Université de POITIERS

Faculté de Médecine et de Pharmacie

ANNÉE 2021 **Thèse n°**

THESE
POUR LE DIPLOME D'ETAT
DE DOCTEUR EN PHARMACIE
(arrêté du 17 juillet 1987)

présentée et soutenue publiquement
le 02, juillet, 2021 à POITIERS
par Monsieur CHATENET Jean
né le 16 mai 1994 à St Jean D'Angély (17)

Mise en œuvre d'un processus de gestion des données pour la création et l'automatisation d'une nouvelle base de données permettant la conduite d'études de Real World Evidence. Étude de cas : l'impact de l'administration des vaccins de la bronchite infectieuse administrés au couvoir comparé à ceux administrés à la ferme chez le poulet de chair.

Composition du jury :

Président : Monsieur le Professeur OLIVIER Jean Christophe

Membres : Madame CHAUZY Alexia, Maître de conférences
Monsieur LEWINER Thomas, Groupe Chief Data Officer, CEVA Santé Animale
Monsieur COUSEIN Etienne, Chef pôle pharmacie, Centre Hospitalier de Valenciennes

Directeur de thèse : Monsieur GREGOIRE Nicolas, Maître de conférences

UNIVERSITE DE POITIERS

Faculté de Médecine et de Pharmacie

Année universitaire 2020-2021

PHARMACIE

Professeurs

- CARATO Pascal, PU, chimie thérapeutique
- COUET William, PU-PH, pharmacie clinique
- DUPUIS Antoine, PU-PH, pharmacie clinique
- FAUCONNEAU Bernard, PU, toxicologie
- GUILLARD Jérôme, PU, pharmacochimie
- IMBERT Christine, PU, parasitologie
- MARCHAND Sandrine, PU-PH, pharmacocinétique
- OLIVIER Jean Christophe, PU, galénique
- PAGE Guylène, PU, biologie cellulaire
- RABOUAN Sylvie, PU, chimie physique, chimie analytique
- RAGOT Stéphanie, PU-PH, santé publique
- SARROUILHE Denis, PU, physiologie
- SEGUIN François, PU, biophysique, biomathématiques

Maîtres de Conférences

- BARRA Anne, MCU-PH, immunologie-hématologie
- BARRIER Laurence, MCU, biochimie
- BODET Charles, MCU, bactériologie (HDR)
- BON Delphine, MCU, biophysique
- BRILLAULT Julien, MCU, pharmacocinétique, biopharmacie
- BUYCK Julien, MCU, microbiologie,
- CHARVET Caroline, MCU, physiologie
- CHAUZY Alexia, MCU, pharmacologie fondamentale et thérapeutique
- DEBORDE-DELAGE Marie, MCU, sciences physico-chimiques
- DELAGE Jacques, MCU, biomathématiques, biophysique
- FAVOT-LAFORGE Laure, MCU, biologie cellulaire et moléculaire (HDR)
- GIRARDOT Marion, MCU, biologie végétale et pharmacognosie
- GREGOIRE Nicolas, MCU, pharmacologie (HDR)
- HUSSAIN Didja, MCU, pharmacie galénique (HDR)
- INGRAND Sabrina, MCU, toxicologie
- MARIVINGT-MOUNIR Cécile, MCU, pharmacochimie
- PAIN Stéphanie, MCU, toxicologie (HDR)
- RIOUX BILAN Agnès, MCU, biochimie
- THEVENOT Sarah, MCU-PH, hygiène et santé publique
- TEWES Frédéric, MCU, chimie et pharmacochimie
- THOREAU Vincent, MCU, biologie cellulaire
- WAHL Anne, MCU, chimie analytique

Maîtres de Conférences Associés - officine

- DELOFFRE Clément, pharmacien
- ELIOT Guillaume, pharmacien
- HOUNKANLIN Lydwin, pharmacien

A.T.E.R. (attaché temporaire d'enseignement et de recherche)

- MIANTEZILA BASILUA Joe, épidémiologie et santé publique

Enseignants d'anglais

- DEBAIL Didier

REMERCIEMENTS

À mes Maîtres et juges :

À Monsieur le Professeur Jean Christophe Olivier : Je vous remercie pour l'honneur que vous me faites en acceptant la présidence de cette thèse. Veuillez recevoir le témoignage de ma reconnaissance et de mon profond respect.

À Monsieur Nicolas Grégoire : Je vous remercie d'avoir accepté de diriger ce travail et de m'avoir accordé de votre temps. Merci pour votre disponibilité, votre réactivité, vos conseils et vos connaissances.

À Madame Alexia Chauzy : Merci de me faire l'honneur d'accepter de juger mon travail ainsi que de m'avoir accordé du temps. Veuillez trouver ici l'expression de ma considération.

À Monsieur Thomas Lewiner : Merci de m'avoir intégré dans ton équipe, d'avoir partagé avec moi tes connaissances et de m'avoir donner l'opportunité d'évoluer professionnellement. Merci également pour ton implication et ta bienveillance. Merci enfin de m'avoir permis d'utiliser les données du laboratoire qui ont permis la réalisation de ce travail.

À Monsieur Etienne Cousein : Je te remercie pour l'honneur que tu me fais en faisant partie de ce jury. Merci de partager avec moi tes connaissances et ta vision de la pharmacie hospitalière de demain.

À tous ceux qui ont contribué à la réalisation de ce travail :

Au laboratoire CEVA Santé Animale : Merci pour la mise à disposition des données qui a permis la réalisation de ce travail.

À Monsieur Higor Cotta : Merci d'avoir partagé avec moi tes connaissances en statistiques et en programmation, d'avoir relu mes analyses et de m'avoir apporté ton aide que ce soit dans le cadre de cette thèse ou lors de mes missions.

A Madame Mathilde Lecoupeur : Merci d'avoir répondu à mes questions sur la filière avicole, d'avoir partagé ton expérience et ta vision de cette filière.

À ma famille et mes amis :

À mes parents, Marie-Josèphe et Laurent : Merci de votre soutien durant toutes ces années d'études, ces longues périodes de révision et ces moments de doute. Vous m'avez toujours poussé à donner le meilleur de moi-même et à me surpasser ; vos conseils m'ont été d'une aide précieuse et je n'en serai pas là sans vous aujourd'hui. Je vous dois ma réussite.

À ma sœur et mon frère, Lucie et Baptiste : Merci à tous les deux de m'avoir soutenu depuis le début, de m'avoir protégé mais aussi fait grandir.

À Anne-Victoire : Merci pour tout ce que tu m'apportes chaque jour. Tu réponds toujours présente pour me soutenir, m'écouter et me motiver. Certaines choses n'auraient pas été possibles sans toi ; je te dois beaucoup.

À Théo : Merci à toi pour ton soutien durant toutes ces années à Poitiers, tes grandes stratégies, projets et tous ces bons moments passés.

À tous mes amis : Je ne peux que vous remercier pour ce que vous m'avez apporté et pour votre présence à mes côtés.

Table des matières

LISTE DES ABRÉVIATIONS

CV	Coefficient de variation
CSV	Comma-Separated Values
ELISA	Enzyme-linked Immunosorbent Assay
FCR	Feed Conversion Ratio
FDA	Food and Drug Administration
IB	Bronchite Infectieuse
IC	Indice de Consommation
RGPD	Règlement Général Protection des Données
RWE	Real World Evidence

LISTE DES ILLUSTRATIONS

LISTE DES TABLEAUX

INTRODUCTION

L'analyse de données de vie réelle permet de générer des preuves issues « de la vraie vie » (real-world evidence, ou RWE) ; c'est une pratique de plus en plus courante dans le domaine de la santé et particulièrement demandée par les agences réglementaires. Les études de Real World Evidence permettent d'obtenir des informations utiles complémentaires aux données des essais cliniques, avec pour ambition d'accroître les connaissances scientifiques sur l'utilisation des médicaments dans la pratique clinique quotidienne. Afin de pouvoir mener à bien ces études, les laboratoires pharmaceutiques doivent se doter de systèmes de gouvernance des données efficaces, permettant de garantir la traçabilité des données mais aussi de rendre les données exploitables.

Depuis plus de quatre ans, Ceva Santé Animale propose à ses clients importants de la filière avicole ses services pour la réalisation d'études ad hoc. Ces études sont réalisées au niveau global, mais les données sont collectées par des équipes locales partout dans le monde.
L'ambition de Ceva Santé Animale est de structurer ces données, permettant ainsi de comparer l'ensemble des données collectées sur le terrain au cours de ces années et celles à venir. Cette ambition permettrait de réaliser des études de preuves empiriques, communément appelées *Real World Evidence* (RWE), pour étudier l'impact de ses vaccins dans diverses conditions et situations. Elle permettrait également d'appliquer des algorithmes d'apprentissage automatique à une base de données commune afin, par exemple, de prédire une baisse de performance ou une éventuelle augmentation de la mortalité dans une exploitation avicole. Cette base de données pourrait donc à terme permettre au laboratoire de compléter ses offres de services existantes, conformément à sa stratégie de diversification de ses activités. Afin de répondre aux attentes, cette base de données doit répondre à plusieurs critères :

- Elle doit pouvoir se mettre à jour automatiquement lorsqu'un nouveau jeu de données est reçu.
- La qualité des nouveaux jeux de données doit être automatiquement évaluée avant leur intégration dans la base de données.
- Elle doit respecter la réglementation relative à la protection des données à caractère personnel.
- Elle doit garantir l'univocité : les variables exprimant une même chose ne peuvent avoir qu'un seul nom.
- Elle doit garantir la traçabilité : le lignage de chaque entrée doit remonter jusqu'aux données brutes d'origine.
- Elle doit être claire : les variables doivent être exprimées en anglais, ainsi que les valeurs qu'elles contiennent.
- Elle doit être suffisamment simple : chaque variable doit, selon le choix de l'utilisateur, être exprimée soit sous forme longue (nom exhaustif de tous les éléments permettant de comprendre la variable ou l'élément), soit sous forme courte (nom courant).
- Elle doit assurer la cohérence : il ne doit y avoir qu'une unité et une échelle par variable ou élément dans un même jeu de données.
- Elle doit permettre la comparaison entre différents systèmes de production : pour les variables et les caractéristiques qui le permettent, l'expression en pourcentage doit être préférée à l'expression en valeur absolue.
- Elle doit servir aux experts : si les données le permettent, elle doit permettre d'établir le programme de vaccination des animaux.

Ces critères sont essentiels pour la bonne utilisation d'une telle base de données. Ils garantissent la qualité des données contenues dans la base et assurent le respect de la législation.

PARTIE I : MISE EN PLACE D'UNE BASE DE DONNÉES

1. Positionnement du problème et de l'approche par rapport à l'état actuel des connaissances :

Le principal problème de la mise en œuvre de cette solution réside dans les origines très diverses des jeux de données collectées. Comme indiqué dans l'introduction, les jeux de données sont récupérés par les équipes locales du laboratoire. Ces données ont été partagées par des clients importants du laboratoire qui ont eux-mêmes récupéré les données sur différents sites (couvoirs, fermes, laboratoires d'analyse...). Ces différentes origines et le grand nombre d'interlocuteurs se traduisent par des données très hétérogènes au niveau global. A ce stade, il n'est pas possible de les regrouper dans une base de données utilisable par l'équipe chargée des données.

Le problème est donc d'étudier et de rechercher la méthode la plus efficace pour :

- Harmoniser tous ces jeux de données en labélisant les variables.
- Harmoniser toutes les valeurs dans les unités souhaitées.
- Qualifier les données manquantes.
- Vérifier la qualité des données avant de les intégrer dans la base de données.
- Rassembler toutes ces données dans une même structure.
- Identifier les vaccins qui ont été administrés et ainsi reconstruire le programme de vaccination.
- Garantir un système de traçabilité afin de retrouver pour chaque variable ou caractéristique les jeux de données qui ont permis de la constituer.

1.1. Approche concernant l'harmonisation des labels

Pour réaliser le premier point, deux approches sont possibles. L'une consiste à labéliser les variables à la main et l'autre à automatiser le processus à l'aide d'algorithmes. Cette dernière méthode est décrite par Xavier Bosch Capblanch(1). Dans cette approche, les

jeux de données sont inclus dans la base de données cible les uns après les autres et trois algorithmes sont utilisés de manière consécutive et itérative :

- Le premier permet d'identifier les cas particuliers qui ne peuvent être résolus automatiquement.
- Le deuxième, qui est effectué si le premier n'a donné aucun résultat, permet d'identifier les mots-clés présents dans le label de la variable étudiée, c'est-à-dire le nom de la variable, et dans les labels des variables déjà présentes dans la base de données cible.
- Le troisième, permet d'identifier les mots-clés présents dans le label de la variable étudiée et dans les labels des variables des autres jeux de données déjà analysés.

Pour chaque label, trois résultats sont possibles :

- Le mot clé est retrouvé dans le label d'une seule variable de la base de données.
- Le mot clé est retrouvé dans le label de plusieurs variables de la base de données.
- Le mot clé n'est retrouvé dans le label d'aucune variable de la base de données.

La méthode que nous avons choisie pour résoudre ce problème est la première, c'est-à-dire la méthode manuelle. Le choix de cette dernière se justifie pour plusieurs raisons. Tout d'abord, nous ne disposions pas d'une base de données référentielle pour exécuter l'algorithme. Nous aurions pu seulement labéliser le premier jeu de données à la main et l'utiliser ensuite comme base de référence pour les jeux de données suivants, mais cette solution n'était pas réalisable en raison de la diversité des données dont nous disposions. De nombreuses variables et caractéristiques que nous possédons sont similaires en termes de nomenclature et la nomenclature seule ne nous permet pas de les identifier. Par exemple, pour une variable dont le label contient le mot-clé "poids", il pourrait s'agir du "poids à l'abattage", du "poids sur 7 jours" ou du "poids de l'ensemble de la production", mais cette variable pourrait tout aussi bien être exprimée en grammes, en kilogrammes ou même en livres ou en pourcentage d'une norme. Seule une analyse rigoureuse du label et des valeurs

de la variable nous permet de l'identifier correctement pour le moment. C'est pourquoi nous avons choisi de labéliser et d'identifier les colonnes manuellement.
Dans un deuxième temps, et grâce au travail déjà réalisé, il sera possible d'implémenter les autres approches.

1.2. Approche visant à garantir la qualité des données

Pour garantir la qualité des données, deux axes peuvent être développés.

- Le premier consiste à agir directement au niveau de la collecte des données par la formation des opérateurs, en définissant à l'avance les paramètres à collecter ou en ayant un contrôle efficace du processus de collecte des données.
- Le second est d'agir directement sur les données collectées, afin d'assurer la cohérence des résultats.

Comme nous ne pouvons pas intervenir sur le premier axe pour la plupart des clients, nous nous sommes concentrés sur le second. Nous avons donc cherché des moyens d'assurer la cohérence des données et de traiter nos valeurs manquantes.
En ce qui concerne les valeurs manquantes, plusieurs possibilités s'offrent à nous. Nous pouvons faire ce qui suit :

- Supprimer les observations incomplètes.
- Remplacer les valeurs manquantes par des valeurs arbitraires (en utilisant la moyenne, la médiane, etc.).
- Essayez de calculer les valeurs à partir des autres données, en laissant éventuellement des valeurs manquantes.

Pour la création de cette base de données, nous avons décidé d'appliquer la troisième méthode. Cette méthode nous a semblé la plus juste afin de ne pas altérer la qualité des données.

1.3. Approche visant à garantir la traçabilité

Tout d'abord, pour cette partie, nous avons dû respecter une contrainte pour interagir avec le projet de gestion des données internes lancé par Ceva Santé Animale en 2019 : l'utilisation de DataGalaxy(2).

DataGalaxy est une plateforme agile permettant aux équipes de faciliter la gestion des données. L'objectif de DataGalaxy est que chaque employé de l'entreprise, qu'il soit issu d'un milieu technique ou non, puisse accéder, explorer et contribuer à une connaissance commune des données de l'entreprise.

DataGalaxy est divisé en plusieurs modules :

- Le dictionnaire de données.
- Le glossaire des termes.
- Le catalogue des traitements (traitement des données).
- Le catalogue des usages.

Pour notre problématique, nous nous concentrerons principalement sur le dictionnaire de données et le catalogue de traitement. Les deux autres modules correspondent aux étapes futures du projet.

Notre approche consiste à garantir un système simple de visualisation des données pour nous permettre de retracer l'origine de nos variables dans notre base de données finale et de savoir quels traitements ont été effectués pour l'obtenir.

