds ratio de X estimé par un modèle logistique en fonction de la spécificité de détection et pour différentes valeurs de prev.X0 et valeurs réelles d'OR(X)

Plus la valeur de prev.X0 augmente, et plus la spécificité pour laquelle l'estimation de l'odds ratio se précise augmente (autour de 90% lorsque prev.X0=0,2 et autour de 95% lorsque prev.X0=0,8). De plus, plus la valeur de prev.X0 augmente, plus le biais dans l'estimation de l'odds ratio tend rapidement vers 0 lorsque ce biais se précise (et ce quelle que soit la valeur réelle de l'odds ratio).

Toulouse, 2017

NOM : COMBELLES **PRENOM : Lisa**

TITRE : Caractérisation des facteurs de risque à partir de données issues d'une surveillance imparfaite : comparaison des modèles de régression logistique et de Poisson enflés en zéro.

Résumé : Les facteurs de risque sont des concepts épidémiologiques permettant d'expliquer l'hétérogénéité de la distribution des maladies. L'imperfection de la détection des maladies génère cependant des observations qui ne représentent pas précisément la situation réelle. Des simulations ont été conduites afin d'évaluer l'impact d'une détection imparfaite sur les résultats des modèles statistiques de régression logistique et de Poisson enflés en zéro. Une situation où la sensibilité de détection est imparfaite et la spécificité parfaite a été simulée, et un facteur de risque influençant la prévalence de la maladie a été introduit ainsi qu'un facteur de confusion influençant la sensibilité de détection. Une situation où la spécificité de détection est imparfaite a aussi été simulée. A la lumière des résultats de simulation, des données de surveillance d'avortements bovins en France ont été analysées avec un modèle de régression logistique et un modèle enflé en zéro. Il apparaît que les modèles logistiques sont plus affectés par l'imperfection des données que les modèles de Poisson enflés en zéro.

Mots-clés : *facteurs de risque, régression logistique, modèle de Poisson enflé en zéro, biais, épidémiosurveillance, avortements bovins*

TITLE: Identification of risk factors with imperfect surveillance data : comparison of logistic models and zero-inflated Poisson models.

Abstract: Risk factors are key epidemiological concepts used to explain the heterogeneity of disease distribution. However, imperfect disease detection implies that the actual disease situation is not accurately depicted. We conducted a simulation study to assess the impact of imperfect detection on the outcome of the logistic regression model and of the zero-inflated Poisson regression model. We simulated a situation where the sensitivity of disease detection is imperfect whereas the specificity is perfect, and we introduced a risk factor affecting the disease prevalence and a confounding factor influencing the sensitivity of disease detection. We also simulated a situation where the specificity of detection is imperfect. Considering the simulation results, we revisited the analysis of the French bovine abortion surveillance data using both a logistic regression model and a zero-inflated Poisson regression model. Our results show that logistic regression models are more affected than zero-inflated Poisson models when applied to imperfect surveillance data.

Keywords: *risk factors, logistic regression, zero-inflated Poisson model, bias, surveillance, bovine abortion*