



Rapport de stage

Recensement et analyse de données santé dans le cadre du socio-écosystème du Rhin Supérieur

Mars – Juillet 2019

Elaboré par : David Witkowski

Encadré par : Florence Le Ber, Erik-André Sauleau, Agnès Braud

Table des matières

Table des figures.....	I
Table des tables.....	II
1 Introduction.....	1
2 La santé dans le contexte des centrales nucléaires.....	2
2.1 Le cancer infantile.....	3
2.2 Les malformations congénitales.....	4
3 Recensement des données de santé.....	5
3.1 Bases de données santé.....	7
3.1.1 France.....	8
3.1.2 Allemagne.....	10
3.1.3 Suisse.....	13
3.2 Demande d'accès aux données du PMSI.....	14
4 Analyse des données.....	16
4.1 Les données ouvertes et développement d'un outil de visualisation des données.....	16
4.2 Les données PMSI avec la MR-005.....	20
4.2.1 Requêtes.....	21
4.2.2 Résultats.....	22
5 Conclusion.....	29
Bibliographie.....	31

Table des figures

Figure 1 : Niveaux d'analyse de la région du Rhin Supérieur.....	6
Figure 2 : Schématisation du circuit des demandes d'accès au SNDS.....	14
Figure 3 : Capture d'écran de l'outil de visualisation.....	20
Figure 5 : Ratio cas observés/attendus par unité géographique.....	28

Table des tables

Tableau 1 : Synthèse des données de santé en France.....	8
Tableau 2 : Synthèse des données de santé en Allemagne	11
Tableau 3 : Synthèse des données de santé en Suisse	13
Tableau 4 : Indicateurs sélectionnés et leurs sources	18
Tableau 5 : Répartition des patients par requête.....	24
Tableau 6 : Répartition des malformations par catégorie.....	25
Tableau 7 : Nombre de naissances et de cas par année	26
Tableau 8 : Caractéristiques des unités géographiques	27

1 Introduction

Le stage « Recensement et analyse de données santé dans le cadre du socio-écosystème du Rhin Supérieur » s'est déroulé du 1 mars 2019 jusqu'au 31 juillet 2019 dans le cadre du projet « Juxta Rhenum ». Ce projet a été lancé par l'Université de Strasbourg et a l'ambition de faire émerger autour du site de la Centrale Nucléaire de Production d'Electricité (CNPE) de Fessenheim un observatoire pluri et interdisciplinaire. Il propose une approche sur du long terme avec des acteurs, à la fois académiques et industriels, structurés autour d'une question scientifique majeure : la capacité de modéliser/anticiper l'évolution future d'un socio-écosystème complexe dans sa globalité en réponse aux modifications qu'induiront les politiques de transition énergétique nationales et européennes.

Dès le départ, la situation de ce socio-écosystème aux frontières de l'Allemagne et de la Suisse renforce une envergure internationale et notamment transfrontalière. Un objectif principal du projet « Juxta Rhenum » est de définir un « état zéro » décrivant de façon globale et la plus exhaustive possible le fonctionnement actuel du socio-écosystème en caractérisant l'impact actuel, tant environnemental que socio-économique de la CNPE Fessenheim. L'établissement et la modélisation de l'état de référence se fait à l'aide d'un ensemble cohérent d'indicateurs.

Dans le cadre du projet « Juxta Rhenum », ce stage vise à analyser la dimension santé et à positionner dans le temps et dans l'espace toute modification du socio-écosystème qui pourrait avoir un impact sanitaire sur tout ou partie de la population. Les objectifs du stage étaient de recenser les données pertinentes existantes en termes d'indicateurs de santé et de commencer la collecte de certaines, de réfléchir à la modélisation d'une base de données pouvant accueillir de telles données et de commencer la mise en œuvre d'analyses statistiques, intégrant des dimensions temporelles et spatiales.

J'ai trouvé ce projet très intéressant dès le début en raison de la perspective internationale. Originaire d'Allemagne, je maîtrise les deux langues présentes dans la région du Rhin Supérieur et il m'est donc possible de faire des recherches des deux côtés du Rhin. De plus, étant stagiaire à l'étranger, j'ai profité d'une bourse Erasmus+, un programme de la Commission européenne visant à donner aux étudiants et aux stagiaires la possibilité de séjourner à l'étranger pour renforcer leurs compétences et accroître leur employabilité. Dans le cadre de mes études de sciences sociales et statistiques appliquées, j'ai suivi des cours universitaires à Bonn, Varsovie, Vancouver et Cologne. Ce stage m'a semblé être un complément approprié à ma carrière

universitaire, à l'approfondissement de mes compétences et à l'acquisition d'une expérience internationale. Après avoir contacté les responsables du stage en octobre 2018, un entretien a eu lieu en novembre 2018, au cours duquel le mois de mars a été fixé comme date de début du stage.

Le stage s'est fait au laboratoire ICube, laboratoire pluri-disciplinaire de l'Université de Strasbourg des sciences de l'ingénieur, de l'informatique et de l'imagerie (UMR7357), le lieu de stage étant l'ENGEES (École nationale du génie de l'eau et de l'environnement de Strasbourg). Dans le cadre du stage, j'ai été encadré par Agnès Braud, maîtresse de conférences en informatique (bases de données, fouille de données), Florence le Ber, directrice de recherche à l'ENGEES (fouille de données, analyses temporelles et spatiales) et Erik-André Sauleau, Professeur des universités – Praticien hospitalier (analyse statistiques, santé publique). Le stage a également bénéficié de l'expertise de Wahida Kihal, chargée de recherches CNRS au LIVE (Laboratoire Image, Ville, Environnement), et de Pierre Tran Ba Loc, praticien hospitalier aux Hôpitaux Universitaires de Strasbourg.

Des réunions organisés par les responsables du projet « Juxta Rhenum » ont permis de faire la connaissance d'autres stagiaires au cours du stage. Pendant les réunions, les stagiaires ont présenté leur sujet de stage et l'état d'avancement des travaux. Le projet étant pluri et interdisciplinaire, des étudiants de différentes disciplines (p.ex. chimie, géographie) se sont rencontrés pour discuter au-delà des frontières de leur discipline.

2 La santé dans le contexte des centrales nucléaires

L'objectif du stage est de recenser et collecter les données de santé existantes dans le Rhin Supérieur. Dans le cadre du projet « Juxta Rhenum », les données sont analysées en particulier dans le contexte de la CNPE Fessenheim. Par conséquent, l'état de la recherche sur le lien entre les rayonnements radioactifs et la santé doit être discuté.

En général, déterminer le rôle des agents environnementaux demeure un défi parce que, dans la plupart des situations, l'environnement impose de nombreux stimuli différents, certains bénéfiques, d'autres neutres et d'autres potentiellement nocifs. Par conséquent, les interactions entre deux ou plusieurs de ces facteurs peuvent avoir des effets différents (Weinhold, 2009). L'exposition d'une personne à certains pesticides et autres produits chimiques peut avoir un effet néfaste sur la santé. Le fait de travailler ou de vivre à proximité des sites de déchets, des fonderies ou des mines peut également constituer un facteur de risque, particulièrement si une personne est

exposée à d'autres facteurs de risque environnementaux ou à des carences nutritionnelles (World Health Organization, 2016). Les substances toxiques provenant de sites pollués (métaux lourds, composés volatils et autres composés organiques) augmentent le risque d'effets néfastes pour la santé (faible poids à la naissance, malformations congénitales, certains types de cancers) (Vrijheid, 2000). La pollution de l'air est un problème majeur de santé publique dans les pays développés et les pays en développement en raison des émissions industrielles et des émissions résultant du transport.

En ce qui concerne le rayonnement radioactif en tant que facteur environnemental, étant donné que les émissions d'une centrale nucléaire n'augmentent que très peu le niveau de rayonnement naturel, aucun effet n'est attendu sur la base des modèles habituels pour les effets des faibles niveaux de rayonnement (IRSN, 2015). Comme l'exposition associée aux rejets des installations nucléaires lors de leur fonctionnement sur les personnes les plus exposées est très faible, d'autres facteurs que la radiation des sites nucléaires sont considérés comme ayant une influence, comme le mélange de populations ou l'exposition à des agents physiques (Sermage-Faure et al., 2012).

Néanmoins, les effets possibles des centrales nucléaires sur l'environnement et la santé publique ont été envisagés depuis l'introduction de la technologie. Depuis de nombreuses années, la question de savoir si l'émission de rayonnements ionisants lors de l'exploitation courante des centrales nucléaires augmente déjà le risque de certaines maladies est controversée. Les maladies les plus couramment étudiées dans la littérature sont la leucémie chez les enfants et des malformations congénitales.

2.1 Le cancer infantile

Le débat public sur les effets des centrales nucléaires a été alimenté principalement par des études sur les cancers infantiles. Les leucémies représentent le premier cancer pédiatrique et de nombreuses études ont été menées sur les effets possibles des radiations ionisantes. Parmi d'autres facteurs de risque contextuels, seules les expositions aux rayonnements ionisants sont reconnues comme facteur de risque de la leucémie (Cancer Environnement, 2018). En particulier les études sur les survivants d'Hiroshima ont montré que les fortes doses de rayonnement radioactif peuvent causer le cancer. Le cancer chez les enfants est souvent étudié parce qu'il est considéré comme l'indicateur le plus approprié pour étudier l'influence du rayonnement radioactif sur l'incidence du cancer. En effet, le cancer qui survient le plus rapidement après l'exposition est la leucémie. De plus, le lien entre les rayonnements radioactifs et le cancer du sang est plus fort que pour les cancers solides. Enfin, le risque de développer un cancer du sang est plus élevé chez

une personne qui est exposée à des radiations durant l'enfance, avant l'âge de 5 ans (Preston, 2004).

De nombreuses études, surtout les études réalisées sur de nombreux sites, ne constatent aucun effet significatif (Evrard et al., 2006; Laurier and Bard, 1999; Mueller and Gilham, 2015; White-Koning et al., 2004). Cependant, d'autres études ont montré un risque accru de cancer chez les enfants à proximité des centrales nucléaires (Forman et al., 1987; Sermage-Faure et al., 2012; Spix et al., 2008). L'étude cas-témoin allemande KiKK (Kinderkrebs und Kernkraftwerk) a reçu beaucoup d'attention et a montré une association entre la leucémie chez les enfants de moins de 5 ans et vivant à moins de 5 km d'une centrale nucléaire (Kaatsch, 2007; Kinlen, 2011; Spix et al., 2008). La zone géographique couverte par l'étude comprend 41 zones situées à proximité de 16 centrales nucléaires allemandes. Le groupe d'étude comprenait tous les enfants jusqu'à l'âge de 5 ans ayant reçu un diagnostic de leucémie entre 1980 à 2003. D'une façon similaire, l'étude française Géocap (Sermage-Faure et al., 2012) a étudié l'association entre le fait de vivre à proximité d'une centrale nucléaire et le risque de développer une leucémie. L'étude montre un risque augmenté chez les enfants vivant dans un rayon de 5 km autour d'une centrale (période 2002-2007 pour 19 centrales en France). Néanmoins, les auteurs précisent que les résultats sont à interpréter avec prudence car ils reposent sur des effectifs faibles et que d'autres facteurs pourraient jouer un rôle.

