Enjeux et place des data sciences dans le champ de la réutilisation secondaire des données massives cliniques : une approche basée sur des cas d'usage

Guillaume Bouzillé

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE RENNES 1

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : Analyse et traitement de l'information et des images médicales

Par
Guillaume BOUZILLÉ

Enjeux et place des data sciences dans le champ de la réutilisation secondaire des données massives cliniques

Une approche basée sur des cas d'usage

Thèse présentée et soutenue à Rennes, le 21 juin 2019
Unité de recherche : Laboratoire Traitement du Signal et de l'Image, Equipe Données Massives en Santé

Rapporteurs avant soutenance :

Nicolas JAY	Professeur d'Université – Praticien Hospitalier, Université de Lorraine
Alexandre MOREAU-GAUDRY	Professeur d'Université – Praticien Hospitalier, Université Grenoble Alpes

Composition du Jury :

Président :	Prénom Nom	Fonction et établissement d'exercice (9)*(à préciser après la soutenance)*
Examinateurs :	Sandra BRINGAY	Professeur, Université de Montpellier
	Anita BURGUN	Professeur d'Université – Praticien Hospitalier, Université Paris Descartes
	Leslie GUILLON	Maître de conférences – Praticien Hospitalier, Université de Tours
	Pascal STACCINI	Professeur d'Université – Praticien Hospitalier, Université de Nice Sophia Antipolis
Dir. de thèse :	Marc CUGGIA	Professeur d'Université – Praticien Hospitalier, Université de Rennes 1

Remerciements

Je tiens à remercier le Professeur Marc Cuggia de m'avoir encadré durant cette thèse, mais surtout pour m'avoir accordé sa confiance en m'intégrant dans son équipe. Tes conseils, ton soutien et ton humanisme sont précieux et nous font tous progresser.

Mes remerciements vont également aux membres du jury. Merci à mes rapporteurs, les Professeurs Nicolas Jay et Alexandre Moreau-Gaudry de me faire l'honneur d'évaluer ce travail. Merci aux Professeurs Anita Burgun, Sandra Bringay et Pascal Staccini, au Docteur Leslie Guillon d'avoir accepté de participer à ce jury de thèse.

Je remercie également l'ensemble de l'équipe Données Massives en Santé, en particulier Pascal, Denis, Christian, Véronique, Pierre, Marie-Lisenn, Julia, sans qui tout le travail présenté dans cette thèse n'aurait pu être réalisé.

Merci également aux doctorants et stagiaires de l'équipe qui ont participé grandement à ce travail. Je pense notamment à Canelle et Richard.

Aux personnes impliquées dans les projets de recherches sur lesquels se base cette thèse, notamment Sahar, Yann et Marie-Noëlle.

Enfin, je remercie chaleureusement ma famille et mes amis pour leur soutien sans failles et leurs encouragements depuis toutes ces années.

Table des matières

Résumé

La dématérialisation des données de santé a permis depuis plusieurs années de constituer un véritable gisement de données provenant de tous les domaines de la santé. Ces données ont pour caractéristiques d'être très hétérogènes et d'être produites à différentes échelles et dans différents domaines. Leur réutilisation dans le cadre de la recherche clinique, de la santé publique ou encore de la prise en charge des patients implique de développer des approches adaptées reposant sur les méthodes issues de la science des données. L'objectif de cette thèse est d'évaluer au travers de trois cas d'usage, quels sont les enjeux actuels ainsi que la place des data sciences pour l'exploitation des données massives en santé.

La démarche utilisée pour répondre à cet objectif consiste dans une première partie à exposer les caractéristiques des données massives en santé et les aspects techniques liés à leur réutilisation. La seconde partie expose les aspects organisationnels permettant l'exploitation et le partage des données massives en santé. La troisième partie décrit les grandes approches méthodologiques en science des données appliquées actuellement au domaine de la santé. Enfin, la quatrième partie illustre au travers de trois exemples l'apport de ces méthodes dans les champs suivant : la surveillance syndromique, la pharmacovigilance et la recherche clinique. Nous discutons enfin les limites et enjeux de la science des données dans le cadre de la réutilisation des données massives en santé.

Mots-clés : Réutilisation secondaire des données, Données massives en santé, Sciences des données, Surveillance syndromique, Recherche clinique, pharmacovigilance

Abstract

The dematerialization of health data, which started several years ago, now generates a huge amount of data produced by all actors of health. These data have the characteristics of being very heterogeneous and of being produced at different scales and in different domains. Their reuse in the context of clinical research, public health or patient care involves developing appropriate approaches based on methods from data science. The aim of this thesis is to evaluate, through three use cases, what are the current issues as well as the place of data sciences regarding the reuse of massive health data.

To meet this objective, the first section exposes the characteristics of health big data and the technical aspects related to their reuse. The second section presents the organizational aspects for the exploitation and sharing of health big data. The third section describes the main methodological approaches in data sciences currently applied in the field of health. Finally, the fourth section illustrates, through three use cases, the contribution of these methods in the following fields: syndromic surveillance, pharmacovigilance and clinical research. Finally, we discuss the limits and challenges of data science in the context of health big data.

Keywords : Data reuse, Health big data, Data sciences, Syndromic surveillance, Clinical research, Drug safety

Productions scientifiques liées à la thèse

1. Bouzillé G, Jouhet V, Turlin B, Clement B, Desille M, Riou C, et al. Integrating Biobank Data into a Clinical Data Research Network: The IBCB Project. Stud Health Technol Inform. 2018;247:16-20.

2. Bouzillé G, Westerlynck R, Defossez G, Bouslimi D, Bayat S, Riou C, et al. Sharing Health Big Data for Research - A Design by Use Cases: The INSHARE Platform Approach. Stud Health Technol Inform. 2017;245:303-7.

3. Bouzillé G, Pan W, Franco-Contreras J, Cuggia M, Coatrieux G. Clinical Data Warehouse Watermarking: Impact on Syndromic Measure. Stud Health Technol Inform. 2017;235:323-7.

4. Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert M-L, Chabot M, Chazard E, et al. Leveraging hospital big data to monitor flu epidemics. Computer Methods and Programs in Biomedicine. 2018;154:153-60.

5. Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, Bouzillé G. Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Study. JMIR Public Health Surveill. 21 déc 2018;4(4):e11361.
6. Claveau V, Oliveira LES, Bouzillé G, Cuggia M, Cabral Moro CM, Grabar N. Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation. In: AIME 2017 - 16th Conference in Artificial Intelligence in Medecine [Internet]. Vienne, Austria: Springer; 2017 [cité 2 mai 2019]. p. 203-8. (LNCS; vol. 10259).

7. Bouzillé G, Osmont M-N, Triquet L, Grabar N, Rochefort-Morel C, Chazard E, et al. Drug safety and big clinical data: Detection of drug-induced anaphylactic shock events. J Eval Clin Pract. juin 2018;24(3):536-44.

8. Bouzillé G, Morival C, Westerlynck R, Lemordant P, Chazar E et al. An automated detection system of drug-drug interactions from electronic patient records using big data analytics. Medinfo 2019 (accepté).

Liste des abréviations

ADN : Acide DésoxyriboNucléique
ANR : Agence National de la Recherche
ARS : Agence Régionale de Santé
CDC : Centre de Données Cliniques
CépiDC : Centre d'épidémiologie sur les causes médicales de décès
CHU : Centre Hospitalier Universitaire
CIM-10 : Classification Internationale des Maladies, 10ème version
DICOM : Digital imaging and COmmunications in Medicine
DMP : Dossier Médical Partagé
eCRF : électronique Case Report Form
eHOP : entrepôt HÔPital
EHR4CR : Electronic Health Record for Clinical Research
ELT : Extract, Load & Transform
ETL : Extract, Transform & Load
GCS : Groupement de Coopération Sanitaire
HUGO : Hôpitaux Universitaires du Grand Ouest
i2b2 : Informatics for Integrating Biology and the Bedside
LTSI : Laboratoire Traitement du Signal et de l'Image
NoSQL : Not Only Structured Query Language
OMOP : Observational Medical Outcomes Partnership
PACS : Picture Archiving and Communication System
PCORnet : Patient-Centered Clinical Research Network
PMSI : Programme de Médicalisation du Système d'Information
RAM : Random Access Memory
RGPD : Règlement Général sur la Protection des Données
RiCDC : Réseau interrégional des Centres de Données Cliniques
ROC : Receiver Operating Curve
SHRINE : Shared Health Research Informatics NEtwork
SNDS : Système National des Données de Santé
SNIIRAM : Système national d'information inter-régimes de l'assurance maladie
STRIDE : Stanford Translational Research Integrated Database Environment
T2A : Tarification liée à l'activité
TF-IDF : Term Frequency-Inverse Document Frequency
XML : eXtensible Markup Language

Avant-propos

Ce travail de thèse a été réalisé au cours de mon assistanat puis depuis ma prise de fonction de praticien hospitalo-universitaire en mai 2018. Ces fonctions m'ont amené à valoriser mon travail sous forme de publications scientifiques en lien avec mon activité de recherche sur les méthodes d'exploitation des données massives en santé. Ces travaux s'inscrivent également de façon plus large dans le cadre de la thématique de recherche de l'équipe Données Massives en Santé du Professeur Marc Cuggia, qui porte sur les méthodes d'intégration et d'exploitation des données massives en santé, auxquelles je participe activement.

Pour cette raison, ce manuscrit est sous la forme d'une thèse sur articles qui illustrent de façon concrète les travaux qui ont été réalisés. Un certain nombre d'éléments complémentaires sont apportés au fil des sections afin de resituer ces travaux dans le contexte de la science des données et de la réutilisation des données massives en santé. Certains aspects spécifiques tels que l'interopérabilité ou la réglementation de la réutilisation des données de santé ne sont en revanche pas détaillés.

La discussion aborde, quant à elle, de façon globale les limites et perspectives liées à ces travaux afin de répondre à l'objectif de cette thèse.

Introduction

La dématérialisation des données de santé a permis depuis plusieurs années de constituer un véritable gisement de données provenant de tous les domaines de la santé (1). Tous les secteurs et acteurs en santé sont dorénavant concernés et participent à la production et l'exploitation de données valorisables. On peut distinguer trois grands secteurs d'intérêt au sein desquels on retrouve l'ensemble des acteurs, à la fois producteurs et consommateurs de données : le soin, la recherche médicale et le domaine médico-administratif. Cette production s'articule autour du patient, lui-même générant des données de plus en plus riches. Du fait de l'intérêt économique que constitue le domaine de la santé, on retrouve aujourd'hui les acteurs traditionnels du secteur public, mais également de nombreux acteurs privés, tels que les grands groupes du domaine médical ou des start-up, qui s'intéressent principalement à l'exploitation de ces données.

Du fait de l'explosion de la quantité de données aujourd'hui produites est apparue la notion de « Big data » ou données massives (2). Cette notion a ouvert un nouveau champ de recherche : la science des données ou « data science ». Cette discipline est en fait issue de la conjonction de plusieurs compétences nécessaires à l'exploitation de ce gisement de données massif : l'analyse de données (statistiques, sciences de l'information, intelligence artificielle), l'informatique (bases de données, algorithmique), mais également des compétences dans le domaine médical pour le cas de la science des données en santé (3). L'objectif sous-jacent est l'extraction d'informations et de connaissances au sens large à partir des données, que ce soit à visée descriptive, diagnostique, prédictive ou prescriptive (4,5). Ainsi, de nombreux travaux s'appuient aujourd'hui sur les data sciences avec pour objectif de répondre à des verrous dans les différents champs de la santé (1).

La science des données est fortement liée au principe de réutilisation secondaire des données ou « data reuse » qui vise à valoriser et à partager des données pour d'autres usages que ceux pour lesquels elles ont été produites (6). Dans ce contexte, les données utilisées sont généralement observationnelles et rétrospectives, ce qui laisse entrevoir de larges volumes de données, mais non recueillies dans le but premier de répondre à l'objectif visé. Néanmoins, leur atout majeur est qu'elles reflètent les événements qui sont observés en pratique lors de la prise en charge des patients ou des individus. On parle de données de vie réelle, par opposition aux données épidémiologiques ou de recherche clinique. Ces données produites pour la recherche sont en effet artificiellement homogénéisées par les critères de sélection des patients inclus dans les études ou les cohortes et par le mode de recueil des données. L'objectif est d'obtenir des

données avec la meilleure qualité possible pour les besoins d'analyse. Par exemple, l'effet d'un traitement en population peut différer de celui observé lors d'un essai clinique notamment du fait d'une population qui peut être différente de celle étudiée lors de l'essai. De même, un traitement mis sur le marché peut voir ses indications élargies de façon empirique sans que ses effets aient été validés scientifiquement (7). Dans cet exemple, l'exploitation des données de vie réelles peut alors permettre d'évaluer de telles pratiques en matière de balance bénéfice/risque pour les patients, mais également d'un point de vue médico-économique (7).

En ce qui concerne la prise en charge des patients, le champ des data sciences s'inscrit dans le concept de système de santé apprenant décrit en 2007, en lien avec la médecine fondée sur les preuves (8). Le principe général est de s'appuyer sur la masse de données produites aujourd'hui par les systèmes de santé pour améliorer la prise en charge des patients. L'objectif sous-jacent est la médecine personnalisée qui nécessite à la fois les données permettant de décrire finement les patients et des outils permettant d'appréhender cette masse d'informations à la lumière des données de la littérature. L'enjeu est notamment de développer des systèmes d'aide à la décision s'appuyant sur les innovations en intelligence artificielle afin de soutenir les médecins dans cette approche de médecine personnalisée (9). Dans ce cadre, l'exploitation des données pour le soin est en pleine émergence, que ce soit dans le milieu académique ou industriel avec notamment de nombreuses start-up qui se positionnent dans ce domaine, pour tirer parti de cette richesse des données médicales (10,11). Cependant, l'utilisation de tels outils implique d'autres contraintes, notamment en matière de validité et de fiabilité des résultats fournis et donc des méthodologies spécifiques (12).

L'exploitation des données massives est rendue aujourd'hui possible par la mise en œuvre d'infrastructures permettant de décloisonner les données. Ces solutions se situent à tous les niveaux : au niveau national, en France, on peut citer le Système National des Données de Santé (SNDS) qui vise à intégrer diverses sources d'intérêt telles que les données de remboursement de l'assurance maladie (SNIIRAM), les données du Programme de Médicalisation des Systèmes d'Information (PMSI) et les données du registre national de décès (13). Cette dynamique s'étend aujourd'hui à l'ensemble des données de santé via l'initiative du Health Data Hub (14). Aux États-Unis, on peut citer le projet PCORnet (Patient-Centered Clinical Research Network) dont l'objectif est de constituer un large réseau de partenaires en recherche clinique permettant de faciliter la conduite d'études par le partage de données (15). Au niveau local, il s'agit essentiellement des entrepôts de données biomédicales qui sont apparus dans les établissements de santé afin de décloisonner les données du dossier patient et les intégrer dans

des bases de données permettant de les exploiter de façon transversale. Les PACS (Picture Archiving and Communication Systems) d'abord créés pour le soin laissent également entrevoir la possibilité de réutiliser de façon massive les données d'imagerie médicale à différentes échelles (16). Des initiatives transnationales ont également vu le jour, par exemple le projet EHR4CR en Europe qui visait à mettre en œuvre un réseau d'entrepôts de données biomédicales pour faciliter les études de faisabilité et le préscreening des patients dans le cadre des essais cliniques à l'échelle européenne (17,18). Ce besoin d'exploitation à différentes échelles illustre le fait que l'accès et le partage des données sont un préalable indispensable à la science des données (8). De ce constat découle un certain nombre de verrous technologiques et méthodologiques en matière de standardisation des données et d'interopérabilité, mais également de sécurité ou de gouvernance autour des données massives en santé.

