APPROCHE METHODOLOGIQUE ET PRESENTATION DU MATERIEL DE RECHERCHE

Ce chapitre a pour objectif de décrire l'approche méthodologique ainsi que l'itinéraire de recherche adopté pendant cette thèse. La recherche a été conduite dans le cadre d'un contrat doctoral au sein du Centre de Gestion Scientifique de l'école Mines ParisTech entre 2015 et 2019. En tant que chercheur en gestion, l'utilisation d'une méthode de recherche est souvent la conséquence d'un choix épistémologique de la part du chercheur. Cette épistémologie permet de cadrer le retour critique que l'on porte sur notre objet de recherche et sur la connaissance en elle-même que celui peut apporter afin *« de décrire, de comprendre, de prédire ou d'expliquer des phénomènes liées aux organisations »* (Ben Aissa, 2001). De manière générale, deux épistémologies sont présentes dans les disciplines de sciences sociales : l'approche positive et l'approche constructiviste. L'approche positiviste a longuement été prédominante comme épistémologie dans les sciences sociales suivant l'influence des travaux d'Auguste Comte pour qui le « mot positif désigne le réel » (Le Moigne, 1995). Dans cette représentation, le réel est régi par un ensemble de lois préexistantes dont le rôle de la science est d'en découvrir l'existence. L'objet de recherche est indépendant du chercheur qui a permis d'arriver à son élaboration. Cette approche implique cependant un certain nombre de principes issu de la logique aristotélicienne comme la notion d'identité, de non contradiction ou de tiers exclus qui ne peuvent être facilement soutenus dans le contexte de la gestion (David, 1999). A la place, les études de cas en science de gestion se basent plutôt sur une approche constructiviste qui considère qu'« un objet existe si on est capable de le construire, d'en exhiber un exemplaire ou de le calculer explicitement » (Largeaut, 1993). Le chercheur n'est plus indépendant de la construction de l'objet qu'il étudie, mais fait partie intégrante de ce processus. David (1999) propose une typologie des différentes démarches de recherche lorsque celles-ci sont basées sur une approche constructiviste en croisant deux critères pour les différencier : l'objectif du chercheur suivant qu'il produit une construction mentale ou concrète de la réalité, et la démarche que celui-ci met en œuvre en fonction de s'il part d'une observation des faits ou d'un projet de transformation ou d'une situation idéalisée.

		Objectif	
		Construction mentale de la réalité	Construction concrète de la réalité
Démarche	Partir de l'observation des faits	**Observation, participante ou non** Elaborer un modèle de fonctionnement du système étudié	**Recherche-action, étude clinique** Aider à transformer le système à partir de sa propre réflexion sur lui-même
	Partir d'un projet de transformation ou d'une situation idéalisée	**Conception de modèles de gestion** Elaborer des outils de gestion potentiels, des modèles possibles de fonctionnement	**Recherche-intervention** Aider à transformer le système à partir d'un projet concret de transformation plus ou moins complètement défini

Dans notre approche, nous partageons plutôt une vision proche de la recherche-action ou de la recherche-intervention selon laquelle la recherche en gestion n'est pas simplement une recherche sur l'action, mais plutôt une recherche dans l'action, « une recherche transformative où le chercheur, participant à la vie de l'organisation, conçoit, met en œuvre, analyse, communique, diffuse les résultats obtenus tant à l'intérieur de l'organisation auprès des praticiens, qu'à l'extérieur en direction des milieux académiques » (Lallé, 2004). Au lieu de se poser la question du « comment » à partir d'un objectif bien défini, le chercheur part de ses études de cas pour se poser à la fois la question du « comment » et du « pourquoi » (Yin, 2003). Le processus de recherche basé sur les études de cas constitue une stratégie de recherche globale, reposant sur de multiples sources de données, les données devant converger tout en bénéficiant du développement au préalable de propositions théoriques pour guider la collecte et l'analyse de données.