2.2 Les malformations congénitales

Les causes des malformations congénitales sont nombreuses et complexes. Pour environ la moitié des malformations congénitales, la cause n'est pas claire et est attribuée à des causes multifactorielles, y compris les facteurs environnementaux (Christianson et al., 2006). La plupart des malformations sont dues à des anomalies du matériel génétique - chromosomes et gènes (World Health Organization, 2016). Au-delà des facteurs génétiques connus, dans la littérature, plusieurs facteurs de risque individuels et contextuels de malformations congénitales sont discutés. En ce qui concerne le rayonnement radioactif, il est établi que l'effet tératogène des rayonnements ionisants peut entraîner la mort intra-utérine du fœtus, des défauts structurels, des arriérations mentales, des troubles fonctionnels et des maladies malignes (De Santis et al., 2005; Doll, 1995; Hinz and Kaul, 1987; Shepard and Lemire, 2004).

L'effet tératogène des rayonnements ionisants sur le fœtus dépend principalement de la dose de rayonnement reçue et du stade de la grossesse. Ces associations sont prouvées pour des doses de rayonnement élevées et font l'objet d'une discussion controversée pour des doses de

rayonnement faibles. Une analyse de la prévalence des anomalies autour des centrales nucléaires avec les données du registre France Centre-Est de malformations ne montre pas des effets significatifs (Gautheron et al., 2005). Également, une étude allemande ne révèle pas une prévalence accrue de malformations congénitales dans les régions entourant les centrales nucléaires comparativement à d'autres régions (Queißer-Luft et al., 2011). Des études menées dans le contexte de Tchernobyl montrent toutefois une relation de cause à effet potentielle entre les rayonnements ionisants persistants et certains taux d'anomalies congénitales. Par exemple, des taux significativement plus élevés de microcéphalie, d'anomalies du tube neural et de microphthalmie ont été identifiés dans les régions ukrainiennes où le taux moyen de césium 137 est également plus élevé que dans d'autres régions (Wertelecki et al., 2017).

3 Recensement des données de santé

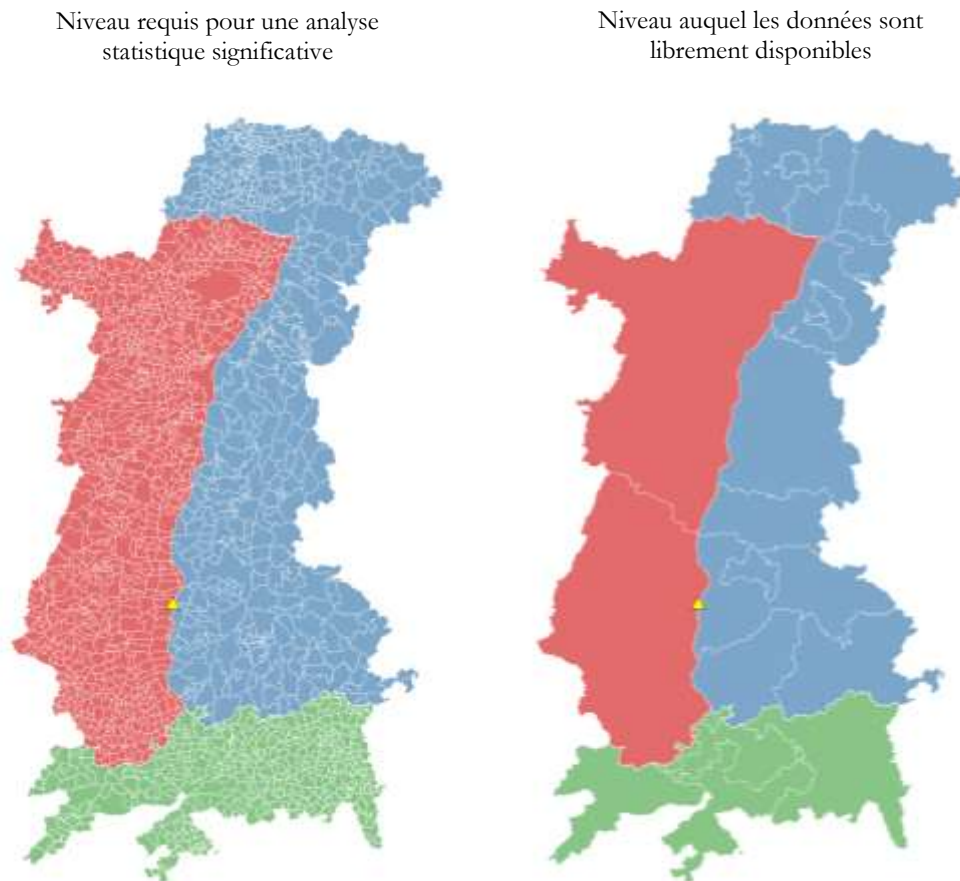
Les données de santé sont des données à caractère personnel et donc soumises à des conditions d'utilisation strictes. La CNIL (Commission nationale de l'informatique et des libertés) définit les données de santé comme « des données relatives à la santé physique ou mentale, passée, présente ou future, 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 »¹. Cette définition large suit le règlement européen sur la protection des données personnelles (RGPD), qui est entré en application le 25 mai 2018. Elle comprend donc les données de santé par nature (maladies, résultats d'examens, handicap, etc.), les données qui lors d'un croisement avec d'autres données deviennent des données de santé (croisement d'une mesure de poids avec une mesure des apports caloriques, etc.), et les données qui deviennent des données de santé en raison de leur destination (utilisation au plan médical).

Compte tenu de cette situation, le recensement des données de santé est associé à des obstacles particuliers. Pour des raisons de protection des libertés et droits fondamentaux des personnes, les données personnelles sur la santé ne sont pas accessibles au public et l'accès est souvent restreint aux projets de recherche qui démontrent un intérêt public. Seules les données agrégées (à différents niveaux temporels et spatiaux), pour lesquelles il n'est plus possible de tirer des conclusions sur les individus, sont proposées ouvertement. A partir de cette situation, deux axes de travail se sont développés dans le cadre du stage : d'une part, les données agrégées en libre accès font l'objet de recensement et collecte, même si, à ce niveau, aucune association avec la CNPE ne peut être établie ; d'autre part, l'accès aux données personnelles issues des statistiques

¹ <https://www.cnil.fr/fr/quest-ce-que-une-donnee-de-sante>, consulté le 08/07/2019.

hospitalières est demandé dans le cadre d'une étude scientifique spécifique dans le contexte de la CNPE Fessenheim.

Figure 1 : Niveaux d'analyse de la région du Rhin Supérieur.



Le triangle jaune représente la CNPE Fessenheim.

La Figure 1 illustre deux niveaux d'analyse différents pour la région du Rhin Supérieur (la partie française en rouge, la partie allemande en bleu et la partie suisse en vert). La carte de gauche montre une carte à l'échelle communale en France, Suisse et Allemagne. A ce niveau, il est possible d'affecter des personnes à des zones géographiques suffisamment fines pour effectuer des analyses statistiques plus profondes (par exemple en tenant compte de la distance entre le domicile et la CNPE Fessenheim). La carte de droite montre une échelle plus agrégée : Kreise ou Verbandsgemeinden en Allemagne, départements en France et cantons en Suisse. C'est à ce niveau que de nombreuses données en accès libre sont proposées. Toutefois, une connexion avec la CNPE Fessenheim ne peut être établie à ce niveau.

Avant que les deux axes de travail ne soient discutés plus en détail, la situation des données de santé en France est présentée, en mettant l'accent sur le contexte régional du Rhin Supérieur. Les sources de données en Allemagne et en Suisse sont également abordées.

3.1 Bases de données santé

Un des objectifs du stage était de recenser les principales bases et jeux de données sur la santé, surtout en France, mais aussi en Allemagne et en Suisse. Ce travail a été effectué par une revue des principaux rapports existants sur les données de santé et une lecture des sites internet des principaux gestionnaires de données.

Globalement, les données de santé peuvent être subdivisées selon leur domaine principal :²

- **santé publique** : état de santé, épidémiologie, environnement, habitudes de vie, inégalités de santé ;
- **offre de soins** : infrastructure, services proposées, personnel, tarifs/honoraires, infrastructure de recherche, informations médicales ;
- **consommation de soins** : activité des établissements de santé, consultations, biologie, transports, médicaments et dispositifs, indemnités et accidents du travail, activité des établissements médicaux-sociaux ;
- **performance et opérations** : performance financière, performance opérationnelle, qualité.

En relation avec le sujet du stage, les recherches se sont concentrées sur le premier type, santé publique et épidémiologie. Les principales bases de données sont présentées ci-dessous. Chaque source de données est décrite à l'aide de plusieurs caractéristiques : le gestionnaire, le type (registre, catalogue, administratif, enquête, etc.), une description des données ainsi que la liberté d'accès (la question de qui peut accéder aux données). Pour certaines sources de données, les conditions d'utilisation doivent également être respectées. Il faut aussi mentionner qu'une même base de données peut fournir aussi bien des données granulaires (données au niveau le plus fin qu'il est possible d'obtenir en fonction de l'origine de la donnée et du système de collecte) que des données agrégées (données obtenues en regroupant des données granulaires selon une caractéristique commune).

² Cette typologie est adaptée de la cartographie des bases de données publiques en santé, <https://www.data.gouv.fr/fr/datasets/cartographie-des-bases-de-donnees-publiques-en-sante/>, consulté le 11/07/2019.

3.1.1 France

Les données de santé en France sont de plus en plus centralisées dans un système unique créé en 2016, le Système National des Données de Santé (SNDS), afin de faciliter l'accès aux données pertinentes pour les études, les recherches et l'évaluation. A ce jour, le SNDS contient et permet de chaîner les données des hôpitaux (PMSI), les données de l'Assurance Maladie (Sniiram) et les données sur les causes médicales de décès (CépiDC). Les données relatives au handicap (en provenance des MDPH, données de la CNSA) seront également intégrées prochainement.

