Test en Data Science de la BNDMR

Ranujan KATHIR

May 1, 2024

1 Abstract

Ce rapport a été conçu après les analyses effectuées sur les patients hospitalisées pour l'épidermolyse bulleuse (EB), une maladie dermatologique rare, à la demande de la filière FIMARAD. Une première partie de ce rapport se repose sur une étude démographique des patients, la seconde partie se focalise sur l'impact du type d'EB dans la concentration de plomb dans le sang.

2 Introduction

2.1 Contexte

D'après la nomenclature Orphanet des maladies rares, l'épidermolyse bulleuse se distingue en plusieurs catégories :

- épidermolyse bulleuse simple, noté 304 :
  - épidermolyse bulleuse simple avec atrésie du pylore, noté 158684;
  - épidermolyse bulleuse simple avec dystrophie musculaire, noté 257;
- épidermolyse bulleuse jonctionnelle, noté 305 :
  - épidermolyse bulleuse jonctionnelle avec atrésie pylorique, noté 79403;
  - épidermolyse bulleuse jonctionnelle localisée, noté 251393.

2.2 Objectifs

Dans cette étude, nous nous intéresserons à :

- étudier les données sur les patients hospitalisées pour l'EB;
- évaluer l'impact du type d'EB sur l'évolution de la concentration de plomb chez les patients traités avec le même médicaments.

3 Management des données

3.1 Présentation des tables

Afin de réaliser cette étude, la direction nous a confiée trois tables :

- *patients* : données démographiques des patients avec la date de naissance des patients, les coordonnées géographiques des patients (numéro de maison, nom de la rue), le statut vital du patient et l'identifiant du patient;
- *diagnostics* : données sur les diagnostics des patients avec le type d'EB atteint par le patient, le niveau associé à cet EB, la date de début de l'hospitalisation et celle de la dernière activité, et l'identifiant des patients;
- *obs_plomb* : données sur l'état des patients traités par le plomb durant 6 semaines avec l'identifiant des patients et la concentration de plomb dans le sang selon 4 semaines

Les 3 tables possèdent comme une variable en commun, l'identifiant des patients. Nous pouvons nous servir de cette variable comme clé primaire pour harmoniser les données.

3.2 Pre-processing des données

L'étape du *pre-processing* permet d'harmoniser les données en procédant par le nettoyage des données (traitement des valeurs manquantes et des doublons) et la création de nouvelles variables à partir des anciennes variables (par exemple, la durée de l'hospitalisation des patients, l'âge de l'inclusion des patients, l'âge lors de la dernière activité ou la redéfinition des catégories de l'EB). Afin d'éliminer toutes valeurs manquantes, nous les supprimerons. Quant aux doublons, nous conserverons la première observation. Il est intéressant de sélectionner uniquement les variables qui peuvent être intéressant dans notre étude. Par exemple, les variables *housenum* et *streetname* possèdent toutes des modalités différentes selon l'observations, donc il n'est pas intéressant de s'intéresser à ces variables après le traitement des doublons. Durant l'étape du pre-processing et l'étude, nous prendrons en compte la variable *level* pour avoir une base de données complètes. Il est évident que la suppression des valeurs manquantes entraîne de la perte d'information.

4 Étude démographique des patients atteints de l'EB

4.1 Analyse des types d'EB

Initialement, on rencense 161 patients hospitalisés pour l'EB, l'harmonisation des données (par suppression des valeurs manquantes) nous permet de compter 124 patients à la place. Cette diminution résulte d'un nombre de valeur manquantes élevées dans la variable *level*.

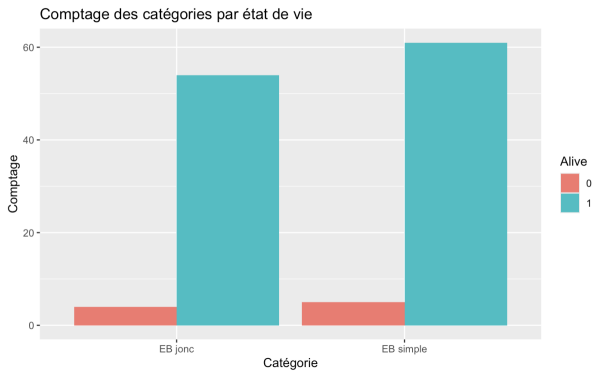

Figure 1: Statut vital des patients atteints de l'EB et dont on a un niveau

Note de lecture: Les valeurs 0 et 1 représentent le statut vital du patient. 0 si le patient est décédé et 1 sinon.