1. ITINERAIRE ET CADRE METHODOLOGIQUE GENERAL DE LA THESE

Dans notre étude, le projet de recherche initial se base sur le questionnement suivant : **quelles sont les logiques de gestion à mettre en place pour s'assurer de l'efficacité systématique des projets de science citoyenne dans le cadre d'un processus data-driven ?** Ce projet de recherche est issu d'observations manifestes dans des cas empiriques observés par le chercheur ou au travers d'exemples issus de la littérature mais dont l'interprétation n'est pas immédiate. Une variété d'approches a été mobilisée pour définir un modèle de gestion adapté aux projets de science citoyenne. Par la suite, chaque méthode est présentée en lien avec la question de recherche identifiée.

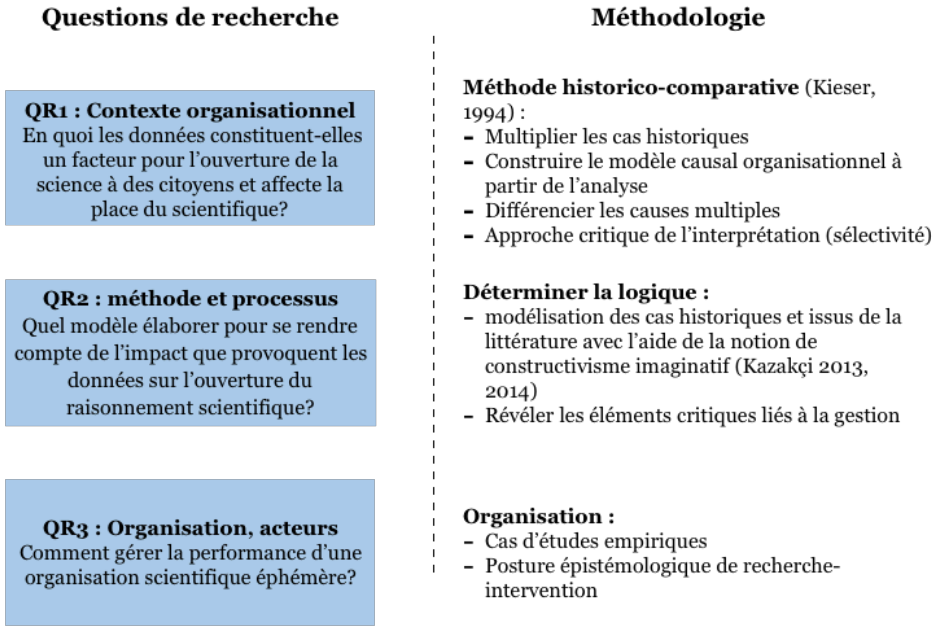

Figure 5. Méthodologie de recherche.

1.1. IDENTIFIER LES CAUSALITES ENTRE TRANSFORMATION DU RAPPORT AUX DONNEES ET ORGANISATION SCIENTIFIQUE

Pour mieux comprendre le contexte organisationnel contemporain et valider les liens supposés entre projets de science citoyenne et science data-driven, notre recherche suit une méthode historico-comparative du phénomène (Kieser, 1994). Cette approche a pour objectif de produire un modèle causal entre transformation du rapport aux données et ouverture de la science qui nous permette de justifier le lien supposé dans notre contexte contemporain. Notre étude se concentrera notamment sur l'apparition de nouveaux acteurs, la redéfinition du rôle des acteurs existants (celui notamment du scientifique), et proposera une ébauche de formalisme des différentes étapes du processus scientifique qui ont été ouvertes à de nouveaux acteurs. Au moins quatre raisons selon Kieser justifient l'intérêt de l'analyse historique pour étudier des phénomènes organisationnels contemporains : 1) L'analyse d'un comportement organisationnel ne peut être séparé d'un effet culturel inclus dans la dimension culturelle. 2) Recontextualiser des problèmes d'organisation contemporains à des situations similaires dans le passé pour éviter des effets idéologiques et des tendances actuelles « à la mode ». 3) Les analyses historiques nous apprennent à interpréter les structures organisationnelles existantes non pas telles que déterminées par les lois, mais comme le résultat de décisions prises dans le cadre d'opportunités de choix passés, certaines intentionnellement et d'autres implicitement. Des possibilités de choix qui n'avaient pas été utilisées à l'avantage des acteurs impliqués peuvent éventuellement se présenter à nouveau ou être restaurées d'une manière ou d'une autre. Les analyses historiques peuvent nous préparer à mieux identifier et à mieux utiliser les opportunités de choix. 4) En confrontant les théories du changement organisationnel aux évolutions historiques, ces théories peuvent être soumises à un test plus radical que celui qu'elles doivent passer lorsqu'elles sont simplement confrontées à des données sur les changements à court terme. En revanche, les analyses historiques au sein des organisations présentent un certain nombre de faiblesses qui peuvent être limitées en établissant un ensemble de bonnes pratiques méthodologiques.