2. Modèles développés

Le modèle de création de base de données que nous avons développé est basé, comme expliqué précédemment, sur l'utilisation de dictionnaires.

La figure ci-dessous donne un aperçu de la structure du modèle, afin d'obtenir une base de données à partir des jeux de données locaux.

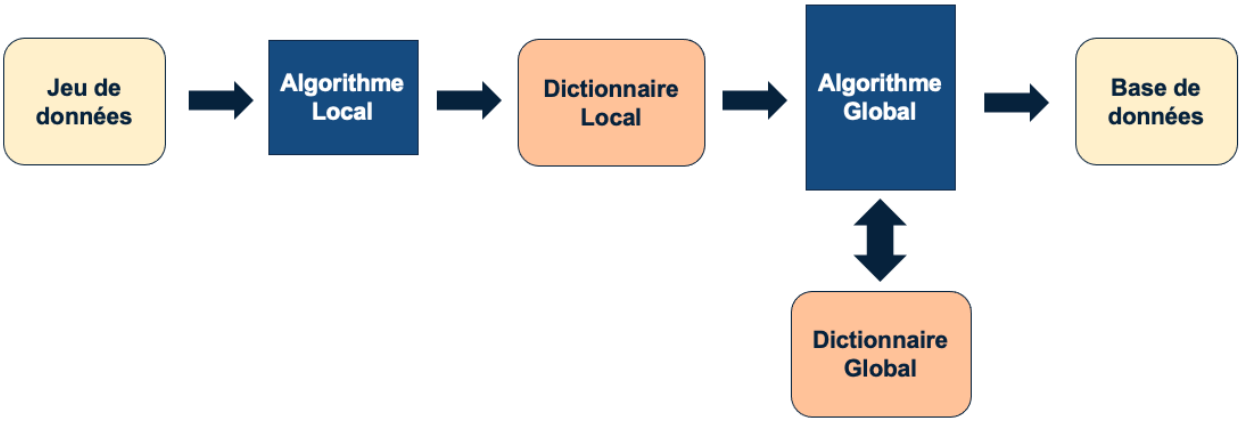


Figure 1 : Structure du modèle

Dans un premier temps, nous appliquons l'algorithme local sur chaque jeu de données. Ce premier algorithme est dit local car il permet, pour chaque jeu de données de constituer une base unique qui permettra de réaliser le dictionnaire local qui, lui aussi, est unique pour chaque jeu de données. Ce dictionnaire local permet de recueillir les informations essentielles contenues dans le jeu de données. Puis nous utilisons l'algorithme global qui permet de relier les dictionnaires locaux tout en appliquant les éléments contenus dans le dictionnaire global. Ce dictionnaire est dit global car il ne dépend pas d'une étude spécifique.

Pour réaliser nos algorithmes, nous avons utilisé le langage de programmation open source Python. Il est très utilisé dans le domaine des Data Sciences, du Machine Learning et de l'Intelligence Artificielle et permet l'automatisation des algorithmes.

Nous allons maintenant examiner plus en détail la composition de ces dictionnaires et leur utilisation, puis nous décrirons les algorithmes qui nous permettent de construire notre base de données finale tout en respectant les contraintes mentionnées dans l'introduction.

2.1. Dictionnaires

2.1.1. Dictionnaire local

Le premier dictionnaire réalisé s'appelle le dictionnaire local. Chaque étude possède son propre dictionnaire local. Il sert à recueillir les informations essentielles contenues dans le jeu de données. Chaque dictionnaire local contient les clés suivantes pour chaque variable :

- name : label de la variable
- count : nombre de valeurs
- unique : nombre de valeurs uniques
- top : valeur la plus courante
- freq : fréquence de la valeur la plus courante
- first : première valeur (uniquement pour les dates)
- last : dernière valeur (uniquement pour les dates)
- mean : moyenne des valeurs
- std : écart-type des valeurs
- min : valeur minimale
- 25% : premier quartile
- 50% : médiane
- 75% : troisième quartile
- max : valeur maximale
- type : catégorie, date ou numérique
- description : petit texte décrivant la variable
- keep : variable à conserver ou non
- place : lieu de collecte
- animal : animal étudié
- scale : échelle d'étude
- objet : ce qui est mesuré, quantifié, répertorié par la variable
- stage : stade d’étude de la variable

- unit : unité de mesure de la variable
- long_name : concaténation des colonnes « place », « animal », « scale », « objet », « stage » et « unit » pour obtenir un identificateur de la variable.

Le label nous donne une première indication de la variable que nous étudions, mais ne nous permet pas de l'identifier avec certitude. Nous devons poursuivre notre enquête avec les informations statistiques de la variable. Ces informations nous permettent d'en savoir plus sur la variable étudiée et nous permettent ainsi d'ajouter de la granularité à nos labels. En effet, à partir de ces informations, il est possible de savoir si la mesure est faite à l'échelle d'un oiseau, d'un troupeau ou même d'identifier l'unité de mesure.
Après avoir examiné ces premiers éléments, nous pouvons remplir manuellement les labels pour « place », « animal », « scale », « objet », « stage » et « unit ». Ces six éléments représentent le cœur de notre nomenclature. Grâce à eux, nous obtenons, en regroupant les valeurs, un « long_name » univoque (mais non cohérent entre les jeux de données) qui sert de clé pour les étapes suivantes.

2.1.2. Dictionnaire global

Ce dictionnaire est dit global car il ne dépend pas d'une étude spécifique. Il est composé de trois sous-dictionnaires, dont nous allons maintenant étudier la composition :

2.1.2.1. Dictionnaire d'opérations

Le dictionnaire d'opérations est divisé en cinq clés :

- operation_order : ordre dans lequel les opérations doivent être exécutées
- long_name : la liste des identificateurs des variables à convertir
- operation : l'opération à effectuer pour convertir la variable
- long_standard_name : le nom de référence de la variable après que l'opération a été effectuée.
- description : petit texte décrivant la variable

Ce dictionnaire a plusieurs fonctions. La première est de convertir nos variables dans l'unité souhaitée et d'introduire ainsi la notion de “long_standard_name" qui, comme son nom l'indique en anglais, est le nom long de référence de cette variable.

Comme les données proviennent de différents endroits du monde, les métriques utilisées pour collecter ou calculer les variables et les caractéristiques ne sont pas les mêmes. Par conséquent, pour pouvoir regrouper plusieurs variables, nous devons disposer d'une unité commune.

Pour chaque " long_name" qui doit être converti, une opération est décrite. Cette opération permet d'appliquer les coefficients nécessaires à sa transformation dans l'unité standard qui aura été préalablement définie et d'obtenir ainsi le nom long standard de la variable (long_standard_name). Par exemple, si nous nous intéressons au poids moyen des poulets de chair à l'abattage, que nous avons cette mesure en kilogrammes, et que l'unité standard est définie comme étant le gramme, nous aurons comme valeur pour le nom long (*long_name*) :

slaughter_broiler_per_flock_weight_average_kg

la clé de l’opération sera :

*df[‘slaughter_broiler_per_flock_weight_average_kg’]*1000*

ce qui nous permettra d’avoir comme nom long standard (*long_standard_name*) :

slaughter_broiler_per_flock_weight_average_g

Ainsi, nous pourrons combiner des variables avec des unités initialement différentes, mais ce dictionnaire a une autre utilité. Il permet, si les données présentes dans le jeu de données le permettent, de calculer les principaux indicateurs de performance et de convertir les valeurs absolues en pourcentages.

Par exemple, pour calculer le « gain moyen quotidien » qui est un indicateur de performance important en zootechnie, si le jeu de données comprend les variables :

- *slaughter_broiler_per_flock_weight_average_g* (le poids moyen des poulets de chair à l'abattage en grammes)
- *slaughter_broiler_per_flock_age__days* (âge du poulet de chair à l'abattage en jours)

Le dictionnaire d'opérations nous permettra de calculer cette variable lors de l'exécution de l'opération :

$$\frac{df['slaughter\_broiler\_per\_flock\_weight\_average\_g']}{df['slaughter\_broiler\_per\_flock\_age\_\_days']},$$

et donc de l'ajouter au jeu de données pour l'intégrer dans la base de données.
Grâce à la nature interprétative de python, ces calculs peuvent être codés directement sous forme de texte dans le dictionnaire.

2.1.2.2. Dictionnaire de contrôle

Le deuxième sous-dictionnaire utilisé est appelé le dictionnaire de contrôle. Ce dictionnaire peut être décrit comme suit :

- long_standard_name : le nom de référence de la variable
- opération : l'opération à effectuer pour vérifier la qualité de la variable
- check_name : le nom de vérification de la variable

Ce dictionnaire est utilisé pour vérifier que les valeurs du jeu de données ainsi que celles calculées à l'aide du dictionnaire des opérations se situent dans la fourchette souhaitée. Par exemple, la variable slaughter_broiler_per_flock_weight_average_g, qui correspond au poids moyen des poulets de chair à l'abattage en grammes, doit être comprise entre 0 et 5000 grammes. Il fonctionne de la même manière que le dictionnaire des opérations. Il applique à la variable correspondant au nom long standard (long_standard_name) une opération prédéfinie et lui attribue un nom de contrôle.

2.1.2.3. Dictionnaire de conversion

Le dictionnaire de conversion nous permet de convertir les noms des labels. Il peut être décrit comme suit :

- keep : variable à conserver ou non dans la base de données finale
- long_standard_name : nom standard long

- short_standard_name : nom standard court
- long_label : label long pour les graphiques
- short_label : label court pour les graphiques
- description : petit texte décrivant la variable

Grâce à ce dictionnaire, nous pouvons facilement sélectionner les variables qui doivent être présentes dans la base de données finale. Par exemple, pour réduire la taille de la base de données finale, nous pouvons uniquement conserver les variables exprimées en pourcentage et supprimer les variables exprimées en valeur absolue.

Il nous permet également de changer facilement le nom de la variable, tant pour des raisons de rapidité comme pour les analyses en sélectionnant le nom court standard (short_standard_name) au lieu du nom long standard (long_standard_name) ou pour la génération de graphiques en ayant le nom des labels (courts ou longs) préalablement défini.

2.1.3. Glossaire de vaccination

Le dernier dictionnaire concerne le programme de vaccination. Il a pour but de regrouper les informations dispersées dans la base de données pour constituer une information ordonnée et harmonisée.

Les informations sur le programme de vaccination sont dispersées dans la base de données parmi 37 variables. Ces variables contiennent des informations telles que :

- les noms des vaccins utilisés
- la ou les maladies concernées par les vaccins
- l'âge d'administration des vaccins
- les voies d'administration des vaccins

Tout d'abord, afin d'identifier les lignes, dans la base de données, avec les mêmes informations sur le programme de vaccination, un identifiant de vaccination est créé. Cet identifiant est créé en regroupant, pour chaque ligne, l'ensemble des valeurs contenues dans les 37 variables.

Une fois l'identifiant créé, pour chaque identifiant unique, cinq informations par vaccin identifié doivent être ordonnées, harmonisées et complétées manuellement, en utilisant les informations collectées :

- vaccine_standard_name : le nom standard du vaccin
- vaccine_disease : la maladie pour laquelle ce vaccin s'applique
- vaccine_route_administration : la voie d'administration du vaccin
- vaccine_days : le jour où le vaccin est administré
- vaccine_company : le laboratoire qui fabrique le vaccin

Une fois que tous les vaccins ont été identifiés, le programme de vaccination peut être construit en regroupant les valeurs "vaccine_standard_name", "vaccine_disease" et "vaccine_days" pour chaque vaccin identifié.
Par la suite, l'ensemble des 37 variables initiales qui ont permis de créer le programme de vaccination sera supprimée, dans la base de données, et remplacée par le programme de vaccination qui contient une information plus ordonnée et harmonisée.

2.2. Algorithmes

2.2.1. Algorithme local

Ce premier algorithme est utilisé sur tous les jeux de données qui seront intégrés dans notre base de données. Son fonctionnement est très simple, il nous permet principalement de définir une structure pour le dictionnaire local, qui sera la même pour toutes les études. Il fonctionne de la manière suivante

- L'algorithme lit le jeu de données.
- Pour chaque variable identifiée, il applique la fonction pandas.DataFrame.describe(3), qui permet d'obtenir les informations statistiques de la variable.

- Il ajoute les clés suivantes au dictionnaire :
    - *type*
    - *description*
    - *keep*
    - *place*
    - *animal*
    - *scale*
    - *object*
    - *stage*
    - *unit*
    - *long_name*
- Enregistre ce nouveau dictionnaire local dans le dossier contenant les autres dictionnaires locaux.

2.2.2. Algorithme global

L'algorithme global permet de relier les jeux de données, les dictionnaires locaux, le dictionnaire global qui comprend les trois sous-dictionnaires (opérations, contrôle et conversion) et le dictionnaire de vaccination.

2.2.2.1. Les fonctions :

L'algorithme est composé de plusieurs fonctions que nous allons présenter maintenant :

2.2.2.1.1. Fonction de normalisation du jeu de données

Cette fonction permet tout d'abord de vérifier que chaque jeu de données possède son dictionnaire local. Si ce n'est pas le cas, un message d'avertissement est envoyé.
Si cette condition est vérifiée, le jeu de données est normalisé. Cela signifie que les majuscules sont remplacées par des minuscules, que les accents potentiels sont supprimés et que les espaces sont remplacés par "_".

Le nom de la variable est alors remplacé par le nom long_name défini dans le dictionnaire local.
Si certaines variables sont vides, la fonction les supprime du jeu de données.

2.2.2.1.2. Fonction d'application du dictionnaire d'opérations

Pour chaque ligne représentant chaque variable, si le long_name est présent dans le dictionnaire des opérations, l'opération de conversion ou de calcul est effectuée.
L'ordre de l'opération est effectué selon le rang défini précédemment dans la clé "operation_order" du dictionnaire.

- Si la fonction a calculé des variables déjà présentes dans le jeu de données, elle effectue alors une comparaison des valeurs du jeu de données et de celles calculées.
- Si ces valeurs diffèrent, un message d'avertissement est envoyé.
- Si les valeurs ne diffèrent pas mais que la variable comporte des données manquantes, la fonction remplace alors les données manquantes par les données calculées.
- Si la variable n'existe pas, la fonction l'ajoute au jeu de données.
- S'il n'est pas possible de calculer la variable, un message d'alerte est envoyé.

2.2.2.1.3. Fonction de contrôle

Cette fonction est utilisée pour effectuer les opérations du dictionnaire de contrôle. Elle effectue l'opération présente dans le dictionnaire et retourne le pourcentage de qualité de la variable. Le pourcentage de qualité est défini comme le nombre de valeurs correspondant à l'intervalle préalablement défini dans le dictionnaire sur l'ensemble des valeurs.

2.2.2.1.4. Fonction de conversion

Utilisant le dictionnaire de conversion, cette fonction permet de convertir le nom des variables dans le format souhaité.

2.2.2.1.5. Fonction d'identification du programme de vaccination

Le but de cette fonction est de mettre en œuvre le programme de vaccination. Elle fait correspondre chaque ligne de la base de données avec l'identifiant de la vaccination.
Ensuite, elle ajoute pour chaque ligne le programme de vaccination et les informations ordonnées pour chaque vaccin.
Enfin, elle supprime les anciennes variables qui ont permis la création de l'identifiant de vaccination, qui sont désormais obsolètes.

2.2.2.2. Comment fonctionne l'algorithme

L'algorithme fonctionne comme suit :

- L'algorithme lit tous les jeux de données et les dictionnaires.
- Applique la fonction de normalisation sur le jeu de données.
- Applique la fonction d'application du dictionnaire d'opérations.
- Applique la fonction de contrôle.
- Applique la fonction de conversion.
- Ne conserve que les variables préalablement définies dans le dictionnaire de conversion et supprime les variables vides.
- Ajoute le jeu de données à la base de données.
- Applique la fonction d'identification du programme de vaccination.
- Crée une nouvelle version de la base de données.

A la fin de l'opération de cet algorithme, nous sommes en possession d'une base de données opérationnelle.