L'étude du statut vital du patient montre que quelque soit le type d'EB du patient, les patients rencensés sont vivants. Il est intéressant de comprendre la répartition des type de l'EB chez les patients. On constate qu'il y a 66 patients atteint de l'EB simple et 58 atteint de l'EB fonctionnelle. Concentrons-nous plus en détail sur les sous-catégorie.

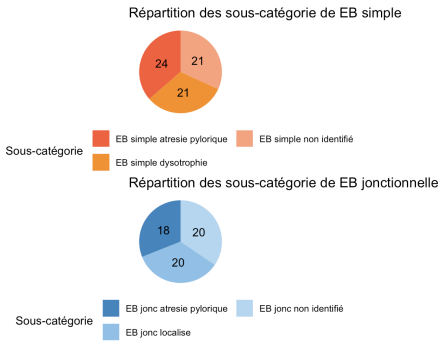

Figure 2: Répartition des sous-catégories

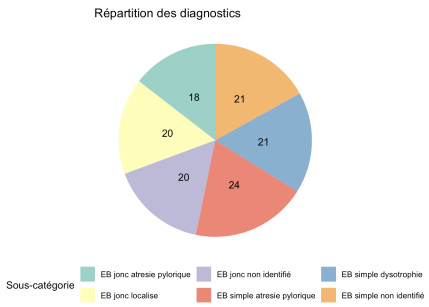

Figure 3: Répartition des diagnostics chez les patients

Note de lecture: Dans ce graphique, on distingue les deux types d'EB principale. Les sous-catégories associées à l'EB jonctionnelle et simple non identifiées font référence aux EB dont on a pu associé à une sous-catégorie.

En général, les données sont plutôt équilibrées. En effet, quelque soit le type d'EB, on constate une répartition plutôt égale dans les sous-catégorie.

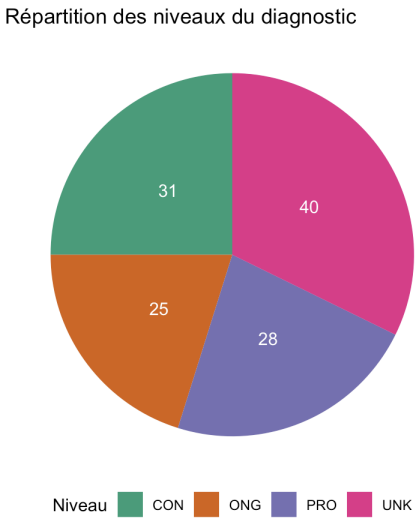

Figure 4: Répartition des niveaux selon les type d'EB

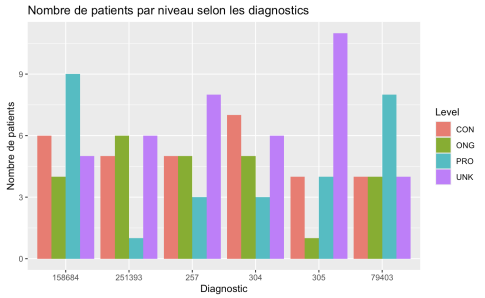

Figure 5: Répartition des niveaux selon les type d'EB

Note de lecture: La variable *level* possède 4 modalités : CON, PRO, UNK et ONG.

Le test du χ^2 entre le diagnostics et le niveau montre une significativité de 20%.

Dans l'ensemble, le niveau UNK prédomine largement. On observe que les patients atteints de l'EB jonctionnelle non identifié en sont les premiers bénéficiaire.