D'abord, une étude restreinte à un seul cas type historique peut mener à des biais interprétatifs. Au contraire, les comparaisons avec d'autres situations similaires permettent d'augmenter la visibilité d'une structure en la contrastant avec une autre. Ainsi, dans notre étude nous allons analyser deux périodes distinctes dans lesquelles nous espérons retrouver des phénomènes similaires. D'abord l'introduction et l'émergence des instruments scientifiques entre le 17 et le 19è siècle au sein de la pratique scientifique. Nous nous appuierons sur les travaux de Christian Licoppe qui a analysé cette période à la fois en France et en Angleterre afin de réduire l'effet culturel possible de notre interprétation. Nous multiplierons les exemples durant la période pour justifier l'existence d'une tendance paradigmatique dans la science ou du moins dans un certain nombre de disciplines scientifiques. Nous étudierons également une deuxième période entre le 19 et le 20è siècle comme l'introduction de la stochastique dans l'analyse des phénomènes naturels dans certaines disciplines scientifiques. Le contraste entre ces deux périodes ponctuées de multiples exemples permettra de réduire ces biais.

Ensuite, Kieser propose une voie stratégique pour étudier les régularités causales que l'on peut extraire dans l'analyse historique. Au lieu de guider l'analyse sur une hypothèse préconçue issue d'une modélisation et d'une réflexion théorique antérieure unique, la construction de la relation causale est constamment modifiée et générée dans un dialogue constant avec les données historiques. Le chercheur n'essaie pas de prouver un modèle existant; au lieu de cela, il s'engage à générer des schémas de causalité capables d'expliquer les développements historiques. Ainsi dans notre approche méthodologique, nous suggérons un lien de causalité entre transformation du rapport aux données et réorganisation de la science sans imposer le schéma causal que nous découvrirons durant notre analyse historique.

Un autre problème est que les événements historiques ont toujours des causes multiples qui ne doivent pas nécessairement s'exclure mutuellement mais peuvent être complémentaires. Ainsi si plusieurs facteurs ont été identifiés en tant que causes possibles d'un événement historique, le chercheur doit alors préciser si une seule cause aurait suffi à en provoquer l'apparition.

Enfin, l'auteur suggère que le contenu historique est inépuisable et donc mène inévitablement à une sélection de la part du chercheur qui doit être justifiée. Dans notre cas, le choix des deux périodes que nous étudions n'est pas anodin. Ils représentent l'étendue de ce que l'on définit comme l'histoire de la science moderne. D'un point de vue organisationnel, ils cherchent à recréer un pont entre un mode de production de la connaissance où on passe du scientifique seul ou accompagné de quelques assistants à notre époque moderne constituée d'acteurs hétérogènes qui ne peuvent être définis comme scientifique.

1.2. Modelisation des projets de science citoyenne : dans quelles conditions les projets sont efficaces ?

Alors que les projets de science citoyenne s'intègrent dans un contexte de transformation du rapport aux données et de systématisation de son utilisation, il devient nécessaire de préciser les conditions qui permettent de s'assurer de son efficacité. Pour identifier quelles sont les caractéristiques nécessaires à la bonne gestion de ce type de projets, nous avons pour objectif de modéliser les différents types de projets de science citoyenne que nous pouvons rencontrer et d'analyser dans quelles conditions ce modèle est suffisant dans le processus data-driven.