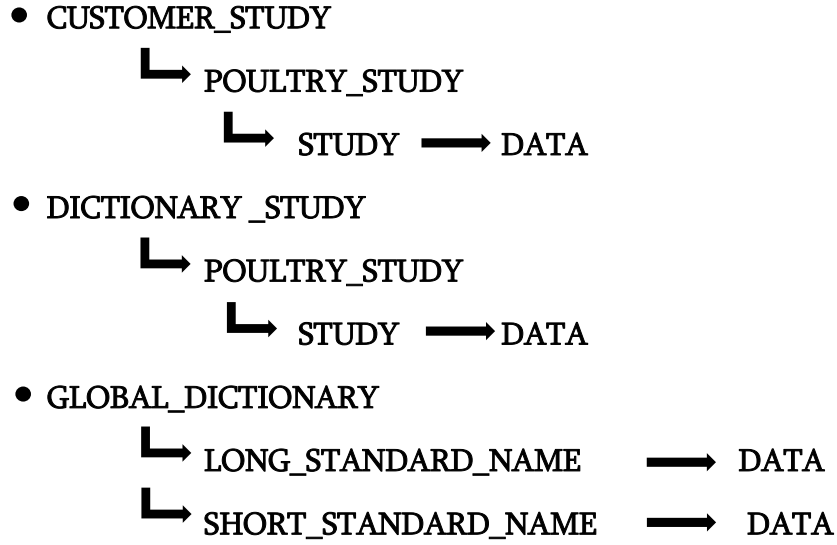
2.2.3. Algorithme de traçabilité

L'objectif de l'algorithme de traçabilité est de répertorier toutes les actions qui ont été entreprises depuis la réception d'un jeu de données jusqu'à son intégration dans la base

de données finale. Il permet aussi de référencer l'ensemble des valeurs contenues dans les dictionnaires locaux afin de faire une recherche rapide sur la plateforme ou de visualiser précisément un sous-ensemble de données. Comme le volume des données est susceptible de croître très fortement, nous devons mettre en œuvre un tel algorithme afin de garantir la qualité de la base de données finale.

L'algorithme est divisé en plusieurs étapes :

- La première étape consiste à collecter toutes les valeurs incluses dans les six colonnes des dictionnaires locaux permettant l'identification de la variable (« place », « animal », « scale », « objet », « stage » et « unit ») afin de créer un « tag » dans DataGalaxy pour chaque valeur identifiée. Ces tags nous permettront d'effectuer une recherche rapide sur la plateforme ou de visualiser précisément un sous-ensemble de données.
- La deuxième étape consiste à créer les "Dictionnaires" qui seront utilisés pour le stockage sur DataGalaxy. Pour l'instant, nous avons identifié trois dictionnaires : CUSTOMER_STUDY, DICTIONARY_STUDY et GLOBAL_STUDY qui ont la structure suivante :



Pour chacun de ces dictionnaires, un script est exécuté pour extraire les informations nécessaires à la création de ces dictionnaires et ainsi permettre leur traçabilité. A la fin de chaque script, un fichier csv, à télécharger sur DataGalaxy, est généré.
La troisième partie consiste à générer les fichiers permettant d'effectuer les « traitements de données », les traitements de données sont les processus permettant d'établir les liens entre les différents dictionnaires. Ici, nous avons généré trois traitements de données :

- harmonization : pour harmoniser les labels
- operation : harmoniser toutes les valeurs
- conversion : passer d'un label long à un label court (et vice versa)

Comme pour les dictionnaires, un script exécuté sur l'ensemble des données permet de collecter les informations nécessaires à ce traitement et de générer des fichiers csv à intégrer dans DataGalaxy.
Finalement, les dictionnaires sont eux-mêmes versionnés et tracés.

2.3. Qualité des données

Dans notre processus, nous avons quatre points de contrôle de la qualité des données distincts.
Le premier point se trouve dans le dictionnaire local avec les informations « keep », qui nous permettent de ne conserver que les variables qui ont été entièrement identifiées, mais aussi d'exclure les variables qui pourraient contenir des données personnelles que nous ne pouvons pas conserver conformément à la législation.
Le deuxième point est effectué à l'aide du dictionnaire de contrôle. Il nous permet de prendre note de la qualité des données ajoutées à la base de données au moment de sa création.
Le troisième point de contrôle est effectué à l'aide du dictionnaire de conversion avec les informations « keep », qui nous permet de ne conserver que les informations pertinentes pour la base de données.
Le dernier point de contrôle est le système de traçabilité qui nous permet d'identifier rapidement les jeux de données problématiques lors de l'utilisation de la base de données.

3. Résultats et vue d'ensemble de la base de données créée

La mise en œuvre de ce processus a été menée sur **42** études ad hoc. Ces études ont été réalisées de 2017 à 2020 et ont porté exclusivement sur la volaille. Les données ont été collectées au niveau de chaque pays et comprennent donc des informations dans la langue de collecte (anglais, français, russe, etc.).

La représentation géographique des données est illustrée dans la figure ci-dessous, où nous avons calculé pour chaque pays sa présence dans la base de données.

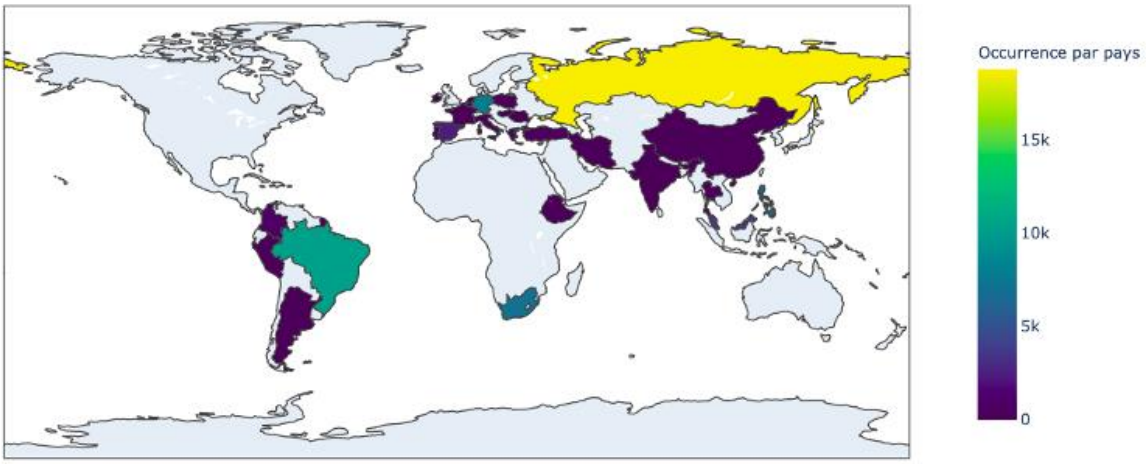


Figure 2 : Répartition géographique des données

3.1. Nombre de labels

Dans toutes ces études, **1 695** labels ont été identifiés. Parmi eux, **1 031** valeurs uniques.

Après la création des **42** dictionnaires locaux, nous avons compté pour chacun des six éléments composant les dictionnaires locaux, le nombre de valeurs uniques. Ces valeurs sont présentées dans le tableau suivant :

Tag	Nb de valeurs uniques	Description
place	5	lieu de collecte
animal	7	animal étudié
scale	2	échelle d'étude
object	309	ce qui est mesuré, quantifié, répertorié par la variable
stage	80	stade d'étude de la variable
unit	74	unité de mesure de la variable

Tableau 1 : Valeurs uniques par tag d'identification

Avec ces six éléments, nous avons obtenu un total de **732** noms longs uniques.
Après cette première étape, nous avons donc réduit le nombre de labels différents de **29** % (1031 labels→732 noms longs).
La deuxième étape consiste à appliquer les opérations du dictionnaire des opérations. Après cette étape, nous avons obtenu **458** noms longs standards. Cela représente une réduction de **55** % par rapport au nombre des labels par défaut et de **38** % par rapport à l'étape précédente.

3.2. Contrôle de la qualité des données

Le processus de contrôle de la qualité des données nous permet d'obtenir les informations nécessaires pour chaque jeu de données. La visualisation est claire et nous permet de mettre en évidence les points faibles ou les erreurs potentielles dans les jeux de données.
La figure 3 montre une visualisation du résultat de la fonction de contrôle de l'algorithme global sur deux jeux de données (France_Client3_2019 et Irlande_Client1_2018). Par exemple, pour le premier jeu de données « France_X_2019 », l'algorithme lit le jeu de données, il cherche ensuite à calculer les variables manquantes. Quand une variable est manquante dans le jeu de données pour réaliser le calcul, l'algorithme affiche un message d'erreur « variable_1 not in data_df to compute variable_2 » (data_df étant le nom attribué par l'algorithme au jeu de données). Ensuite, l'algorithme effectue la fonction de contrôle

et retourne le pourcentage de qualité qui est défini comme le nombre de valeurs correspondant à l'intervalle préalablement défini dans le dictionnaire sur l'ensemble des valeurs. On peut lire par exemple : *check_farm_broiler_per_flock_adg_final_g_day quality* = 100 %.

Figure 3 : Résultat de la fonction de contrôle de l'algorithme global.

3.3. Programme de vaccination

Un autre objectif était d'identifier les vaccins administrés aux troupeaux et ainsi de retracer le programme de vaccination. Ce programme de vaccination, qui est important pour Ceva Santé Animale, permet d'étudier l'effet de ses vaccins sur les données de production. Après avoir compilé l'ensemble des données, 37 variables contenant des informations sur le programme de vaccination ont été identifiées. Après agrégation de ces variables, **631** identifiants uniques de vaccination ont été générés pour l'ensemble des données.

Pour rappel, une fois qu'un vaccin a été identifié, cinq informations doivent être recherchées :

- vaccine_standard_name : le nom standard du vaccin
- vaccine_disease : la maladie pour laquelle ce vaccin est appliqué
- vaccine_route_administration : la voie d'administration du vaccin
- vaccine_days : le jour où le vaccin est administré
- vaccine_company : le laboratoire qui fabrique le vaccin

Dans certaines études, on a constaté que jusqu'à huit vaccins étaient référencés par troupeau.

Après analyse, le nombre de programmes de vaccination répertoriés est de **359**, ce qui représente une baisse de **43** % du nombre de programmes de vaccination initiaux.

3.4. Intégration à DataGalaxy

DataGalaxy assure la traçabilité des traitements effectués sur les jeux de données pour obtenir la base de données mais la plateforme permet également une représentation visuelle de ces traitements.

La figure ci-dessous représente cette visualisation, dans un ensemble réduit, en présentant uniquement les jeux de données, sans les données et le détail des opérations. En lisant de gauche à droite, on a tout d'abord le jeu de données (première colonne), on lui applique l'algorithme local (colonne 2) qui permet d'obtenir le dictionnaire local (colonne 3). À partir de là, on peut appliquer l'algorithme global (colonne 4) qui va constituer le dictionnaire global (colonne 5). Pour finir la fonction de conversion (colonne 6) va permettre de convertir les noms des variables dans le format souhaité au niveau du dictionnaire global. En suivant chaque ligne, on obtient précisément tous les traitements effectués sur chaque jeu de données. Sur la figure, les variables contenues dans les jeux de données et dans la base de données globale ne sont pas visibles (par manque d'espace) mais si on clique sur une variable dans la base de données globale, les jeux de données, et leurs variables, qui ont permis de réaliser cette variable sont mis en évidence.

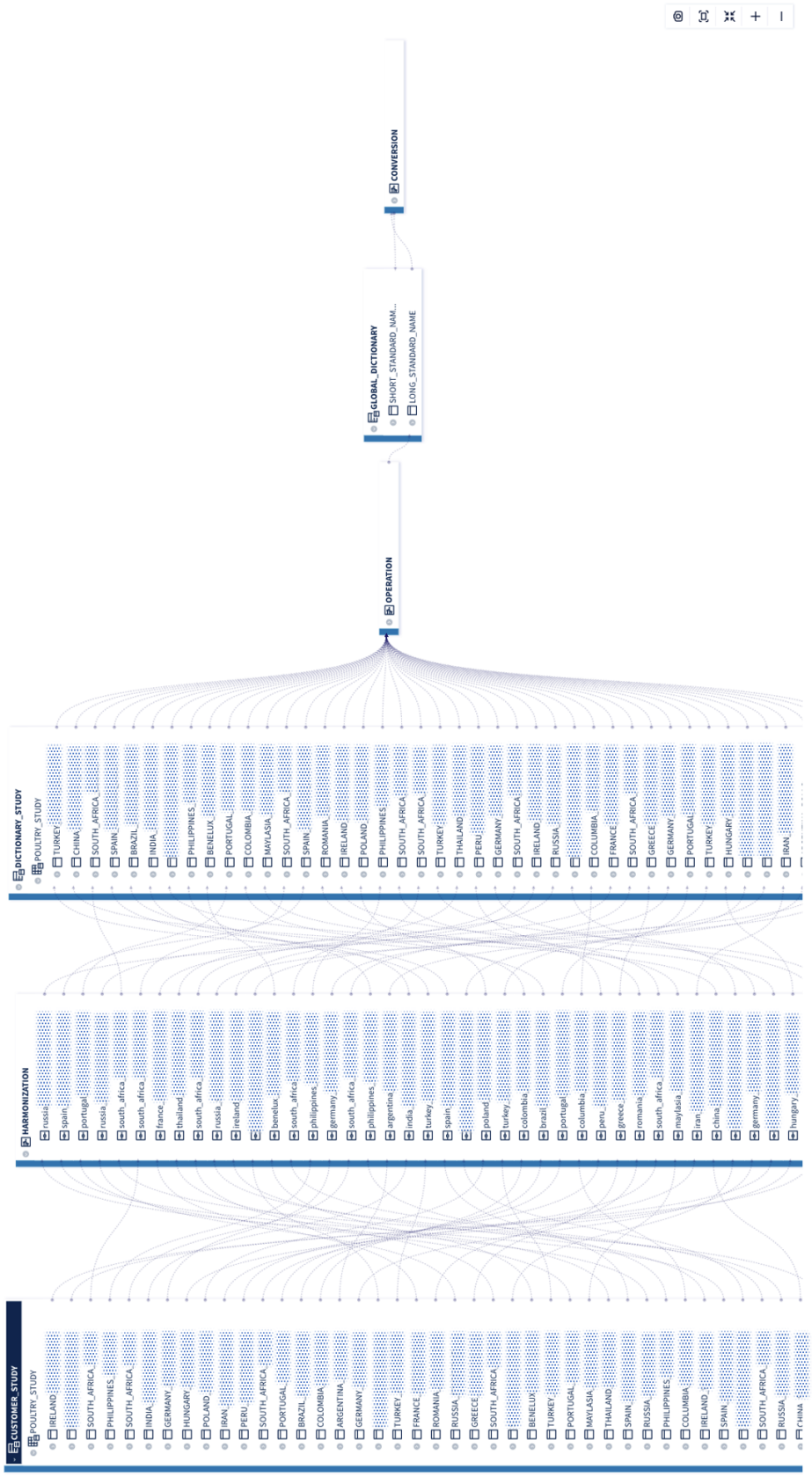


Figure 4 : Représentation des dictionnaires et des traitements des données sous DataGalaxy (vue réduite)

PARTIE II : ÉTUDE DE CAS

A partir de la base de données que nous avons générée dans la première partie, nous allons maintenant conduire une étude de preuves empiriques sur l'impact de l'administration des vaccins de la bronchite infectieuse administrés au couvoir comparé à ceux administrés à la ferme chez le poulet de chair.

1. La bronchite infectieuse

La bronchite infectieuse est probablement la plus commune des infections respiratoires infectant le poulet. Cette maladie est l'une des principales causes de perte économique dans le secteur avicole, affectant les performances des oiseaux de chair et des oiseaux de ponte. Le virus, un coronavirus, se reproduit non seulement dans l'épithélium des tissus des voies respiratoires supérieures et inférieures, mais aussi dans de nombreux tissus le long du tube digestif et ailleurs, par exemple dans les reins, les oviductes et les testicules. Il peut être détecté dans les matières respiratoires et fécales (4).

1.1. Mode de transmission

Le virus de la bronchite infectieuse se propage rapidement parmi les poulets d'un troupeau. La maladie est très contagieuse et a une période d'incubation très courte. Les oiseaux placés avec des poulets infectés développent généralement des signes cliniques dans les 24-48 heures. Le virus peut se retrouver pendant 10 jours dans les sécrétions respiratoires et pendant 20 semaines dans les fientes des poulets contaminés (5). Le virus peut donc se transmettre par inhalation avec des gouttelettes contaminées, par ingestion après consommation d'eau ou de nourriture contaminée ou par contact avec du personnel ou du matériel infecté.