Dans ce contexte, l'objectif de cette thèse est d'évaluer au travers de trois cas d'usage, quels sont les enjeux actuels ainsi que la place des data sciences pour l'exploitation des données massives en santé.

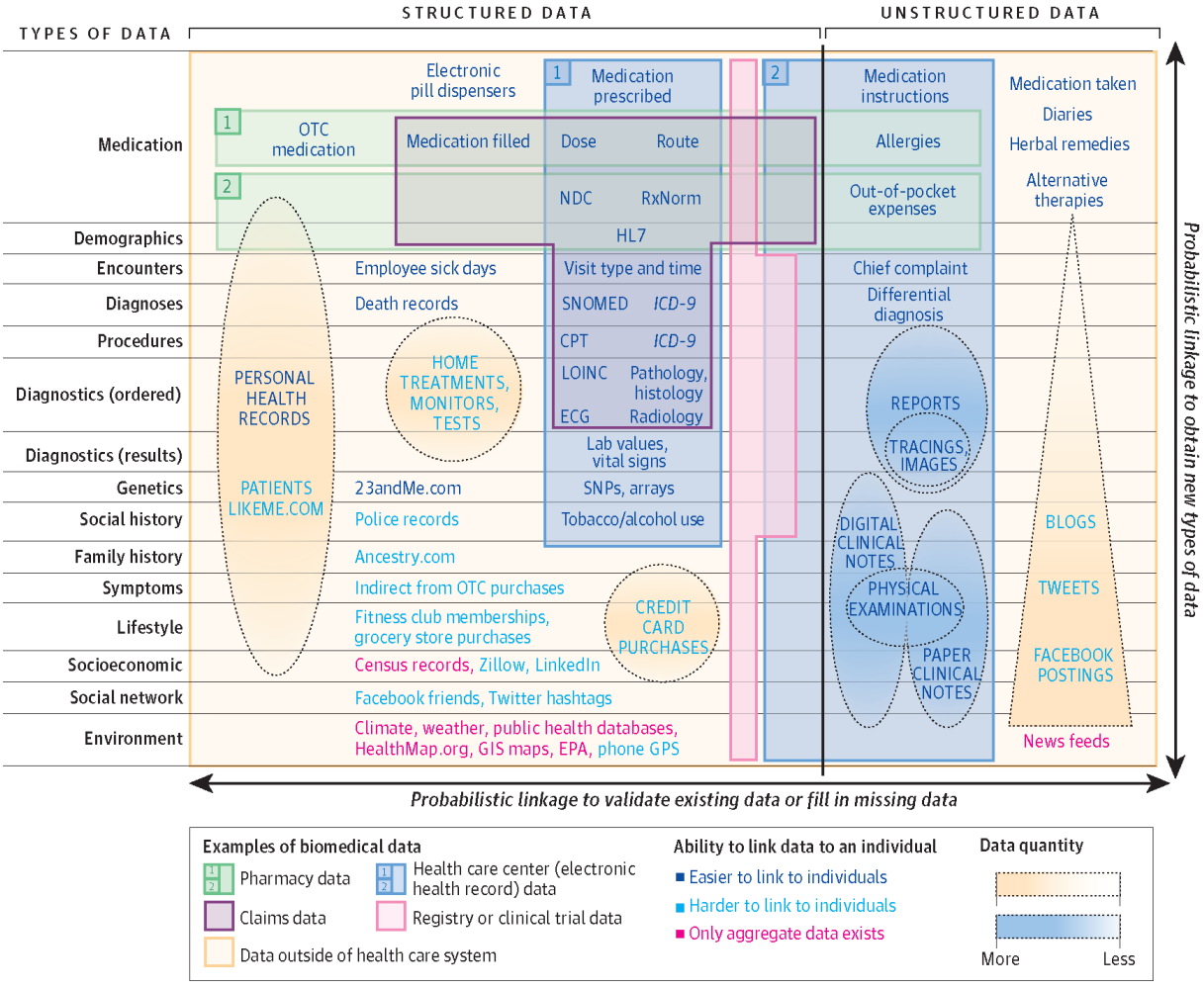
La démarche utilisée pour répondre à cet objectif consiste dans une première partie à exposer les caractéristiques des données massives en santé et les aspects techniques liés à leur réutilisation. La seconde partie expose les aspects organisationnels permettant l'exploitation et le partage des données massives en santé. La troisième partie décrit les grandes approches méthodologiques en science des données appliquées actuellement au domaine de la santé. Enfin, la quatrième partie illustre au travers de trois exemples l'apport de ces méthodes dans les champs suivant : la surveillance syndromique, la pharmacovigilance et la recherche clinique. Nous discutons enfin les limites et enjeux de la science des données dans le cadre de la réutilisation des données massives en santé.

Première partie : les données du big data en santé

I. Les sources de données

On peut distinguer quatre grands cadres de productions de données de santé : la recherche médicale, le domaine médico-administratif, le soin et les patients eux-mêmes, comme illustré sur la figure de Weber, Mandy & Kohane (2014) (Figure 1) (19). À cela s'ajoutent des données non personnelles telles que les données environnementales ou les données agrégées issues de la surveillance en santé publique qui sont généralement disponibles en open data et peuvent s'avérer primordiales dans le contexte des données massives en santé (20).

Figure 1 : Panorama des données massives en Santé. Source : Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. JAMA. 25 juin 2014;311(24):2479-80.



Par nature, les données de santé sont multi-échelles, c'est-à-dire qu'elles peuvent aussi bien concerner l'individu que des populations à l'échelle d'un établissement, à un niveau régional, national, voire international (21).

Ainsi, en recherche, aussi bien les cohortes épidémiologiques que les essais cliniques s'intéressent à des populations dépassant largement celles d'un seul centre. Les essais cliniques requièrent un nombre de sujets à inclure toujours plus important pour démontrer une efficacité ou une équivalence des produits de santé évalués, avec des critères d'éligibilité toujours plus stricts. Ces besoins sont dès lors incompatibles avec des essais monocentriques. Ces essais ciblent pourtant des patients toujours plus spécifiques, rendant difficile leur identification parmi l'ensemble des patients. Il en est de même pour les registres épidémiologiques où la démarche consiste à tendre vers une exhaustivité des cas recensés. L'étude des maladies rares implique également de pouvoir constituer de grandes cohortes permettant d'obtenir la puissance statistique nécessaire pour étudier ces pathologies. C'est dans ce cadre que par exemple, la banque de données nationale sur les maladies rares s'est constitué (22).

Ces collections de données issues de la recherche clinique peuvent donc se situer à une échelle locale, multisite, nationale ou internationale. Ces données sont généralement de bonne qualité, mais leur exploitation obéit à une finalité précise. Ces données sont susceptibles d'être réutilisées et permettent d'identifier de nouvelles hypothèses de recherche en s'appuyant sur des données de bonne qualité (23).

Dans le domaine médico-administratif, les principales données d'intérêt sont collectées à un niveau national, par le biais de la base du Programme de Médicalisation des Systèmes d'Information (PMSI) ou par celui de la base du Système National Inter-Régime de Remboursement de l'Assurance Maladie (SNIIRAM), via des producteurs locaux tels que les établissements de santé, les médecins ou les pharmaciens. Ces bases sont aujourd'hui disponibles grâce au le Système Nationale des Données de Santé (SNDS) (13). Elles sont par exemple exploitées à un niveau régional par les Agences Régionales de Santé (ARS) pour l'organisation de l'offre de soin, ou à un niveau local au sein des établissements de santé, pour leur pilotage.

Les données du soin demeurent, quant à elles, majoritairement cloisonnées à un niveau local, car leur utilisation première concerne le patient. Dès lors, il n'y a pas a priori d'intérêt à les colliger de façon systématique à une échelle plus large, dans le contexte de la prise en charge des patients. Elles sont donc le plus souvent stockées au sein des systèmes d'information des

établissements ou des cabinets médicaux. L'accessibilité technique des données de soin varie en fonction du domaine et du pays. Par exemple, en France, contrairement aux pays de l'Europe du Nord, les données de soins primaires qui sont produites à partir de « logiciels de cabinet » très hétérogènes et donc peu interopérables sont parmi les sources les plus difficiles à mobiliser (24). Le Dossier Médical Partagé (DMP) qui vise à faciliter la continuité des soins entre la médecine de ville et la médecine de spécialité pourrait devenir une source de données en vie réelle intéressante à utiliser, car ce dispositif collectera à terme l'ensemble des informations de trajectoire de soin des patients (24). L'hétérogénéité des données et leur manque de standardisation risquent cependant d'être un frein à leur exploitation.

Plus récemment, des données produites par les patients eux-mêmes sont apparues avec le développement des dispositifs médicaux implantables, des objets connectés ou des réseaux sociaux (25,26). Ces sources permettent de capturer des phénomènes et des comportements jusqu'alors non mesurables avec les sources de données médicales classiques. Enfin, il existe également des sources de données de spécialité, par exemple les données d'échographie en cardiologie ou en obstétrique, les données de capteurs en réanimation, ou encore les données OMICs (27,28).

L'exploitation des données de soin implique donc une étape essentielle de décloisonnement des données depuis leurs sources locales avant de pouvoir envisager leur exploitation. Par décloisonnement, nous entendons (i) l'extraction des données depuis les systèmes sources qui les génèrent (par exemple un DPI, un logiciel de gestion de laboratoire) (ii) l'harmonisation des données afin d'assurer leurs interopérabilités syntaxique et sémantique (iii) leurs intégrations dans des systèmes permettant une exploitation transversale des données (c'est-à-dire avec d'autres sources de données), mais également leur partage à une échelle plus large.

Article 1 : Integrating Biobank Data into a Clinical Data Research Network: The IBCB Project

L'article présenté ci-après illustre ce processus de décloisonnement appliqué aux données de biobanques. L'objectif de ce travail était d'une part de décloisonner les données de biobanques de deux établissements hospitaliers (les CHU de Rennes et de Bordeaux) afin de les intégrer dans des entrepôts de données biomédicales contenant par ailleurs les données clinico-biologiques des patients. Ce travail avait en outre l'objectif de répondre aux enjeux de standardisation et d'interopérabilité pour le partage de ces données afin de faciliter la conduite d'essais cliniques multicentriques nécessitant la disponibilité d'échantillons biologiques.

L’article expose les approches méthodologiques et techniques choisies pour répondre à ces besoins.

Ma contribution a porté sur l’alignement terminologique des données de biobanques et de biologie sur un référentiel commun avec le CHU de Bordeaux puis de réaliser les requêtes sur l’entrepôt de données i2b2 ou eHOP de Rennes permettant de retrouver les patients d’intérêt pour les utilisateurs finaux.

Integrating Biobank Data into a Clinical Data Research Network: The IBCB Project

Guillaume BOUZILLE[a,1], Vianney JOUHET[b], Bruno TURLIN[a], Bruno CLEMENT[a]
Mireille DESILLE[a], Christine RIOU[a], Moufid HAJJAR[b], Denis DELAMARRE[a],
Danielle LE QUILLEUC[a], Frantz THIESSARD[b], Marc CUGGIA[a]

a) Inserm, Univ Rennes, CHU Rennes, Inra, Laboratoire Traitement du Signal et de l'Image (LTSI-U1099), CIC-1414, Centre de Données Cliniques, Nutrition Metabolisms and Cancer (NuMeCan), The liver biobanks network, CRB-Santé, Biosit, Biogenouest, Rennes, France.

b) Inserm, Univ Bordeaux, CHU Bordeaux, Pôle de santé publique, Service d'information médicale, unit IAM, Bordeaux, France

Abstract. Development of biobanks is still hampered by difficulty to collect high quality sample annotations using patient clinical information. The IBCB project evaluated the feasibility of a nationwide clinical data research network for this purpose. Method: the infrastructure, based on eHOP and I2B2 technologies, was interfaced with the legacy IT components of 3 hospitals. The evaluation focused on the data management process and tested 5 expert queries in Hepatocarcinoma. Results: the integration of biobank data was comprehensive and easy. Five out of 5 queries were successfully performed and shown consistent results with the data sources excepted one query which required to search in unstructured data. The platform was designed to be scalable and showed that with few effort biobank data and clinical data can be integrated and leveraged between hospitals. Clinical or phenotyping concept extraction techniques from free text could significantly improve the sample annotation with fine granularity information.

Keywords. biobank, data integration, interoperability, big data, data sharing

1. Introduction

Biobanks operate at the interface between patient care in hospitals and clinical, translational or basic researches. The cooperation and extensive collaboration through a network of multiple organizations are encouraged to enable the streamlined exchange of bio specimens and associated data. As such, biobanks are central for the development of both academic and industrial R&D, which requires an easy access to biological resources and associated data, to generate innovative drugs and biomarkers related to specific diseases [1]. With the development of high throughput genomics and big data systems, even a single experiment with human samples may give rise to huge numbers of hits, whose interest and specificity strictly depends on the quality of the original data linked to the samples. However, the development of biobanks is still hampered by the difficulty to collect and to process human bio specimens based on standards that support quality, regarding storage, phenotyping and clinical annotations including medical, genealogical, and lifestyle information in a biobank. In parallel, Clinical Data warehouse (CDW) technologies and now Clinical Data Research Networks (CDRN) are

coming forward as one of the solutions to address bio clinical data exploitation and data sharing at multiple scales. In these networks, stakeholders provide to the research community a part of their data while maintaining a data-sharing control at any time. In this context, the main objective of the iBCB project (**i**ntegrating **B**iological and **C**linical data for **B**iobank) was designed to explore, through a proof of concept, how semantic integration and CDW technologies could enrich biobanks data and facilitate sharing sparse information, that was up to now compartmentalized into clinical information systems. The aim of IBCB was to design a multi-site platform prototype capable to provide bio clinical information for biobanks in an efficient and secure way. In this paper, we present the technical infrastructure and the evaluation of the platform Hepato-Cellular Carcinoma (HCC), which is a critical research field. Indeed, Chronic liver disease incidence leading to liver cancers is increasing dramatically in the last 20 years worldwide. In France, the incidence of chronic liver disease has increased by 2.5-fold [2].

2. Material and methods:

Definition of the use case: The project involved two academic hospitals (Rennes and Bordeaux) and a Cancer Center (CLCC Bergonié of Bordeaux). To define the road map of the IBCB platform, we interviewed the potential end users (Pathologists and physicians) of the three hospitals, to identify their needs and their functional requirements regarding data reuse. We collected the functionalities and the different queries they would like to perform on the platform to conduct their research. A main requirement was to get the count of patients meeting specific clinical and biological criteria and having one or several samples, in a multi-site fashion. Five examples of queries were provided by users:

Q1: *all* p*atients ha*v*ing sample*(*s*) *A*ND *with* H*CC A*ND *with other non-he*p*atic tumor*
Q2: *all samples o*f *liver tumor o*f p*atient with* H*CC A*ND *with other non-he*p*atic tumor*
Q3: *all* p*atients ha*v*ing liver sample*(*s*) *with* H*CC A*ND *with non-cirrhotic liver*
Q4: *all* p*atients ha*v*ing liver sample*(*s*) *with* H*CC A*ND N*A*SH *syndrome*
Q5: *all* p*atients ha*v*ing liver sample*(*s*) *with* H*CC A*ND *with cirrhotic liver and A*ST > 800 U*I*/L *or ALT > 800 UI/L*

Infrastructure design and technical aspects: One challenge of the project was to design a scalable architecture that could be extended to several hospitals. As a proof of concept we build the architecture (fig. 1) on different CDW technologies. Two types of CDW (eHOP and I2B2) were used: eHOP is a CDW developed by the team of Rennes, which is used in 8 hospitals within the western CDRN of France [3]. I2B2 is an Open Source CDW developed by the Boston university to facilitate the use of the clinical patients' data in the translational informatics [4]. I2B2 is used in Bordeaux as a legative CDW and in Rennes to share structured data coming from eHOP. SHRINE was the third technical component that allowed to distribute query and to compute counts of patients from I2B2 endpoints. All technical components of the infrastructure were hosted within the information systems of the 3 hospitals.