Durant notre revue de littérature, nous avons montré que le modèle dominant pour étudier les projets de science citoyenne est celui du crowdsourcing. Or, celui-ci se base essentiellement sur l'augmentation de la diversité et n'est pas suffisant dans le cas où les participants peuvent interagir entre eux ainsi que pour la génération des hypothèses basées sur les données. Le but est donc de construire un modèle qui prenne en compte la systématisation des tâches afin de déterminer les critères qui permettent d'assurer l'efficacité des projets. Notre modélisation démarre initialement sur les exemples que nous avons rencontrés dans notre analyse historique ainsi que sur les cas contemporains de projets de science citoyenne. Nous nous appuyons sur les

théories sous-jacentes du modèle du crowdsourcing pour construire un modèle de tâche basé sur le principe de résolution de problème (Simon & Newell, 1971). Ensuite, nous analysons un ensemble de cas de modélisation informatique de la génération des hypothèses scientifiques pour montrer que le modèle de résolution de problème n'est pas suffisant pour interpréter cette tâche (e.g. King et al., 2009; Kulkarni & Simon, 1988). Nous proposons alors d'étendre le modèle initial en nous basant sur une théorie issue de la conception appelée « constructivisme imaginatif » (Kazakçi; 2013, 2014).

A partir de ce modèle formel, nous tentons de relier la systématisation des projets dans une organisation et les critères qui vont établir le lien avec son efficacité. Il est important de souligner que ce modèle n'a pas pour vocation à être une représentation robuste et exhaustive des projets de science citoyenne, mais plutôt de fournir les caractéristiques nécessaires pour s'assurer de l'efficacité des projets dans le cas où ils sont systématisés.

1.3. DEUX CAS D'ETUDE D'ORGANISATION POUR IDENTIFIER UNE FIGURE MANAGERIALE ET DES LOGIQUES ORGANISATIONNELLES

La troisième étape de notre recherche consiste à réinterroger notre modèle et ses implications théoriques dans un contexte empirique. L'enjeu principal est de comprendre quels sont les principes qui permettent de gérer les tâches qui n'apparaissent pas dans les modèles de gestion traditionnels. Nous adopterons dans notre approche une posture épistémologique de « recherche collaborative » (Shani et al., 2008) ou « recherche-intervention » (David, Hatchuel, & Laufer, 2012) menée par des chercheurs et des praticiens afin de créer des connaissances concrètes pour l'organisation et de nouveaux modèles théoriques pour la recherche en sciences de gestion (David & Hatchuel, 2008). Nous avons choisi d'inscrire notre recherche expérimentale dans deux contextes particuliers.

Nous étudierons d'abord le cas d'un outil de gestion, le RAMP (pour Rapid Analytics and Model Prototyping), développé par le Centre de Data Science de Paris-Saclay. Cet outil propose de développer des projets basés sur des problématiques et des bases de données fournies par des scientifiques de disciplines variées (économie, biologie, physique des particules,…). Chaque problème est formalisé comme un problème d'optimisation d'algorithme d'analyse de données et soumis à une foule de participants. Le RAMP est utilisé comme plateforme de compétition et de collaboration où les spécialistes des données travaillent sur un problème pour des délais relativement courts (généralement un ou deux jours). Cela peut être considéré comme une forme de Hackathon avec la principale différence que l'objectif est d'optimiser une métrique claire. L'intérêt de la plateforme est qu'elle a été conçue également comme outil d'observation du travail des data scientists. Des mesures sont réalisées durant les challenges pour lesquelles il est possible d'extraire des informations quantitatives sur les trajectoires des participants. En collaboration

avec l'équipe organisatrice de la plateforme, nous avons mené un travail d'analyse des données à partir d'outils statistiques ainsi que de méthodes de visualisation. Cette analyse a permis de mettre en avant les caractéristiques critiques sur l'efficacité de la plateforme et donc a aidé à aiguiller dans les choix organisationnels mis en œuvre.