1.2. Signes cliniques

Les signes cliniques dépendent du sérotype et de son tropisme. Souvent, il y a peu de signes, et les animaux guérissent spontanément. Les signes sont plus sévères chez les jeunes, avec une mortalité d'origine primaire. Chez les adultes, la mortalité est souvent causée par des infections secondaires (6).

1.2.1 Signes respiratoires

Les signes respiratoires de la maladie sont :

- Toux.
- Râles trachéaux humides ou bruit de pompe chez les jeunes.
- Éternuements.
- Écoulement nasal séro-muqueux jamais hémorragique.
- Parfois sinus enflés et conjonctivite séreuse avec yeux humides.

Ces signes peuvent être accompagnés de symptômes généraux chez les jeunes poulets. La guérison souvent spontanée en deux semaines s'accompagne d'un retard de croissance marqué.

1.2.2. Signes reproducteurs

Les signes reproducteurs sont :

- Chute de ponte (10-50%).
- Œufs de mauvaise qualité (coquille mince, molle ou absente, pâle ou rugueuse, albumen trop liquide, œufs déformés).
- Lésions à l'oviducte.

L'infection des poules pondeuses de moins de deux semaines aura des conséquences désastreuses sur la ponte.

1.2.3. Signes Généraux

Seulement avec certaines souches virales :

- Dépression.
- Soif intense.
- Fèces humides.
- Mortalité.

1.3. Diagnostic

Le seul diagnostic clinique ne suffit pas. La confirmation en laboratoire est nécessaire pour le diagnostic des formes respiratoires de bronchite infectieuse en raison des similitudes de la maladie avec des agents tels que le virus de la maladie de Newcastle, le métapneumovirus aviaire, le virus de la laryngotrachéite infectieuse et les mycoplasmes. La preuve d'une séroconversion ou d'une augmentation du titre d'anticorps contre la bronchite infectieuse (IB) par ELISA, ou des tests d'inhibition de l'hémagglutination ou de neutralisation du virus peuvent être utilisés pour le diagnostic lorsqu'il y a des antécédents de maladie respiratoire ou de réduction de la production d'œufs.

1.4. Prévention

En suivant les règles de prophylaxie suivantes, l'infection des troupeaux par la maladie ou du moins la diffusion de la propagation peut être limitée :

- Effectuer un vide sanitaire entre les différents troupeaux d'au moins 14 jours avec désinfection en profondeur des locaux et du matériel utilisé pour l'élevage (le virus est sensible à la majorité des désinfectants).
- Ne pas regrouper d'oiseaux d'origines différentes.
- Maintenir une température adéquate et une ventilation adaptée.
- Éviter la surpopulation.
- Vacciner le troupeau (7).

2. Cohorte de l'étude

A partir de notre base de données, nous avons sélectionné les troupeaux de poulets de chair et plus particulièrement des poulets ayant la génétique ROSS 308, qui ont reçu au moins un vaccin contre la bronchite infectieuse. Cette étape a été effectuée à l'aide du programme de vaccination que nous avons vu précédemment dans la partie I. Les poulets de chair ROSS 308 ont la particularité d'avoir une croissance rapide, pour une même quantité de nourriture et d'avoir de hautes performances. Afin de nous assurer que nous ne traitions pas un même troupeau deux fois, nous avons supprimé les lignes ayant le même :

- Identifiant de la ferme
- Identifiant du bâtiment
- Identifiant du troupeau
- Date de mise en place
- Date d'éclosion
- Programme de vaccination.

En supprimant l'ensemble des doublons pour les lignes ayant ces mêmes informations, nous nous assurons de ne pas prendre en compte plusieurs fois les données d'un même troupeau dans notre étude. Nous obtenons ainsi **2 462** troupeaux uniques pour conduire notre étude. La figure ci-dessous représente la répartition géographique de l'ensemble de ces troupeaux.

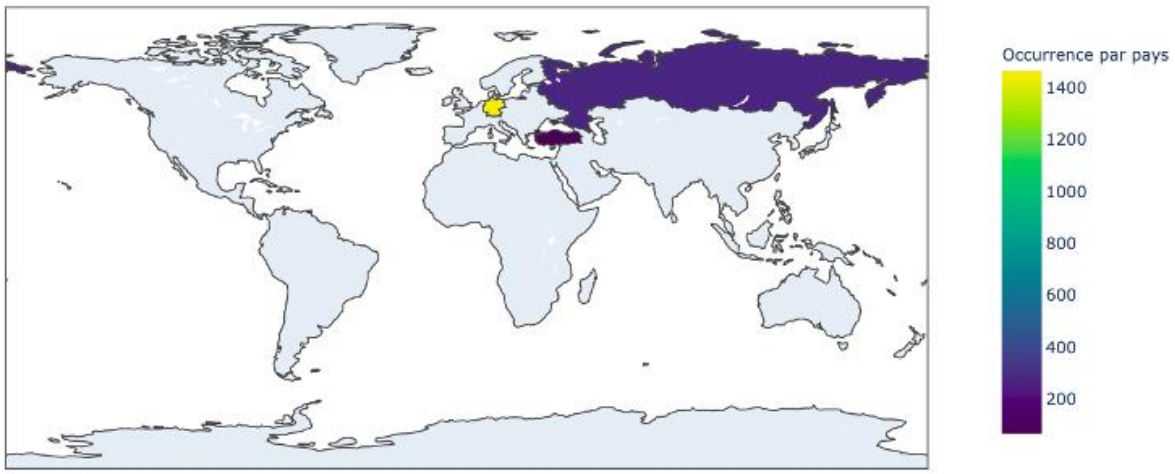


Figure 5 : Répartition géographique de l'ensemble des troupeaux de l'étude.

L'objectif de l'étude est d'étudier l'impact de l'administration des vaccins de la bronchite infectieuse administrés au couvoir comparé à ceux administrés à la ferme chez le poulet de chair sur les données cliniques et sub-cliniques.

Pour constituer le groupe couvoir, nous avons sélectionné les troupeaux ayant seulement reçu un ou des vaccins de la bronchite infectieuse in-ovo ou par pulvérisation à 1 jour. En ce qui concerne le groupe vacciné à la ferme, nous avons sélectionné les troupeaux ayant seulement reçu un ou des vaccins après un jour de vie. Les vaccins à la ferme sont généralement administrés via l'eau de boisson distribuée dans les élevages. Certains troupeaux sont à la fois vaccinés au couvoir et à la ferme, ils constituent alors le troisième groupe. Le tableau ci-dessous répertorie la taille de nos trois groupes.

Groupe	Nombre de troupeaux uniques
Couvoir	1810
Ferme	421
Couvoir et Ferme	231

Tableau 2 : Tailles des échantillons

Les figures ci-dessous s'intéressent à la répartition géographique des traitements en couvoir et à la ferme.

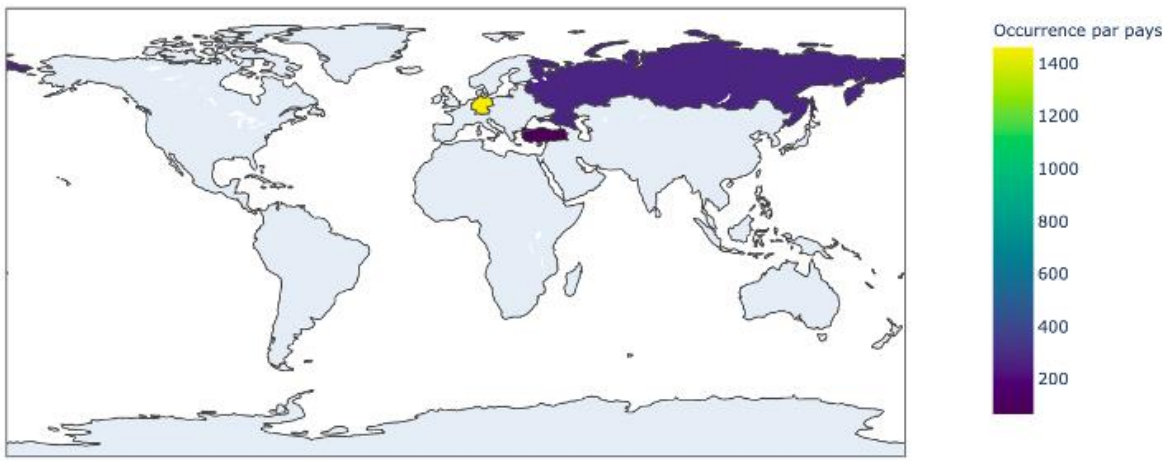


Figure 6 : Répartition géographique du groupe couvoir.

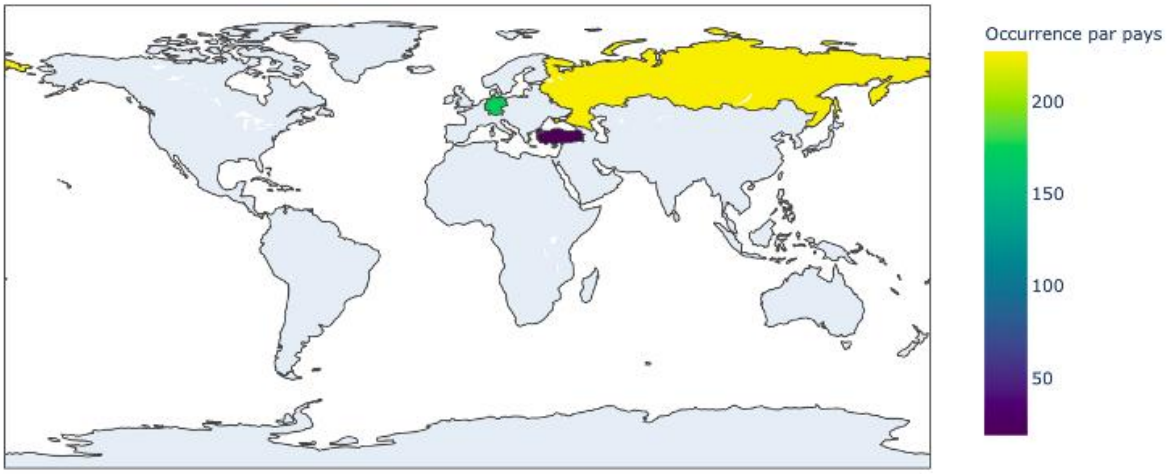


Figure 7 : Répartition géographique du groupe ferme.

Pour les groupes couvoir et ferme, le nombre d'occurrences par pays est résumé dans le tableau ci-dessous.

Pays	**Couvoir**	**Ferme**	***Total***
Allemagne	1 465	174	*1 639*
Turquie	68	19	*87*
Russie	277	228	*505*
Total	*1 810*	*421*	*2 231*

Tableau 3 : Répartition des échantillons par pays

On remarque que les poulets vaccinés au couvoir représentent 89% de notre échantillon en Allemagne, contre 78% en Turquie et 55% en Russie.

3. Méthode

3.1. Préparation des données

Dans nos jeux de données récoltés, il est possible de retrouver des valeurs aberrantes. On peut expliquer ces valeurs par le fait que les jeux de données sont généralement produits en complétant des tableaux Excel manuellement. Il peut arriver que l'opérateur se trompe en rentrant la valeur en ajoutant une virgule, un point ou un zéro par inadvertance. Pour résoudre ce problème, nous avons décidé de supprimer de chacune de nos analyses statistiques les valeurs extrêmes, en supprimant systématiquement, le premier et le dernier centile.

3.2. Constitution des sous-groupes

Comme nos données proviennent d'une agrégation de plusieurs jeux de données, nous n'avons pas pour chaque troupeau les mêmes informations. Chaque analyse est donc faite sur l'ensemble des troupeaux disposant de la variable étudiée, il en résulte que la taille des différents échantillons pour chaque analyse est différente.

3.3. Analyses statistiques

Les données utilisées pour créer cette base de données proviennent de différents pays. Pour chaque pays, il existe un scénario différent, néanmoins, nous avons décidé de les regrouper. Dans cette étude, nous ne nous intéresserons pas aux différences entre les pays, mais à la comparaison des vaccins de la bronchite infectieuse administrés au couvoir et à la ferme. Pour chaque analyse, nous avons décidé d'effectuer un test de Student car nos variables sont indépendantes et identiquement distribuées, et que $n > 30$ pour nos deux échantillons.

4. Résultats

Dans cette partie, nous verrons dans un premier temps les résultats de l'étude sur les données sérologiques puis sur les données cliniques et de production.

4.1. Données sérologiques.

Tout d'abord, nous nous sommes intéressés à la moyenne des titres sérologiques pour ces deux groupes. Cette analyse a été menée sur 280 troupeaux, 170 troupeaux vaccinés à la ferme et 110 vaccinés au couvoir. Les graphiques ci-dessous représentent la distribution de la variable pour les deux groupes étudiés sous forme d'histogramme et de boîte à moustache. Ces graphiques sont complétés par un tableau regroupant les informations sur les moyennes, les écarts types, les valeurs maximales et minimales ainsi que les quartiles.

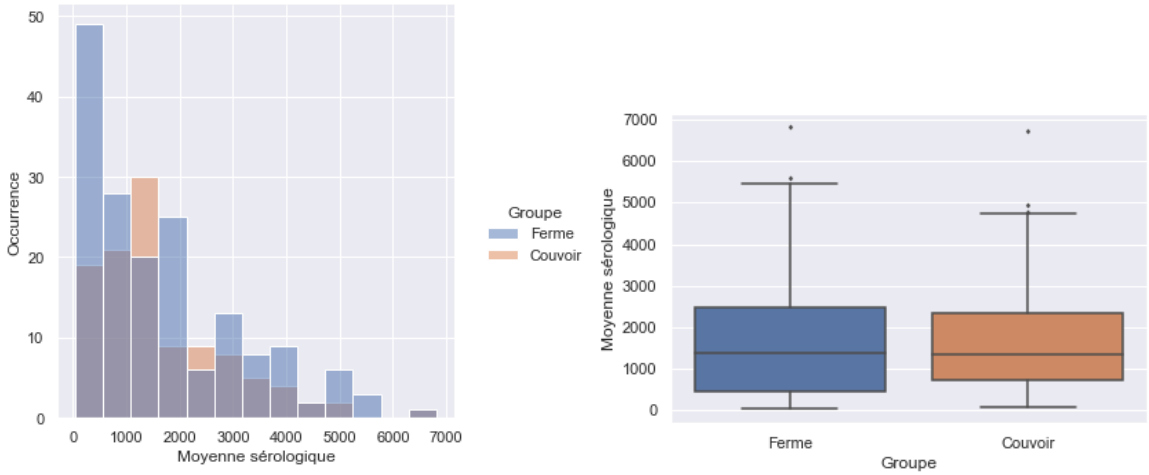


Figure 8 : Distribution moyenne des titres sérologiques

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	110	1 668	1 246	75	744	1 363	2 350	6 713
Ferme	170	1 679	1 475	45	453	1 382	2 489	6 811

Tableau 4 : Valeurs statistiques moyennes des titres sérologiques (titres)

Comme nos variables sont indépendantes et identiquement distribuées, et que n > 30 pour nos deux échantillons, nous avons donc effectué un test de Student pour comparer nos deux groupes avec $\alpha = 0,05$. Le résultat de ce test nous donne une valeur de $p = 0,952$, comme $p > 0,05$, nous rejetons l'hypothèse d'une différence significative entre les moyennes des titres sérologiques des deux groupes.

Nous nous sommes alors intéressés au coefficient de variation des titres sérologiques. Selon l'INSEE, le coefficient de variation (CV) est le rapport de l'écart-type à la moyenne. Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la moyenne est grande. La distribution du CV pour nos deux groupes est représentée dans les deux graphes ci-dessous.