4.2 Analyse des séjours d'hospitalisation

Nous analyserons les séjours des patients atteint de l'EB durant leur hospitalisation. Nous distinguerons deux périodes : l'inclusion du patient à l'hôpital et sa dernière activité à l'hôpital. L'âge d'inclusion et de la dernière activité minimale est de 20 ans dans les deux cas, en revanche, dans l'âge maximale sont respectivement 102 et 103 ans. L'âge moyen de l'inclusion du patient à l'hôpital est d'environ 60 ans et celui lors de la dernière activité est d'environ 61 ans.

Pour cela, il est intéressant de créer une tranche de classe avec les patients de plus de 50 ans et ceux de moins de 50 ans.

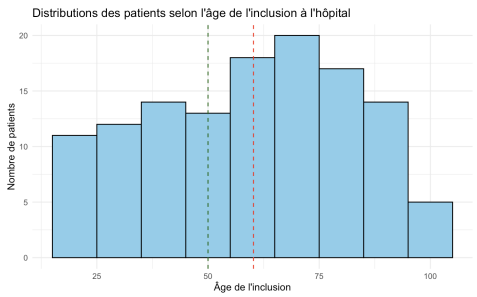

Figure 6: Distribution de l'âge d'inclusion

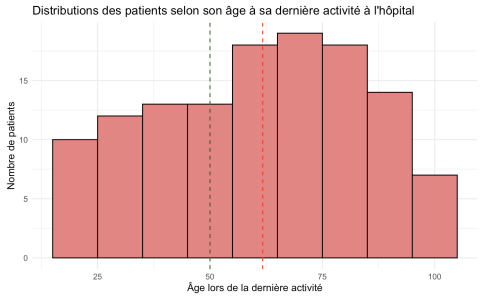

Figure 7: Distribution de l'âge lors de la dernière activité

Note de lecture: Le trait pointillé en rouge représente la moyenne de l'âge des patients hospi-

talisés pour l'EB au moment de l'inclusion, soit 60.23 ou de la dernière activité, soit 61.72. Le trait pointillé en vert fait référence à 50. Durant notre étude, nous différencierons les patients agés de moins de 50 ans et plus de 50 ans.

On constate que beaucoup des patients hospitalisés sont agés de plus de 50 ans. Concernant la répartition des types d'EB selon les patients, on constate une domination des patients de plus de 50 ans. En effet, on obtient un pic vers 70-75 ans au niveau de l'âge d'inclusion à l'hôpital et la dernière activité à l'hôpital. Il s'avère qu'on obtient 20 patients.

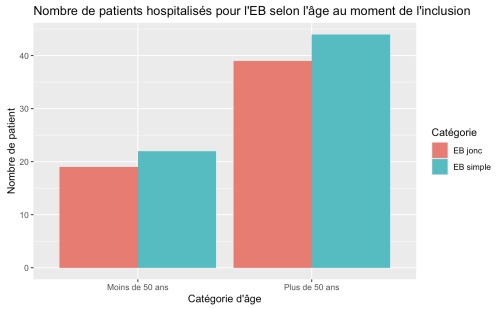

Figure 8: Nombre de patients hospitalisés lors de l'inclusion selon le type d'EB

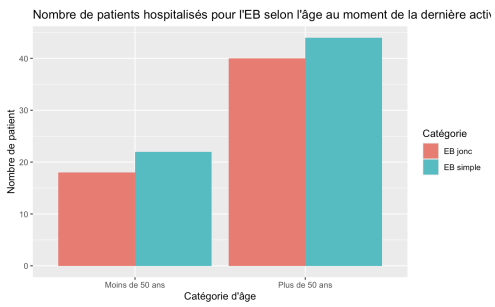

Figure 9: Nombre de patients hospitalisés lors de la dernière activité selon le type d'EB

Note de lecture: À droite, le graphique représente le nombre de patients hospitalisés pour l'EB lors de la dernière activité. À gauche, le graphique représente le nombre de patients hospitalisés lors de son arrivée.

Quelque soit la classe d'âge des patients, les patients atteints de l'EB simple sont majoritaires.