Au-delà du SNDS, il existe une initiative pour mettre en place un hub de données (« Health Data Hub ») afin de faciliter les interactions entre les producteurs des données et les utilisateurs ainsi que l'exploitation des données contenues dans le SNDS. Pour ce faire, des projets innovants dans le domaine de la santé seront sélectionnés et accompagnés. A long terme, le Health Data Hub servira comme point d'entrée pour mieux comprendre les conditions d'accès aux bases de données. L'objectif est de multiplier les possibilités d'exploitation des données en développant des nouvelles techniques, notamment celles fondées sur les méthodes d'intelligence artificielle.

Outre le SNDS, il existe à ce jour plusieurs fournisseurs des données de santé dont les plus pertinents sont présentés dans le Tableau 1.

Tableau 1 : Synthèse des données de santé en France

SNDS – Système National des Données de Santé	<p>Gestionnaire : INDS (Institut National des Données de Santé)</p> <p>Accès : restreint</p> <p>Contient les données de l'Assurance Maladie (base SNIIRAM) ; les données des hôpitaux (base PMSI) ; les causes médicales de décès (base du CépiDC de l'Inserm) - depuis juin 2018; les données relatives au handicap (en provenance des MDPH - données de la CNSA) ; un échantillon de données en provenance des organismes d'Assurance Maladie complémentaire.</p>
Health Data Hub	<p>Gestionnaire : DREES (Direction de la recherche, des études, de l'évaluation et des statistiques)</p> <p>Accès : restreint</p> <p>A pour objectif de favoriser l'utilisation et de multiplier les possibilités d'exploitation des données de santé, en particulier dans les domaines de la recherche, de l'appui au personnel de santé, du pilotage du système de santé, du suivi et de l'information des patients.</p>
Score-Santé	<p>Gestionnaire : FNORS (Fédération Nationale des Observatoires Régionaux de la Santé)</p> <p>Accès : libre et gratuit</p> <p>Met à disposition des informations sur la santé de la population et ses déterminants. Présente des séries historiques, parfois de plus de trente ans où les indicateurs sont déclinés à plusieurs échelons géographiques. Entrées thématiques : population générale, populations spécifiques, déterminants de santé, facteurs de risques, prévention, offre de</p>

	soins, pathologies, recours aux soins, économie de la santé, etc.
Data.Drees	<p>Gestionnaire : DREES (Direction de la recherche, des études, de l'évaluation et des statistiques)</p> <p>Accès : libre et gratuit</p> <p>Permet la mise à disposition de tableaux de données dynamiques abordant les thèmes de la santé et de la protection sociale. Les restitutions proposées mobilisent largement les données du SNDS, notamment celles portant sur le champ de la santé.</p>
PMSI - Programme de médicalisation des systèmes d'information	<p>Gestionnaire : ATIH (Agence technique de l'information sur l'hospitalisation)</p> <p>Accès : restreint</p> <p>L'objectif principal du PMSI est de procéder à l'analyse de l'activité médicale des établissements hospitaliers à des fins d'allocation budgétaire. Dans le cadre du PMSI, tout séjour dans un établissement de santé, public ou privé, fait l'objet d'un recueil systématique et minimal d'informations administratives et médicales. Depuis 2017, l'accès se fait via le SNDS, géré par l'INDS (Institut National des Données de Santé).</p>
ScanSanté	<p>Gestionnaire : ATIH (Agence technique de l'information sur l'hospitalisation)</p> <p>Accès : libre et gratuit</p> <p>Plateforme de restitution des données des établissements de santé, principalement issues du PMSI (activité médicale, médicaments et DMI en sus), mais également des données financières et de coûts, des données de ressources humaines, des indicateurs du bilan social et d'autres sources (HAPI, DREES, INSEE).</p>
Sniiram	<p>Gestionnaire : CNAMTS (Caisse Nationale d'Assurance Maladie et de Travailleurs Salariés)</p> <p>Accès : restreint</p> <p>Base de données nationale dont les objectifs sont de contribuer à une meilleure gestion de l'Assurance Maladie et des politiques de santé, d'améliorer la qualité des soins et de transmettre aux professionnels de santé les informations pertinentes sur leur activité. Les données du Sniiram permettent de réaliser des études explorant les parcours de soins des patients et les modalités de prise en charge des maladies. Depuis 2017, l'accès se fait via le SNDS, géré par l'INDS (Institut National des Données de Santé).</p>
Base des causes de mortalité (CépiDC)	<p>Gestionnaire : INSERM (Institut National de la Santé et de la Recherche Médicale)</p> <p>Accès : restreint aux données granulaires</p> <p>Statistiques sur les causes médicales de décès. Celle-ci est un outil essentiel pour le pilotage des politiques de santé publique, au niveau national et international.</p>
Portail « Epidémiologie – France »	<p>Gestionnaire : INSERM (Institut National de la Santé et de la Recherche Médicale)</p> <p>Accès : libre et gratuit</p> <p>Catalogue en ligne des principales bases de données en santé de source française. Le portail n'entend pas être une plateforme de mise à disposition de données en santé, mais bien un catalogue des sources de données existantes.</p>
Base documentaire « BDSP »	<p>Gestionnaire : EHESP (École des hautes études en santé publique)</p> <p>Accès : libre et gratuit</p> <p>La Banque de données en santé publique (BDSP) est un réseau documentaire d'informations en santé publique et couvre la littérature scientifique et technique française et étrangère en santé publique. Archivée le 01/07/2019.</p>
ORS - Observatoire Régional de la Santé	Gestionnaire : ORS Grand Est

Grand Est	<p>Accès : libre et gratuit</p> <p>Présente l'information sur l'état de santé et sur les besoins des populations dans la région Grand Est, dans le cadre d'une mission d'aide à la décision. La principale mission des ORS consiste en l'observation, l'analyse et la mise à disposition des informations ayant trait à la santé des populations, aux besoins, offres et déterminants de santé sur les territoires.</p>
Les Données du Cancer	<p>Gestionnaire : INCa (Institut national du cancer)</p> <p>Accès : libre aux données agrégées</p> <p>Présente une rubrique dédiée aux données du cancer permettant un accès direct, simple et dynamique aux principaux indicateurs sur le cancer et la cancérologie. Les informations mises à disposition, validées par les producteurs de données, présentent une vision transversale du cancer et de la cancérologie, en traitant de l'épidémiologie, des facteurs de risques, des dépistages, des soins, de la prise en charge du patient.</p>

Outre les sources de données énumérées, il existe d'autres données qui ne sont pas librement accessibles, comme les études et enquêtes réalisées par des entreprises privées ou les données sur les cohortes issues de la recherche épidémiologique.

3.1.2 Allemagne

La « Gesundheitsberichterstattung (GBE) des Bundes » (la surveillance fédérale de la santé) fournit des informations sur la situation sanitaire et les soins de santé de la population en Allemagne. Elle s'appuie sur des descriptions et des analyses fondées sur des données et des indicateurs. Les thèmes de la GBE du gouvernement fédéral sont choisis de telle sorte qu'ils couvrent tous les domaines du service de santé : conditions cadres du système de soins de santé, situation sanitaire, comportement et risques pour la santé, problèmes de santé et maladies, médecine, dépenses de santé, coûts et financement des services de santé.

En ce qui concerne les données en libre accès, le Statistisches Bundesamt (*l'office fédéral de la statistique*) présente des chiffres clés pour différents sujets autour de la santé.³ Les informations couvrent à un niveau agrégé les établissements de soins de santé, les coûts des soins de santé, l'état de santé de la population ainsi que les causes de décès. Pour des comparaisons régionales, la Regionaldatenbank (*base de données régionale*) peut être consultée. Elle offre la possibilité de comparer les données sur la santé au niveau régional - mais souvent uniquement au niveau des Länder. Au-delà des structures fédérales, il existe dans chaque Bundesland des bureaux de statistiques. Pour pouvoir procéder à des comparaisons encore plus approfondies, il faut utiliser les données des offices statistiques des Länder. Pour le Land de Bade-Wurtemberg, l'office régional de la santé fournit également des faits et des chiffres sur la santé et les thèmes liés à la

³ <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/inhalt.html>, consulté le 15/07/2019.

santé au niveau Kreise. En Rhénanie-Palatinat, la Gesundheitsberichterstattung présente des informations sur l'état de santé de la population, dans certains cas jusqu'au niveau du Kreise.

Un système central similaire au SNDS existe avec le Forschungsdatenzentrum⁴ (*centre de données de recherche*) qui a été créé en 2001/2002 pour améliorer l'accès des chercheurs aux données granulaires des statistiques officielles. Le Forschungsdatanzentrum non seulement offre des données sur la santé, mais aussi des données sur d'autres domaines de la statistique officielle, tels que les finances, l'environnement, la démographie, l'éducation etc. Dans le domaine de la santé, des données sur les causes de décès, les statistiques hospitalières et les statistiques sur les soins infirmiers sont incluses. L'accès aux données est soumis à certaines conditions : l'utilisation n'est possible que pour les projets de recherche scientifique et elle est payante (la montant de la redevance dépend du nombre de statistiques demandées, des années d'enquête et des voies d'accès).

Une synthèse des questionnaires des données de santé les plus pertinents en Allemagne est présentée au Tableau 2 : Synthèse des données de santé en Allemagne.

Tableau 2 : Synthèse des données de santé en Allemagne

Gesundheitsbericht erstattung des Bundes	<p>Gestionnaires : RKI (Robert-Koch-Institut), DESTATIS (Statistisches Bundesamt)</p> <p>Accès : restreint</p> <p>Fournit des informations aux politiciens, aux scientifiques, aux acteurs du système de santé et au public intéressé. Les rapports sur la santé présentent et interprètent des aspects liés à la santé qui ont souvent une référence démographique ou régionale. Offre plus de trois milliards de nombres et de ratios. La base de données rassemble des données et des informations sur la santé provenant de plus de 100 sources différentes en un lieu central, y compris de nombreuses enquêtes réalisées par les services statistiques du gouvernement fédéral et des Länder, mais aussi par de nombreuses autres institutions du secteur de la santé.</p>
Gesundheit in Deutschland	<p>Gestionnaire : Statistisches Bundesamt</p> <p>Accès : libre aux données agrégées</p> <p>Présente des chiffres clés sur la santé. Les informations couvrent à un niveau agrégé les établissements de soins de santé, les coûts des soins de santé, l'état de santé de la population ainsi que les causes de décès.</p>
Regionaldatenbank	<p>Gestionnaire : Statistische Ämter des Bundes und der Länder</p> <p>Accès : libre</p> <p>La base de données régionale Allemagne fournit des résultats détaillés (Gemeinde, Kreis, Bundesland) des statistiques officielles fédérales et régionales pour des sujets différents. Dans le domaine de la santé, des données des hôpitaux et des installations de prévention ou de réadaptation ainsi que des diagnostics des patients hospitalisés sont présentées.</p>

⁴ <https://www.forschungsdatenzentrum.de/de>, consulté le 15/07/2019.