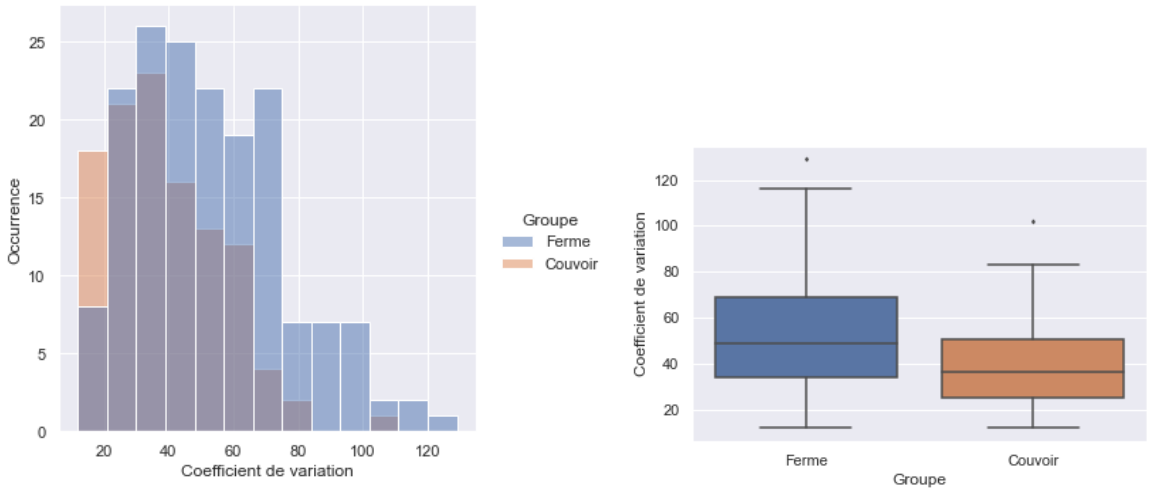


Figure 9 : Coefficient de variation des titres sérologiques

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	110	38,5	17,4	12	25,3	36,5	50,8	102
Ferme	170	52,2	23,7	12	34,3	49,0	69	129

Tableau 5 : Valeurs statistiques du coefficient de variation des titres sérologiques

Nous nous sommes donc retrouvés dans les mêmes conditions que pour l'analyse précédente et avons effectué à nouveau un test de Student avec $\alpha = 0,05$. Cette fois-ci, $p < 0,05$, nous

pouvons donc déduire qu'il y a significativement moins de dispersion autour de la moyenne des titres sérologiques dans le groupe vacciné au couvoir que dans celui vacciné à la ferme. A noter que la détection d'anticorps peut être tardive par rapport à l'âge d'abattage. La présence de quelques oiseaux avec des titres élevés (et donc un coefficient de variation élevé) est déjà un signe alarmant.

4.2. Données cliniques et de production

4.2.1. Impact sur la mortalité

Nous avons voulu étudier l'impact de la voie d'administration des vaccins sur la mortalité des poulets de chair. Nous nous sommes dans un premier temps intéressés à la mortalité totale. La mortalité totale est définie comme l'ensemble des décès répertoriés dans un troupeau depuis la mise en place à la ferme jusqu'à l'abattoir en pourcentage du nombre de poussins mis en place. La distribution de cette variable est représentée à l'aide des deux graphes ci-dessous.

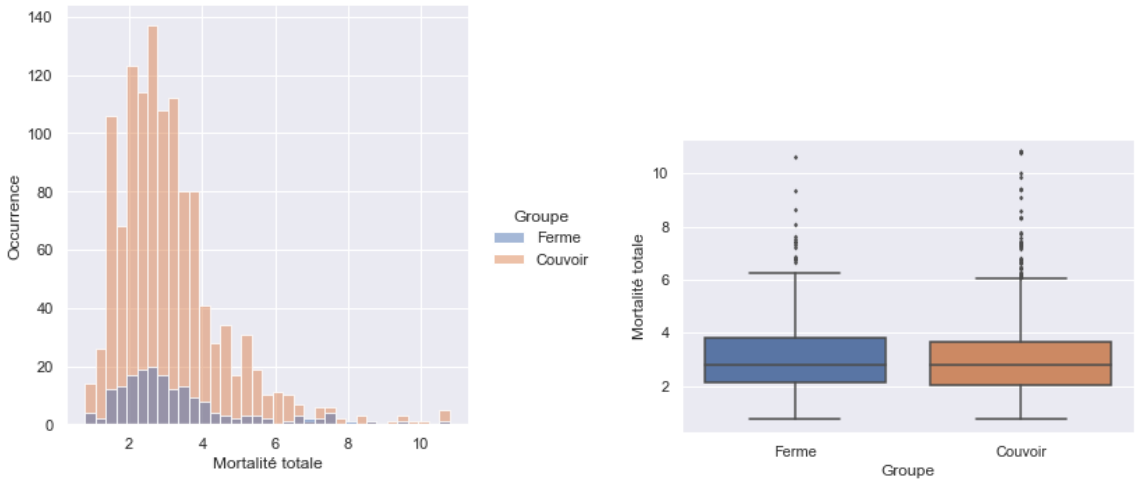


Figure 10 : Distribution de la mortalité totale (%)

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	1 206	3,09	1,47	0,77	2,06	2,80	3,67	10,80
Ferme	179	3,27	1,72	0,77	2,15	2,83	3,81	10,62

Tableau 6 : Valeurs statistiques de la mortalité totale (%)

Puis nous nous sommes intéressés à la mortalité précoce. Nous avons défini la mortalité précoce comme l'ensemble des décès survenus dans un troupeau depuis la mise en place à la ferme jusqu'à la septième journée de vie. Autrement dit, l'ensemble des décès survenus la première semaine. Nous avons obtenu pour cette variable la distribution suivante :

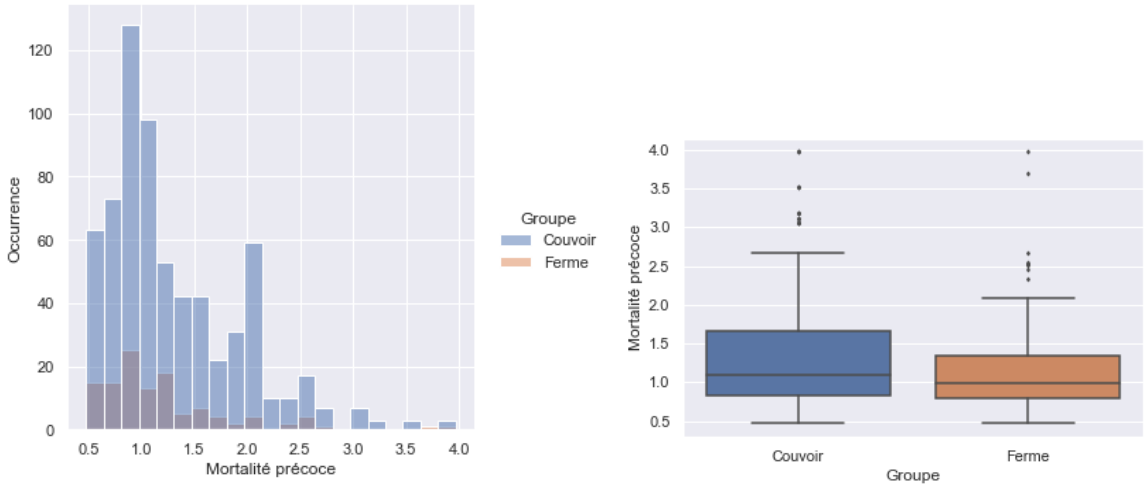


Figure 11 : Distribution de la mortalité précoce

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	671	1,30	0,63	0,48	0,84	1,10	1,67	3,97
Ferme	117	1,18	0,62	0,48	0,80	0,99	1,35	3,97

Tableau 7 : Valeurs statistiques de la mortalité précoce (%)

Comme les deux variables vues précédemment sont indépendantes et identiquement distribuées, nous avons effectué un test de Student avec $\alpha = 0,05$. Pour ces deux variables, $p > 0,05$, nous n'avons pas trouvé de différence significative entre les deux groupes étudiés.

4.2.2. Impact sur la production

Nous définissons les données de production comme l'ensemble des données récoltées sur les troupeaux, du couvoir à l'abattoir, et permettant de suivre l'évolution du troupeau au fil du temps. Parmi ces données, nous nous sommes particulièrement intéressés au poids à l'abattage, l'âge à l'abattage, au gain moyen quotidien, à l'indice de consommation et au taux de saisie à l'abattoir. Ces données sont des indicateurs clés de performance pour les éleveurs de volailles.

4.2.2.1. Impact sur le poids à l'abattoir

Le premier indicateur de production que nous avons analysé est le poids moyen à l'abattage. Pour un éleveur ou un industriel, plus le poids moyen à l'abattage est élevé, plus le profit sera fort. La distribution de cette variable est répartie comme telle :

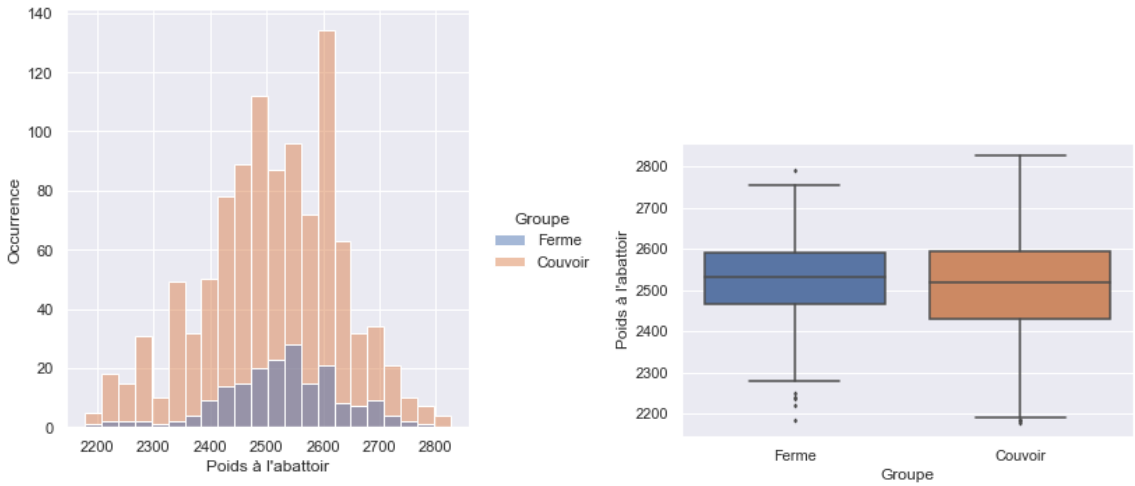


Figure 12 : Distribution du poids à l'abattoir (en grammes)

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	1 049	2 507	122	2 178	2 431	2 518	2 595	2 826
Ferme	190	2 525	106	2 185	2 466	2 531	2 593	2 792

Tableau 8 : Valeurs statistiques poids à l'abattoir (en grammes)

En effectuant un test de Student avec $\alpha = 0,05$, nous n'obtenons pas de différence significative. Comme le poids varie en fonction de l'âge, nous avons décidé de refaire notre analyse, mais cette fois en ajustant le poids à l'abattoir par l'âge de référence des poulets de chair Ross 308 (annexe 1). Par exemple, si on a comme valeur mesurée, un poids moyen de 3,12 kg à 42 jours. On remplace cette valeur brute par une donnée relative à la référence de la génétique. Dans ce cas, la référence est de 2,918 kg (Annexe 1), donc le poids moyen mesuré est de 0,20 kg au-dessus de la référence. On compare ces écarts aux références entre eux, ce qui nous permet d'éliminer l'effet de l'âge d'abattage. En faisant ainsi et en effectuant une nouvelle fois un test de Student, nous trouvons que pour $\alpha = 0,05$, $p < 0,05$. Nous avons donc une différence significative entre les deux groupes, c'est-à-dire que lorsque le poids à l'abattoir est ajusté sur l'âge de référence, les poulets du groupe couvoir sont significativement plus gros que ceux du groupe ferme.

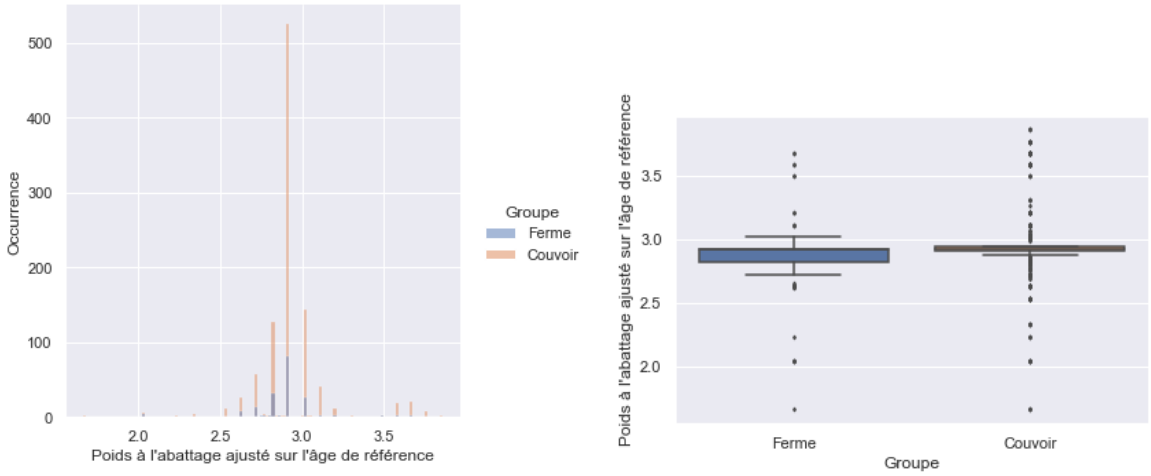


Figure 13 : Distribution du poids à l'abattoir ajusté sur l'âge de référence des poulets de chair Ross 308 (en grammes)

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	1 049	2,933	0,23	1,664	2,907	2,918	2,937	3,856
Ferme	190	2,879	0,23	1,664	2,821	2,218	2,918	3,675

Tableau 9 : Valeurs statistiques du poids à l'abattoir ajusté sur l'âge de référence des poulets de chair Ross 308 (en grammes)

4.2.2.2. *Impact sur l'âge à l'abattoir*

L'âge à l'abattoir est un indicateur important pour les éleveurs. Plus l'âge est élevé, plus le coût de production d'un poulet augmente. Il est donc favorable d'avoir un âge à l'abattoir faible. Comme pour le poids à l'abattage, nous nous sommes d'abord intéressées à l'âge à l'abattoir brut puis nous l'avons par la suite ajusté. Cette fois-ci, nous avons ajusté notre variable par le poids de référence des poulets de chair Ross 308 (annexe 1).

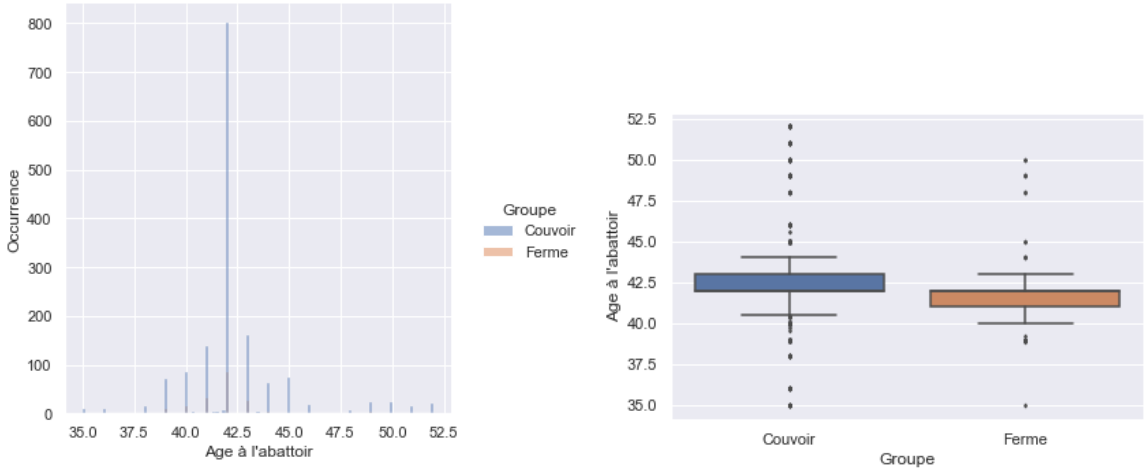


Figure 14 : Distribution de l'âge à l'abattoir (jours)

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	1 568	42,37	2,44	35	42	42	43	52
Ferme	198	41,90	1,85	35	41	42	42	50

Tableau 10 : Valeurs statistiques de l'âge à l'abattoir (jours)

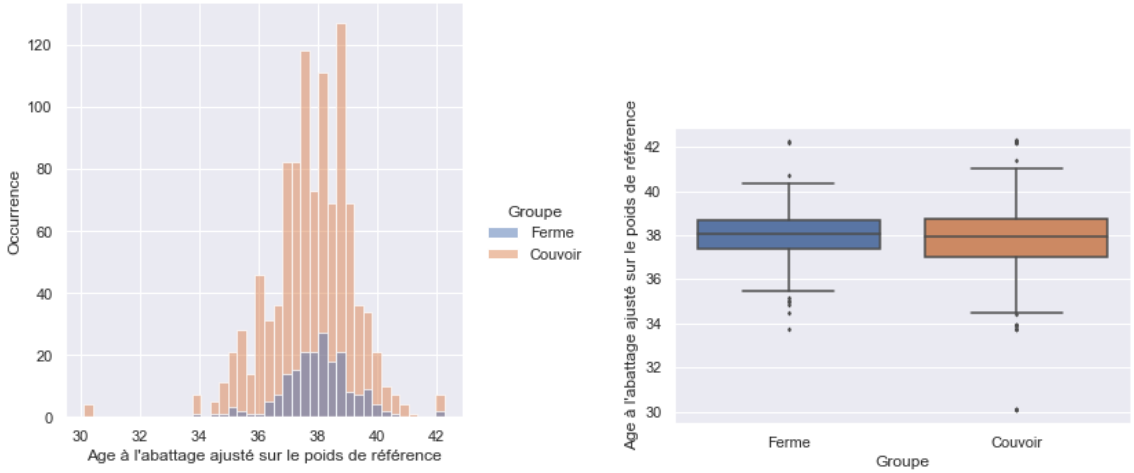


Figure 15 : Distribution de l'âge à l'abattoir ajusté sur le poids de référence des poulets de chair Ross 308 (jours)

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	1 054	37,79	1,42	30,11	37,01	37,91	38,72	42,29
Ferme	192	38,01	1,20	33,77	37,37	38,06	38,67	42,26

Tableau 11 : Valeurs statistiques de l'âge à l'abattoir ajusté sur le poids de référence des poulets de chair Ross 308 (jours)

Dans les deux analyses, on effectue un test de Student avec $\alpha = 0,05$. Lorsque l'âge n'est pas ajusté, on obtient une valeur de $p < 0,05$, l'âge à l'abattoir est alors significativement plus élevé dans le groupe vacciné au couvoir, mais lorsqu'on ajuste l'âge par le poids de référence des poulets de chair Ross 308 (annexe 1), la différence est toujours significative avec une valeur de $p < 0,05$; cette fois ci, c'est l'âge à l'abattoir des poulets vaccinés à la ferme qui est significativement plus élevé. Ce résultat peut être expliqué par le fait qu'en ajustant l'âge à l'abattoir sur le poids de référence des poulets de chair Ross 308, on élimine l'effet du type et du pays de production.