Table 1: Répartition des patients selon l'âge et le type de l'EB lors de l'inclusion

Type de l'EB	**Moins de 50 ans**	**Plus de 50 ans**
EB jonctionnelle atrésie pylorique	5	13
EB jonctionnelle localise	7	13
EB jonctionnelle non identifié	7	13
EB simple atrésie pylorique	6	18
EB simple dysotrophie	7	14
EB simple non identifié	9	12

Table 2: Répartition des patients selon l'âge et le type de l'EB lors de la dernière activité

Type de l'EB	**Moins de 50 ans**	**Plus de 50 ans**
EB jonctionnelle atrésie pylorique	5	13
EB jonctionnelle localise	6	14
EB jojonctionnellenc non identifié	7	13
EB simple atrésie pylorique	6	18
EB simple dysotrophie	7	14
EB simple non identifié	9	12

La durée moyenne des patients à l'hôpital est d'environ 551.1 jours, soit un peu plus de 1 ans et 6 mois.

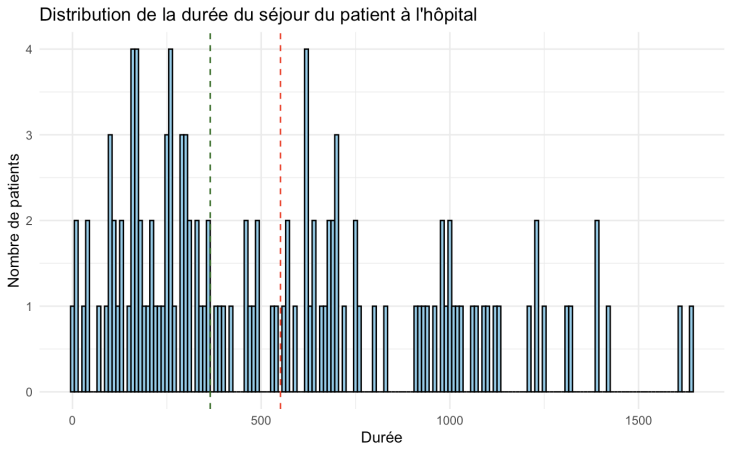

Figure 10: Nombre de patients selon la durée du séjour

Note de lecture: Le trait vert pointillé représente 365 jours, soit 1 an, le trait rouge pointillé représente la durée moyenne de séjour.

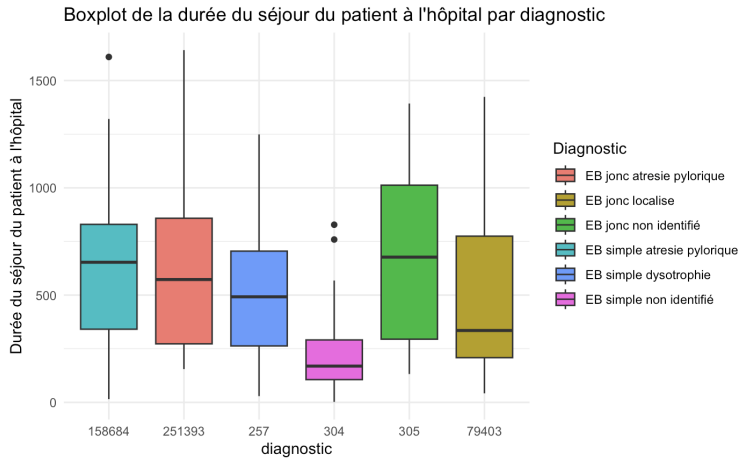

Figure 11: Répartition de la durée de séjour selon le diagnostics

Pour beaucoup de patients, la durée de séjour ne dépasse pas une année. Il s'avère que suivant le diagnostic, la durée de séjour varie (cf.figure 10)

Concernant les patients suivis pour la concentration de plomb dans le sang, on trouve 83 patients atteints d'un type d'EB et dont nous avons des informations sur le niveau.

5 Étude de l'impact du type EB sur l'évolution de la concentration de plomb

Dans cette partie, nous nous intéresserons à l'impact de l'EB sur la concentration de plomb dans le sang.