Forschungsdatenzentrum	<p>Gestionnaire : Statistische Ämter des Bundes und der Länder</p> <p>Accès : restreint</p> <p>Offre des données sur la santé, mais aussi des données sur d'autres domaines de la statistique officielle, tels que les finances, l'environnement, la démographie, l'éducation etc. Dans le domaine de la santé, des données sur les causes de décès, les statistiques hospitalières et les statistiques sur les soins infirmiers sont incluses.</p>
Gesundheit in Baden-Württemberg	<p>Gestionnaire : Statistisches Landesamt Baden-Württemberg</p> <p>Accès : libre aux données agrégées, accès payant aux données plus fines</p> <p>Fournit au public des données complètes sur le développement économique et social de la région et offre des conseils individuels sur la façon de traiter les données. Dans le domaine de la santé, des données agrégées sur les médecins et les institutions, les traitements et les taux de maladie, les causes de décès ainsi que les personnes handicapées et dépendantes sont proposées.</p>
Gesundheitsberichterstattung Rheinland-Pfalz	<p>Gestionnaire : Statistisches Landesamt Rheinland-Pfalz</p> <p>Accès : libre aux données agrégées, accès payant aux données plus fines</p> <p>Fournit des données statistiques de base sélectionnées à partir du système de surveillance de la santé de la région. Le service s'adresse en premier lieu au public spécialisé, mais il offre également un large éventail d'informations à toute personne intéressée par les questions de santé.</p>
Gesundheitsatlas Baden-Württemberg	<p>Gestionnaire : Landesgesundheitsamt Baden-Württemberg</p> <p>Accès : libre et gratuit</p> <p>Chiffres sur la santé et les sujets liés à la santé au niveau des Länder et des Kreise, disponibles sous forme de tableaux, de cartes et de rapports. Permet de consulter les chiffres actuels sur des thèmes relatifs à la santé de la population du Bade-Wurtemberg, y compris des données sur la structure de la population, la fréquence des maladies et les soins médicaux.</p>
Zentrum für Krebsregisterdaten	<p>Gestionnaire : Robert-Koch-Institut (RKI)</p> <p>Accès : restreint</p> <p>Regroupe les données des registres épidémiologiques des cancers des Länder au niveau fédéral. L'objectif principal est de fournir à la recherche scientifique des données fiables sur les cancers en Allemagne et de décrire les événements cancéreux en Allemagne au public intéressé.</p>
Krebsregister Baden-Württemberg	<p>Gestionnaire : Deutschen Krebsforschungszentrum Heidelberg</p> <p>Accès: restreint</p> <p>Le registre épidémiologique des cancers du Bade-Wurtemberg décrit le paysage cancéreux du Bade-Wurtemberg et soutient la recherche sur les causes du cancer ainsi que l'assurance qualité pour la détection précoce du cancer.</p>
Krebsregister Rheinland-Pfalz	<p>Gestionnaire : Krebsregister Rheinland-Pfalz</p> <p>Accès: restreint</p> <p>Le registre clinique et épidémiologique des cancers de Rhénanie-Palatinat recueille des données sur l'apparition, le traitement et l'évolution des maladies tumorales chez les patients hospitalisés et les patients externes.</p>

3.1.3 Suisse

Les données de santé en Suisse sont principalement accessibles via l'Office fédéral de la statistique. Il présente des chiffres clés pour différents sujets autour de la santé et permet de faire des comparaisons temporelles et régionales. Similairement à l'Allemagne, au-delà des structures fédérales, il existe des bureaux de statistiques cantonales qui fournissent souvent des données plus précises. Outre les bureaux de statistiques, il existe également des observatoires de la santé publique, comme l'Office fédéral de la santé publique et l'Observatoire suisse de la santé.

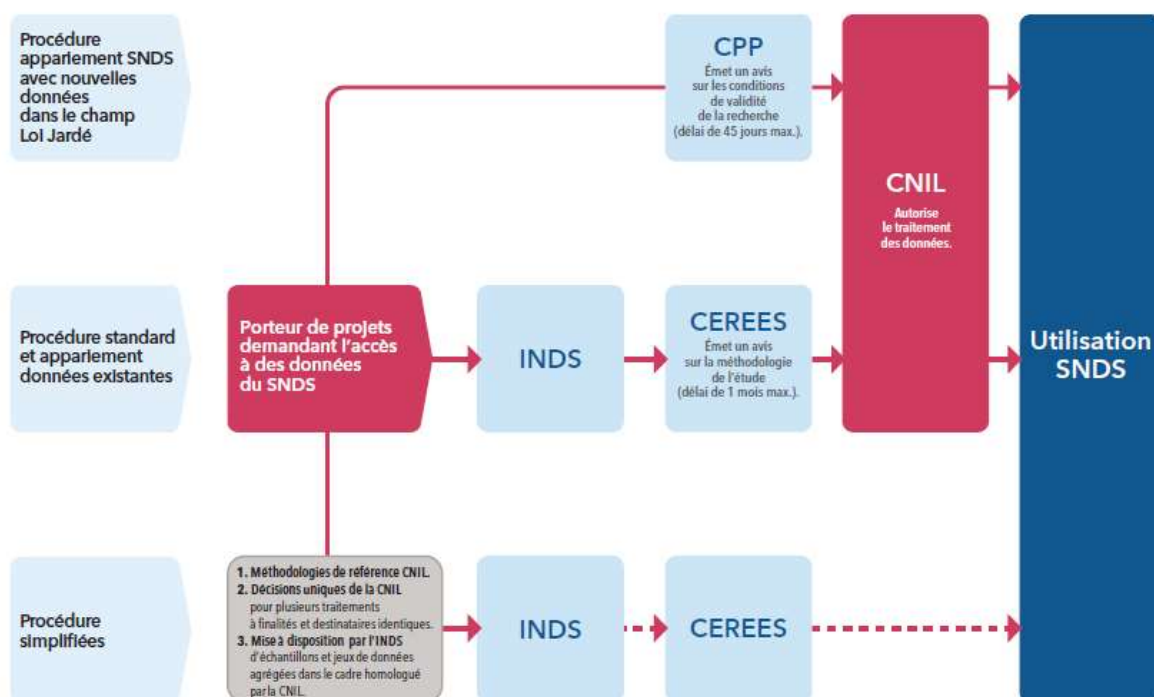
Tableau 3 : Synthèse des données de santé en Suisse

Office fédéral de la statistique	<p>Gestionnaire : Office fédérale de la statistique</p> <p>Accès : libre aux données agrégées</p> <p>L'Office fédéral de la statistique est un office fédéral suisse responsable des services de la statistique. Il est le centre national de compétences et de services de la statistique et fait fonction d'observateur de secteurs clés de l'État et de la société, de l'économie et de l'environnement. Sur le thème de la santé sont proposées des statistiques sur les déterminants de la santé, l'état de santé, les soins de santé et le coût/financement.</p>
Office fédéral de la santé publique (OFSP)	<p>Gestionnaire : Département fédéral de l'intérieur</p> <p>Accès : libre aux données agrégées</p> <p>L'Office fédéral de la santé publique (OFSP) fait partie du Département fédéral de l'intérieur. De concert avec les cantons, il assume la responsabilité des domaines touchant à la santé publique ainsi que la mise en œuvre de la politique sanitaire. Autorité à vocation nationale, il représente les intérêts sanitaires de la Suisse dans les organisations internationales et auprès d'autres Etats.</p>
Observatoire suisse de la santé (Obsan)	<p>Gestionnaires : Confédération et cantons</p> <p>Accès : libre aux données agrégées</p> <p>L'Obsan traite les six thèmes de la santé suivants: « Santé de la population », « Santé mentale », « Âge et soins de longue durée », « Système de santé », « Personnel de santé » et « Coûts et financement ». L'Obsan met ses résultats à la disposition de la Confédération, des cantons et d'autres institutions de la santé. Il fournit également des conseils personnalisés et procède à des analyses sur mesure pour ces partenaires.</p>
Statistisches Amt der Stadt Basel	<p>Gestionnaire : Präsidialdepartement des Kantons Basel-Stadt</p> <p>Accès : libre</p> <p>Le Bureau de statistiques de Bâle offre une vue d'ensemble des faits et développements importants dans le secteur de la santé et permet de tirer des conclusions sur l'état de santé de la population. Les questions concernant les décès et leurs causes, les soins médicaux, la durée des soins hospitaliers ou les taux de couverture vaccinale pour certaines maladies sont traitées.</p>
Zahlenfenster	<p>Gestionnaire : Office de la statistique de Bâle-Campagne</p> <p>Accès : libre</p> <p>Le <i>Zahlenfenster</i> contient des données cantonales sur plus de vingt domaines de la statistique publique, le sujet de santé inclus. Des graphiques et des cartes interactifs s'ajoutent progressivement au service.</p>

3.2 Demande d'accès aux données du PMSI

Afin de pouvoir effectuer des analyses plus approfondies, l'accès aux données hospitalières du PMSI a été demandé dans le cadre du stage. Comme les données PMSI sont intégrées dans le SNDS depuis 2017, l'accès est réglementé de façon centralisée par le gestionnaire du SNDS, l'Institut National des Données de Santé (INDS). Il existe plusieurs procédures pour obtenir un accès aux données SNDS (cf. Figure 2).

Figure 2 : Schématisation du circuit des demandes d'accès au SNDS



Pour la procédure standard (centrée sur la Figure 2), les demandes d'accès sont à déposer auprès de l'INDS. L'INDS assure le traitement des demandes et accompagne les utilisateurs du SNDS dans la prise en main des données. Il est en lien avec le Comité d'Expertise pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé (CEREES). Le CEREES est chargé d'examiner les demandes du point de vue méthodologique, afin de fournir un avis à la

Commission Nationale de l'Informatique et des Libertés (CNIL) sur la cohérence entre la finalité de l'étude proposée et la méthodologie présentée.

Pour déposer un dossier auprès de l'INDS selon la procédure standard, plusieurs documents sont requis : un résumé de l'étude, un protocole scientifique précisant la question de recherche et la méthodologie, des déclarations d'intérêts du responsable de traitement et du responsable de la recherche ainsi qu'une demande d'autorisation pour la CNIL. A la suite du dépôt du dossier, l'INDS vérifie sa complétude et transmet le document au CEREES pour évaluation et avis. Si l'avis du CEREES est favorable, l'INDS dépose le dossier automatiquement à la CNIL pour autorisation. Si l'avis du CEREES est réservé ou défavorable, le demandeur peut compléter le dossier en vue d'un nouvel examen par le comité.