4.2.2.3. Impact sur le gain moyen quotidien

Le gain moyen quotidien est un indicateur important en zootechnie, il se calcule en divisant le poids moyen par l'âge moyen des poulets. On peut observer sur la figure 16, que la distribution n'est pas normale. Quand on s'intéresse à cette distribution par pays (figure 17), on s'aperçoit que le gain moyen quotidien est corrélé avec le pays de production. En effet, sur notre graphique, on voit que le gain moyen quotidien est plus important en Allemagne qu'en Russie. Cette différence peut être expliquée par les objectifs différents que peuvent avoir les pays et les contraintes auxquelles ils doivent faire face. Pour éliminer l'effet du pays sur nos résultats, nous avons donc décidé d'ajuster nos résultats sur le poids de référence des poulets de chair Ross 308 (annexe 1) et ainsi, nous obtenons les résultats de la figure 18 et du tableau 13.

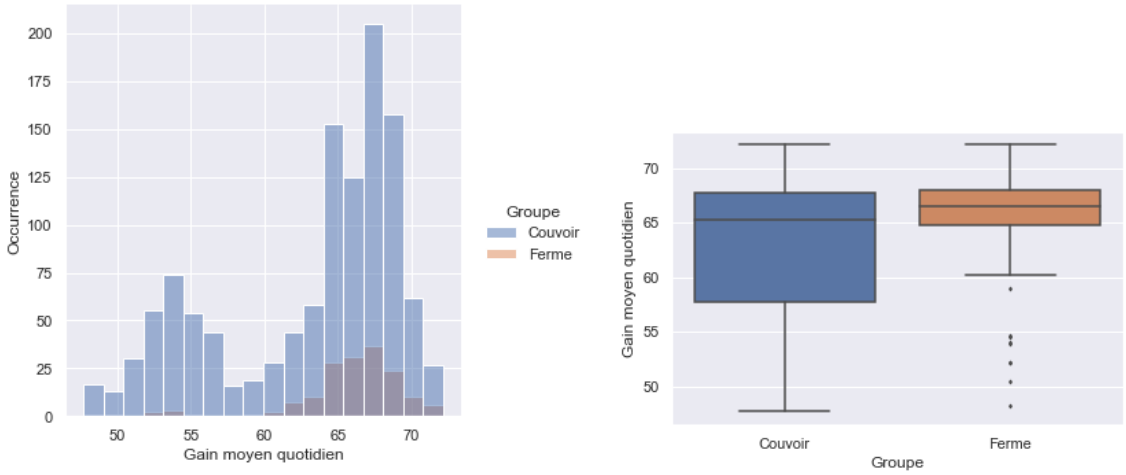


Figure 16 : Distribution du gain moyen quotidien (grammes par jour)

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	1 022	64,64	5,14	47,70	63,32	66,07	68,05	72,15
Ferme	328	59,66	6,67	48,16	53,69	56,40	66,49	72,15

Tableau 12 : Valeurs statistiques du gain moyen quotidien (grammes par jour)

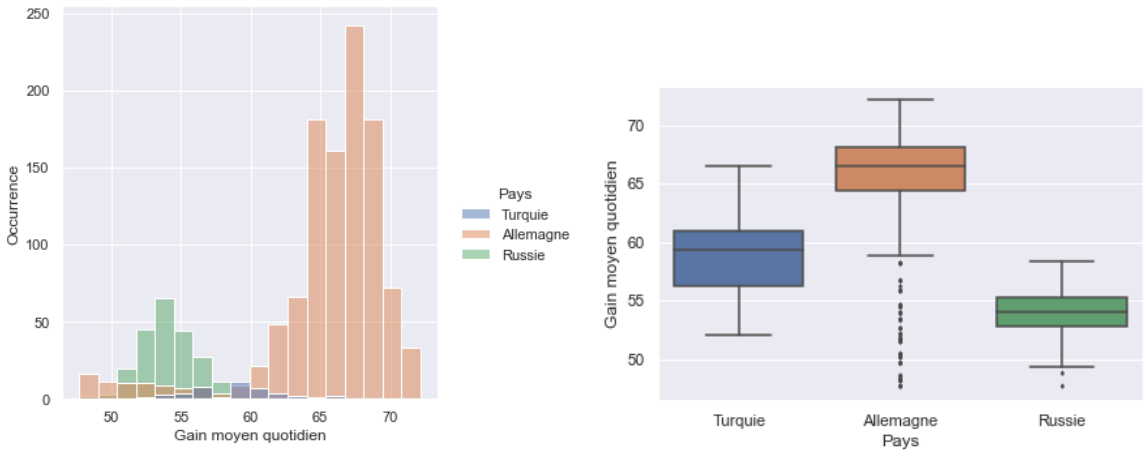


Figure 17 : Distribution du gain moyen quotidien par pays (grammes par jour)

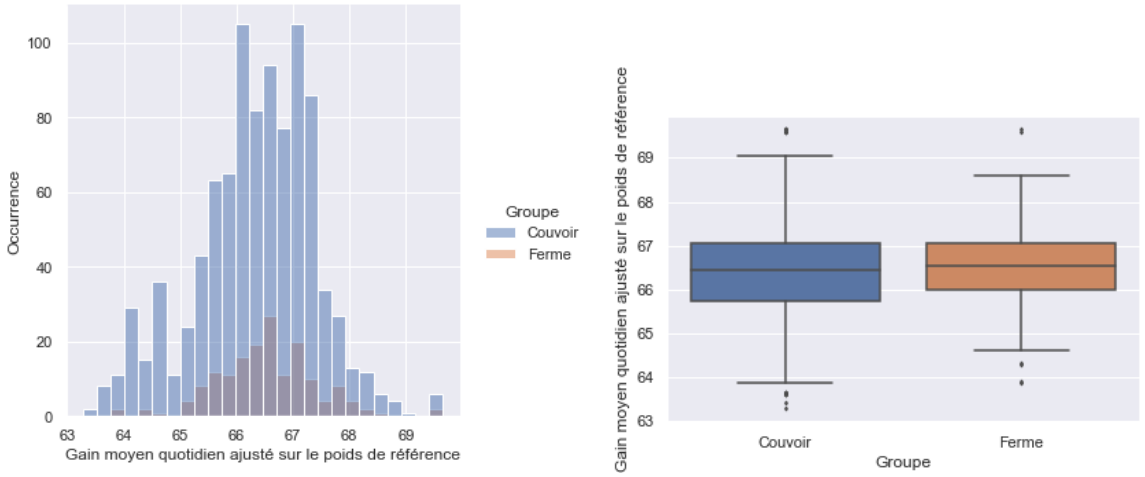


Figure 18 : Distribution du gain moyen quotidien ajusté sur le poids de référence des poulets de chair Ross 308 (grammes par jour)

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	968	66,35	1,04	63,28	65,73	66,43	67,06	69,64
Ferme	159	66,46	0,99	63,61	65,97	66,52	67,03	69,64

Tableau 13 : Valeurs statistiques du gain moyen quotidien ajusté sur le poids de référence des poulets de chair Ross 308 (grammes par jour)

En effectuant un test de Student avec $\alpha = 0,05$, on obtient une valeur de $p = 0,245$, comme $p>0,05$; nous n'avons donc aucune différence significative entre les deux groupes.

4.2.2.4. Impact sur l'indice de consommation

L'indice de consommation (IC) est un ratio utilisé pour mesurer l'efficacité de l'alimentation. Il mesure le nombre de kg d'aliments nécessaire pour produire 1 kg de poids vif. Tout comme le gain moyen quotidien, il est corrélé au poids de l'animal. Comme on peut le voir sur la figure 20, il y a de grandes différences entre les pays. Nous avons donc encore une fois décidé d'ajuster cette variable sur le poids de référence des poulets de chair Ross 308 (annexe 1).

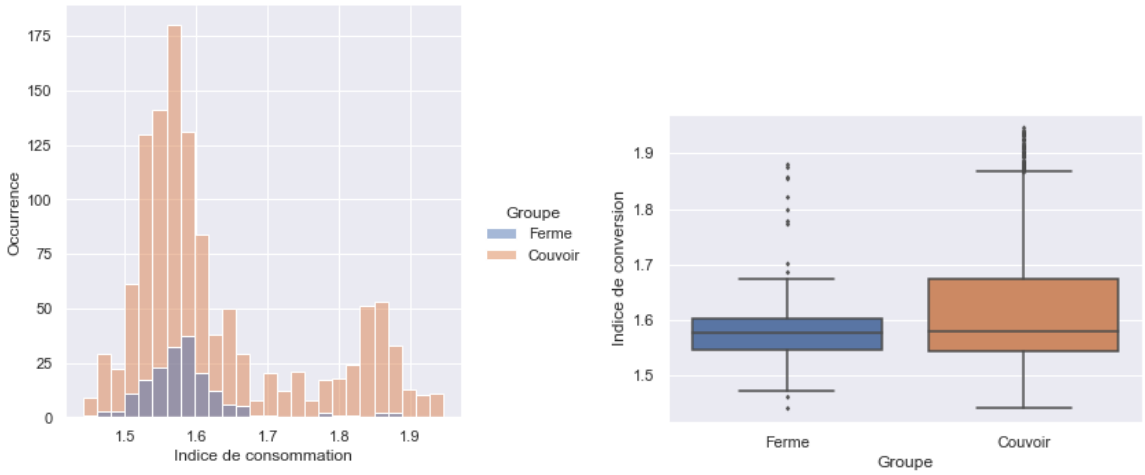


Figure 19 : Distribution de l'indice de consommation

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	1054	1,59	0,09	1,44	1,54	1,57	1,61	1,94
Ferme	333	1,72	0,15	1,44	1,58	1,58	1,86	1,96

Tableau 14 : Valeurs statistiques de l'indice de consommation

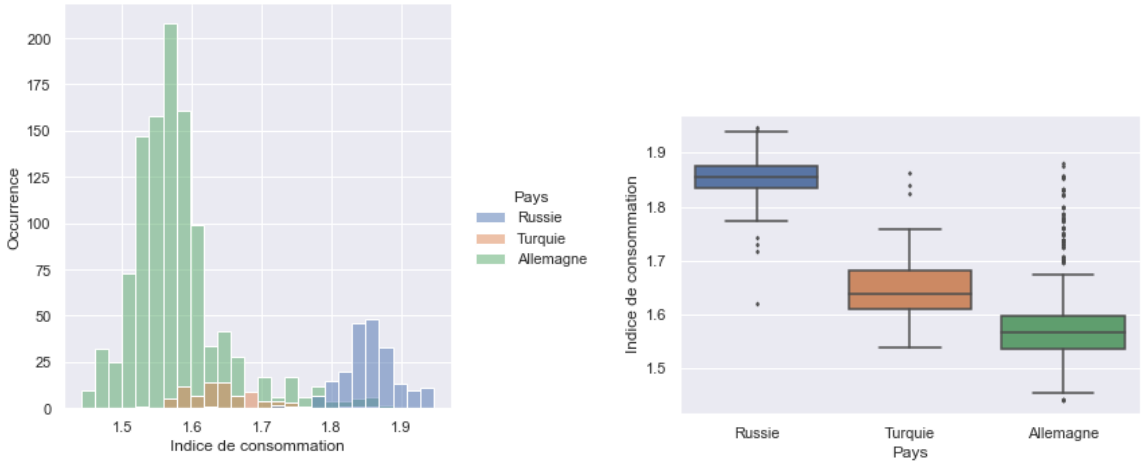


Figure 20 : Distribution de l'indice de consommation par pays

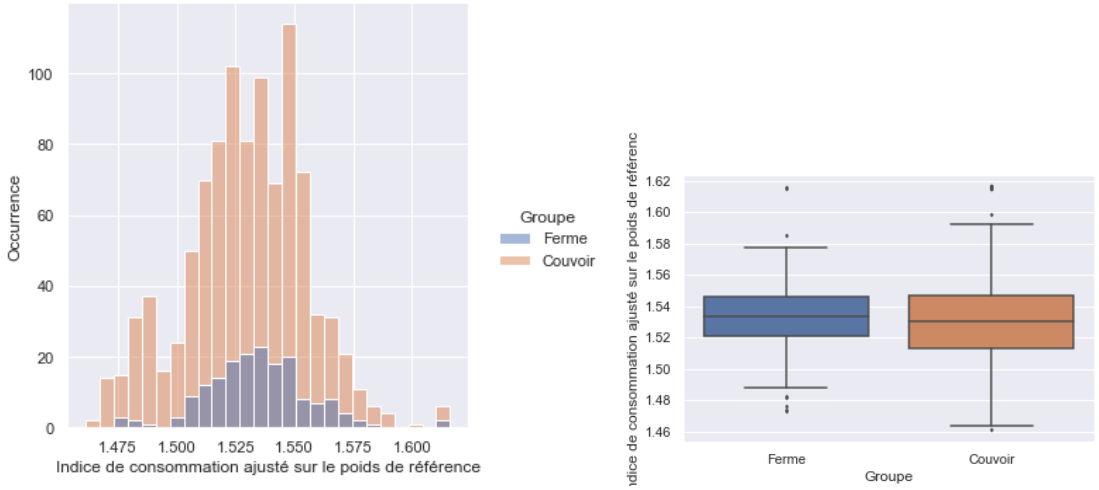


Figure 21 : Distribution de l'indice de consommation quotidien ajusté sur le poids de référence des poulets de chair Ross 308

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	989	1,529	0,03	1,461	1,513	1,530	1,546	1,616
Ferme	177	1,534	0,02	1,473	1,520	1,533	1,546	1,616

Tableau 15 : Valeurs statistiques de l'indice de consommation quotidien ajusté sur le poids de référence des poulets de chair Ross 308

Après analyse statistique par un test de Student avec $\alpha = 0,05$, on obtient une valeur de $p = 0,022$. Il y a donc une différence significative entre ces deux groupes. Le groupe vacciné au couvoir nécessite moins d'aliments pour produire un kilogramme de poids vif.