Le deuxième terrain expérimental est Epidemium, un programme de recherche collaboratif basé sur l'épidémiologie, qui s'est déroulé entre novembre 2015 et mars 2018. Epidemium, financé en partie par les Laboratoires Roche, a pour mission de rassembler une communauté autour de de bases de données massives afin d'explorer ces bases de données pour générer des hypothèses scientifiques. L'étude expérimentale a été menée en partie avec Olga Kokshagina, une collègue chercheuse ayant travaillé au sein du laboratoire du CGS. Elle est basée à la fois sur un ensemble de documents, une participation active pour le déroulement et l'organisation de certains évènements, une communication avec les organisateurs et les participants, et la recherche de mise en place d'outils pour piloter l'efficacité du projet.

2. CONTEXTE DES TERRAINS DE RECHERCHE DANS LEURS DOMAINES DE SCIENCE

L'analyse et les principaux résultats organisationnels de la thèse sont basés sur les deux terrains que nous avons présentés ci-dessus. Ils présentent plusieurs caractéristiques qui justifient de leur intérêt méthodologique pour notre analyse. D'abord, la plateforme RAMP nous permet d'avoir accès à des données uniques pour étudier le comportement de participants durant un projet de type science citoyenne. En plus d'un accès privilégié aux résultats et aux solutions de tous les participants, les organisateurs ont intégré des métriques qui permettent d'analyser le déroulement de la collaboration entre les participants et d'estimer son efficacité. Ce type de métriques est unique dans les projets de type challenge. Un autre élément est que la philosophie gestionnaire portée par la plateforme suppose l'émergence de l'importance des méthodes d'analyse des données regroupées sous le terme de « sciences de données ». Nous présentons dans cette section le contexte du RAMP et de sa construction.

Ensuite, le programme Epidemium est un projet unique d'ouverture de la génération d'hypothèses basées sur les données à une foule de participants. Ce programme est caractéristique d'un besoin grandissant dans de plus en plus d'organisations scientifiques qui cherchent à valoriser les larges bases de données qui leurs sont disponibles. Pourtant, il n'existe pas de méthodologie pour construire des hypothèses scientifiques à partir de bases de données massives. Nous montrons dans cette section que notre étude se place dans un cadre plus large d'épidémiologie populaire qui inclut les acteurs de la société dans le processus scientifique. Ce processus est amplifié dans le cadre d'une explosion des données de santé qui pousse les scientifiques à chercher des solutions pour générer de la valeur à partir de ces bases de données.

2.1. CADRE D'ETUDE POUR LE RAMP : LES « SCIENCES DES DONNEES »

Des quantités massives de données sont produites dans l'environnement scientifique actuel (Miller, 2010). Au-delà des problèmes d'infrastructure et d'ingénierie, l'extraction d'informations à partir de données est devenue un défi majeur pour les universités et exige de repousser les limites des techniques d'analyse actuelles et de développer des avancées radicales. Cet impératif a donné lieu à la notion de « science des données » dans les milieux universitaires (Agarwal & Dhar, 2014; Davenport & Patil, 2012). La science des données peut être définie au sens large comme la conception de méthodes automatisées pour analyser des données massives et complexes afin d'extraire des informations utiles. La science des données et le Big data sont étroitement liés mais ne concernent pas les mêmes problématiques. Alors que le Big data couvre un large éventail de thèmes sur la capture, le transfert, le stockage, la recherche, le partage sécurisé, l'archivage et l'analyse de données massives, la science des données se concentre sur les aspects algorithmiques et mathématiques de l'extraction de nouvelles connaissances à partir de données. En tant que telle, la science des données se situe à la croisée de l'informatique, des mathématiques appliquées et des statistiques. Le débat académique sur la notion de science des données s'accompagne d'initiatives institutionnelles à travers le monde. Par exemple, après l'annonce de l'Initiative nationale de R&D sur les données massives de la Maison Blanche en 2012, les agences de financement nationales (NSF, NIH et DARPA) et les universités ont mené des actions à grande échelle pour promouvoir la science des données et la recherche sur le Big data. Les cas suivants méritent d'être mentionnés. La Research Data Alliance (RDA) est créée pour accélérer l'innovation axée sur les données dans le monde entier grâce au partage et à l'échange de données de recherche. L'Université de New York a inauguré son Centre for Data Science. L'Université de Washington a fondé son institut eScience. Berkeley a lancé son Institute for Data Science. Les fondations Moore et Sloan ont annoncé une initiative interinstitutionnelle de 37,8 millions de dollars sur cinq ans pour soutenir les trois instituts précédents. Harvard a créé le Harvard Data Science Initiative pour accélérer la collaboration entre la recherche et l'enseignement. En Europe, l'Université d'Amsterdam a annoncé la création de son centre de recherche Data Science. L'Université d'Edimbourg a lancé son Centre de formation doctorale en science des données. L'Imperial College London s'est associé à l'université de Zhejiang pour lancer une collaboration en science des données. Depuis 2016, l'école Polytechnique en France a créé l'Initiative Data Science qui regroupe une équipe de chercheurs et d'enseignants afin de fournir des réponses concrètes à des questions qui émanent des milieux industriels.