Biobank and clinical data integration: The first step of data integration consisted in developing a specific ETL job to extract, transform and load sample data coming from the legacy biobank softwares (that were different in the 3 sites) in each CDW: TD biobank for Rennes and TumoroteK for Bordeaux. An ontology of data elements taking into account the comprehensive content of information available in the biobank software databases was

commonly defined by the 2 hospitals: the clinical datasets included relevant structured data: ICD-10 and ICD-O diagnosis codes, Procedures Codes (CCAM) and pathology codes (French ADICAP terminology). For lab tests, each site had its own interface terminology, so we used LOINC to map a subset of relevant data elements.

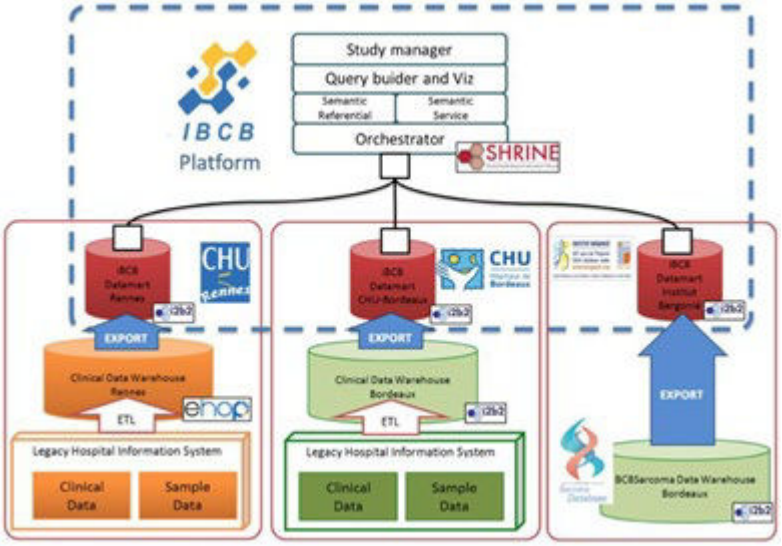


Figure 1: IBCB platform architecture

The second step was to export from the legacy CDWs, clinical and samples data in the I2B2 DataMart intended to be shared between hospitals through the SHRINE query orchestrator.

Validation and evaluation methods: Data management study: this first study was carried out to assess the ETL processes and the data integration at each level of the platform. Counts of data elements were compared from the sources to the target Datamart. Biobank data managers were solicited to evaluate the data quality before and after integration in the legacy CDW and the target Datamart. A random sample of 100 records was compared by the data managers from the two sites. Functional evaluation of the platform: A second study consisted into executing queries provided by the users to test the platform. Such queries were performed on the legacy CDW and on the IBCB datamarts. The objective was to compare the capability of each component to provide consistent and complementary information.

3. Results

Data management study: Table 1 compares from the 2 sites the count of patients and data elements integrated in the CDW:

		Rennes Site	Bordeaux Site
Available Bioclinical data collection in CDW:	- Nb of patients - Nb of bioclinical documents - Nb of related data elements - Period of time	1.2 millions 38 millions 299 millions 1995 to 2017	140,000 10 millions 235 millions 2010 to 2017
Biobank data integrated in CDW:	- Nb of samples / Nb of patient - Nb of data elements	33,074 / 4,958 708,323	18,086 / 13,535 257,552

	- % integrated / biobank software - Period of time	100% 2010 to 2017	100% 2006 to 2017
Data exported to I2B2 shared DataMar	- Nb of patients - Nb of data elements - Period of time	4,958 7,428,426 2010 to 2017	13,535 33,061,726 2006 to 2017

Functional evaluation of the platform: We performed a set of queries to compare the results from the IBCB infrastructure with those coming from the legacy CDW. Table 2 shows the results of the queries at the different stages of the platform.

Query executed on :	Q 1	Q 2	Q 3	Q 4	Q 5
CDW Rennes (eHOP) CDW Bordeaux (I2B2)	34 patients 84 patients	34 samples 128 samples	84 patients 170 patients	3 patients -	30 patients 71 patients
DataMart Rennes (I2B2) DataMart Bordeaux (I2B2)	34 patients 84 patients	34 samples 128 samples	84 patients 170 patients	- -	30 patients 71 patients

The Figures 2 and 3 show the Biobank data representations into eHOP and I2B2 user interfaces. Specifically, eHOP enables visualization of documents and not only data elements. Integration of samples information was fully validated by data managers of biobank software.

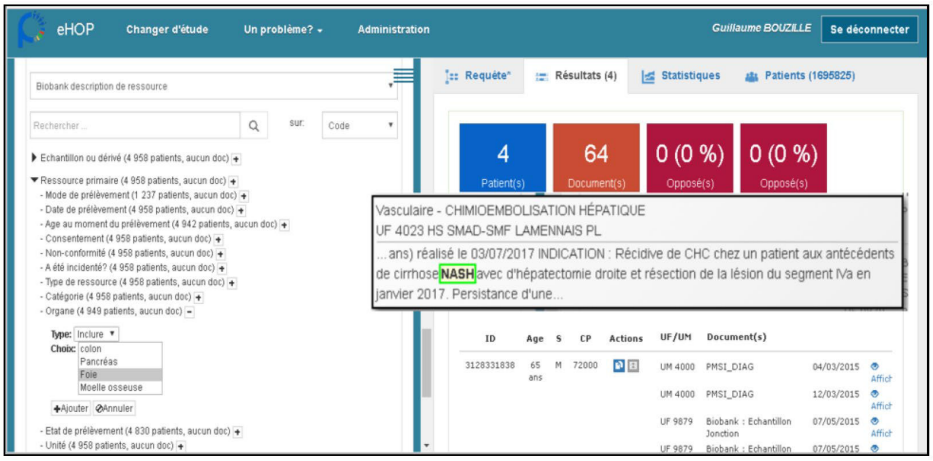


Figure 2: eHOP user interface: hierarchy of biobank data elements. Result including free text research

4. Discussion and conclusion

The aim of the IBCB project was to investigate whether biobank data combined with bio clinical data could be shared with the researcher community at a nationwide level through a flexible and scalable infrastructure. This first attempt successfully showed that such approach is feasible and could leverage existing technologies. This needed few efforts regarding data integration, since biobanking items are quite standardized from one site to the other. The limited scope of bioclinical data used in the project also helped data integration. The collection of data elements was natively encoded with reference terminologies excepted lab tests, which required a manual mapping with the LOINC terminology.

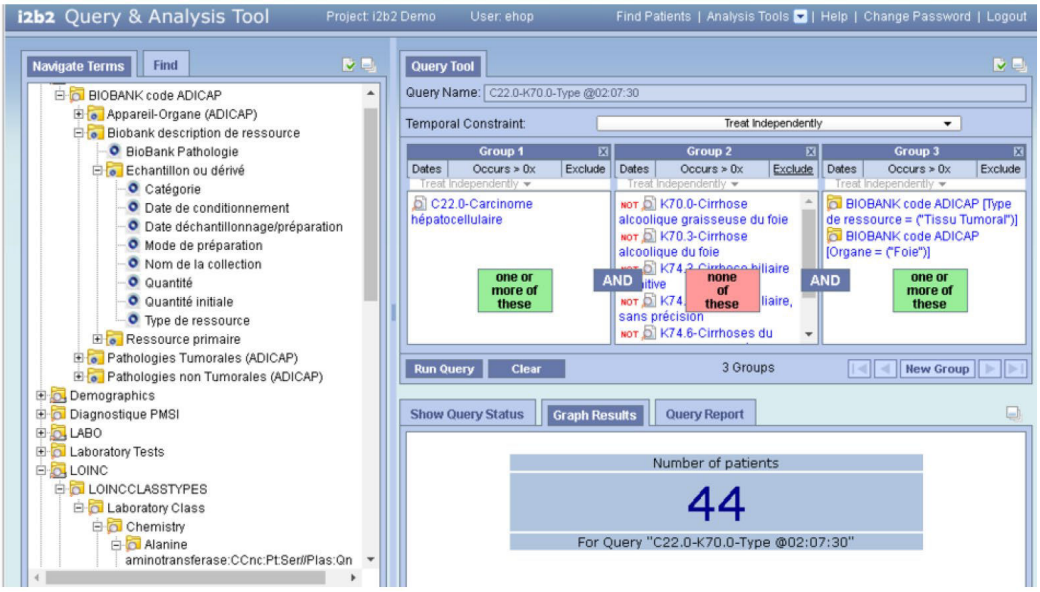


Figure 3: I2B2 User interface with hierarchy of biobank data elements

As a limit, our first experiment queried and shared data from only two sites. However, we used scalable and open source components (I2B2 and Shrine). Query 4 focused to find patient with a NASH syndrome, which turned out to be not currently coded with existing reference terminology. As only structured data was transferred to Datamart, query 4 failed on I2B2 but succeeded with eHOP (eHOP has the advantage to natively enable advanced information retrieval on both structured data and free text documents). Even if free text queries generate noise, from the user's perspective, the workload to manually review the cases returned was negligible compared to the benefit. This shows that inferring phenotypes from unstructured data is crucial to answer user needs with technologies like I2B2. Future works will focus to deploy the IBCB platform on a larger number of hospitals and to provide at a national level the existing and new services such as pre-screening functionalities, deep phenotyping [5], and data export to populate target databases such as epidemiologic registries or cohort databases.

Acknowledgements

The IBCB project was funded by the following French national infrastructures: IBiSA and BIOBANQUES

References

[1] S. Sarojini, A. Goy, A. Pecora, and K.S. Suh, Proactive Biobanking to Improve Research and Health Care, *J. Tissue Sci. Eng.* **3** (2012). doi:10.4172/2157-7552.1000116.

[2] Santé publique France. Etat de santé de la population en France : rapport 2017, (http://www.santepubliquefrance.fr/Actualites/Etat-de-sante-de-la-population-en-France-rapport-2017 (accessed February 22, 2018).

[3] D. Delamarre, G. Bouzille, K. Dalleau, D. Courtel, and M. Cuggia, Semantic integration of medication data into the EHOP Clinical Data Warehouse, *Stud. Health Technol. Inform.* **210** (2015) 702–706.

[4] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, and H.C. Chueh, Integration of clinical and genetic data in the i2b2 architecture, *AMIA Annu. Symp. Proc. AMIA Symp.* (2006) 1040.

[5] W.-Q. Wei, and J.C. Denny, Extracting research-quality phenotypes from electronic health records to support precision medicine, *Genome Med.* **7** (2015) 41.

II. Caractéristiques des données

Du fait de la très grande diversité des sources de données et des différents contextes dans lesquels elles sont produites, les données de santé sont extrêmement hétérogènes. Il est intéressant de faire une analyse de leurs caractéristiques selon différents axes afin de guider la façon dont ces données doivent être intégrées en vue de leur réutilisation.

A. Types de données

De façon générale, on distingue les données non structurées, les données semi-structurées et les données structurées. Le premier type correspond par exemple aux données textuelles telles que les comptes rendus d'hospitalisation, de consultation, d'imagerie. Bien qu'il n'y ait pas d'étude permettant de l'affirmer, il est généralement admis que 80 % des données patients informatisées d'un établissement sont non structurées (29). Un autre exemple de données non structurées très répandues en médecine est les images. Les données non structurées peuvent cependant être accompagnées de métadonnées qui permettent d'appréhender le contexte de la donnée. Pour l'imagerie médicale, le standard DICOM a vocation à remplir ce rôle (30). Les données semi-structurées correspondent généralement à des données représentées dans un langage à base de balises tel que XML (eXtensible Markup Language). De ce fait, les données peuvent être décrites par des attributs qui peuvent faciliter leur structuration. Des exemples de fichiers semi-structurés sont les questionnaires médicaux ou tout autre document stockés au format Clinical Document Architecture (CDA) du standard HL7 (31). Enfin, les données structurées sont, elles, décrites avec un référentiel permettant de leur apporter une sémantique et ainsi faciliter leur exploitation ou analyse. Cette description peut être standard et alors partagée par plusieurs producteurs ou locale, ce qui complique l'interopérabilité des systèmes les produisant les données (32).

B. Volume des données

Les données peuvent également être décrites selon leur volume, qui dépend de plusieurs paramètres :

- le type de données concernées, c'est-à-dire un élément atomique (entier, flottant) ou un ensemble d'éléments atomique (image, texte, séquence ADN, etc.).
- Le rythme auquel les données sont collectées
- le nombre d'individus concernés par cette collecte.

C. Variété des données

La variété des données tient au fait que pour une même source de données, celles-ci peuvent être de formats très différents. Par exemple, un séquençage ADN peut avoir été fait avec différents séquenceurs, les données d'imagerie peuvent avoir été acquises de façon très différente. Les textes peuvent être dans des formats différents ou décrire de façon différente une même chose. La variété s'explique également par l'extrême diversité des domaines les produisant et leur format de production.

D. Temporalité des données

Un autre axe décrivant les données est leur temporalité : un recueil répété de données peut parfois permettre de les représenter sous forme de séries chronologiques. C'est notamment le cas pour les mesures physiologiques pouvant être réalisées chez les patients. On parle alors de données de signaux qui peuvent donc être définies par leur fréquence d'acquisition. Les données de biologie ont également une temporalité qu'il peut être intéressant d'exploiter afin d'évaluer l'évolution des paramètres biologiques. La notion de temporalité peut également intéresser une échelle plus large par exemple pour la constitution de trajectoires de soin. Cette temporalité peut parfois se trouver dans des données non structurées et être décrite de façon relative. La reconstitution des trajectoires de soins est donc une tâche relativement complexe.

E. Finalité des données

La finalité pour laquelle les données sont produites a également un impact sur leurs caractéristiques et conditionne leur qualité. En effet, le niveau d'exigence en matière de qualité n'est pas le même dans le cas d'un essai clinique ou du soin, pour une même information. De manière générale, dès lors qu'une source de données est constituée dans un but d'analyse, les données stockées sont majoritairement structurées : les données d'essais cliniques ou de registres en recherche, les données du PMSI dans le domaine médico-administratif. À l'inverse, la précision des données peut différer en fonction des besoins. Ainsi, les données à visée médico-administrative n'ont pas les mêmes besoins en matière de description médicale des patients que le soin. Dans le contexte de la réutilisation de ces données, il est primordial de prendre en compte ces aspects, car des données décrivant de façons différentes une même information vont devoir potentiellement être réconciliées.