En plus de la procédure standard, des procédures simplifiées sont proposées (Figure 2). Par exemple, la CNIL a défini plusieurs méthodologies de référence qui s'appliquent à des types d'études spécifiques. Un engagement par un demandeur à se conformer à une méthodologie de référence vaut autorisation de la CNIL et une déclaration simple du projet de recherche auprès de l'INDS est suffisante pour commencer un projet. Il existe six méthodologies de référence, parmi d'autres la MR-005, relative au traitement de données nécessitant l'accès par les établissements de santé et des fédérations aux données du PMSI.⁵ Si un projet vise à utiliser non seulement les données existantes du SNDS, mais aussi le recueil d'autres données via un questionnaire ou un examen, la recherche ne peut être mise en œuvre qu'après avis favorable d'un comité de protection des personnes (CPP) (ci-dessus sur la Figure 2).

Dans le cadre du stage, un dossier a été élaboré selon la procédure standard. La préparation du dossier a été soutenue par Sarah Piquette-Muramatsu, déléguée à la protection des données au sein de l'Université de Strasbourg. Le protocole scientifique traite des malformations congénitales dans le Rhin Supérieur. La question de recherche est de savoir si plus d'enfants atteints de malformations congénitales naissent à proximité de la CNPE Fessenheim qu'à distance du site. L'objectif est aussi d'examiner s'il existe un effet négatif lié à la distance (hypothèse de travail : Plus le domicile est près du réacteur, plus le risque qu'un enfant naisse avec une malformation est grand). A partir des données sanitaires et environnementales collectées sur l'Alsace, des analyses sont prévues pour mettre en relation la distribution spatiale des taux estimés à l'échelle du code PMSI et la proximité résidentielle aux sites nucléaires, tout en prenant en compte les autres

⁵ Les différentes méthodologies peuvent être consultés sur le site de la CNIL : <https://www.cnil.fr/fr/sante> (consulté le 15/07/2019).

nuisances environnementales qui caractérisent l'environnement de vie des femmes enceintes. Le dossier a été déposé le 22 mai 2019, un avis réservé du CEREES a été reçu le 22 juin 2019 (suite à la session du 13 juin 2019). Le CEREES recommande de se rapprocher d'équipes spécialisées comme celles des réseaux régionaux de périnatalogies et de Santé Publique Française. Il a été décidé de prendre contact avec des réseaux de périnatalogie et de malformations, avant que le dossier ne soit transmis à la CNIL.

4 Analyse des données

4.1 Les données ouvertes et développement d'un outil de visualisation des données

Pour la région du Rhin supérieur, il existe des données en accès libre sur différents indicateurs comme l'état de santé de la population et l'offre de soins. Toutefois, les données à un niveau fin ne sont généralement spécifiques qu'à un pays. Afin de permettre des comparaisons au-delà des frontières nationales, il est nécessaire de rassembler des données provenant de différentes sources. Dans le cadre de ce stage, un outil de visualisation a été développé afin de visualiser les données en accès libre. Le but était de pouvoir comparer interactivement différents indicateurs dans les zones du Rhin Supérieur sur un niveau assez fin (départements en France, *Kreise* en Allemagne, cantons en Suisse). Pour la région du Rhin Supérieur, il n'existe jusqu'à présent pas d'applications qui rendent possible des comparaisons transfrontalières à ce niveau géographique.

L'outil a été développé en R en utilisant le package *Shiny*⁶, un projet open source développé par RStudio. *Shiny* est un package qui permet de créer des applications dynamiques et interactives pour le web, tout en utilisant R (sans coder en HTML, CSS ou JavaScript). En *Shiny*, les applications consistent en deux éléments : une partie *ui* (pour *user interface*) et une partie *server*. La partie *ui* définit tous les éléments graphiques de l'interface de l'application, c'est-à-dire tout ce qui est visible à l'utilisateur : la mise en page (éléments de division, etc.) et des composantes à afficher qui communiquent avec la partie serveur (graphiques, tableaux, etc.). Dans la partie *server* sont appelées les composantes de l'interface qui sont créés suite aux choix de l'utilisateur. C'est ici qu'est entré le code R à exécuter pour créer des visualisations ou des tableaux.

Un autre package principal pour le développement de l'outil de visualisation est le package *leaflet*⁷. Originellement, *leaflet* est une bibliothèque JavaScript libre pour la cartographie interactive, qui est

⁶ <https://shiny.rstudio.com>, consulté le 10/07/2019.

⁷ <https://rstudio.github.io/leaflet/>, consulté le 10/07/2019.

utilisée par *OpenStreetmap*, entre autres. Une fois qu'une carte créée avec *leaflet*, il est possible de zoomer et de se déplacer sur la carte, et des éléments comme des unités géographiques ou des cercles sont cliquables. Le package *leaflet* adapté pour R permet de réaliser des cartes dynamiques tout en codant en R, avec la diversité offerte par le package JavaScript original. De plus, le package *leaflet* comprend des fonctionnalités pratiques pour s'intégrer avec les applications *Shiny*.

R dispose également de packages puissants dans le domaine des systèmes d'information géographique. Le quasi standard moderne est le package *sf*⁸ (pour *simple features*).⁹ Avec *sf*, toutes les opérations classiques sur des données spatialisées peuvent être réalisées, comme lire et explorer des données géographiques, manipuler des attributs (création, sélection) et appliquer des traitements géomatiques (intersection, jointure, calcul de surface). Les objets *sf* ne sont que des *data frames* qui sont des collections d'objets spatiaux. Chaque ligne est un objet spatial (par exemple un point ou un polygone), qui peut être associé à des données (par exemple un code identifiant) et une variable *geometry* qui contient les coordonnées.

Le développement de l'outil de visualisation se divise en quatre étapes :

1. récupérer des données à visualiser,
2. récupérer et traiter des *shapefiles* pour chaque pays,
3. joindre les données à visualiser avec le *shapefile* combiné,
4. développer l'application en *Shiny*.

Ces étapes sont décrites ci-dessous plus en détail.

Récupérer des données à visualiser

Avant de développer l'application elle-même, les différents ensembles de données provenant de différentes sources en différents formats (principalement .csv et .xlsx) sont fusionnés. Pour la plupart des indicateurs, les données se trouvent dans des fichiers séparés et nécessitent donc un traitement d'harmonisation. Pour pouvoir afficher les données sur une carte, un géoréférencement sous forme de codes géographiques est nécessaire (par exemple le numéro du département). Le jeu de données final inclut des variables sur des indicateurs démographiques (population, densité), ainsi que des indicateurs de santé publique (tels que la mortalité, l'espérance de vie, ...). Tableau 4 présente une description des données sélectionnées et de leurs sources.

⁸ <https://r-spatial.github.io/sf/>, consulté le 10/07/2019.

⁹ Traditionnellement, le package *sp*, qui fournissait des structures de données spatiales et quelques fonctions utilitaires était essentiel pour presque toutes les analyses spatiales en R. *sf* a à peu près les mêmes fonctionnalités que *sp*, mais est beaucoup plus pratique et flexible, grâce à la représentation des données spatiales en tant qu'objets R natifs (*list*, *matrix*, *vector*).

Tableau 4 : Indicateurs sélectionnés et leurs sources

	France	Suisse	Bade-Wurtemberg	Rhénanie-Palatinat
Démographie				
Population	INSEE	Office fédéral de la statistique	Regionaldatenbank	Regionaldatenbank
Densité	INSEE	Office fédéral de la statistique	Regionaldatenbank	Regionaldatenbank
Taux de chômage	INSEE	Office fédéral de la statistique	Regionaldatenbank	Regionaldatenbank
État de santé				
Mortalité	Score Santé	Office fédéral de la statistique	Gesundheitsatlas Baden-Württemberg	-
Espérance de vie	Score Santé	Office fédéral de la statistique	Gesundheitsatlas Baden-Württemberg	-
Offre de soins				
Nombre de lits	Eco-Santé	Office fédéral de la statistique	Landatlas (Ministère fédéral de l'Agriculture)	Landatlas (Ministère fédéral de l'Agriculture)

Récupérer et traiter des *shapefiles* pour les différents pays

Les données spatialisées sont fournies sous la forme de *shapefiles* provenant de différentes sources pour les différents pays.¹⁰ Les *shapefiles* couvrant plus qu'un pays ne sont pas adaptés pour des niveaux plus fins. Pour chaque pays, un *shapefile* est utilisé au niveau administratif le plus fin disponible afin de conserver la flexibilité nécessaire à la poursuite des analyses. Le package *sf* sert à filtrer les données pour les régions requises (Alsace, Bade, la partie sud de Rhénanie-Palatinat, cinq cantons en Suisse), et à les agréger au niveau approprié, puis à les fusionner.

Combiner les données à visualiser avec le *shapefile*

Enfin, le jeu de données contenant le *shapefile* est combiné avec le jeu de données contenant les indicateurs à afficher sur la carte. Dans chacun des deux ensembles de données, le géocode a été harmonisé afin que les données puissent être fusionnées par une simple jointure.

¹⁰ Allemagne (Gemeinden) :

http://www.geodatenzentrum.de/geodaten/gdz_rahmen.gdz_div?gdz_spr=deu&gdz_akt_zeile=5&gdz_anz_zeile=1&gdz_unt_zeile=0&gdz_user_id=0,

France (Iris) : <http://professionnels.ign.fr/contoursiris>

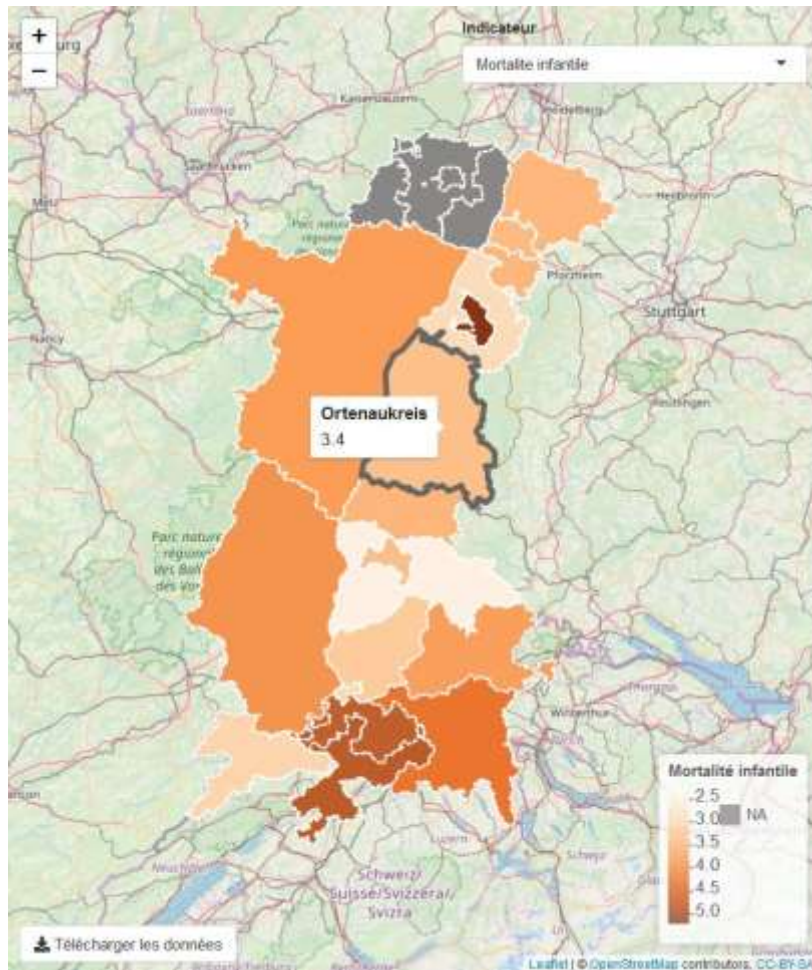
Suisse (Communes) : <https://opendata.swiss/fr/dataset/swissboundaries3d-gemeindegrenzen>