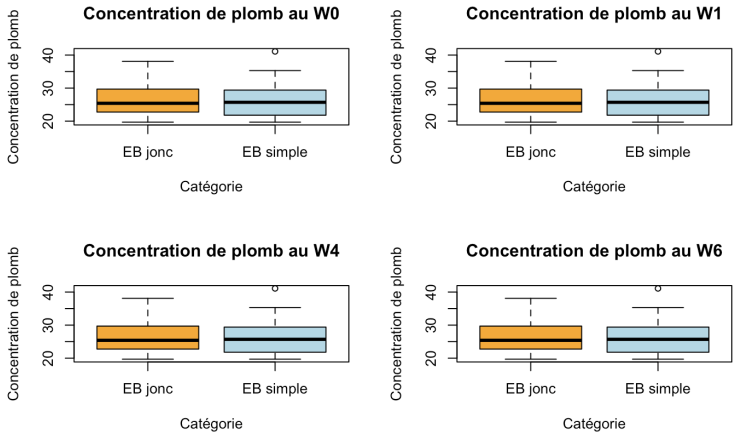

Figure 12: Répartition de la la concentration de plomb selon les semaines en fonction du type EB

D'une manière générale, on constate que la répartition de la concentration de plomb selon les types d'EB est la même quelque soit la semaine.

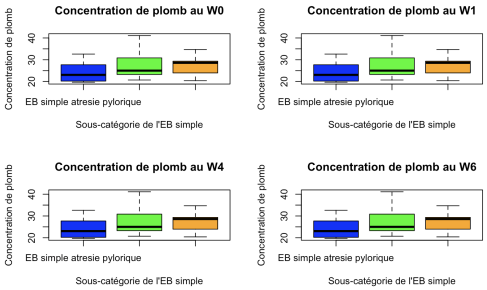

Figure 13: Répartition de la concentration de plomb selon les types d'EB simple

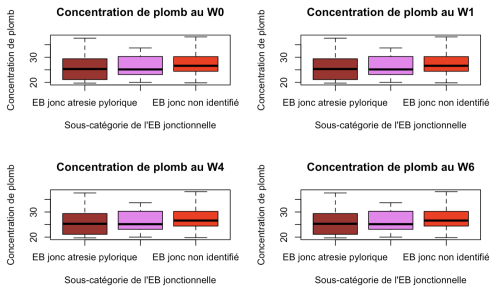

Figure 14: Répartition de la concentration de plomb selon les types d'EB jonctionnelle

Note de lecture: À droite, nous avons la répartition de la concentration de plomb selon les types d'EB jonctionnelle. À gauche, nous avons la répartition de la concentration de plomb selon les types d'EB simple.

En revanche, si l'on se focalise sur les sous-catégories en particulier, on constate que la concentration de plomb pour les patients atteint de l'EB atrésie pylorique est moins élevés que les autres. On retrouve la même chose du côté de l'EB atrésie pylorique dont la répartition est plutôt inégale.

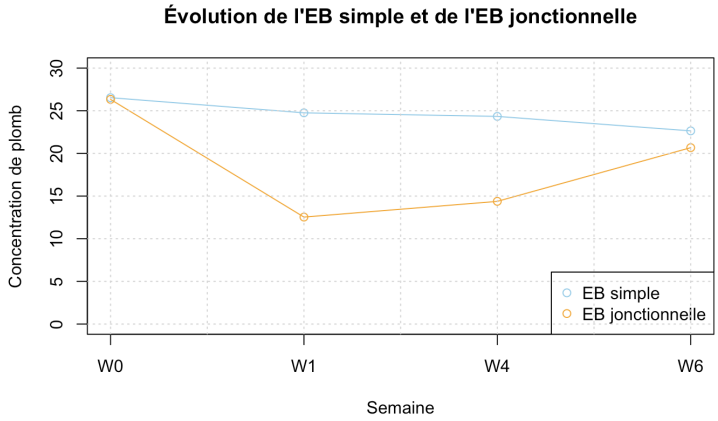

Figure 15: Évolution de la concentration de plomb moyenne hebdomadaire selon le type d'EB

Note de lecture: On observe que, la concentration de plomb dans le sang diminue lentement pour l'EB simple alors que celle de l'EB jonctionnelle connaît un pic à la première semaine avant de redresser lentement.