F. Qualité des données

Plusieurs définitions existent pour définir la qualité des données. La plus simple et la plus communément admise est « fitness for use » (33) : la qualité dépend des exigences vis-à-vis de ce que l'on veut en faire. De nombreuses dimensions existent pour caractériser la qualité des données. Ces dimensions sont étroitement liées à la finalité et les données ne peuvent raisonnablement pas être de qualité sur l'ensemble des dimensions et donc pour l'ensemble des besoins (34). Des exemples fréquemment rencontrés sont les données manquantes, des données en doublon, le délai de production de la donnée ou encore l'invalidité de la donnée. Dans le contexte de la réutilisation secondaire des données, les usages sont définis après que les données aient été produites. Outre les critères dépendants de l'usage, les données pour être de bonne qualité doivent a minima respecter certains critères de base décrits par les principes « FAIR » : Foundable, Accessible, Interoperable, Reusable (Existantes, Accessibles, Interopérables et Réutilisables) (35). Plusieurs moyens permettent par ailleurs d'agir sur la qualité des données qui seront réutilisées :

- Mettre en œuvre des mesures de surveillance de la qualité tout au long du processus d'intégration des données, afin de s'assurer de ne pas dégrader les données brutes durant le processus d'intégration depuis les sources.
- Mettre en œuvre des méthodes d'analyse pour corriger les problèmes de qualité de données (réconciliation, dédoublonnage, etc.)
- Appliquer des actions correctrices à la source, ce qui est parfois facilité par le fait que les utilisateurs finaux sont également les producteurs de données.

La réutilisation secondaire implique de définir les dimensions d'intérêt en matière de qualité des données, par rapport aux usages prévus afin de mettre en place les indicateurs permettant d'évaluer et de surveiller la qualité des données.

III. Données massives

Par leurs caractéristiques, les données de santé peuvent répondre à la définition classique des données massives ou « Big Data » : volumétrie, variabilité, véracité, vélocité ou encore leur valeur (2). Les données massives peuvent également se définir par les moyens technologiques nécessaires pour les traiter, c'est-à-dire que les moyens classiques de stockage (bases de données relationnelles) et de calcul ne sont plus suffisants et que l'emploi de technologies de stockage et de calcul distribués ou de supercalculateurs s'impose (36). Une définition originale

basée sur une revue de la littérature médicale a également été proposée par Baro, Degoul, Beuscart et Chazard (2015). Ils définissent les données massives en santé ou « Health Big Data » comme étant des jeux de données respectant la formule : $Log(n \times p) \geqslant 7$ (avec n le nombre d'individus statistiques et p le nombre de caractéristiques les décrivant) (37).

On peut considérer que les critères de véracité, de vélocité ou de valeur concernent potentiellement tout type de données et sont surtout dépendants de l'usage qui est prévu. La volumétrie et la variabilité s'appliquent différemment selon le type de données concernées. Par exemple, les données d'imagerie ou « OMICs » respectent le critère de volumétrie, mais moins la variabilité. À l'inverse, les données du dossier patient électronique sont extrêmement variables, mais ne représentent qu'un volume modéré à l'échelle d'un établissement. Quoi qu'il en soit, les méthodes de stockage et d'analyse sont à adapter en fonction du caractère massif des données qui seront à exploiter.

IV. Stockage des données

En matière de réutilisation de données, deux types de stockage sont rencontrés aujourd'hui. Il existe d'une part les architectures classiques d'entrepôt de données avec un modèle de données en étoile et reposant donc par définition sur des technologies classiques de bases de données relationnelles, qui sont apparues dans le domaine médical à partir des années 2000. Le modèle de données en étoile est simple avec une table dite de faits contenant les données observées et des tables de dimensions toutes reliées à la table de faits et permettant l'analyse des faits selon différents axes. La technologie d'entrepôt de données la plus répandue est i2b2, développée par l'université de Harvard et aujourd'hui utilisée partout dans le monde (22). D'autres initiatives existent cependant : STRIDE (23) de la Maio Clinic ou eHOP (24) qui est une technologie d'entrepôt de données biomédicales développée par notre équipe « Données Massives en Santé » du Laboratoire Traitement du Signal et de l'Image (LTSI). L'entrepôt eHOP est orienté « Document », c'est-à-dire que la table des faits contient l'ensemble des documents des patients. Les données structurées provenant des documents correspondent à une table de dimension tout comme les caractéristiques patients ou les séjours. Ces solutions demeurent efficaces notamment pour les données structurées ou les données non structurées dès lors qu'une indexation plein texte des documents est possible et lorsque la volumétrie de données demeure raisonnable. À titre d'exemple, l'entrepôt eHOP du CHU de Rennes contient aujourd'hui les données de 1,3 million de patients, c'est-à-dire 60 millions de documents et 360 millions

d'éléments de données structurées. Le temps de réponse des requêtes est le plus souvent de moins de 1 minute.

En revanche, le passage à l'échelle au-delà d'un certain volume de données à stocker et surtout interroger, ou encore le stockage de données de type imagerie ou d'autres flux de données volumineuses en temps réel peut s'avérer compliqué. De ce fait, les approches de réutilisation des données biomédicales tendent à exploiter les nouvelles technologies de stockage issues du Big Data. Le concept de base est la notion de « Data Lake » qui permet d'entreposer les données de façon massive dans un format brut (38). Les « Data Lake » s'appuient généralement sur les technologies Apache Hadoop, notamment le format de stockage de fichiers distribué HDFS (39). Dans cette approche, l'interrogation et l'exploitation des données doivent quant à elles être réalisées via d'autres briques logicielles, en fonction des besoins. Ainsi, dans l'approche entrepôt de données, on parle d'approche « schema on write » puisque les données sont chargées dans un modèle prédéfini tandis que dans l'approche lac de données, on parle d'approche « schema on read », car les relations entre les données sont définies, si nécessaire, au moment de l'accès aux données. Il en résulte certaines problématiques liées au temps de réponses nécessaires pour résoudre ces jointures, mais qui peuvent être résolues par l'emploi de technologies adaptées. De même les procédures d'intégration de données diffèrent puisque l'on parle de processus « ETL » (Extract, Transform and Load) dans le cas des entrepôts de données alors qu'il s'agit de processus « ELT » (Extract, Load and Transform) pour les « Data Lake » (40).

Cependant, les approches qui semblent les plus prometteuses sont les approches hybrides, qui s'appuient lorsque cela est adapté sur des technologies d'entrepôt de données, notamment pour le stockage d'index, permettant d'accélérer l'interrogation des données. Ces bases peuvent être relationnelles ou issues des technologies « NoSQL » (Not Only SQL) et peuvent éventuellement être stockées en RAM pour accélérer encore l'interrogation de données (41). Les bases de données orientées « Graphe » paraissent également avoir leur place dans le domaine médical du fait du nombre important de terminologies décrivant les données (42). Ces bases de données permettent de précalculer les relations entre les concepts ou encore de gérer les versions des terminologies.

Deuxième partie : organisations pour le partage et l'exploitation des données massives en santé

Comme nous l'avons vu, les données massives en santé et leur exploitation sont multiéchelles. Les initiatives de structuration de l'activité d'exploitation des données fleurissent à la fois au niveau local jusqu'au niveau national. L'objectif de cette partie est donc de décrire les différents modes d'organisation pour l'exploitation des données de santé qui ont pu être explorés au cours de cette thèse.

I. À l'échelle d'un établissement

L'avantage majeur d'un établissement est la proximité des producteurs de données et donc de l'expertise nécessaire pour structurer, exploiter de façon intelligente ces dernières, mais également pour fournir les problématiques pouvant être étudiées par l'usage de ces données. Les producteurs sont également souvent les utilisateurs finaux ce qui facilite leur implication dans un projet de réutilisation des données et permet de mieux cerner les usages et besoins vis-à-vis des données. Il s'agit du premier niveau de décloisonnement à mettre en place, le plus souvent indispensable. Dans ce contexte, l'implication de la direction des systèmes d'information de l'établissement est indispensable afin de mettre en place les flux de données depuis les applications métiers vers l'entrepôt de données cibles.

Au-delà des outils technologiques, l'exploitation des données d'un établissement ne peut s'envisager qu'avec la mise en place d'une gouvernance et d'une organisation permettant d'assurer et de gérer l'activité nécessaire à l'exploitation des données. Au CHU de Rennes, nous avons structuré l'activité d'exploitation des données de l'établissement au sein du Centre de Données Cliniques (CDC), qui vise à rassembler les compétences à la fois techniques et méthodologiques nécessaires à cette activité. Les CDC qui ont été initialement décrits au CHU de Brest visent également à proposer une offre de service aux cliniciens et chercheurs de l'établissement pour faciliter leur activité de recherche.

II. À l'échelle de plusieurs établissements

Le mode d'organisation sous forme de CDC s'est aujourd'hui diffusé dans les 6 CHU de la région Grand-Ouest et 2 centres de lutte contre le cancer. Ceci a permis de constituer à l'échelle de l'interrégion le réseau des CDC (RiCDC) qui partagent donc les mêmes outils technologiques d'entrepôt de données et la même organisation.

Ce réseau permet aujourd'hui d'envisager la réalisation d'études multicentriques, mais également le partage de données. En effet l'uniformisation des technologies résout une part de l'hétérogénéité des données et facilite donc la mise en commun de celle-ci, bien que des verrous liés aux référentiels terminologiques demeurent. Le RiCDC bénéficie également d'un pilotage interrégional par le GCS HUGO ce qui permet à l'ensemble des acteurs du réseau d'avoir une ligne directrice commune pour l'exploitation des données des établissements.

III. Partage de données

L'exploitation des données au niveau local présente plusieurs avantages indéniables tels que la proximité des utilisateurs. Cependant deux limites peuvent être discutées.

Premièrement, le volume de données produit au niveau local ne permet pas de répondre à toutes les problématiques. Les essais cliniques multicentriques sont aujourd'hui la norme. Les études monocentriques, bien qu'encore très répandues, ne constituent le plus souvent que des études pilotes ou exploratoires. En pharmaco-épidémiologie, les signaux d'intérêt ne peuvent être mis en évidence que sur de larges populations. De façon plus générale, les méthodes développées en data science, telles que l'apprentissage profond, requièrent un nombre d'individus conséquent qui vont bien au delà de ceux rencontrés au sein d'un unique établissement. Une initiative de partage de données permet également de croiser des données multidomaines qui ne sont plus restreintes à celles d'établissement de santé, par exemple les registres épidémiologiques ou encore les données du SNDS.

Deuxièmement, il existe une potentielle redondance des initiatives locales sur le territoire. Les différents besoins d'exploitation sont partagés et les efforts réalisés pour y répondre peuvent être mutualisés à la fois en matière de partage de données et de développement de méthodes d'analyse. Le coût de ce partage est évidemment un effort en matière d'interopérabilité avec

par exemple la création ou la mise en place de terminologies communes, l'alignement vers des terminologies standards ou l'utilisation d'un modèle commun de données tel que OMOP (43).

Plusieurs arguments permettent d'envisager une exploitation des données à une échelle multicentrique (6,44) :

- Des solutions techniques ont déjà été développées à l'échelle internationale. On peut citer notamment SHRINE qui permet de connecter et d'interroger de façon distante un réseau d'entrepôts de données biomédicales reposant sur la technologie i2b2 (45). Cela reste cependant une approche fédératrice qui ne permet pas forcément la mutualisation des moyens.
- Une proximité géographique des centres dans le cas d'une organisation régionale ou interrégionale permettant de respecter une logique de territoire.
- L'implication des acteurs dans les sociétés savantes tant en informatique que dans les différentes spécialités médicales ce qui permet de faciliter les synergies pour le développement d'initiatives d'exploitation des données. On peut citer par exemple l'effort en cours, de partage des données radiologiques par la communauté de radiologie française (46).

L'intérêt du partage de données peut également résider dans l'intégration de données provenant d'autres sources que les établissements hospitaliers, qu'elles soient d'ordre médico-économique, de la recherche ou issues du soin.

Plusieurs aspects doivent cependant être pris en compte avant d'envisager ce type de partage :

- une nécessité de réconciliation des identités par des méthodes d'appariement ce qui peut nécessiter le recours à des méthodes probabilistes (47).
- Assurer un haut niveau de sécurité autour de données toujours plus sensibles, ce qui est défini aujourd'hui par le Règlement Général de la protection des données (RGPD) comme une bulle de sécurité.

Article 2 : Sharing health big data for research - A design by use cases: the INSHARE platform approach

Cet article décrit un prototype d'architecture permettant une intégration et un partage de données hétérogènes multisources multiéchelles pour la recherche, développé dans le cadre du projet ANR Inshare. Il vise dans un premier temps à proposer l'architecture permettant de lever

les verrous liés à l'exploitation de ces données ainsi que des méthodes innovantes en lien avec la traçabilité de l'exploitation des données. Dans un second, une évaluation de l'architecture est réalisée dans le cadre de plusieurs cas d'usage dans les domaines de l'épidémiologie, de la surveillance syndromique et de la pharmaco-épidémiologie.

Ma contribution dans ce projet a été de participer à la définition de l'architecture de la plateforme, de décrire les méthodologies pour répondre aux cas d'usage concernant la surveillance syndromique et la pharmaco-épidémiologie et de réaliser les traitements de données sur la plateforme Inshare.

Sharing health big data for research - A design by use cases: the INSHARE platform approach

Guillaume Bouzillé[abc], Richard Westerlynck[cd], Gautier Defossez[e], Dalel Bouslimi[fg], Sahar Bayat[h], Christine Riou[ac], Yann Busnel[d], Clara Le Guillou[g], Jean-Michel Cauvin[g], Christian Jacquelinet[j], Cécile Vigneau[ikl], Patrick Pladys[b], Emmanuel Oger[h], Eric Stindel[g], Pierre Ingrand[e] , Gouenou Coatrieux[fg], Marc Cuggia[abc]

[a] *INSERM U1099, LTSI, Université de Rennes 1, Rennes, F-35000, France*
[b] *CIC Inserm 1414, Rennes F-35000, France,*
[c] *Centre de Données Clinique, CHU Rennes, F-35000, France*
[d] *Institut Mines-Telecom Loire Atlantique, IRISA, Université Bretagne Loire, France*
[e] *Registre général des cancers de la région Poitou-Charentes, Poitiers, France.*
[f] *Institut Mines-Télécom; Télécom Bretagne,*
[g] *INSERM 650 Latim, Brest, France,*
[h] *EHESP Rennes, Sorbonne Paris Cité, France,*
[i] *Cellule regionale d'appui épidemiologique du registre REIN, Rennes, F-35000 France,*
[j] *Agence de la biomédecine, Saint Denis, F-93212 France.*
[k] *CHU Pontchaillou, service de nephrologie, Rennes, F-35000, France,*
[l] *INSERM UMR 1085-IRSET, Rennes, F-35000 France.*

Abstract

CONTEXT: Sharing and exploiting efficiently Health Big Data (HBD) lead to tackle great challenges: (1) data protection and governance taking into account legal, ethical and deontological aspects which enables a trust, transparent and win-to-win relationship between researchers, citizen and data providers (2) lack of interoperability: data are compartmentalized and are so syntactically and semantically heterogeneous (3) variable data quality with a great impact on data management and statistical analysis. The objective of the INSHARE project is to explore, through an experimental proof of concept, how recent technologies could overcome such issues. It aims at demonstrating the feasibility and the added value of an IT platform based on Clinical Data Warehouse, dedicated to collaborative HBD sharing for medical research. **METHOD:** The consortium includes 6 data providers: 2 academic hospitals, the SNIIRAM (the French national reimbursement database) and 3 national or regional registries. The platform is designed following a 3 steps approach: (1) to analyze use cases, needs and requirements (2) to define data sharing governance and secure access to the platform (3) to define the platform specifications. **RESULTS:** 3 use cases (healthcare trajectory analysis, epidemiologic registry enrichment, signal detection), corresponding to 5 studies and 11 data sources, were analyzed to design the platform. The governance was derived from the SCANNER model and adapted to data sharing. As a result, the platform architecture integrates the following tools and services: Data repository and hosting, semantic integration services, data processing, aggregate computing, data quality and integrity monitoring, Id linking, multisource query builder, visualization and data export services, data governance, study management service and security including data watermarking.