Développer l'application en *Shiny*

Dans l'application *Shiny*, seule la carte interactive est créée et modifiée après saisie par l'utilisateur. Le jeu de données final est simplement chargé dans l'application sans traitement supplémentaire pour éviter des temps de chargement trop longs. La plupart des objets *Shiny* sont incorporés dans une application en incluant une fonction pour définir la sortie (par exemple *plotOutput*) dans la partie interface, et en utilisant une fonction de rendu (par exemple *renderPlot*) dans la partie serveur. Similairement, pour incorporer les cartes de *leaflet* dans *Shiny*, dans la partie interface, la fonction *leafletOutput* est appelée, et du côté serveur un appel *renderLeaflet* est assigné à la sortie. A l'intérieur de l'expression *renderLeaflet*, un objet *Leafletmap* est retourné. Les entrées de l'utilisateur entraînent que l'ensemble de la carte est complètement redessinée ; elles réinitialisent la position de la carte et le niveau de zoom. Pour avoir un contrôle plus fin sur une carte qui est déjà en cours d'exécution sur la page (p.ex. changer les couleurs des polygones selon un indicateur choisi), la fonction *leafletProxy* est utilisée. Ainsi, les éléments statiques de la carte sont créés avec *leaflet*, et les éléments dynamiques sont gérés avec *leafletProxy*.

Figure 3 illustre l'outil de visualisation. Le fond de carte est chargé par *OpenStreetMap*. Les polygones sont basés sur les données préparées qui sont chargées dans l'application. En haut à droite, on peut sélectionner un indicateur à visualiser sur la carte. La légende en bas à droite est automatiquement mise à jour. Lorsqu'on passe la souris sur une région, le nom et la valeur correspondante de l'indicateur sélectionné apparaissent. En cliquant sur « Télécharger les données » l'ensemble de données sous-jacent peut être téléchargé en format csv.

Figure 3 : Capture d'écran de l'outil de visualisation



Cette version de l'application peut bien naturellement être étendue à d'autres indicateurs, selon le sujet à visualiser. Par exemple, il pourrait être utile d'afficher aussi des facteurs environnementaux dans le contexte de santé publique. Il est également possible d'ajouter d'autres sélecteurs, par exemple pour permettre de choisir différentes échelles géographiques selon les échelles disponibles dans l'ensemble de données. Aussi, on pourrait ajouter un *slider* pour sélectionner l'année ou le mois afin de permettre des comparaisons temporelles. Enfin, d'autres éléments géographiques peuvent être ajoutés à la carte, par exemple des points indiquant des sites polluants.

4.2 Les données PMSI avec la MR-005

En tant que stagiaire dans le projet « Juxta Rhenum », j'ai bénéficié d'une coopération avec les Hôpitaux Universitaires de Strasbourg qui disposent d'un accès à la base nationale du PMSI (limité sur les dix dernières années) selon la Méthodologie de référence MR-005 (études nécessitant l'accès aux données du PMSI par les établissements de santé et les fédérations

hospitalières). Afin d'étudier la faisabilité du projet lié aux données SNDS, des données du PMSI ont été obtenues via la MR-005 et de premières analyses descriptives menées. Les données correspondantes sont mises à disposition sur la plateforme sécurisée de l'Agence technique de l'information sur l'hospitalisation (ATIH). La plateforme permet de tracer tous les accès et traitements réalisés ; l'authentification se fait via un jeton (*token*). Sur la plateforme, différents programmes sont offerts pour récupérer et analyser les données. Dans le cadre de ce travail, le logiciel SAS est utilisé pour interroger et extraire des données de la base de données (en utilisant des requêtes SQL), et le logiciel R pour traiter et analyser les données. L'accès aux données et de premiers résultats sont présentés ci-dessous.

4.2.1 Requêtes

Le but général était d'extraire l'ensemble des séjours concernant la naissance d'enfants nés-vivants ou mort-nés présentant au moins une malformation. Suivant le protocole scientifique préparé pour le dossier SNDS, quatre requêtes ont été créées pour obtenir les données nécessaires. La base de données consiste en plusieurs tables différentes pour chaque année (respectivement 2010-2018). Pour les requêtes, seule la table « fixe », contenant des informations essentielles de la partie fixe des RSA, et la table « diag », contenant tous les diagnostics du séjour (DP, DR ou DAS), étaient requises.¹¹ Dans le PMSI, la variable « ident » sert de clé de jointure entre les différentes tables. Le numéro « anonyme » est censé être unique pour chaque individu et permet le chaînage des séjours d'un même patient à travers les années, quel que soit le lieu de prise en charge hospitalière. Il est créé à partir du numéro de sécurité sociale, du sexe et de la date de naissance. La variable « ano_retour » signale un problème de chaînage (p.ex. format incorrect ou valeur manquante). Tous les séjours avec un mauvais code « ano_retour » ne peuvent pas être traités.

Dans une première requête (requête 1), les séjours pour naissance avec malformation de mères résidant dans les départements du Haut-Rhin et du Bas-Rhin sont extraits. Les naissances avec malformations sont les séjours mentionnant une naissance (Z38.- ou P95) et contiennent un code de malformation congénitale (Q00.- à Q99). Le code de résidence de ces séjours doit être 67- ou 68-. En plus des variables identifiantes (ident, anonyme, ano_retour), sont récupérés le diagnostic

¹¹ Tout séjour effectué dans la partie court-séjour d'un hôpital fait l'objet d'un résumé de sortie standardisé (RSS). L'anonymisation du RSS a pour résultat la production d'un résumé de sortie anonyme (RSA). Le diagnostic principal (DP) est le problème de santé qui a motivé l'admission du patient dans l'unité médicale (UM), déterminé à la sortie de l'UM par le médecin responsable du patient. Le diagnostic relié (DR) permet de préciser le contexte pathologique lorsque le DP est codé avec un code en Z dans la CIM-10. Un diagnostic associé significatif (DAS) est un problème de santé distinct supplémentaire, pris en charge au cours du séjour (comorbidité, complication...).

(diag), le mois du séjour (mois), l'année du séjour (annee) et le code géographique du domicile (codeGeo).

La deuxième requête (requête 2) extrait les séjours pour interventions sur des malformations, dont le code de résidence peut-être n'importe où en France mais qui sont liées à un séjour concernant une naissance dont le code de résidence est dans le Haut-Rhin ou le Bas-Rhin. L'hypothèse sous-jacente à cette méthode est que les règles de codage PMSI peuvent entraîner le sous-codage de malformations jugées minimales au moment de la naissance. Les séjours concernant une malformation contiennent un code Q00.- à Q99 en DP, DR ou DAS et sont chaînés avec un séjour de naissance (code Z38.- ou P95 en DP, DR ou DAS) dont le code géographique de résidence est 67- ou 68-.

A ce corpus de séjours sont ajoutées avec une autre requête (requête 3) les interruptions médicales de grossesses (IMG) avant 22 semaines (les interruptions au-delà de 22 semaines sont des enfants mort-nés dans la première requête), repérées par des séjours dont le DP est O04.- et le DAS O35.- (pour ne retenir que les causes fœtales). Les ré-hospitalisations pour complications d'IMG sont exclues car ne mentionnant pas de O35.-.

Les séjours extraits par les trois différentes requêtes sont rassemblés afin de créer un jeu de données contenant tous les séjours pertinents pour l'analyse. Une dernière requête (requête 4) sert à extraire le nombre de naissances par unité géographique et année pour pouvoir mettre en relation le nombre de malformations.

4.2.2 Résultats

Cas des mauvais ano_retour et des naissances multiples

Le nombre de patients avec une clé de chaînage mal générée (ano_retour différent de 0) s'élève à 317. Ces patients sont exclus de l'analyse. Un autre groupe dont il faut tenir compte avant l'analyse est celui des naissances multiples. Pour des naissances multiples de même sexe, il n'est pas possible de distinguer les patients car le même code anonyme est attribué (même numéro d'assuré social, même date de naissance, même sexe). Cette problématique implique des approches différentes pour les requêtes 1 et 2.

Dans la requête 1, deux séjours avec des codes « ident » différents pour le même numéro « anonyme », avec code Z38.3-8 indiquent une naissance multiple. On pourrait soit les éliminer, soit les garder, et dédoubler le numéro anonyme en créant une variable « anonyme 2 », qui

permette de séparer ces séjours en autant de cas que nécessaire pour bien compter le nombre de patients différents.

Dans la requête 2, il existe des cas où un séjour avec une malformation présente ultérieurement dans le PMSI est rattaché à deux séjours de naissance. Ainsi, certains séjours avec malformations sont reliés à plusieurs naissances (même numéro anonyme pour deux variables identifiées différentes). Dans ce contexte, il y a deux solutions à traiter différemment :

1. Les deux séjours naissances n'ont aucun code de malformation.

Solution : Conserver un seul cas. On ne sait pas lequel des deux enfants est touché, mais on sait qu'il y a un cas, dont on connaît le code Q. C'est bien une nouvelle identification de cas, car les naissances avec code Q sont dans la requête 1.

2. Au moins un des deux séjours naissances correspondant à déjà un code Q de malformation (et ce(s) séjour(s) est(sont) normalement identifié(s) dans la requête 1), donc on ne sait pas si le cas « à distance de la naissance » correspond à un cas déjà identifié à la naissance (dont il complèterait l'information avec des codes Q peut-être différents), ou à l'autre jumeau également identifié à la naissance, ou à un cas non identifié à la naissance (situation où un seul des deux séjours naissances a un code Q, et l'autre n'en a aucun : dans ce cas il s'agirait d'un nouveau cas identifié grâce à la requête 2).