Ces initiatives sur les sciences de données sont généralement conçues et organisées de manière à trouver et recruter temporairement un chercheur spécialiste en intelligence artificielle et en analyse de données, dans l'espoir que des progrès significatifs seront réalisés durant cette période. En effet, un principe largement partagé dans les communautés scientifiques considère que le manque de spécialistes des données constitue un des problèmes majeurs des projets basés sur les données complexes (Davenport & Patil, 2012). Cependant, les retours d'expérience sur les projets

sur les données tendent à montrer que cette vision est limitée, et que le problème essentiel est plutôt organisationnel (Kazakçi, 2015; Kégl et al., 2018). En général la pratique du processus scientifique s'organise dans un système où les chercheurs sont incités à faire carrière par le biais de publications qu'ils produisent et la recherche d'une reconnaissance par les pairs (Merton, 1957; Merton & Storer, 1973). Cette conception de la science néglige le fait qu'une grande partie de la science se fait aujourd'hui dans des communautés dispersées aux sein de différentes organisations de R&D (von Zedtwitz, Gassmann, & Boutellier, 2004). La principale limite de ces processus scientifiques basés sur les données est cette « incapacité de la communauté scientifique dans son ensemble à comparer rapidement et à évaluer la valeur scientifique d'un jeu de données et de sa configuration analytique » (Kégl et al., 2018). Entre les années 1990 et les années 2000 la recherche en intelligence artificielle était construite autour de jeux de données de référence qui permettaient une comparaison entre les différents algorithmes. Cependant la situation a grandement évolué et les scientifiques sont maintenant dotés de plusieurs algorithmes très performants sur des critères de performance établis. Ainsi le problème n'est plus d'obtenir la performance d'un algorithme suivant une configuration donnée, mais plutôt de se concentrer sur la configuration en elle-même (Kégl et al., 2018).

Depuis une dizaine d'années, la communauté en intelligence artificielle et machine learning a réalisé des efforts importants pour standardiser les algorithmes de modèles prédictifs par le biais de plateformes tel que Scikit-learn (Vanderplas, J. et al., 2011) ou Keras. Ces travaux ont permis de grandement faciliter la réutilisation des modèles existants et de distinguer la phase d'expérimentation de la phase d'optimisation dans la construction de modèles prédictifs. En effet, séparer la conception des processus de la partie optimisation garantit que l'optimisation ne démarre que lorsque l'expérience est entièrement spécifiée. La recherche en intelligence artificielle et en machine learning s'est organisée sous forme de défis appelés « data challenges », dans lesquelles des communautés ou des personnes indépendantes acceptent de travailler sur des problèmes scientifiques majeurs afin d'optimiser une métrique bien définie. La configuration standard des data challenges est un concours de compétition pur - les participants opèrent en parallèle sans communication et le ou les gagnants obtiennent une récompense. La principale hypothèse de ces challenges est qu'un grand nombre de participants augmenteront les chances de trouver des solutions exceptionnelles. La littérature montre que, lorsqu'une forte incitation est présente (Boudreau, Lacetera, & Lakhani, 2011), cette configuration est susceptible de stimuler les efforts fournis par les participants et d'augmenter la qualité globale des solutions (Afuah & Tucci, 2012). D'un autre côté, cette configuration présente l'inconvénient de ne pas capitaliser pleinement sur la production entière de la foule, puisque seules les solutions gagnantes sont divulguées à la fin (Kazakçi, 2015). Des idées potentiellement bonnes qui ne donnent pas de succès immédiat sont perdues (Boudreau & Lakhani, 2015). C'est un handicap important pour les problèmes nécessitant une collaboration étroite entre les domaines (par exemple, la physique) et les spécialistes de l'analyse de données. Ce constat a été l'un des principaux moteurs du Center for Data Science (CDS) de l'université Paris-Saclay, une des initiatives de data science, qui a souhaité