Keywords: Data sharing, Information Storage and Retrieval, Registries/statistics & numerical data

Introduction

In its broad definition, Health Big Data (HBD) is more than just a very large amount of data or a large number of data sources. It also refers to the complexity, the challenges and the new opportunities presented by the combined analysis of data. Health data collected or produced are now potentially sharable and reusable. They can be exploited at different levels and across different domains, especially concerning questions related to multidisciplinary research. This huge amount of data holds the promise of supporting a wide range of medical and health care functions, including among others clinical decision support, disease surveillance or population health management [1]. This explains the incentive policy of opening HBD around health data science being supported by different public authorities and scientific communities, such as OpenData, AVIESAN or Inserm initiatives, as well as European research programs like IMI or Horizon 2020. Recently, strong initiatives have been launched in the U.S to enhance the utility of health Big Data and finally to enter in the next level of knowledge discovery [2].

In this context, clinical data warehouse (CDW) technology comes forward as one of the solutions to address HBD exploitation. CDW, which are becoming increasingly widespread in the US, are being put to use for different purposes such as cohort discovery, biomarker detection, feasibility studies or the enrolment of patients in clinical trials. Research communities are currently connecting CDW to one another with the aim of creating Clinical Data Research Networks (e.g. PCORNET [2]) or biomedical research network (e.g. Data to Knowledge).

In these networks, data providers such as researchers, health facilities, research agencies or institutions provide a part of their data available to the research community while maintaining data-sharing control at all time. Thus, these trusted third-party platforms integrate and open-up scientific or potentially scientific health data [3]. This makes the use of these data at a large-scale possible. In France, such a platform, that would be able to integrate and share multisource and multiscale big and small health data produced by health institutions for research purposes, does not exist.

This is the aim of the INSHARE French national project,in which different and very actual key issues like the governance as well as the organizational and technical factors to perform a such data sharing will be explored and addressed. The absolute goal is to facilitate access to data and to foster collaborative research and data sharing between researchers and data providers. In this paper, we present and discuss these key issues as well as the approach we are following so as to design the platform. This approach is driven by real research use cases of high interest for individuals as well as for states.

Background

The data to share: In its large acceptance, HBD sources comprise various types of data from structured information such as OMICS data, administrative or billing data, drug prescription data consisting of dates and dosages that are captured through a standardized ePrescription

system, to unstructured and textual data such as clinical narratives that describe the medical reasoning behind prescriptions [4]. Beyond the data generated by hospitals, several health data sources come from health registries or insurance databases, which are a valuable source of standardized, longitudinal, population-wide data. For instance, the French health reimbursement database (Système National d'Information Inter-Régimes de l'Assurance Maladie, SNIIR-AM) contains individualized, anonymous and comprehensive data for all health spending reimbursements received by affiliated subjects. This includes basic patient demographic data such as age, gender and medical drugs, as well as outpatient medical cares, prescribed or performed by health professionals from both public and private practices. The SNIIR-AM is also linked via a unique personal health number to the French hospital discharge database (PMSI), which contains diagnostic codes, medical procedures and admission dates for all hospitalizations. Data from the SNIIR-AM is increasingly used for research projects, especially related to the detection of adverse drug effects in epidemiology or in clinical research. The platform governance models was derived from the SCANNER model and adapted to the data sharing.

The sharing barriers: The regulatory hurdles obstructing optimal use of data for research have been extensively discussed within specialised literature [5]. The identified factors are characterized by (i) an over-cautious approach among data custodians, many of whom are unwilling to link or share data, (ii) legislators' failure to consider the flexibility required to allow and support such linking and sharing and (iii) the incorporation of 'good governance' models or intelligent design of working instances thereof are not contemplated within the regulatory framework, nor is there any reflection on the subject [5-6]. Sethi proposes a model for data-sharing governance including (i) guiding principles and best practices (ii) a safe, effective and proportionate governance, (iii) an articulation of the roles and responsibilities of data controllers and data processors and (iv) the development of a training program for researchers that covers appropriate vetting procedures prior to sharing valuable data.

The cornerstone of data sharing and reuse is trust. Therefore, implementing a trustworthy process for handling citizens' and patients' health data is a pivotal goal. Based on the definition of a trusted relationship, one party (trustor) is willing to rely on the actions of another party. In addition, the trustor abandons (full) control over the actions performed by the trustee. As a consequence, a trustworthy system is that in which the trustor can "place his/her trust and rest assured that the trust will not be betrayed". A system for data reuse should thus prove its trustworthiness by fulfilling the responsibility of dealing with data within the limits of a social contract regulated by policies between citizens and the organizations handling the system.

The technological components behind a trustworthy system involve designing and implementing IT tools and services capable of guaranteeing data quality and security while providing interoperability, adaptability and scalability. Specific projects funded by the EU and by the IMI initiative [7], such as EHR4CR, are dealing with such challenges, with the prospect of defining use cases, tools, technologies and a business model for data reuse. In particular, the EHR4CR business model includes accreditation and certification plans for EHR systems that can be integrated within a system for data reuse. The purpose of data reuse has implications that belong to the realm of policies and regulations, which are essential aspects for establishing trust.

How to manage informed consent is one of the key aspects connected to this issue. In fact, the current regulations in many European countries, which are similar to the US, with the HIPAA act, assume that consent (implied or explicit) for use of data is strictly limited to the purpose for which data were collected. This may seriously limit the scope of data analysis.

This theme needs to be reconsidered in the light of the existence of a proper, trustworthy system based on an agreement between citizens and healthcare organizations. Specific practical examples of policies for handling data reuse are provided by regional initiatives in Europe, two such cases being the United Kingdom and Catalonia. ISO/TS 14265:2011 provides a classification of different purposes for processing personal health information that can help make policy formulation more granular.

Methods

To design the organizational and technical dimension of the INSHARE platform, an iterative and 4 steps bottom-to-top approach has been adopted, by analyzing on the ground, existing needs, use cases and actual difficulties encountered by the project partners. Five partners are involved in the project as data providers: 2 academic hospitals (CHU Rennes and Brest), which provide datamarts from their Clinical Data Warehouse (eHOP-CDW), 3 epidemiologic registries at a regional or national scale.

This approach aims at defining technical and functional specifications, data protection policies and governance for an efficient and valuable data sharing. Furthermore, this approach takes into account the fact that some technological issues have to be addressed and especially the evolution needs for data analysis and security tools in the scaling-up to HBD.

Step 1 - To Describe use cases and user needs: The aim of this step is to define precisely scenarios from an operational perspective, the information workflow and the system/actor interfaces that relate to the exploitation of health and research data via the INSHARE platform. The relevant scenarios leverage the richness and the variability of the data sources hosted by the platform in terms of domain, quality and origin. Herein, the objective is to identify the functional needs, which are expected by the different users of the platform: researchers-users, data providers as well as the internal operators of the platform.

Step 2 - To define data sharing governance and secure access to the platform: Regarding the ethical, legal and deontological aspects, a focus group composed of domain experts and representatives of patient associations conducts this study. According to the state-of-the-art step and the specified use cases defined at the first step, the objective is to establish governance guidelines guaranteeing data protection and individuals' privacy rights. This step includes the submission of these guidelines for validation to institutional and regulatory authorities such as the Comité consultatif sur le traitement de l'information en matière de recherche (CCTIRS) and the Commission Nationale de l'Informatique et des Libertés (CNIL), two French authorities in charge of such regulation aspect.

Step 3 - To define INSHARE platform specifications: The aim of this step is to define a comprehensive description of the intended purpose and environment of the platform. These

specifications describe what the software does and how it will be expected to perform, taking into accounts the operational scenarios, security aspects and the stakeholders (users, data providers, and data managers) inputs. It also addresses some key issues in relation with data analysis and security. In terms of data protection in the scaling up, a special interest is given to the data traceability and on how to give back some control to data providers on the data they make available to researchers. On one hand, users have to know of their action accountability and, in another hand, patient or data provider consent for data exploitation duration has to be guaranteed. Database watermarking, a very recent solution [8], is one of the technology actually explored for those purposes.

Each data provider is part of the INSHARE project to bring their knowledge and experience with their respective data. Data providers are thus responsible for supplying necessary data to the platform in order to answer to the use cases. They have to supply as well, all necessary information about data to correctly perform their integration and to subsequently give the capability to the platform to provide the best-suited data for each user request.

Results

Use Cases: Three main use cases corresponding to 5 studies and application domains have been identified and chosen to be performed on the platform. Being able to ensure one of them will be of great interest for cares of individuals as well as for populations. Table 1 illustrates for each use case the sources of data which will be shared and used in the different INSHARE platform studies.

Use case	Study & application domain	Data sources
Health care Trajectory analysis	Pre and post-dialysis care trajectory of end-stage renal disease patients starting dialysis in emergency Characterizing the healthcare trajectories of children (and their mother) included in Birth Defect Registry	Kidney Failure registry (REIN) SNIIR-AM Birth defect registry eHOP-CDW
Registry enrichment	Assessment of association between cancer incidence and diabetes in end-stage renal disease patients	Kidney Failure registry (REIN) Cancer registry SNIIR-AM
Signal detection	Influenza surveillance Adverse drug effect surveillance	eHOP-CDW SNIIR-AM Sentinel Network Open Data

Table 1. Use cases and application domain

For instance, regarding the study on health care trajectories of end-stage renal disease patients, comorbidities are currently collected and registered in the REIN database at the initiation of the

renal replacement therapy (RRT). But the occurrence of comorbidities after RRT started is not a mandatory field of REIN [9]. Moreover, no information on prescribed treatments is available in the REIN database. Through the INSHARE platform, the hospital CDWs (Rennes and Brest) will be used to collect comorbidities and expensive drug prescriptions while drug exposures will be extracted from SNIIR-AM in order to enrich the REIN registry with accurate comorbidities and medications including standards and dates of occurrence.

Platform governance: The governance determines how the range of controls and procedures with contractual obligations work together so as to ensure an end-to-end secure and trustfully platform, where the security and reliability of data is guaranteed. Indeed, INSHARE use cases imply to perform Id linkage processing and require the access to identified data (e.g., Epidemiologic registry enrichment). Moreover, all use cases require to aggregate data coming from multiscale institutions (local academic hospitals, regional to national registries, and for the SNIIR-AM, data issued from a nationwide database). Some of them imply intensive computation on big volume of data (e.g., signal detection). All these constraints have led to define a model of governance for the platform adapted to big data sharing. The model we propose is derived and adapted from the Distributed Scalable National Network for Effectiveness Research (SCANNER). It consists in the identification of 10 basic requirements: platform and data provider information, institution information, study information, ethical agreements (coming from an independent IRB-like comity), Data Sharing Agreement (from data providers who are involved in the study), approved users (external users and internal operators of the platform), authentication and access, data use, audit and accounting, patient rights, data segregation). In addition, according to the type of personal data, it includes de-identification, data watermarking, individual access, correction, openness/transparency, individual choice, use and disclosure limitation, integrity, accountability, safeguards. Some requirements can be met by technology and some others by contracts, attestation of users, or management supervision.

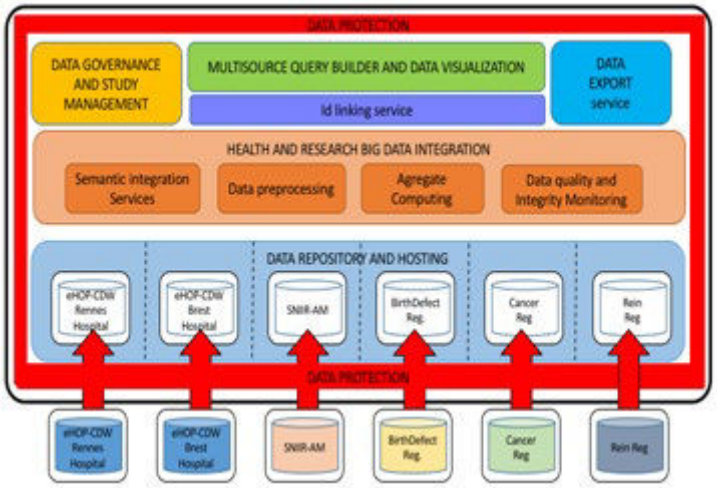


Figure 1: Uses case under the scope of the Inshare Platform

Platform Design: As a result, we designed the platform architecture shown in figure 1. This architecture is oriented to meet the different use cases under the scope of the project, and to perform the expected data processing while respecting the governance framework mentioned

above. The platform encompasses services of several weakly coupled components, the whole in a cloud oriented architecture. Hereby, we detail some of the key components and services:

The data repository and hosting component is a buffer zone where data providers make available the required datasets or datamart to share. The core idea is to host in the platform the data with the finest granularity and in their most original form, (i.e., with the less transformation possible).

The health big data integration layer comprises the components and services dedicated to data integration and processing. The semantic integration service (SIS) contains information models (i.e., database schemas of the data sources such as eHOP or SNIIR-AM) as well as semantic resources either used in the sources or required for semantic integration (reference or interface terminologies, ontologies and mappings) and ensures standards' interoperability (such as HL7, PN13, HPRIM). SIS provides tools and methods to the other standard components.

Data preprocessing service is devoted to data enrichment. It includes NLP tasks and data indexing to make easier extraction of useful information from large-scale data stored across the different INSHARE sources. A core functionality of these services is to execute data processing tasks and to query the data virtualization layer in order to access stored data. The developed engine is also responsible for the planning, coordination and execution of queries to the data virtualization layer in a distributed manner commercial data processing frameworks and parallel relational database management systems.

Aggregate computing service is designed for building online auxiliary indexing and summarization structures based on the incoming data processing tasks and their data requirements. Based on profiling and statistic information of the submitted processing tasks. For instance this service is used to compute from CPOE data, aggregates of drug dose per day, week, stay or globally for a patient or a population.

Data quality and integrity monitoring: The INSHARE platform deals with data sources having heterogeneous data quality, from EHRs to epidemiologic registries. Integration process has thus to manage such quality disparities. This component is dedicated to compute metrics to monitor data quality during the integration process. These metrics are useful (i) to alert data providers and to take corrective actions at the data source, (ii) to perform more accurate analysis taken into account possible bias due to data quality issues, (iii) to improve data quality within the INSHARE platform, each source bringing complementary information. For instance, for a same patient, in and out hospital drug information come from different sources and are registered in different ways (structured and coded data from CPOE or SNIIR-AM, text for clinical charts, forms and notes). One source can provide more accurate or exhaustive information to the others.

Id Linking service: For security reasons, in France, as in most countries, there are currently no patient identifiers that can be used to directly link data from different data sources. Nonetheless, several national programs or initiatives provide researchers either trusted third-party linkage services, or big, pre-linked datasets. For instance, this is the case of the French hospital discharge database (used as part of the hospital billing system) that matches data coming from all hospitals in France. The SNIIR-AM is arguably one of the most noteworthy linked data

sources recently opened to the research community. The Id Linking service reuse and provide methods to link the data sources using deterministic and probabilistic approaches on common data elements. For instance DRG data coming from hospital are already linked with the other data of the SNIIR-AM. EHOP CDW includes the DRG data. Even without specific common Id, linking can be performed using dates, groups of diagnosis and procedure codes, and ADT mode.