Solution : Supprimer ce cas et ne pas le compter comme un cas supplémentaire, ni comme un complément d'information d'un cas déjà identifié dans la requête 1.

Il faut faire attention à ne pas identifier les cas problématiques de la requête 2 avec la variable « anonyme 2 », car cette dernière sépare les jumeaux. Dans le cadre d'une analyse plus approfondie, ces cas devraient être examinés sous la forme des solutions présentées ci-dessus. Pour l'analyse préliminaire dans le cadre de ce rapport de stage, cependant, tous les cas des naissances multiples sont éliminés.

Enfin, les cas IMG pour d'autres raisons que des malformations congénitales sont à exclure du jeu de données. Il s'agit de 11 cas pour maladie héréditaire du fœtus (O352), 7 cas pour lésions fœtales dues à d'autres actes à visée diagnostique et thérapeutique (O357), 1 cas pour lésions fœtales résultant d'une maladie virale de la mère (O353), 1 cas pour lésions fœtales médicamenteuses (O355) et 1 cas pour lésions fœtales dues à une irradiation (O356).

Chiffres clés du jeu de données

Le nombre total des patients est 14.363, dont 13.835 naissances uniques, 486 jumeaux, 36 triplés et 6 sextuplés. Alors, 3,7% de tous les patients sont issus d'une naissance multiple et ne sont pas inclus dans les analyses suivantes. Après l'exclusion des 21 cas IMG sans rapport aux malformations congénitales, le jeu de données d'analyse contient 13.814 patients. Le nombre de patients différents repérés par chaque requête est présenté dans le Tableau 5 :

Tableau 5 : Répartition des patients par requête

Requête	Nombre de patients	Pourcentage [%]
1	9.303	67,3
2	3.746	27,1
3	765	5,5
Total	13.814	100,0

La plupart des patients dans l'ensemble de données a reçu un diagnostic de malformation congénitale à la naissance et éventuellement aussi ultérieurement (n = 9.303). Un tiers de tous les patients ont été traités pour une malformation à un moment donné sans qu'une malformation ne soit détectée à la naissance (n = 3.746). Enfin, on dénombre 765 cas d'IMG en Alsace sur la période 2010-2018.

Parmi les patients nés avec une malformation (requête 1), 1.477 sont hospitalisés avec un code de malformation dans les 0 à 8 années qui suivent la naissance (NB : aucun recul pour les naissances en décembre 2018). Il s'agit de cas qui sont nés avec une malformation et qui ont été hospitalisés plus tard avec un code de malformation dans le codage PMSI (que ce soi(en)t le(s) même(s) code(s) ou non).

Répartition des codes CIM/malformations congénitales

Parmi tous les patients inclus dans le jeu de données sont codés 18.530 malformations, dont 502 malformations différentes. La majorité des patients (81,9%) souffre de seulement une forme de malformation congénitale. Seulement 1,3% de tous les patients identifiés ont plus que 5 malformations, mais il existe aussi un cas avec 21 malformations différentes.

Les malformations les plus fréquentes sont l'ankyloglossie (Q381, 12,2% parmi tous les diagnostics), des anomalies morphologiques congénitales du pied en varus (Q663, 4,0% parmi tous les diagnostics) et des testicules en situation anormale (Q530, 3,7% parmi tous les

diagnostics). Parmi les malformations les moins fréquentes se trouvent 75 formes qui n'apparaissent qu'une seule fois dans l'ensemble des données.

La répartition des malformations entre différents sous-groupes de malformations est présentée dans Tableau 6 :

Tableau 6 : Répartition des malformations par catégorie

Code CIM	Nombre de diagnostics	% diagnostics
Q00-Q07 - Malformations congénitales du système nerveux	652	3,5
Q10-Q18 - Malformations congénitales de l'œil, de l'oreille, de la face et du cou	915	4,9
Q20-Q28 - Malformations congénitales de l'appareil circulatoire	3.097	16,7
Q30-Q34 - Malformations congénitales de l'appareil respiratoire	358	1,9
Q35-Q37 - Fente labiale et fente palatine	399	2,2
Q38-Q45 - Autres malformations congénitales de l'appareil digestif	2.977	16,1
Q50-Q56 - Malformations congénitales des organes génitaux	2.363	12,8
Q60-Q64 - Malformations congénitales de l'appareil urinaire	1.421	7,7
Q65-Q79 - Malformations congénitales du système ostéo-articulaire et des muscles	4.346	23,5
Q80-Q89 - Autres malformations congénitales	1.002	5,4
Q90-Q99 - Anomalies chromosomiques, non classées ailleurs	775	4,2
Anomalies et lésions fœtales autres et sans précision (IMG)	225	1,2
Total	18.530	100,0

Les sous-groupes avec le plus de diagnostics sont les malformations congénitales du système ostéo-articulaire et des muscles (23,5%) et d'autres malformations congénitales de l'appareil digestif (16,1%).

Nombre de naissances et patients par année

En total, en Alsace, il y avait 188.795 naissances et 13.814 patients avec au moins une malformation (malformation diagnostiquée à la naissance, ultérieurement ou IMG liée à une malformation) entre 2010 et 2018. Le taux de malformations s'élève à 7,3% et reste relativement

stable le long de la période couverte (avec une exception pour 2018, la dernière année incluse dans la base de données) (Tableau 7).

Tableau 7 : Nombre de naissances et de cas par année

Année	Nombre de naissances	Nombre de patients	% naissances avec malformation
2010	21.703	1.606	7,4
2011	21.219	1.602	7,6
2012	21.422	1.643	7,7
2013	21.459	1.563	7,3
2014	21.727	1.626	7,5
2015	20.885	1.631	7,8
2016	20.519	1.535	7,5
2017	20.061	1.446	7,2
2018	19.800	1.162	5,9
Total	188.795	13.814	7,3

Nombre de patients par rapport à la distance de la CNPE Fessenheim

Pour définir les unités géographiques étant proche de la CNPE Fessenheim, il s'agit de déterminer le centroïde de chaque unité géographique et de calculer la distance entre les centroïdes et la CNPE Fessenheim.¹² Ensuite, les unités sont regroupées selon quatre niveaux de proximité (moins de 10 km, entre 10 et 15 km, entre 15 et 20 km et plus que 20 km). Il existe seulement une unité géographique dont le centroïde est à moins de 10 km (Fessenheim), trois unités sont dans un rayon de 15 km (Biesheim, Reguisheim, Ottmarsheim) et quatre autres unités dans un rayon de 20 km (Ste Croix en Plaine, Sausheim, Ensisheim, Andolsheim).¹³

Le Tableau 8 décrit la répartition des naissances, des malformations et certaines caractéristiques sociodémographiques. Une première analyse présentée ci-dessus utilise le rapport du nombre de cas observés au nombre attendu, sans ajustements sur des facteurs de confusion. En prenant comme référence toutes les unités géographiques incluses dans l'analyse, on calcule des nombres attendus. Il est à noter que dans des analyses approfondies, les nombres attendus devraient être

¹² Deux codes géographiques dans les données PMSI sont incorrects (« 67999 » et « 68999 ») ; ils n'apparaissent pas dans la table de correspondance entre les codes postaux et les codes PMSI de l'ATIH (<https://www.atih.sante.fr/mise-jour-2018-de-la-liste-de-correspondance-codes-postaux-codes-geographiques-pmsi>, consulté le 22/07/2019). Les 10 cas correspondants doivent être exclus de l'analyse.

¹³ La détermination de la distance entre les unités géographiques et la CNPE Fessenheim au moyen de centroïdes des polygones n'est qu'une première approche. Il vaudrait mieux, par exemple, recalculer les centroïdes selon la dispersion réelle de la population.

calculés avec ajustements sur des facteurs de confusion, tels que l'année de naissance, le département et l'âge de la mère. Il est nécessaire d'examiner la composition de la population ainsi que les conditions géographiques et économiques afin de pouvoir comparer les taux des malformations. Il apparaît que les zones plus proches de la centrale sont moins densément peuplées (<10km : 78,3 hab./km² ; >20km : 232,9 hab./km²), ce qui indique qu'il s'agit plus de communes rurales. En ce qui concerne le revenu, il existe également des différences entre les groupes (<10km : 24.526 € ; >20km : 23.525 €) ce qui justifie un ajustement sur cette variable.

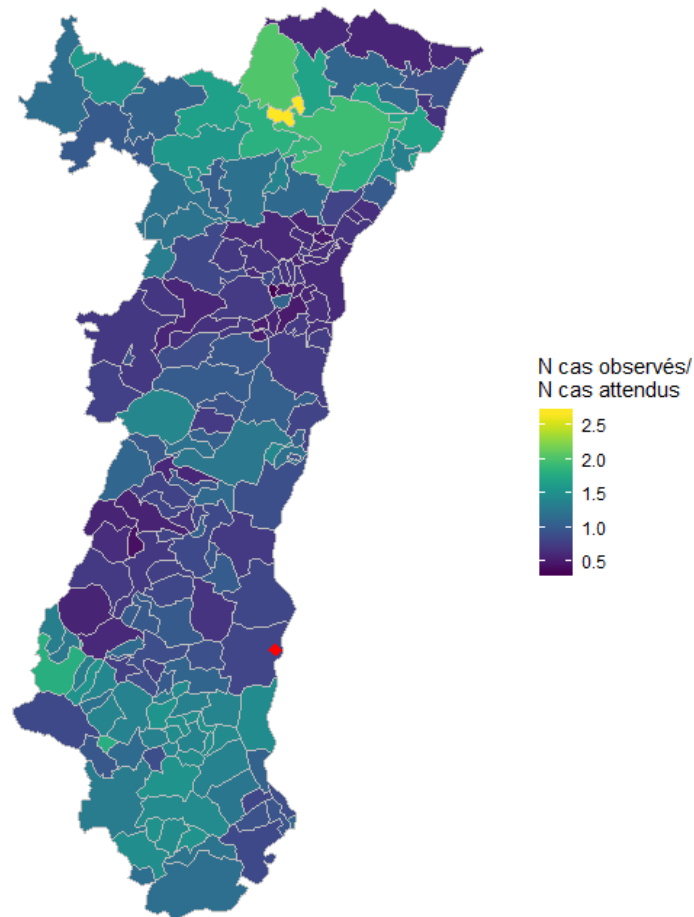
Tableau 8 : Caractéristiques des unités géographiques

	<10 km	10-15 km	15-20 km	>20 km
Nombre d'unités géographiques	1	3	4	168
Population en 2014	10.966	24.370	32.800	1.811.129
% population en 2014	0,58	1,30	1,75	96,37
Densité de population (hab./km ²)	78,3	98,8	148,0	232,9
Revenu familial net médian (euros/an)	24.526	24.072	24.200	23.525
Nombre de naissances (2010-2018)	1.001	2.052	2.993	182.663
% naissances (2010-2018)	0,53	1,09	1,59	96,80
Nombre de patients observés (2010-2018)	60	148	231	13.365
% nombre de patients observés (2010-2018)	0,43	1,06	1,62	96,88
Nombre de patients attendus (2010-2018)	73,22	150,10	218,94	13.361,73
% nombre de patients attendus (2010-2018)	0,53	1,09	1,59	96,80
Ratio nombre de patients observés/ nombre de patients attendus	0,82	0,99	1,06	1,00

Dans l'unité géographique située à moins de 10 km, 60 cas avec au moins une malformation sont observés mais selon le taux d'incidence pour toute l'Alsace, 73 sont attendus (correspondant à un ratio de 0,82). Pour les autres unités géographiques, il n'y a pas de différences majeures entre le nombre de cas attendus et observés. Il y a donc moins de cas que prévu près de la CNPE Fessenheim, ce qui contredit l'hypothèse de l'étude principale (selon laquelle plus le domicile est près du réacteur, plus le risque qu'un enfant naisse avec une malformation est élevé).