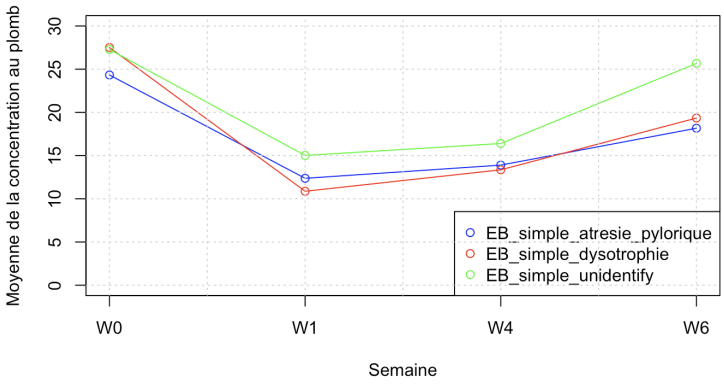

Figure 16: Évolution de la concentration de plomb moyenne hebdomadaire selon les types d'EB simple

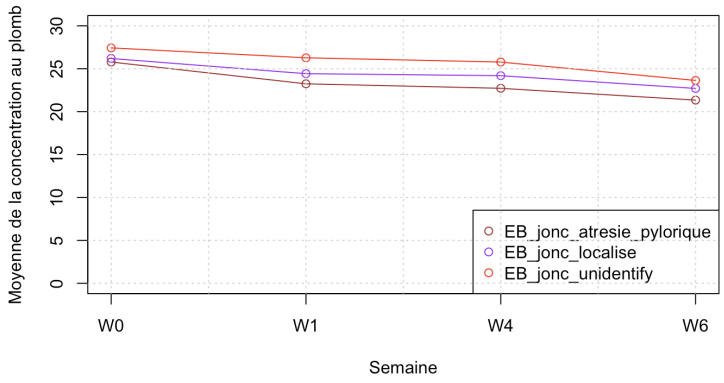

Figure 17: Évolution de la concentration de plomb moyenne hebdomadaire selon les types d'EB jonctionnelle

Note de lecture: En bas, nous avons l'évolution de la concentration de plomb selon les types d'EB jonctionnelle. En haut, nous avons l'évolution de la concentration de plomb selon les types d'EB simple.

On observe que, pour les patients atteints de l'EB jonctionnelle, quelques soient le types en particulier, ont leur concentration de plomb qui diminuent lentement et ont la même tendance. De même, on retrouve ce même cas avec les types d'EB simple.

Annexe

Les valeurs manquantes

Afin d'expliquer le traitement des valeurs, nous nous concentrerons sur les variables *level* et *diagnostics*. Le traitement des valeurs manquantes est primordiale afin d'avoir une analyse de données de qualités. Toutefois, celle-ci entraîne une perte d'information. Dans notre cas, il y a matière à se poser la question avec l'intérêt de la variable *level* qui avait 4 modalités distinctes.

Type EB	**Quantité de "NA" (%)**
EB simple avec atrésie du pylore	10.0
EB jonctionnelle localisée	17.5
EB simple avec dystrophie musculaire	15.0
EB simple	15.0
EB jonctionnelle	20.0
EB jonctionnelle avec atrésie pylorique	22.5

Table 3: Proportion de valeurs manquantes dans la variable *level* pour chaque modalité de la variable *diagnostics*

On constate malgré tout que, beaucoup de patients associés aux types EB jonctionnelle ont des valeurs manquantes pour la variable *level*.

Test du χ^2

Le test du χ^2 est une méthode statistique utilisée pour évaluer l'indépendance entre deux variables catégorielles. Il compare les fréquences observées avec les fréquences attendues si les variables étaient indépendantes. Le test du χ^2 repose sur l'hypothèse nulle selon laquelle il n'y a pas de relation entre les variables, et il génère une statistique de test qui mesure la divergence entre les observations réelles et celles attendues sous l'hypothèse nulle. Si la statistique de test est significative, cela suggère que les variables ne sont pas indépendantes. Dans notre cas, nous avons volontairement mis une significativité au seuil de 20% compte tenu de nombre d'observations assez faible.