4.2.2.5. Impact sur le taux de saisie à l'abattoir

A l'abattoir, certaines carcasses ou certaines parties de l'animale peuvent être saisies suite à la constatation d'états anormaux. Ces saisies peuvent être justifiées par trois raisons :

- Présence de risque pour l'homme et les animaux (risque d'infection).
- Aspect de la viande (couleur, odeur, répugnance, etc.).
- Composition ou propriétés physico-chimique anormales.

Pour un éleveur, plus le taux de saisie est important, plus la perte et donc le coût de production est fort.

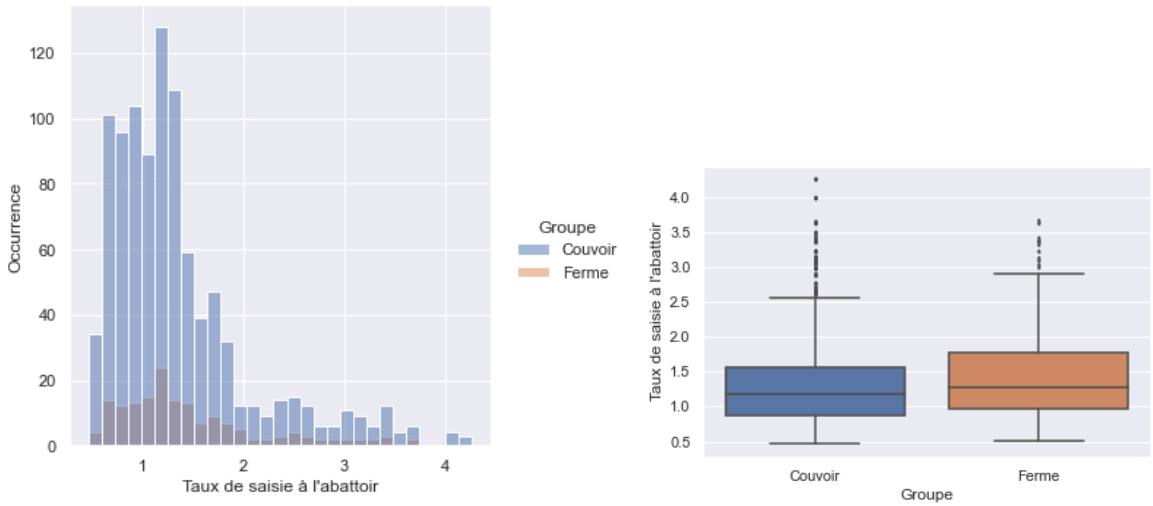


Figure 22 : Distribution du taux de saisie à l'abattoir

Groupe	n	Moyenne	Écart-type	Min	25%	50%	75%	Max
Couvoir	979	1,36	0,69	0,47	0,88	1,18	1,57	4,25
Ferme	166	1,47	0 ,72	0,52	0,96	1,27	1,77	3,66

Tableau 16 : Valeurs statistiques du taux de saisie à l'abattoir (%)

Après analyse statistique de nos données par un test du Student avec $\alpha = 0,05$, on obtient une valeur de $p = 0,042$ donc $p<0,05$. Il y a donc une différence significative entre les deux groupes. Le taux de saisie est plus important dans le groupe vacciné à la ferme que dans celui vacciné au couvoir.

5. Conclusion

D'après les résultats de notre étude, nous avons vu qu'il n'y a pas de différence significative entre les moyennes des titres sérologiques dans les deux groupes. En revanche, lorsque l'on s'intéresse aux coefficients de variation des titres sérologiques, on observe que les titres sérologiques sont plus dispersés dans le groupe vacciné à la ferme que dans celui vacciné au couvoir. On peut expliquer ce résultat par le fait que dans la vaccination au couvoir, certains animaux sont vaccinés de façon individuelle alors que dans la vaccination à la ferme, tous les animaux sont vaccinés de manière collective. L'avantage, et l'importance, d'avoir un coefficient de variation faible réside dans le fait que plus le coefficient de variation est faible, plus les titres sérologiques dans les troupeaux sont stables et reproductibles, c'est-à-dire que nous observons moins de différences entre les différents troupeaux et qu'il est plus rare d'avoir un troupeau avec des résultats sérologiques dispersés. D'après ce résultat, la vaccination au couvoir permet donc une distribution plus uniforme des vaccins.

Nous avons par la suite étudié la mortalité totale et la mortalité précoce. L'administration des vaccins au couvoir in-ovo étant plus invasive que l'administration à la ferme via l'eau de boisson, nous aurions donc pu penser que cette méthode allait avoir un impact négatif sur la mortalité et plus particulièrement sur la mortalité précoce. Or, les données ne démontrent aucune différence significative entre les deux groupes.

Concernant les analyses sur les données de production, nous avons vu qu'en analysant les données brutes ou en les ajustant sur les références des poulets de chair Ross

308, les résultats étaient différents. Comme expliqué dans l'introduction, les données proviennent de différents endroits dans le monde. Il faut savoir que le type, les contraintes et les impératifs de production ne sont pas les mêmes partout. Il est alors difficile de comparer entre elles les données brutes. En ajustant les données avec les standards de référence, on peut limiter l'impact de certains biais sur nos résultats. Ainsi, nous pouvons en déduire, d'après les résultats de notre étude, que la vaccination au couvoir a un impact positivement significatif sur le poids et l'âge à l'abattage, mais aussi sur l'indice de consommation.

Enfin, le taux de saisie peut, lui, être analysé sans ajustement car il repose sur des impératifs d'hygiène générale. Les données de notre étude nous ont montré que le taux de saisie était significativement plus faible pour les poulets de chair vaccinés au couvoir.

Pour conclure, la voie d'administration des vaccins de la bronchite infectieuse ne semble pas avoir d'impact sur la mortalité des poulets de chair. En revanche, la vaccination au couvoir, d'après notre étude, semble donner de meilleurs résultats concernant les données de production et du taux de saisie, mais aussi la variabilité des titres sérologiques de la maladie.

PARTIE III : PROBLÉMATIQUES ET DIFFICULTÉS LIÉES À LA MISE EN PLACE D'UNE BASE DE DONNÉES

1. Problématiques et difficultés liées à l'origine des données

1.1. Diversité des origines

Nous avons relevé trois points principaux pouvant être sources d'erreurs et de difficultés lors de la constitution d'une base de données concernant l'origine des données.

Tout d'abord, les données recueillies par le laboratoire peuvent provenir de sources de données liées à la production comme le couvoir, la ferme d'élevage, l'abattoir et les laboratoires d'analyses médicales, mais aussi de sources de données internes au laboratoire, comme d'anciens jeux de données. Si ces jeux de données ne sont pas clairement identifiés et que les variables ne sont pas explicitement définies, cette diversité d'origine peut être source de problèmes lors du rassemblement des fichiers dans une base commune.

En outre, ces jeux de données proviennent de plusieurs pays, ce qui donne pour résultat d'avoir des jeux de données récoltés dans différents alphabets et langages. Comme les Data Scientist ou les analystes ne sont pas experts dans toutes les langues utilisées, cette diversité peut entraîner des problèmes dans la gestion et l'utilisation des jeux de données.

Enfin, dans les pays avec une production diversifiée et surtout en Europe, notre façon de consommer évolue. Pour s'adapter à cette évolution, les industriels créent des labels de production (*Label Rouge*, *Agriculture Biologique*) afin de répondre au mieux aux attentes des consommateurs. Volaille Française, recense ainsi 4 000 labels différents en France. Si ces labels sont mal définis et identifiés, ils peuvent être source d'erreurs pour le traitement des données. Par exemple, tous les labels n'ont pas les mêmes normes en matière de densité de poulets au mètre carré, d'âge ou de poids minimum d'abattage.

1.2. Nécessité d'établir un système de traçabilité fiable

Comme expliqué précédemment, le principal problème de la mise en œuvre de cette solution réside dans les origines très diverses des jeux de données collectés. Pour que ces données puissent être exploitables lorsqu'elles sont rassemblées, il est essentiel de mettre en place des systèmes de traçabilité, qu'ils soient externes ou internes au laboratoire, afin de garantir une exploitation fiable de la donnée.

1.2.1. Traçabilité externe au laboratoire

La traçabilité externe doit être garante de l'affiliation des mesures récoltées à un animal ou un troupeau, mais aussi à un centre de production. Cette traçabilité doit permettre au laboratoire de récolter toutes les mesures qui ont été faites sur un troupeau en étant certain que ces mesures proviennent de ce troupeau et qu'à aucun moment, elles n'ont pu être altérées ou modifiées par un tiers. Cette traçabilité nécessite la mise en place de numéro d'identification unique pour les centres de production (couvoir, ferme, abattoir, etc.) et pour les troupeaux (numéro de lot). En mettant en place ce système d'identification unique, le laboratoire, mais aussi les autres acteurs la production avicole, pourront ainsi être certains de l'origine et du sujet des mesures récoltées.

1.2.2. Traçabilité interne au laboratoire

En plus de la traçabilité externe au laboratoire, qui se doit d'être garant de la véracité des données, ce dernier doit mettre en place une traçabilité interne. Elle doit, tout d'abord, permettre de répertorier l'origine des jeux de données récoltés. En effet, dans une multinationale, le nombre de collaborateurs et la rotation des postes sont importants. En répertoriant ainsi l'origine des jeux de données recueillis, le laboratoire se dote d'un moyen efficace, pour ses responsables ou futurs collaborateurs, de retrouver la provenance ainsi que l'ensemble des informations du jeu de données même si le collaborateur ayant effectué

le recueil ne travaille plus dans l'entreprise. Ce système de traçabilité doit également répertorier l'ensemble des actions entreprises sur ce jeu de données afin de pouvoir revenir à la source d'une potentielle erreur si nécessaire. Enfin, il permet aussi de retrouver facilement et rapidement un jeu de données particulier sur les serveurs de l'entreprise.

L'équipe de Data Management de Ceva Santé Animale, travaille actuellement sur cette problématique et a mis en place, via l'outil DataGalaxy, cette traçabilité interne.
DataGalaxy est une plateforme agile permettant aux équipes de gérer la gouvernance de leurs données. L'objectif de DataGalaxy est que chaque employé de l'entreprise, qu'il soit issu d'un milieu technique ou non, puisse accéder, explorer et contribuer à une connaissance commune des données de l'entreprise. Ce projet fondamental et ambitieux a été lancé en 2019 par Ceva Santé Animale et doit, à terme, permettre au laboratoire de valoriser les données produites et récoltées.

2. Problématiques et difficultés liées à la mesure des variables

Les différences de mesure des variables représentent la deuxième grande difficulté que nous avons répertoriée lors de la création de notre base de données.

2.1. Des mesures de variable très hétérogènes

Beaucoup de variables différentes, que nous avons répertoriées, ont pour objectif de mesurer une seule et même chose mais elles diffèrent sur certains points. Ces différences peuvent nous limiter dans le regroupement de nos fichiers et ainsi dans la constitution de la base de données.

2.1.1. Sur le temps

La première différence que nous avons relevée concerne le temps de mesure de nos variables. Effectivement, nous avons remarqué que beaucoup de nos mesures avaient pour objectif de mesurer une seule et même chose, mais qu'elles divergeaient sur le temps de prise de la mesure. Par exemple, pour mesurer la mortalité précoce, nous avons répertorié des mesures de la mortalité, au deuxième jour, au septième jour et au dixième jour. Ces différences de temps, nous empêchent de regrouper nos variables dans notre nouvelle base de données.

2.1.2. Sur les unités

La deuxième différence que nous avons dû prendre en compte et la différence d'unité utilisée dans les différents pays. Ces différences d'unités ne sont pas un obstacle aux regroupements des variables, car, en effectuant les bons facteurs de conversion, nous pouvons facilement passer d'une unité à l'autre. Ce qui pose problème avec le fait d'avoir des fichiers dans différentes unités, c'est le risque accru d'erreur d'interprétation ou de conversion des fichiers. En effet, si les unités sont mal renseignées sur le fichier d'origine, nous sommes obligés de déduire ou de faire des suppositions sur les unités utilisées. Cette déduction peut être très simple comme dans le cas d'un poids moyen exprimé en grammes ou en kilogrammes, mais, elle peut être plus compliquée comme lorsque l'on parle de système monétaire ou de variables moins fréquemment étudiées.

2.1.3. Sur les indicateurs

La troisième différence que nous avons remarquée concerne les indicateurs utilisés pour mesurer une variable. Ces indicateurs ne sont pas les mêmes dans les différents jeux de données. Par exemple, pour calculer la performance d'efficacité de production entre différents troupeaux, nous avons rencontré, dans nos jeux de données, différents indicateurs

comme le Feed Converting Ratio (FCR), le FCR corrigé, le FCR ajusté ou bien le FCR économique. Ces différents indicateurs ne peuvent pas, en l'état, être regroupés, mais si les données le permettent, en effectuant les bons calculs de conversion, ils pourront l'être par la suite. La diversité des indicateurs peut engendrer un risque accru d'erreur lors de la manipulation des jeux de données mais peut aussi empêcher le regroupement des données sur ces indicateurs.

2.2. Nécessité d'établir des références internationales

Cette hétérogénéité de la mesure des variables qui peut conduire à un risque d'erreur accru et à l'incapacité de les regrouper nous amène à réfléchir à la possibilité et aux avantages d'établir des normes de références internationales pour le recueil et la mesure des variables dans les centres de production des poulets de chair. La mise en place de telles normes ne pourrait se faire sans mettre autour de la table les différents acteurs de la filière qu'ils soient liés au médical, à la production, mais aussi à la distribution. En définissant communément des références et en établissant des règles de mesure, chaque acteur de la filière pourrait en tirer parti. Bien évidemment, pour que tous les protagonistes puissent en profiter de manière pérenne, chacun se doit d'être transparent sur les données renseignées.

L'avantage principal pour les laboratoires pharmaceutiques, et les acteurs transnationaux en général, d'établir des références internationales serait de pouvoir récupérer des jeux de données plus homogènes chez leurs clients, afin de réaliser des études de preuves empiriques sur les produits qu'ils proposent. Ainsi, il serait plus facile et efficace pour eux de redéfinir leurs positions, qu'elles soient économiques ou commerciales ou même de cibler de nouveaux marchés ou de nouvelles opportunités.

3. Problématiques et difficultés liées à des différences d'intérêts économiques

Les différences d'intérêts économiques représentent la troisième grande difficulté que nous avons eue à aborder lors de la constitution et de l'analyse de notre base de données.

3.1. En fonction du type de production du pays

3.1.1. Pays avec une production standardisée

Dans les pays avec une production standardisée, le but principal de la production avicole est encore de nourrir les habitants. Dans ces pays, il existe peu de labels différents et la réduction de l'usage des antibiotiques et le bien-être animal ne sont pas une priorité. On observe alors, par exemple, des fermes de production avec des densités de poulets au mètre carré plus élevées que dans les fermes des pays avec une production diversifiée. Or, cette forte densité peut avoir un impact sur la propagation de maladies dans le troupeau ; il nous est donc difficile de comparer directement les données recueillies dans ces fermes avec des fermes ayant un type de production différent. Ces observations de différences entre pays avec une production diversifiée et pays avec une production standardisée peuvent se faire sur d'autres variables comme le poids ou l'âge à l'abattoir.

3.1.2. Pays avec une production diversifiée

Dans les pays avec une production diversifiée, le but principal de la production est toujours de nourrir la population, mais tout en prenant en compte l'évolution de la consommation et du marché. En effet, dans ces pays, la très grande partie de la population ne meurt pas de faim et de nouvelles problématiques rentrent en compte comme l'antibiorésistance et le bien-être animal. Pour répondre à ces nouvelles façons de consommer, les industriels ont développé de nouveaux labels censés répondre aux attentes des consommateurs. Ces poulets de chair, issus, par exemple, de labels de production du type agriculture biologique ou en

plein air, semblent difficiles à comparer avec des poulets produits dans des pays avec des types de production différents.

3.2. Nécessité d'ajuster certaines variables par des références de production

Pour limiter les biais d'analyse liées à ces contraintes, nous pouvons, comme nous l'avons fait dans l'étude de cas, ajuster certaines variables par les références de production de la race étudiée. L'ajustement statistique consiste alors à corriger le résultat de la variable en prenant en compte les valeurs d'une variable de référence et en s'appuyant sur le référentiel de production de la variable étudiée. Par exemple, si on a comme valeur mesurée, un poids moyen de 3,12 kg à 42 jours. On remplace cette valeur brute par une donnée relative à la référence de la génétique. Dans ce cas, la référence est de 2,918 kg (Annexe 1), donc le poids moyen mesuré est de 0,20 kg au-dessus de la référence. On compare ces écarts aux références entre eux, ce qui nous permet d'éliminer l'effet de l'âge d'abattage. Ainsi, nous pouvons comparer certaines variables issues de différents pays avec différents types de production.