Figure 4 rend compte du rapport entre le nombre de cas observés et le nombre de cas attendus pour chaque unité géographique en Alsace. Le rectangle rouge indique la CNPE Fessenheim. Il existe des différences géographiques dans l'incidence des malformations congénitales, mais aucune tendance n'est associée à la CNPE Fessenheim.

Figure 4 : Ratio cas observés/attendus par unité géographique



Les malformations semblent plus fréquentes dans le nord de l'Alsace. Les communes de Mertzwiller (77 cas observés, 29 attendus) et Reichshoffen (221 cas observés, 109 attendus) ont les valeurs les plus élevées. En revanche, des communes près de Strasbourg et des communes vosgiennes semblent moins affectées. Les communes ayant les taux les plus bas sont Hangenbieten (3 cas observés, 9 attendus) et Labaroche (5 cas observés, 12 attendus). Ces exemples montrent le petit nombre de cas dans certaines unités géographiques qui peuvent mener à des valeurs aberrantes. Il convient aussi de noter encore une fois que les chiffres sont des données brutes et ne tiennent pas compte des conditions locales (composition de la population, géographie).

A ce stade, d'autres limites aux résultats devraient être discutées. Il s'agit simplement d'une analyse descriptive, des tests de signification statistique et des analyses telles que la régression de Poisson sont nécessaires afin de permettre un ajustement supplémentaire des facteurs de confusion et l'évaluation du lien entre les risques de malformations congénitales et la distance du domicile à la CNPE Fessenheim. Aussi, il apparaît que les données géographiques du PMSI peuvent ne pas être suffisamment précises. D'autres études menées sur le lien entre le rayonnement radioactif et les effets négatifs sur la santé sont plus graduelles et analysent en particulier un rayon de 5 km autour d'une centrale nucléaire (p.ex. Evrard et al., 2006; Gautheron et al., 2005) ce qui n'est pas possible avec les codes géographiques dans le PMSI. Une autre solution consisterait à utiliser les données des registres des malformations qui enregistrent l'adresse exacte de la mère. Enfin, d'autres études utilisent non pas un seul site, mais plusieurs centrales nucléaires pour inclure davantage de cas dans l'analyse (dans un rayon de 5 km autour d'une centrale nucléaire, il n'y a que peu de naissances) et pour comparer les taux de malformations autour des différents sites. Pour une analyse plus approfondie du site de Fessenheim, on pourrait peut-être aussi s'appuyer sur des données plus anciennes. Le SNDS contient des données PMSI des 20 dernières années (comparativement aux 10 années de la présente analyse). Les registres des malformations peuvent contenir des données provenant d'un passé encore plus lointain. Enfin, dans les analyses ultérieures, les types de malformations doivent être différenciés plus précisément (p.ex. malformations mineures, malformations géniques/chromosomiques, malformations majeures non syndromiques), car des différences d'incidence sont probables.

5 Conclusion

Le stage sur les données santé dans le Rhin supérieur, effectué dans le cadre du projet « Juxta Rhenum », a offert plusieurs axes de travail : d'une part, des recherches générales sur la situation des données santé dans la région ainsi que sur le lien entre la radiation ionisante et des événements sanitaires ; d'autre part, la préparation d'un dossier pour obtenir l'accès aux données sensibles et le travail avec les statistiques hospitalières en préparant des requêtes pour accéder aux données de la base nationales. Tout le long du stage, il s'agissait de se familiariser avec de nouveaux objets d'étude et d'acquérir de nouvelles compétences pour traiter des données spatialisées et temporelles. Dans le cadre de ce stage, j'ai acquis des nouvelles connaissances dans le domaine de la santé et de l'épidémiologie ainsi que dans le traitement des données.

Pour moi, ce fut une expérience enrichissante, non seulement en raison du sujet et des méthodes déployées, mais aussi car il s'agissait d'un stage à l'étranger. J'ai bénéficié de cette opportunité de travailler à l'Université de Strasbourg en faisant des progrès linguistiques. Des réunions et présentations régulières y ont contribué ainsi que la rédaction de documents en français. En ce qui concerne les résultats du stage, il convient de préciser qu'il s'agit de résultats préliminaires. Ils peuvent servir de base à la poursuite des travaux dans ce domaine. Notamment, la demande aux données PMSI au sein du SNDS est en cours de traitement ; en cas de réponse positive, on pourra utiliser le travail effectué pendant ce stage comme base pour des analyses complémentaires avec une base de données étendue.

Bibliographie

- Cancer Environnement, 2018. Les hémopathies malignes de l'enfant.
- Christianson, A., Howson, C.P., Modell, B., 2006. Global Report on Birth Defects. The Hidden Toll of Dying and Disabled Children.
- De Santis, M., Di Gianantonio, E., Straface, G., Cavaliere, A.F., Caruso, A., Schiavon, F., Berletti, R., Clementi, M., 2005. Ionizing radiations in pregnancy and teratogenesis: a review of literature. *Reproductive Toxicology* 20, 323–329.
- Doll, R., 1995. Hazards of ionising radiation: 100 years of observations on man. *British Journal of Cancer* 72, 1339–1349.
- Evrard, A.-S., Hémon, D., Morin, A., Laurier, D., Tirmarche, M., Backe, J.-C., Chartier, M., Clavel, J., 2006. Childhood leukaemia incidence around French nuclear installations using geographic zoning based on gaseous discharge dose estimates. *British Journal of Cancer* 94, 1342–1347. <https://doi.org/10.1038/sj.bjc.6603111>
- Forman, D., Cook-Mozaffari, P., Darby, S., Davey, G., Stratton, I., Doll, R., Pike, M., 1987. Cancer near nuclear installations. *Nature* 329, 499. <https://doi.org/10.1038/329499a0>
- Gautheron, S., Chevrier, C., Laborier, J.-C., Robert-Gnansia, É., 2005. Prévalence des malformations congénitales autour des centrales nucléaires/utilisation des données du registre France Centre-Est. *Environnement, Risques & Santé* 4, 179–186.
- Hinz, G., Kaul, A., 1987. Radiation exposure and pregnancy. *Gynakologe* 20, 137–143.
- IRSN, 2015. Rapport IRSN 2015-00001 - Exposition de la population française aux rayonnements ionisants.
- Kaatsch, P., 2007. Epidemiologische Studie zu Kinderkrebs in der Umgebung von Kernkraftwerken (KiKK-Studie).
- Kinlen, L., 2011. A German storm affecting Britain: childhood leukaemia and nuclear power plants. *Journal of Radiological Protection* 31, 279–284. <https://doi.org/10.1088/0952-4746/31/3/E02>
- Laurier, D., Bard, D., 1999. Epidemiologic Studies of Leukemia among Persons under 25 Years of Age Living Near Nuclear Sites. *Epidemiol Rev* 21, 188–206. <https://doi.org/10.1093/oxfordjournals.epirev.a017996>
- Mueller, W., Gilham, C., 2015. Childhood leukemia and proximity to nuclear power plants: A systematic review and meta-analysis. *Journal of Cancer Policy* 6, 44–56. <https://doi.org/10.1016/j.jcpo.2015.10.003>
- Preston, R.J., 2004. Children as a sensitive subpopulation for the risk assessment process. *Toxicology and Applied Pharmacology* 199, 132–141. <https://doi.org/10.1016/j.taap.2003.12.030>
- Queißer-Luft, A., Wiesel, A., Stolz, G., Mergenthaler, A., Kaiser, M., Schlaefel, K., Wahrendorf, J., Blettner, M., Spix, C., 2011. Birth defects in the vicinity of nuclear power plants in Germany. *Radiation and Environmental Biophysics* 50, 313–323.
- Sermage-Faure, C., Laurier, D., Goujon-Bellec, S., Chartier, M., Guyot-Goubin, A., Rudant, J., Hémon, D., Clavel, J., 2012. Childhood leukemia around French nuclear power plants-The geocap study, 2002-2007. *International Journal of Cancer* 131, E769–E780.
- Shepard, T.H., Lemire, R.J., 2004. *Catalog of Teratogenic Agents*. JHU Press.

- Spix, C., Schmiedel, S., Kaatsch, P., Schulze-Rath, R., Blettner, M., 2008. Case–control study on childhood cancer in the vicinity of nuclear power plants in Germany 1980–2003. *European Journal of Cancer* 44, 275–284. <https://doi.org/10.1016/j.ejca.2007.10.024>
- Weinhold, B., 2009. Environmental Factors in Birth Defects: What We Need to Know. *Environ Health Perspect* 117, A440–A447.
- Wertelecki, W., Chambers, C.D., Yevtushok, L., Zymak-Zakutnya, N., Sosyniuk, Z., Lapchenko, S., Ievtushok, B., Akhmedzhanova, D., Komov, O., 2017. Chernobyl 30 years later: Radiation, pregnancies, and developmental anomalies in Rivne, Ukraine. *European Journal of Medical Genetics* 60, 2–11.
- White-Koning, M.L., Hémon, D., Laurier, D., Tirmarche, M., Jouglà, E., Goubin, A., Clavel, J., 2004. Incidence of childhood leukaemia in the vicinity of nuclear sites in France, 1990–1998. *British Journal of Cancer* 91, 916–922. <https://doi.org/10.1038/sj.bjc.6602068>
- World Health Organization, 2016. Congenital Anomalies Fact Sheet.