fournir une plateforme plus flexible où les configurations collaboratives et compétitives étaient possibles. C'est dans ce cadre que la plateforme RAMP a été créée.

En conclusion, le besoin grandissant de compétences spécifiques pour appliquer les méthodes de sciences de données couplées à une standardisation de ces méthodes ont grandement facilité et poussé les scientifiques à exploiter des ressources externes notamment sous la forme de projets d'analyse des données. La performance et de l'efficacité de la répétition de ces projets est donc une question fondamentale.

2.2. L'EPIDEMIOLOGIE COMME CADRE D'ETUDE POUR LE PROGRAMME EPIDEMIUM

2.2.1. Avalanches de données dans le domaine médical

Les scientifiques du domaine médical et les acteurs du domaine de l'épidémiologie font face à une explosion du nombre de données accessibles et de leur variété (Chiolero, 2013). Alors qu'une grande partie des données relatives à la santé tel que la tenue de registres administratifs, les données relatives aux patients (notes et ordonnances écrites du médecin, imagerie médicale, laboratoire, pharmacie, assurance et autres données administratives) étaient généralement stockées sous forme de papier, la tendance actuelle est à la numérisation en grandes bases de données (Raghupathi & Raghupathi, 2014). Par ailleurs, des instituts internationaux compilent des bases de données massives sur des sujets relatifs à la santé et les rendent accessibles au grand public. A titre d'exemple, l'Organisation Mondiale de la Santé a mis en place depuis 2012 une base de données de veille sanitaire accessible aux habitants des 194 Etats membres comprenant plus de 1000 indicateurs. Les grandes bases de données que constitue l'ensemble des publications scientifiques sont maintenant stockées numériquement et peuvent être plus facilement analysées et traitées afin de faire émerger des problématiques scientifiques (Bohannon, 2017; Gray, 2009; Sybrandt, Shtutman, & Safro, 2017). De la même manière, l'analyse des résultats d'essais cliniques prend également une toute autre dimension à l'ère des données massives (DerSimonian & Laird, 2015). Enfin, les données issues des plateformes de réseaux sociaux tel que Twitter ou Facebook peuvent potentiellement être une source d'information pour la pharmacovigilance (Bian, Topaloglu, & Yu, 2012).

Les rapports indiquent que les données du seul système de santé américain ont atteint, en 2011, 150 exaoctets et atteindront bientôt l'échelle des zettaoctets (1021 gigaoctets) et, peu après, le yottabyte (1024 gigaoctets) (Cottle & Hoover, 2013). Pour les spécialistes du Big Data, il existe, parmi cette vaste quantité de données, une opportunité. En découvrant les associations entre les variables que constituent les bases de données, l'analyse du Big Data a le potentiel d'améliorer les soins, de sauver des vies et de réduire les coûts en augmentant le rôle de la médecine préventive.