Data Governance, study management service and security: These services encompass tools, procedures [10] and workflow to cover the governance requirements and to provide continuous data protection, from their acquisition to their outsourcing and mutualization within the INSHARE platform and beyond (e.g., when exported). The idea is to complement current data protection, which mainly relies on the security of the information system, which do not make possible to know if data are used for the purposes originally foreseen, especially when data are outsourced. The protection of digital content we deploy is based on watermarking [8] and crypto-watermarking solutions (i.e., mechanisms that combine encryption and watermarking [11]) that fulfill different security objectives, especially in terms of integrity and traceability (identification of information-leak sources or end-user misbehavior). If data are protected as long as they are not decrypted, watermarking leaves free access to them and maintains them protected by means of security attributes (e.g., digital signatures, users' ID, access rights) invisibly inserted or embedded into the data themselves. Moreover, watermarking protection is independent of the data storage format. These data protection tools are designed to take into account strong interoperability constraints so as to: i) provide security resilient to information-processing; ii) make the protection on the data provider's side compliant with the one used by the INSHARE platform and beyond.

Multisource Query Builder, Visualization and data export services: These services are intended to: design and perform complex queries on multisource data; visualize results with different modalities; and, export processed dataset to the end users. From the end user point of view, the interaction with the platform consists in submitting a request for a study to the platform. Only certified and authorized operators of the platform will be able to access to the query workbench for data exploitation and eventually to export the required datasets to the end user.

Figure 2 and 3 illustrate the workflow for two scenarios. In the first one (fig 2), a targeted research database is fed by data extracted both from two registries (REIN and Cancer playing the role of data provider) according to a study protocol about the association between kidney failure and occurrence of cancer. This protocol, which defines criteria for data selection, variables to be extracted from the source, and user agreement are submitted to the platform. Figure 3 illustrates how the platform is able to enrich a data source (here the Rein registry) by collecting from a list of patient ID, missing or required data (e.g., comorbidities) from the different sources. In this scenario Rein registry is a user of the platform and recipient of the data. All along this process, health data is maintained secured by means of digital content protection tools, with a special interest for data traceability and audit trails.

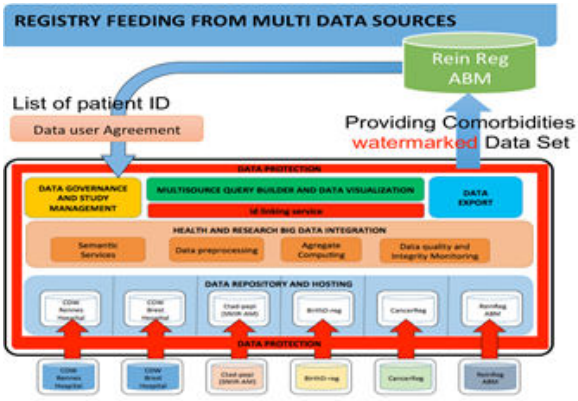


Figure 2: workflow for ad-hoc study and register enrichement

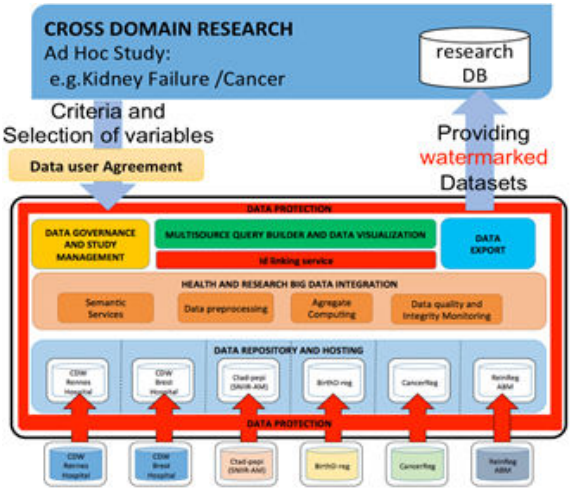


Figure 3 : workflow for register enrichement

Discussion and conclusion

The INSHARE project's consortium made the choice to focus on some of today's crucial challenges, which, in our opinion, are still not resolved: data quality assessment for research purposes, scalability issues when integrating heterogeneous health "big data" or patient data privacy and data protection. Moreover, adoption of electronic health data is still an active process and the way to exploit these data is rapidly evolving. In a second time, the use of health data still requires developing to reach maturity: the aforementioned barriers have to be raised before we can obtain more significant knowledge and practical consequences. In any case, the project's use-cases were selected in view of their potentially important impact on clinical knowledge and health research, and we believe they could help demonstrate the benefits of secondary use of data. From a methodological point of view, our approach goes beyond the today's most cutting-edge secondary-use technologies. We propose a combination of innovative algorithms for cryptography and watermarking supported by big data technologies, with the aim of enhancing content digital protection. Compared to other projects mainly oriented to distributed architecture, we also follow a different strategy based on a trusted third-party oriented on health-data sharing within a community of users.

Lead by the large amount of heterogeneous data available among the consortium, traditional approaches for data organization and analysis (using classical SQL server such as I2B2) are no longer efficient and quite overwhelmed. For instance, the Rennes CDW contains by itself raw data of 1.6 million patients. Recent Big Data technologies are filling the gap of this evolution that requires real time analysis, low latency, data mining, and heterogeneous unstructured data treatment. Born from the needs of scalability, fault tolerance and interoperability, challenging frameworks come out as Hadoop or Cassandra. However, big data technologies are an evolving landscape. The INSHARE project is a great opportunity to test and evaluate combination of disruptive technologies and provide improved analysis performance in the perspective to meet

real-world use cases and users needs, on real and massive data. For instance, preliminary tests on adverse drug effect detection have been carried out. In this example, the combination of OrientDB (which is a graph oriented database) with SPARK for aggregate computing have shown promising performance for intensive computing. Nonetheless, applying existent solutions should not be sufficient in the background of the Inshare project. Indeed, starting from the available massive datasets, a second objective aims to design innovative algorithms and techniques in a prospective way (using a data sciences approach, sharing the statistical and computer sciences skills

Acknowledgements

We would like to thank the French National Research Agency (ANR), for funding this work inside the INSHARE (INtegrating and Sharing Health dAta for Research) project (grant no. ANR-15-CE19-0024).

References

[1] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Health Inf. Sci. Syst., vol. 2, no. 1, p. 3, 2014.

[2] Patient-Centered Outcomes Research Institute (PCORI), "PCORnet | National Patient-Centered Clinical Research Network." [Online]. Available: http://www.pcornet.org/. [Accessed: 23-Apr-2015].

[3] M. D. Natter et al., "An i2b2-based, generalizable, open source, self-scaling chronic disease registry," J. Am. Med. Inform. Assoc., vol. 20, no. 1, pp. 172–179, Jan. 2013.

[4] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," Nat. Rev. Genet., vol. 13, no. 6, pp. 395–405, May 2012.

[5] J. Peto, O. Fletcher, and C. Gilham, "Data protection, informed consent, and research," BMJ, vol. 328, no. 7447, pp. 1029–1030, May 2004.

[6] N. Sethi and G. T. Laurie, "Delivering proportionate governance in the era of eHealth," Med. Law Int., vol. 13, no. 2–3, pp. 168–204, Jun. 2013.

[7] M. Goldman, "The Innovative Medicines Initiative: A European Response to the Innovation Challenge," Clin. Pharmacol. Ther., vol. 91, no. 3, pp. 418–425, Mar. 2012.

[8] J. Franco-Contreras and G. Coatrieux, "Robust Watermarking of Relational Databases with Ontology-Guided Distortion Control," IEEE Trans. Inf. Forensics Secur., vol. accepted, 2015.

[9] C. Couchoud et al., "The renal epidemiology and information network (REIN): a new registry for end-stage renal disease in France," Nephrol. Dial. Transplant., vol. 21, no. 2, pp. 411–418, Feb. 2006.

[10] G. Bouzillé et al., "An Integrated Workflow For Secondary Use of Patient Data for Clinical Research," Stud. Health Technol. Inform., vol. 216, p. 913, 2015.

[11] D. Bouslimi and G. Coatrieux, "A crypto-watermarking system for ensuring reliability control and traceability of medical images," Signal Process. Image Commun., vol. 47, pp. 160–169, Sep. 2016.

Address for correspondence:

Pr Marc Cuggia,

UMR LTSI (Inserm), Faculté de médecine. Rue du Pr Léon Bernard. 35043 Rennes Cedex. marc.cuggia@univ-rennes1.fr

IV. Centralisation des données

La centralisation des données peut avoir des avantages indéniables. L'harmonisation des procédures, la constitution de larges bases de données sont des atouts majeurs de cette approche. Plusieurs exemples illustrent les succès de cette approche. Par exemple, la base nationale du PMSI, le Système National des Données de Santé ou encore les registres nationaux en épidémiologie qui outre leurs objectifs premiers permettent de mener de nombreuses études scientifiques. Les succès précédemment cités s'expliquent par l'incitation politique dans lesquels ils s'inscrivent (T2A, Sécurité sociale, Politiques de santé), ainsi que le champ limité des données qui sont intégrées.

Néanmoins, on peut observer que les données une fois centralisées deviennent extrêmement difficiles d'accès pour les experts en capacité de les exploiter, alors même que certains d'entres eux en sont les producteurs. Cela est dû d'une part aux contraintes réglementaires et de sécurité nécessaires pour ces sources de données sensibles et d'autre part à une gouvernance centralisée qui s'avère détachée des problématiques de terrain.

Les données massives s'inscrivent dans une dynamique différente puisqu'il s'agit de ne pas restreindre le périmètre des données d'intérêt, mais au contraire d'élargir le champ des données pouvant être exploitées. Cette dynamique est soutenue au niveau national avec l'annonce du projet de « Health Data Hub » dont l'objectif est de répondre aux besoins de partage de données, de rassemblement des expertises à la fois publiques et privées et d'implication des citoyens (14). La politique de gouvernance et l'évolution des contraintes réglementaires vont donc être décisives pour que ce dispositif puisse répondre aux enjeux de l'exploitation des données massives en santé.

Troisième partie : méthodologie en data sciences pour l'exploitation des données massives en santé

I. Méthodes de traitement des données

Le traitement des données intervient à toutes les étapes de la vie des données. Deux démarches peuvent être distinguées :

- le prétraitement des données, qui consiste à traiter des données plus ou moins brutes, mais dans un état ne répondant pas en l'état à un besoin défini. Cela consiste donc à transformer, synthétiser, et préparer la donnée. C'est finalement une étape qui produit de nouvelles données utilisables que l'on peut assimiler à une information consolidée. Ce processus est à distinguer des processus ETL qui ne produit pas de données nouvelles, mais rend compatibles les données avec un format cible. Néanmoins, les processus ETL peuvent embarquer un certain nombre de procédures permettant de consolider les données, mais sans être guidés par un besoin précis. Ces prétraitements s'avèrent primordiaux, car ils répondent aux problématiques de variabilité et/ou de volumétrie en réduisant généralement ces dimensions, mais aussi en palliant dans la mesure du possible les éventuels problèmes de qualité. Par exemple, les méthodes de traitement automatique du langage peuvent être utilisées pour passer de données non structurées textuelles à des données structurées pouvant être exploitées par d'autres méthodes. Un autre exemple est le phénotypage à partir des données qui visent à produire une information consolidée à partir de sources de données hétérogènes au sein desquelles l'information peut être morcelée, redondante, incohérente (48).
- Le traitement des données répond quant à lui, selon les termes réglementaires, à une finalité. Il s'agit de traiter les données de façon à répondre à un objectif (une question de recherche par exemple, la production d'un modèle prédictif) et donc de tirer une connaissance à partir de l'information. Il n'en reste pas moins que les résultats de ces traitements peuvent également produire des données, ayant leur cycle de vie et donc suivre à leur tour un circuit de traitement de données dans le cadre de la réutilisation secondaire des données.

Le traitement de données est donc un processus itératif générant de nouvelles données à chaque étape. Conserver les données originelles apparaît essentiel afin d'être en mesure par exemple d'adapter la chaîne de traitement des données à toutes les étapes. Le concept de « Data Lake » prend ici tout son sens. Cela permet à la fois d'assurer la reproductibilité des traitements, mais également l'amélioration éventuelle de l'information produite par l'application de nouvelles méthodes.

Les traitements peuvent concerner différents niveaux de granularité allant de la donnée elle-même jusqu'à une population d'individus et qui définissent le niveau d'agrégation qui sera utilisé. La finalité peut tout aussi bien concerner la médecine personnalisée qu'un objectif de santé publique, le point commun étant que les méthodes utilisées nécessitent un ensemble d'individus en entrée. Les méthodes employées peuvent donc être retrouvées indifféremment dans différents cas d'usage, à partir du moment où les données auront été préparées de façon adaptée pour répondre à la finalité. Nous développons ici les grands types de méthodes pouvant être rencontrées dans la réutilisation secondaire des données, quelle que soit la finalité.

A. Recherche d'information

Les méthodes de recherche d'information consistent à ramener à l'utilisateur l'information pertinente considérée comme déjà présente dans les données. Ce sont des méthodes dites expertes, puissantes et rapides pour peu que l'interrogation des données soit faite de façon pertinente. Il s'agit de constituer une requête pouvant mobiliser des données (ou métadonnées) structurées (via des terminologies) et/ou non structurées (via des mots-clés). L'objectif en recherche d'information est d'aboutir à des requêtes précises et exhaustives (minimiser le bruit et le silence). Pour aller plus loin, il est également possible de hiérarchiser les documents retournés en fonction de leur pertinence, mesurée par des indicateurs de pertinence des documents tels que le tf-idf ou la c-value. Cependant, dans le cas de recherche s'intéressant à des patients à partir de documents ou nécessitant de mêler différents types de données, il est souvent plus difficile de hiérarchiser les résultats.

Au final, la recherche d'information permet de mobiliser aisément l'information potentielle contenue dans les données par une indexation pertinente relativement agnostique de la finalité. Ainsi, l'indexation des documents permet indirectement d'identifier des patients pertinents lorsqu'ils sont décrits par des documents contenant l'information d'intérêt. La recherche d'information peut aussi tirer avantage à la fois de l'indexation sémantique des données et du

traitement automatique du langage qui permettent de prendre en compte par exemple la synonymie ou la certitude associée aux mots-clés (49,50).

Un exemple concret de l'intérêt de la recherche d'information est l'identification de cas similaires dont l'objectif est d'être en mesure d'identifier des patients présentant des symptômes similaires à un individu dont on ne maîtrise pas soit le diagnostic soit la prise en charge thérapeutique. Dans ce contexte, ces cas similaires peuvent permettre de faciliter la prise en charge de ce patient non pas par rapport aux données de la littérature, mais par rapport à des cas similaires dignes d'intérêt (51). On voit ici apparaître l'intérêt du partage de données puisqu'il permet d'accéder à des cas similaires qui ne seraient pas présents dans l'établissement qui prend en charge le patient.

B. Analyse statistique et fouille de données

Il existe un continuum concernant les méthodes issues du champ des statistiques, en passant par l'analyse de données jusqu'à la fouille de données. L'objectif commun est de mettre en évidence de nouvelles connaissances à partir de jeux de données plus ou moins volumineux et hétérogènes. On parle généralement de processus de découverte de connaissances à partir des données (Knowledge discovery in databases) (52,53). Ces méthodes reposent sur une analyse numérique des données qui doivent donc nécessairement être structurées. On peut distinguer deux grands types d'approches : les méthodes non supervisées et les méthodes supervisées. Dans le premier cas, il n'y a pas de connaissance a priori de l'appartenance des individus statistiques à des groupes prédéfinis. Il s'agit alors d'employer des méthodes permettant d'identifier de tels groupes à partir des données décrivant les individus. En ce qui concerne les approches supervisées, des groupes ou une mesure quantitative d'intérêt sont considérés comme une cible à expliquer. Il s'agit alors d'identifier les liens (ou corrélations) entre les variables décrivant les individus et cette cible. En fonction du type de cible, on parlera de classement (classification en anglais) ou de régression. Comme en épidémiologie, ces associations ne traduisent pas nécessairement de relations causales, mais permettent de générer des hypothèses.