4. Contraintes similaires pour une application dans la santé humaine

En effectuant une recherche sur PubMed avec les mots-clefs « real world data », on obtient 28 247 résultats depuis 2001 dont 25 295 résultats depuis 2011. Cette recherche nous montre bien que l'engouement pour les real world data est un phénomène récent. Actuellement, il existe de nombreuses bases de données contenant des real world data comme les bases de données d'IQVIA et du Système National des Données de Santé (SNDS). Ces données sont principalement utilisées pour réaliser des études comparatives comme nous l'avons fait dans la partie deux de cette thèse et contiennent principalement des données dites « données de santé ».

Les données de santé sont définies par le règlement général sur la protection des données (RGPD) (8) comme « *les données à caractère personnel relatives à la santé physique ou*

mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent des informations sur l'état de santé de cette personne ».

En utilisant les données personnelles et de santé, nous pourrions nous servir du travail effectué dans cette thèse pour mettre en place une solution de recueil et d'automatisation de l'harmonisation des données personnelles et de santé afin de réaliser des études de Real World Evidence sur les médicaments humains. Les études de Real World Evidence peuvent servir à :

- Surveiller la sécurité et les effets indésirables des médicaments après la mise sur le marché afin de prendre des décisions réglementaires.
- Élaborer des directives et des outils d'aide à la décision pour la pratique clinique.
- Étayer la conception des essais cliniques et des études observationnelles afin de générer des approches thérapeutiques innovantes et nouvelles.

En mettant en place une telle solution, nous pourrions rencontrer des contraintes similaires à celles auxquelles nous avons dû faire face pour constituer notre solution mais aussi d'autres contraintes liées à la réglementation.

4.1. Diversité des sources de recueil des données

Les données de santé peuvent être produites dans différents lieux comme les hôpitaux et les pharmacies. Au sein de ces derniers, elles peuvent être regroupées sous différentes formes comme dans des bases de données médico-administratives, des cohortes et des registres, des dossiers médicaux. En plus de ces sources de données établies depuis longtemps, nous retrouvons dorénavant de nouvelles sources de données comme les objets connectés, les smartphones, les sites internet, mais aussi les réseaux sociaux. Cette multitude de sources est très intéressante pour réaliser des études de preuves empiriques, mais elle rend le regroupement des données bien plus difficile.

4.2. Diversité des formats

Les données produites par l'ensemble des sources de recueil citées précédemment ne peuvent être rassemblées en l'état. Chacune de ces sources produit des données qui ont pour but de répondre à une utilisation précise mais le fait de les rassembler dans une seule et même base n'a pas été pensé lors de leur conception et de leur génération. Pour pouvoir regrouper l'ensemble des données dans une seule base, deux solutions s'offrent à nous :

- Revoir la conception de ces sources de données de manière à produire des données utilisables à la fois de façon séparée (but initial de la création de ces données) et à la fois directement utilisables dans une base de données.
- Concevoir un système d'harmonisation des différents jeux de données issus de l'ensemble de ces sources, comme nous l'avons fait dans cette thèse, qui intervient après la génération des données par les différentes sources.

La solution la plus fiable serait de réaliser la première solution, mais, en réalité, il existe une multitude de sources de données ; il est donc plus simple pour l'utilisateur voulant coupler des bases de données entre elles de réaliser la deuxième solution.

4.3. Contrainte supplémentaire avec les données personnelles et de santé

En plus des contraintes similaires à l'implémentation de notre base de données, la mise en place d'une telle base de données avec des données personnelles et de santé devrait faire face à une autre contrainte majeure : le respect de la réglementation définie par le RGPD (8). Les entreprises qui souhaitent travailler sur les données de santé doivent obtenir le consentement de la personne. Le consentement est donné pour un traitement précis de la donnée ; donné pour un but et un objectif précis de traitement, il est donc difficile d'utiliser ce même consentement pour recouper à nouveau les données avec d'autres bases. De plus, dans certains cas comme les maladies rares, les données de santé peuvent facilement servir

à identifier une personne. Or d'après le RGPD, les données à caractère personnel ne doivent pas permettre l'identification de la personne.
Afin de travailler en accord avec la réglementation, la réflexion sur une base de données doit, en amont, tenir compte de ces contraintes pour obtenir un consentement libre et éclairé de la personne et définir les bons niveaux d'agrégation et de pseudonymisation.

5. Conclusion

En constituant notre base de données, nous avons rencontré plusieurs points de difficultés comme l'origine très diverse des jeux de données. En effet, les jeux de données proviennent de différentes sources de données, provenant elles-mêmes de différents pays. Si ces jeux de données ne sont pas clairement identifiés et que les variables ne sont pas explicitement définies, cette diversité d'origine peut être source de problèmes lors du rassemblement des fichiers dans une base commune. Pour résoudre ce problème, nous proposons de mettre en place deux systèmes de traçabilité. Le premier, externe au laboratoire, a pour but de garantir l'affiliation des mesures récoltées à un animal ou un troupeau, mais aussi à un centre de production. Le deuxième, interne au laboratoire, permet de répertorier les jeux de données et classifier les opérations conduites dessus.

En plus de cette première contrainte, nous avons constaté que beaucoup de variables avaient pour but de mesurer une seule et même chose. Ces variables ne différaient seulement que par le temps de mesure, les unités choisies ou les indicateurs utilisés. Nous proposons alors pour résoudre ce problème pouvant conduire à un risque d'erreur accru et à l'incapacité de regrouper les données de réfléchir à établir des normes et des références internationales pour recueillir et mesurer les variables dans les centres de production.

Enfin, nous nous sommes rendu compte, qu'en fonction de l'état de développement du pays, les intérêts économiques n'étant pas les mêmes, les systèmes de production des poulets de chair étaient différents. Pour certaines variables, il n'est pas pertinent de comparer des

données rassemblées en l'état. Il est préférable alors d'ajuster ces variables par des références de production de la race étudiée pour limiter les biais d'analyse.

Pour conclure, il est possible d'étendre nos travaux pour mettre en place une solution de recueil et d'automatisation de l'harmonisation des données personnelles et de santé afin de réaliser des études de preuves empiriques sur les médicaments humains. La conception de ces nouvelles bases devra faire face à des problématiques similaires aux nôtres, mais devra en plus tenir compte des contraintes réglementaires sur les données personnelles et de santé.

CONCLUSION :

La méthodologie appliquée a permis d'obtenir une base de données compréhensible et utilisable mais il semble que les étapes qui la composent ne soient que la première d'une longue série. Les processus mis en place pour obtenir cette première base de données doivent être améliorés, notamment en ce qui concerne le processus de vérification de la qualité des données afin de s'assurer que les données incluses dans la base de données répondent aux critères de qualité requis. Certains aspects concernant la collecte des jeux de données n'ont pas encore été abordés, c'est-à-dire qu'un processus doit être mis en place pour récupérer tous les jeux de données tout en ayant la possibilité d'interagir avec le collecteur ou l'analyste de données pour s'assurer que les variables et les caractéristiques sont comprises.

Cependant, la mise en œuvre de ce processus de gouvernance des données pour la création et l'automatisation d'une nouvelle base de données pour les études de preuves empiriques a permis plusieurs avancées pour Ceva Santé Animale.

Tout d'abord, ce travail a permis à l'entreprise de structurer ses "données clients". En effet, comme expliqué précédemment, ces données proviennent de différentes sources ; en les harmonisant, il a donc été possible de les rassembler dans une base de données claire.

Ce travail a également donné à l'entreprise l'occasion de réfléchir à un nouveau système de gouvernance des données pour les jeux de données clients. Plus de 100 études sur la volaille ont été réalisées depuis 2017, mais près de 60% de ces données, principalement les plus anciennes, sont difficilement exploitables. La création d'un processus de récupération de ces données semble absolument nécessaire afin de ne pas perdre ces données dans un premier temps, et dans un deuxième temps, de les valoriser.

Ce travail a aussi mis en évidence la disparité des mesures prises selon les pays mais aussi selon les sites de production. Une tâche intéressante serait d'établir un consensus sur les mesures à prendre sur un site de production mais aussi sur le moment où ces mesures doivent être prises.

La réalisation de ces travaux peut également servir de base à d'autres projets, en lien avec la volonté de Ceva Santé Animale de diversifier ses activités, notamment en intégrant davantage de données et de services numériques dans son offre. Il est tout à fait envisageable que cette base puisse servir à la mise en place d'algorithmes permettant de prédire les indicateurs clés de performance de production en temps réel.

Ce travail a été réalisé sur des études ad hoc relatives à la volaille mais il est destiné à être généralisé aux secteurs bovin et porcin bien que Ceva Santé Animale ait moins centralisé ses données pour ces secteurs. Il peut même être très intéressant de le mener avant de disposer d'un volume important de données afin de mettre en place des processus permettant une bonne collecte de données.

Enfin, la demande des industriels pour la réalisation d'études de preuves empiriques est croissante ; il serait donc intéressant d'appliquer ce travail au secteur de la santé humaine. Ce faisant, nous ferions face aux mêmes contraintes sur les données personnelles et de santé qu'à celles rencontrées et nous serions confrontés à d'autres liées à la réglementation.

LEXIQUE

Real World Evidence :

Les preuves empiriques en médecine, ou de Real World Evidence en anglais, désignent, d'après la Food and Drug Administration (FDA), les preuves obtenues à partir des données du monde réel, qui sont des données d'observation obtenues en dehors du contexte des essais randomisés contrôlés et générées au cours de la pratique clinique de routine. Elles peuvent servir à :

- Surveiller la sécurité et les effets indésirables des médicaments après la mise sur le marché afin de prendre des décisions réglementaires.
- Élaborer des directives et des outils d'aide à la décision pour la pratique clinique.
- Étayer la conception des essais cliniques et des études observationnelles afin de générer des approches thérapeutiques innovantes et nouvelles.

Labéliser :

D'après le dictionnaire Larousse, le terme labéliser définit le fait d'attribuer un label à un article, à un produit. Dans notre cas, le fait d'attribuer un label à une variable.

LISTE DES ANNEXES

Annexe 1 : Référence de production pour les poulets de chair Ross 308

Age (jour)	Poids moyen (Kg)	gain moyen quotidien (g/jour)	Indice consommation (ratio)
0	0,043		
1	0,061	61,000	0,206
2	0,079	39,500	0,370
3	0,099	33,000	0,502
4	0,122	30,500	0,607
5	0,148	29,600	0,693
6	0,176	29,333	0,763
7	0,208	29,714	0,821
8	0,242	30,250	0,869
9	0,280	31,111	0,911
10	0,321	32,100	0,947
11	0,366	33,273	0,979
12	0,414	34,500	1,007
13	0,465	35,769	1,033
14	0,519	37,071	1,057
15	0,576	38,400	1,080
16	0,637	39,813	1,101
17	0,701	41,235	1,122
18	0,768	42,667	1,142
19	0,837	44,053	1,162
20	0,910	45,500	1,182
21	0,985	46,905	1,201
22	1,062	48,273	1,221
23	1,142	49,652	1,240
24	1,225	51,042	1,259
25	1,309	52,360	1,278
26	1,395	53,654	1,297
27	1,483	54,926	1,317
28	1,573	56,179	1,336
29	1,664	57,379	1,355
30	1,757	58,567	1,375
31	1,851	59,710	1,394
32	1,946	60,813	1,414
33	2,041	61,848	1,433
34	2,138	62,882	1,453
35	2,235	63,857	1,473

Age (jour)	Poids moyen (Kg)	gain moyen quotidien (g/jour)	Indice consommation (ratio)
36	2,332	64,778	1,492
37	2,430	65,676	1,512
38	2,527	66,500	1,532
39	2,625	67,308	1,552
40	2,723	68,075	1,571
41	2,821	68,805	1,591
42	2,918	69,476	1,611
43	3,015	70,116	1,631
44	3,112	70,727	1,651
45	3,207	71,267	1,671
46	3,303	71,804	1,690
47	3,397	72,277	1,710
48	3,491	72,729	1,730
49	3,583	73,122	1,750
50	3,675	73,500	1,770
51	3,766	73,843	1,789
52	3,856	74,154	1,809
53	3,944	74,415	1,829
54	4,032	74,667	1,848
55	4,118	74,873	1,868
56	4,203	75,054	1,887
57	4,286	75,193	1,907
58	4,369	75,328	1,926
59	4,450	75,424	1,945
60	4,530	75,500	1,965
61	4,608	75,541	1,984
62	4,685	75,565	2,003
63	4,760	75,556	2,022
64	4,835	75,547	2,041
65	4,907	75,492	2,060
66	4,979	75,439	2,079
67	5,049	75,358	2,098
68	5,117	75,250	2,116
69	5,184	75,130	2,135
70	5,250	75,000	2,154

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Bosch-Capblanch X. Harmonisation of variables names prior to conducting statistical analyses with multiple datasets: an automated approach. BMC Med Inform Decis Mak. 19 mai 2011;11:33.

2. DataGalaxy- Le Data Catalog 360° pour la datagovernance [Internet]. [cité 11 déc 2020]. Disponible sur: https://www.datagalaxy.com/

3. pandas.DataFrame.describe — pandas 1.1.4 documentation [Internet]. [cité 4 déc 2020]. Disponible sur: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html

4. Cavanagh D. Coronavirus avian infectious bronchitis virus. Vet Res. mars 2007;38(2):281-97.

5. Saif YM, Barnes HJ, éditeurs. Diseases of poultry. 12th ed. Ames, Iowa: Blackwell Pub. Professional; 2008. 1324 p.

6. Guérin J-L, Boisseau C. La bronchite infectieuse. AVIcampus; 2008.

7. Association des vétérinaires en industrie animale. Bronchite Infectieuse. 2013.

8. Règlement (UE) 2016/679 du parlement Européen et du conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractçre personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE. Journal officiel de l'Union européenne; 2016.

RÉSUMÉ

Cette thèse décrit le processus de gouvernance des données mis en place permettant la collecte et l'harmonisation de la nomenclature des jeux de données, partagés avec Ceva Santé Animale par ses clients pour des études ad hoc, afin de créer et d'automatiser la constitution d'une nouvelle base de données permettant de conduire des études de preuves empiriques. Cette harmonisation a permis de réaliser une étude transversale sur la performance des vaccins administrés au couvoir comparé à ceux administrés à la ferme. Enfin, ce travail a permis de mettre en évidence les problématiques et difficultés que nous avons rencontrées tout au long de ce processus de mise en place d'une base de données obtenues en conditions réelles d'utilisation (« Real world data ») dans le domaine de la santé.

MOTS-CLÉS

Industrie pharmaceutique ; Médicaments vétérinaire ; Vaccins ; Gestion des données ; Real World Evidence ; Data Management

Université de Poitiers

Faculté de Médecine et Pharmacie

SERMENT DE GALIEN

En présence des Maitres de la Faculté, je fais le serment :

D'honorer ceux qui m'ont instruit(e) dans les préceptes de mon art et de leur témoigner ma reconnaissance en restant fidèle aux principes qui m'ont été enseignés et d'actualiser mes connaissances,

D'exercer, dans l'intérêt de la santé publique, ma profession avec conscience et de respecter non seulement la législation en vigueur, mais aussi les règles de Déontologie, de l'honneur, de la probité et du désintéressement,

De ne jamais oublier ma responsabilité et mes devoirs envers la personne humaine et sa dignité,

De ne dévoiler à personne les secrets qui m'auraient été confiés ou dont j'aurais eu connaissance dans l'exercice de ma profession,

De faire preuve de loyauté et de solidarité envers mes collègues pharmaciens,

De coopérer avec les autres professionnels de santé.

En aucun cas, je ne consentirai à utiliser mes connaissances et mon état pour corrompre les mœurs et favoriser des actes criminels.

Que les Hommes m'accordent leur estime si je suis fidèle à mes promesses. Que je sois couvert(e) d'opprobre et méprisé(e) de mes confrères si j'y manque.

Signature de l'étudiant	du Président du jury
Nom :	Nom :
Prénom :	Prénom :

Version validée par la conférence des Doyens de facultés de Pharmacie le 7 février 2018

faculté de médecine et de pharmacie

1431 Université de Poitiers