2.2.2. L'épidémiologie entre médecine, statistique et science populaire

2.2.2.1. Avalanche de données en épidémiologie

L'épidémiologie a un rôle fondamental à jouer dans cette transformation. Cette discipline impliquant à la fois les médicaments et les statistiques étudie principalement les facteurs de risque associés à l'incidence ou à la mortalité des maladies. Depuis les années 1950, les études épidémiologiques ont utilisé des méthodes statistiques qui leur permettent d'extrapoler les résultats obtenus sur des échantillons à des populations beaucoup plus importantes. Cette approche a conduit à l'émergence de nombreuses études sur les facteurs de risque comportementaux tels que l'exposition à l'alcool, le tabagisme ou la nutrition. Les biais statistiques dans l'échantillonnage affectent cependant l'extrapolation des phénomènes locaux et plusieurs études ont mis en évidence que les résultats sont parfois contradictoires sur des facteurs de risque similaires : par exemple, un aliment peut prévenir et favoriser le cancer selon différentes études (Schoenfeld & Ioannidis, 2013). L'émergence récente de bases de données massives sur l'incidence et la mortalité des maladies est considérée dans la communauté de l'épidémiologie comme une opportunité pour des études épidémiologiques susceptibles de réduire les problèmes actuels et les limites des approches existantes, et de récentes études montrent comment les méthodes de machine learning peuvent être utilisées en épidémiologie (Szymczak et al., 2009).

2.2.2.2. Une tradition récente de « l'épidémiologie populaire »

La recherche épidémiologique est généralement réalisée en laboratoire par des épidémiologistes ou des spécialistes de la santé comme des oncologues ou des cliniciens. Cependant, une forme d'épidémiologie fondée sur une large participation du public telle que rencontrée dans les sciences citoyenne fait l'objet d'un courant de recherche depuis la fin des années 1980 sous le terme « épidémiologie populaire ». L'épidémiologie populaire désigne le processus par lequel des citoyens collectent eux-mêmes des données et mobilisent des connaissances scientifiques pour comprendre la distribution et les causes d'une maladie (Barthe, 2013). Cette notion a été créée par le sociologue Phil Brown pour qualifier le travail d'enquête réalisé par les riverains d'un site contaminé afin d'établir l'origine des leucémies infantiles qui frappaient leur communauté (Brown, 1987, 1992). Par la suite, la notion a été mobilisée dans un certain nombre de champs scientifiques relatifs à la santé environnementale (Barthe, 2013). Si cette forme de pratique scientifique intéresse les chercheurs en sciences sociales, elle crée également un certain nombre de débats entre les professionnels de la santé pour reconnaître ce type d'enquête comme viable d'un point de vue scientifique. Ainsi, certains chercheurs militent pour trouver des moyens d'intégrer cette pratique en créant de nouveaux protocoles de recherche, tandis que d'autres chercheurs y voient l'apparition d'une pratique dangereuse car pouvant intégrer au débat public des résultats sans valeur scientifique.

Notre étude se place dans cette tendance d'une épidémiologie populaire, à la différence que les données n'ont pas été collectées par les participants mais par un ensemble d'acteurs hétérogènes,

puis ont été choisies et compilées par des équipes de scientifiques. De plus, la validation d'un nouveau résultat scientifique ne provient pas uniquement du travail des participants, mais est mesuré, évalué et reconnu comme tel en accord avec des spécialistes de la santé et des méthodes épidémiologiques. La tâche que nous analysons ici, à savoir la formulation d'hypothèses basées sur les données, n'est pas synonyme de ce qui pourrait être considéré comme un retour aux méthodes épidémiologiques de la première heure qui consistaient à proposer des analyses descriptives des objets étudiés sans se baser sur des méthodes de statistiques inférentielles (Schwartz in Lechopier, 2010). En effet notre démarche n'est pas purement inductive mais cherche au contraire à relier des effets à des causes. Par le fait, notre étude porte sur la formulation des hypothèses scientifiques et la vérification de ces hypothèses grâce aux données qui sont disponibles. La validation ensuite de cette construction comme résultat scientifique ou non dépendra ensuite de la fiabilité des données collectées ainsi que de la pertinence des méthodes employées. Suivant la pertinence des résultats obtenus, l'hypothèse pourra faire l'objet dans un cadre ultérieur d'une étude plus approfondie grâce à la collecte de données spécifiques.

3. SYNTHESE DE L'ITINERAIRE DE RECHERCHE ET DES METHODES CHOISIES