En pratique, des situations intermédiaires peuvent se rencontrer : il arrive fréquemment en médecine que l'on connaisse l'appartenance à un groupe pour une fraction d'individus seulement, les individus restants étant de groupes inconnus. Des approches semi-supervisées peuvent alors être employées pour permettre de classer les individus restants.

C. Apprentissage automatique

Les méthodes d'apprentissage automatique peuvent paraître se chevaucher avec les méthodes de fouille de données. Il y a peu de données dans la littérature définissant clairement les différences entre ces deux approches. La principale distinction est le fait que la fouille de données est étroitement liée à l'utilisateur qui effectue cette démarche, généralement exploratoire. Il s'agit généralement d'un processus itératif mêlant algorithmes et interprétation. Autrement dit, ce qui relève de la fouille de données est davantage la mise en évidence de relations que la prédiction d'événements.

L'apprentissage automatique va plus loin dans l'autonomie des algorithmes et leur place dans la prise de décision, quitte à reposer sur des relations complexes pour aboutir aux conclusions. L'objectif est avant tout de développer des méthodes prédictives que ce soit pour des tâches de classement ou de régression. Les avancées récentes en matière d'apprentissage automatique laissent une place privilégiée à l'apprentissage automatique reposant sur des modèles « boîte noire » et en particulier l'apprentissage profond basé sur les réseaux de neurones. L'apprentissage profond se distingue principalement de l'apprentissage automatique classique par sa capacité à détecter automatiquement les caractéristiques d'intérêt par rapport à la tâche d'apprentissage (54).

Ces méthodes prennent un essor important aujourd'hui avec de nombreux domaines d'applications que ce soit pour le prétraitement des documents par des méthodes de traitement automatique de la langue ou pour répondre à une finalité médicale tels que la détection de pathologies, la prédiction du risque de décès ou la prédiction de la durée de séjour.

Un vaste champ en apprentissage automatique concerne cependant le développement de méthodes permettant d'interpréter et de comprendre comment un algorithme a abouti à sa conclusion. Dans le domaine médical, ces méthodes sont primordiales, car les professionnels de santé sont dans l'obligation de pouvoir être en mesure de comprendre les propositions de l'algorithme afin d'expliquer les décisions aux patients (55).

Un autre champ très important est l'implication de l'expert dans le processus d'apprentissage machine, c'est-à-dire d'intégrer l'avis de l'expert dans le processus d'apprentissage. Cela permet de faciliter les situations où l'apprentissage est complexe, mais également de faciliter les situations où le nombre de données disponibles est limité. On parle alors d'apprentissage par renforcement.

II. Échelle du traitement

L'échelle du traitement dépend elle aussi de l'objectif du traitement. Le plus souvent, l'unité d'intérêt est le patient. D'autres situations impliquent une unité différente, soit à une échelle plus fine, soit à une échelle plus large. Par exemple en recherche clinique, trois cas de figure peuvent se présenter :

1. À l'échelle de l'élément de données : l'alimentation d'eCRF (electronic Case Report Form) porte sur l'extraction et la structuration de données à partir du dossier patient pour remplir les champs du formulaire. Le traitement va donc porter sur la recherche d'information à l'échelle de l'élément de données, sur un sous-ensemble de patients.
2. À l'échelle du patient : cette fois, les résultats attendus correspondent bien à des patients avec l'ensemble des données les caractérisant, par exemple dans le cadre de l'identification de patients éligibles à un essai clinique.
3. À l'échelle populationnelle : il s'agit ici de l'étude de faisabilité, correspondant dans ce cas à une agrégation simple des résultats pouvant être rendus à l'étape précédente.

En fonction des cas d'usage, d'autres unités peuvent être pertinentes : en traitement automatique du langage, on s'intéressera à des documents, dans le cadre du PMSI à des séjours.

Le traitement peut également nécessiter d'agréger les données. Il peut s'agir d'agrégation simple pour passer de documents à des patients par exemple, mais il est parfois nécessaire d'ajouter d'autres axes d'agrégation notamment spatio-temporelle. Ainsi en surveillance syndromique l'objectif est la production d'indicateurs d'incidence hebdomadaire d'activité épidémique.

L'échelle du traitement implique un écueil potentiel important de la réutilisation des données. Partants du patient, nous pouvons descendre vers une unité plus fine tels que le séjour, le document ou la donnée élémentaire qui peuvent faire l'objet d'un traitement. Nombre d'analyses s'intéressant à des données volumineuses concernent un nombre restreint de patients : données d'imagerie, données de séquençage. Pourtant la finalité est bien la prise en charge du patient. Ce type d'analyse se trouve nécessairement limité quant à leur capacité de généralisation en population. On entrevoit ici tout l'intérêt d'un partage de données multicentriques pour pallier ce manque de puissance statistique et appréhender la variabilité qui peut exister entre ces sites en matière de données.

III. Sécurité et traçabilité

Les gisements de données pouvant être traitées sont du fait de leur richesse extrêmement sensibles. La description fine des patients apportée par les données font qu'elles constituent rapidement un marqueur individuel et ne peuvent jamais être considérées comme anonymes, sauf à les dénaturer de telle sorte qu'elles perdraient leur intérêt et ne permettraient plus de répondre à la finalité du traitement. L'agrégation des données à un niveau supra-individuel, quand elle peut répondre au besoin, peut permettre de s'affranchir de cette sensibilité. Il faut toutefois rester vigilant, car les méthodes d'agrégation peuvent parfois rapporter des agrégats individuels notamment quand on s'intéresse à des phénomènes rares. Un exemple est la représentation géospatiale des patients en clusters qui peut permettre d'identifier des patients lorsque ceux-ci sont situés dans des zones peu habitées.

Des mesures doivent donc être mises en œuvre dans tous les cas pour assurer le respect de l'utilisation des données dans le cadre réglementaire adéquat.

En matière de sécurité, la réutilisation des données implique tout d'abord un stockage sécurisé des données. Cela sous-entend un stockage au sein des établissements de santé ou chez des hébergeurs agréés de données de santé.

La seconde mesure consiste en un contrôle de l'accès à ces données de façon stricte grâce à des méthodes d'authentification et l'usage d'infrastructures de gestion de clés.

Les aspects de traçabilité des accès et des traitements des données sont également essentiels. Il s'agit de pouvoir à tout moment connaître qui a accédé à quelle donnée, quel traitement a été fait et dans quel contexte. Le rôle des entrepôts de données ou des plateformes de réutilisation des données est ici essentiel grâce à la centralisation des accès aux données, ce qui facilite la traçabilité. À l'inverse, la centralisation des données engendre des conséquences toujours plus importantes si les mesures de protection des données sont violées. Les méthodes de tatouage de données prennent leur placent ici puisqu'elles permettent d'assurer la traçabilité des données dans le cas d'accès légitime aux données afin de s'assurer que celles-ci ne sont pas exploitées à d'autres fins que la finalité prévue et surtout par d'autres personnes à qui les données auraient été transmises.

Article 3 : Clinical Data Warehouse Watermarking: Impact on Syndromic Measure

L'article ci-après a pour objectif d'évaluer l'impact des méthodes de tatouage de données sur les résultats de l'exploitation de données massives hospitalières dans le cadre de la surveillance des épidémies grippales. Ces méthodes reposent sur une altération mineure des données afin d'y insérer un motif permettant de connaître l'origine des données. L'amplitude de cette altération conduit à des biais plus ou moins importants concernant les indicateurs produits à partir des données, qu'il s'agit de quantifier.

Ma contribution a été de définir les données devants être tatouées afin d'assurer une traçabilité satisfaisante puis de réaliser les traitements de données en fonction de différents paramètres de tatouage afin d'en évaluer les impacts.

Clinical Data Warehouse Watermarking: Impact on Syndromic Measure

Guillaume BOUZILLE[abcd], Wei PAN[e], Javier FRANCO-CONTRERAS[ef], Marc CUGGIA[abcd], and Gouenou COATRIEUX[e]

[a]*INSERM, U1099, Rennes, F-35000, France* [b]*Université de Rennes 1, LTSI, Rennes, F-35000, France* [c]*CHU Rennes, CIC Inserm 1414, Rennes, F-35000, France*

[d]*CHU Rennes, Centre de Données Cliniques, Rennes, F-35000, France*

[e] *Institut Mines-Telecom;Telecom Bretagne; Latim Inserm UMR1101, Brest, France.*

[f]*WaToo, Brest, France*

Abstract. Watermarking appears as a promising tool for the traceability of shared medical databases as it allows hiding the traceability information into the database itself. However, it is necessary to ensure that the distortion resulting from this process does not hinder subsequent data analysis. In this paper, we present the preliminary results of a study on the impact of watermarking in the estimation of flu activities. These results show that flu epidemics periods can be estimated without significant perturbation even when considering a moderate watermark distortion.

Keywords. Traceability, Clinical data warehouse, Watermarking

1. Introduction

The widespread adoption of Electronic Health Records (EHRs) and the emergence of clinical data warehouses (CDWs) offer nowadays great opportunities to share and reuse patients' data for secondary purposes, such as medical research [1]. This implies using appropriate infrastructures that comply with the policies of privacy and data protection dedicated to patients' data. In this context, secondary use of health data requires a governance dedicated to regulate access to data and technologies ensuring reliable access matching this governance. A typical organization offering this kind of service is made of two parts: first, a CDW that embodies the infrastructure allowing to efficiently reuse data; a regulatory board that supervises ethical and legal aspects of studies that aim to reuse patient's data. The conditions to be fulfilled to access patient data for secondary purposes depend on countries' legislation but it remains on several unshakable criteria: parsimony, deidentification, authorization and traceability.

Nowadays, most CDWs technologies, such as i2b2, STRIDE or other custom systems are consistent with these requirements [2-4]. However, organizations are now in the way to share data in order to reach a new scale concerning sample size or geographical coverage. Such exploitation requires new centralized or distributed large scale platform (e.g SHRINE or EHR4CR [5-6]) complying with the regulatory. In this context, new levels of requirements for data security and patient privacy have to be provided. Indeed, patient's datasets are likely to be exposed outside healthcare organizations increasing the risk for malicious data collection. One key security measure consists in ensuring the traceability of

datasets and data processing. Watermarking is a promising approach allowing the embedding of a message containing traceability information (e.g., user identity, date of access) into the database by slightly modifying some of its attributes' values. Watermarking provides free access to the data while keeping it protected through the embedded message and it is complementary to other security mechanisms already deployed. Nevertheless, it is necessary to ensure that the distortion resulting from the watermark does not impair the interpretation or any subsequent data analysis [7].

In this paper, we evaluate a database watermarking method to ensure traceability data sharing in the context of epidemiologic trends analysis. More precisely, we evaluate the impact of watermarked data on the production of flu activity estimates.

2. Methods

2.1. Dataset to be Watermarked

We extracted from eHOP (the CDW of the academic hospital of Rennes [8]) all flu PCR tests that were performed on patients between January 1, 2011 and February 13, 2016. We considered all PCR tests carried out, regardless of whether the result was negative or positive. The aim was to get a signal connected with influenza-like illness (ILI) symptoms and not only with the flu. This datasource is known for having a high correlation with standard ILI estimates provided by the french Sentinel network, at a regional and national scale.

2.2. Watermarking Algorithm and Parameterization

In order to deploy this test, we implemented the scheme developed in [9], the advantage of which is that it injects a constant distortion. More clearly, this one embeds a sequence of bits into the values of an integer attribute *At* of dynamic range [*min*, *max*] by adding the quantity Δ to approximately a half of its values and $-\Delta$ to the others. The length of the message and its robustness to database alterations (i.e., the capability to retrieve the message) depend on Δ. The greater Δ, the more robust or the longer the message can be (for more details see [9]). This watermarking algorithm was applied on the date of PCR tests' realization with 28 different Δ, ranging from 1 to 28 days. For each Δ, we replicated the watermarking procedure 27 times to assess the variability of its induced distortion. We thus created 756 different watermarked datasets to be compared with the original one. The embedded message held 100 bits.

2.3. Statistical Analyses

For each dataset (original and watermarked ones), we computed weekly ILI activity incidences: every PCR date was modified to match the monday of the same week, according to the ISO 8601 for representation of dates. We then counted PCR realisations according to their modified dates.

The evaluation of the distortion induced by the watermark procedure relied on two indicators. First we used the Pearson's correlation coefficient (PCC) to assess how watermarked data could produce ILI activity indexes associated with the original one. Second, we used the Normalized Root Mean Squared Error (NRMSE) to measure how ILI incidences computed with watermarked data deviate from the original ones. For each Δ, we performed a non-parametric bootstrap procedure with 1,000 replicates of the original sample of 27 watermarked datasets. We used the bootstrap replicates to compute 95% confidence intervals for PCC and NRMSE estimates.

To assess the effect of watermarking on the decision making process, we also applied the statistical model used in routine practice by the french sentinel network to detect influenza epidemic periods: the Serfling periodic regression model [10]. We assessed differences in epidemics periods detected with watermarked data and reference data. All analyses were performed on R (version 3.3.1).

3. Results

We extracted 10,555 PCR tests between January 1, 2011 and February 13, 2016, from the CDW of CHU-RENNES. They were performed on 2,965 different patients. The watermarking process applied on this dataset took approximately 2 minutes, a time that strongly depends on database server read/write performance. The entire set used in the study contained 756 watermarked datasets. The dates of reference data were distributed on 267 weeks, that is to say we produced 267 weekly ILI activity estimates. Watermarked data produced 267 to 271 weekly ILI activity estimates depending on the strength of the distortion parameter. We calculated PCCs and NRMSE on the period shared by the reference and watermarked data. PCC decreased from 0.99 (95% CI: 0.99; 0.99) to 0.64 (95% CI: 0.64 ; 0.64) for a time shift between 1 day and 28 days (Table. 1). NRMSE increased from 3.5 (95% CI: 3.5; 3.5) to 19.4 (95% CI: 19.3 ; 19.5) respectively for a Δ value of 1 day to 28 days (Table 1).

Table 1. Pearson's correlation coefficients and Normalized Root Mean Squared Errors for Δ=[1,7,14,21,28]

Δ (days)	Pearson's correlation coefficient (95% CI)	Normalized Root Mean Squared Error (95% CI)
1	0.99 [0.99;0.99]	3.5 [3.5; 3.5]
7	0.90 [0.89;0.90]	10.1 [10.1;10.2]
14	0.76 [0.76;0.76]	15.7 [15.6;15.8]
21	0.67 [0.67;0.67]	18.4 [18.4;18.5]
28	0.64 [0.64;0.64]	19.4 [19.3;19.5]

We used Serfling's periodic regression to detect epidemics. All bootstrap replicates with a Δ below 7 days allowed to detect the 6 epidemics periods, which were present in the original data. With Δ up to 8 days, all epidemics were not detected: a Δ of 8 days resulted in 9% of bootstrap replicates with 7 detected epidemics ; a Δ of 21 days resulted in a detection of 5 epidemics in 613 out of 1,000 replicates and a Δ of 28 days resulted in 5 epidemics detected

in 99% of the replicates. For replicates which correctly detected the 6 epidemics, delays to detect epidemics compared to the true epidemics periods are depicted in Figure 1.

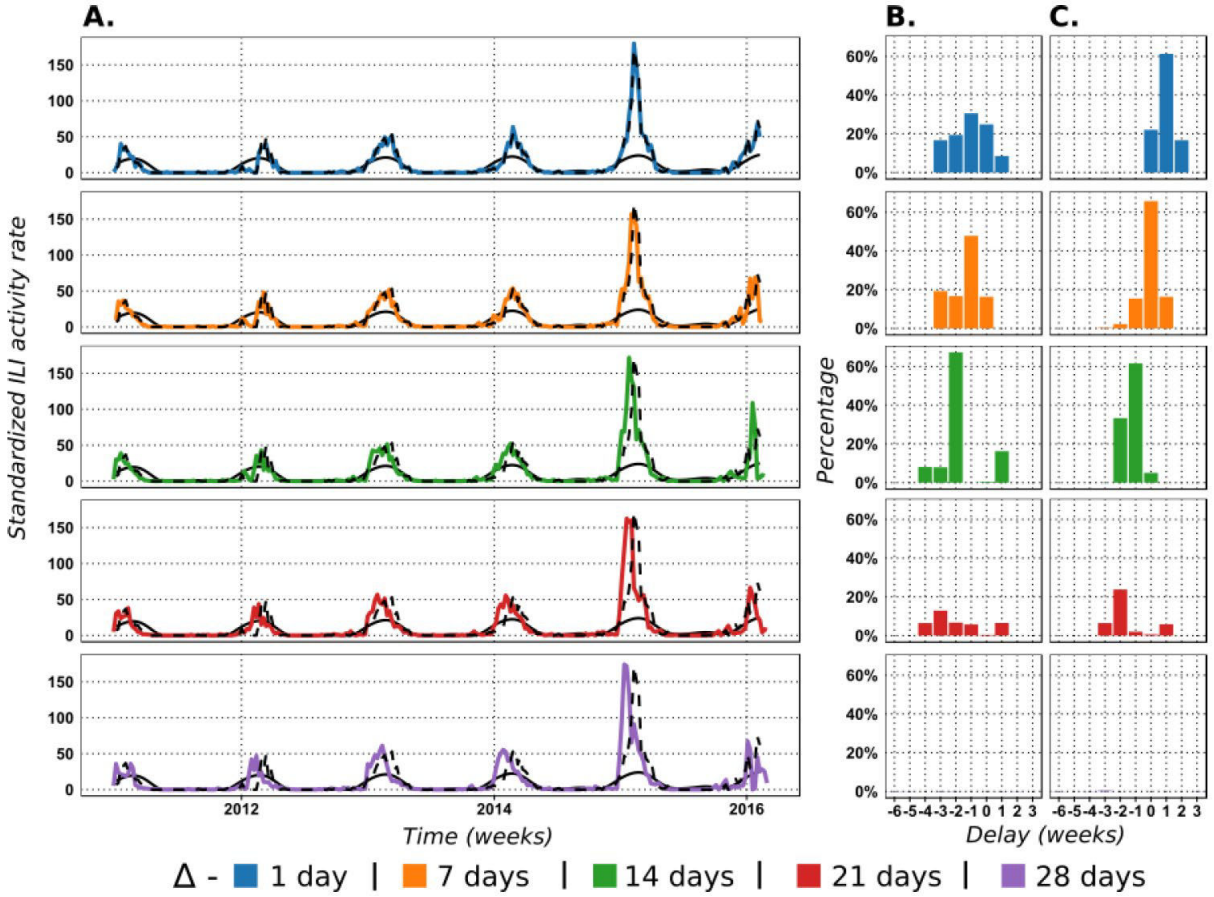


Figure 1. A - Weekly ILI activity estimates from watermarked data. Black dashed lines represent ILI activity computed from original data ; black solid lines are Serfling's model fitted on original data. **B and C -** Delays to detect respectively start and end of epidemics from watermarked data. Percentages are computed on bootstrap replicates which detected the 6 true epidemic periods.

4. Discussion

Watermarking of clinical data appears to be a fairly simple and fast process to ensure traceability of health big data. Traceability represents an additional security level that is currently rarely considered when reusing patients' data but of great concern with the expansion of health data sharing. We assessed a watermark of a datamart –stemming from our CDW– which was dedicated to influenza surveillance. We chose to watermark dates of PCR laboratory results. Indeed, watermarking dates seems to be a rational choice when producing time series of disease activity, because this field is mandatory to produce the time-dependent measures and therefore may not to be deleted. We showed that moderate distortion impacts the reliability of produced incidences in a way that will not drastically change their meaning. For instance, all flu epidemics periods were detected up to a distortion of 7 days, with slight enlargement of epidemics periods. However, with stronger distortion, several epidemics are not detected or present wrong periods. This suggests that watermark procedures have to be carefully tuned so as to preserve the quality of data analyses. Syndromic surveillance also implies forecasting, which can rely on predictive algorithms. It seems interesting to conduct

further work to assess how watermarking disrupt learning process, for instance to forecast disease activity.

Watermarking is of particular concern for syndromic surveillance, which requires sharing data from different sources distributed on a geographic area, before to produce aggregated data. This necessarily implies exposing data outside infrastructures generating data, which typically are CDWs. However, CDWs are dedicated to store data for secondary use with multi purposes. Yet, it seems hard to perform a watermark of a whole CDW, which could be compatible to all use cases for which they are intended. As a consequence, we believe that, in the context of health data traceability, watermarking must be performed at the final step of the data sharing process, when the data are exported for a given use case. Watermarking parameters can be tuned according to the use case, to ensure distortion which will be acceptable, without jeopardizing other use cases of CDWs.

Acknowledgements

This work was supported in part by the French National Research Agency-Project ANR inside the INSHARE (INtegrating and Sharing Health dAta for Research) project (grant no. ANR-15-CE19-0024).

References

[1] Cohen B, Vawdrey DK, Liu J, Caplan D, Furuya EY, Mis FW, et al. Challenges Associated With Using Large Data Sets for Quality Assessment and Research in Clinical Settings. *Policy Polit Nurs Pract*. (2015) 16(0):117–24.

[2] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. (2010);17(2):124–30.

[3] Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – An Integrated Standards-Based Translational Research Informatics Platform. *AMIA Annu Symp Proc*. 2009;2009:391–5.

[4] Holmes JH, Elliott TE, Brown JS, Raebel MA, Davidson A, Nelson AF, Chung A, La Chance P, Steiner JF, "Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature.", *J Amer Med Inf Assoc*. 21(4), (2014), 730-736

[5] Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. *J Amer Med Inf Assoc*. (2009) 16(5):624–30..

[6] De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. *J Biomed Inform*. (2015) 53:162–73.

[7] Franco-Contreras J, Coatrieux G. Robust Watermarking of Relational Databases With Ontology- Guided Distortion Control. *IEEE Trans. Inf. Forens. Sec*. 2015 Sep;10(9):1939–52.
[8] Delamarre D, Bouzille G, Dalleau K, Courtel D, Cuggia M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. *Stud Health Technol Inform*. 2015;210:702–6.
[9] Franco-Contreras J, Coatrieux G, Cuppens F, Cuppens-Boulahia N, Roux C. Robust Lossless Watermarking of Relational Databases Based on Circular Histogram Modulation. *IEEE Trans. Inf. Forens. Sec*.. 2014 **9**(3):397–410.
[10] Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Rep*. 1963 **78**(6):494–506.

Quatrième partie : applications de l'exploitation des données massives en santé

Les cas d'usage des données massives sont aujourd'hui bien identifiés et proviennent des usages historiques des données de santé et des grands enjeux en médecine (32).

I. Veille sanitaire et surveillance syndromique

La veille sanitaire et la surveillance syndromique sont des domaines d'application naturels des data sciences puisqu'il y a de nombreuses similitudes avec le monde financier, notamment en ce qui concerne les objectifs de prévision des indicateurs de suivi des épidémies.

La surveillance syndromique est organisée en France par Santé Publique France (56). Cet organisme dispose de différents outils pour assurer cette surveillance telle que le réseau de médecines généralistes Sentinelles ou le réseau des services d'urgences OSCOUR qui permettent d'obtenir des données de terrain pour mesurer les phénomènes épidémiologiques (57,58). Par ailleurs, de nombreux travaux se sont intéressés à l'utilisation de sources de données alternatives comme les données du web pour la surveillance syndromique notamment, car elles permettent de capturer l'information depuis la population (59–61).

Dans ce contexte, les données massives hospitalières paraissent intéressantes dans la mesure où elles intègrent des données issues des prises en charge de patients concernés par les pathologies surveillées par la surveillance syndromique.

Article 4 : Leveraging hospital big data to monitor flu epidemics

L'article ci-après visait à démontrer l'apport des données hospitalières dans le cadre de la surveillance des épidémies grippales. Pour répondre à cet objectif, des méthodes de recherche d'information ont été utilisées afin d'évaluer l'association entre des signaux extraits à partir de l'entrepôt de données eHOP du CHU de Rennes et les signaux de référence produits par le réseau Sentinelles.

Ma contribution pour ce travail a été de définir les données pertinentes afin de calculer les signaux liés aux syndromes grippaux et de superviser les stagiaires de Master 1 ayant réalisé les analyses.

Leveraging hospital big data to monitor flu epidemics

Guillaume Bouzillé[1234*], Canelle Poirier[126], Boris Campillo-Gimenez[12], Marie-Laure Aubert[6], Mélanie Chabot[6], Emmanuel Chazard[7], Audrey Lavenu[35], Marc Cuggia[1234]

[1]INSERM, U1099, Rennes, F-35000, France
[2]Université de Rennes 1, LTSI, Rennes, F-35000, France
[3]CHU Rennes, CIC Inserm 1414, Rennes, F-35000, France
[4]CHU Rennes, Centre de Données Cliniques, Rennes, F-35000, France
[5]Université Rennes 1, Rennes, F-35000, France
[6]Université de Rennes 2, IRMAR, Rennes, F-35000, France
[7]Département de Santé Publique, Université de Lille EA 2694, CHU Lille, F-59000 Lille, France

* Corresponding author

E-mail: guillaume.bouzille@chu-rennes.fr

LTSI - UMR Inserm - Université de Rennes 1

Equipe-Projet Données massives en santé (DMS)

Campus de Villejean - Bât. 6

35043 Rennes Cedex, France

Abstract

Background and Objective

Influenza epidemics are a major public health concern and require a costly and time-consuming surveillance system at different geographical scales. The main challenge is being able to predict epidemics. Besides traditional surveillance systems, such as the French Sentinel network, several studies proposed prediction models based on internet-user activity. Here, we assessed the potential of hospital big data to monitor influenza epidemics.

Methods

We used the clinical data warehouse of the Academic Hospital of Rennes (France) and then built different queries to retrieve relevant information from electronic health records to gather weekly influenza-like illness activity.

Results

We found that the query most highly correlated with Sentinel network estimates was based on emergency reports concerning discharged patients with a final diagnosis of influenza (Pearson's correlation coefficient (PCC) of 0.931). The other tested queries were based on structured data (ICD-10 codes of influenza in Diagnosis-related Groups, and influenza PCR tests) and performed best (PCC of 0.981 and 0.953, respectively) during the flu season 2014-15. This suggests that both ICD-10 codes and PCR results are associated with severe epidemics. Finally, our approach allowed us to obtain additional patients' characteristics, such as the sex ratio or age groups, comparable with those from the Sentinel network.

Conclusions

Hospital big data seem to have a great potential for monitoring influenza epidemics in near real-time. Such a method could constitute a complementary tool to standard surveillance systems by providing additional characteristics on the concerned population or by providing information earlier. This system could also be easily extended to other diseases with possible activity changes. Additional work is needed to assess the real efficacy of predictive models based on hospital big data to predict flu epidemics.

Keywords: Health Big Data; Clinical Data Warehouse; Information Retrieval System; Health Information Systems; Influenza; Sentinel surveillance

1 Introduction

Currently, flu activity monitoring remains challenging and is a costly and time-consuming task [1]. Flu epidemics are a major public health issue because each year, they cause 250,000 to 500,000 deaths worldwide and they destabilize health care systems, resulting in overcrowding

of primary care centers and emergency departments [2–4]. Many actors are involved in influenza monitoring, at the local, regional, national and international level. National surveillance systems are the cornerstone of this system. For instance, the US influenza Sentinel Provider Surveillance Network, belonging to the Center for Disease Control and Prevention (CDC), in the United States of America, and the Sentinel network in France, both provide weekly flu activity reports based on data collected from general practitioners [5,6].

Such national flu surveillance systems provide a fine-grained description of what happens at the regional or national level and allow researchers to observe inter-annual epidemic variations. However, these reports are usually available with a delay of one to two weeks and need to be refreshed until all data from a given week have been reported. This delay in data availability limits their use for real-time monitoring purposes. Moreover, data reported by the Sentinel network provide very few details about patients, beside age or sex. Yet, it would be of great interest to better describe, for instance, the comorbidities (e.g., International Classification of Diseases, 10th revision, ICD-10, codes), or to identify subgroups of patients who are more likely to catch influenza or to develop influenza-related complications.

For these reasons, influenza surveillance now relies also on other data sources that gather additional information, such as self-reporting from patients, viral surveillance or data from emergency departments (ED) [2,7,8]. In France, the French Public Health Agency launched an additional monitoring system based on data collected from 86% of all French EDs, thus covering most of the French territory [9]. This project provides a better understanding of flu epidemic severity, especially in relation to cases that require hospitalization.

There is also a growing interest in finding other ways that rely on alternative data sources to achieve near real-time monitoring. Many studies have assessed the use of internet-user activity data because they can produce real-time indicators [10–18]. Several data sources have been explored, including Wikipedia, Twitter or Google search-engine data. For instance, Google created a project dedicated to influenza monitoring: Google Flu Trends (GFT). This project uses search queries connected with influenza-like illnesses (ILI) from Google.com to produce influenza activity estimates [2]. Since its launch in the United States in 2008, GFT predictions have proven to be very accurate when compared to CDC reports. Moreover, GFT data are available 7-10 days before those of the CDC [12]. GFT was extended to other countries and its estimates confirmed to be accurate. However, GFT yielded inaccurate data during several periods [19,20]. In 2009, it produced lower estimates at the start of the H1N1 pandemic; in 2013 its estimates were almost twice those from the CDC. As a result, GFT is currently closed to the public. GFT appeared to be sensitive to uncommon flu epidemics, to media coverage, to changes in the internet users' habits and to modifications of the algorithm in the Google search engine [11,20]. Consequently, other studies proposed to combine traditional surveillance systems and web data, to benefit from the advantages of both systems. One example is the recently published work on the ARGO model that could be considered to be a GFT update. It combine Google and CDC ILI activity data with a dynamic statistical model (least absolute shrinkage and selection operator, LASSO) to weekly redefine the best predictors for the current week and readjust their coefficients [11]. This model seems very promising because it can produce near real-time flu activity indexes that are very accurate compared with those produced

by the CDC, with a correlation coefficient of predicted values for the flu seasons of the 2010-2014 period ranging from 0.928 to 0.993.

However, neither standard systems nor the current web-based models are designed to monitor flu activity at a smaller scale, such as that of a hospital. Yet, flu epidemics strongly contribute to the overcrowding of adult and pediatric EDs. A study by Dugas et al, showed a high correlation between city-level GFT data (Baltimore) and the number of patients visiting adult ($r = 0.885$) and pediatric EDs ($r = 0.652$). Specifically, GFT data correlation with standard overcrowding measures was high for pediatric EDs ($r = 0.641$ to 0.649) and moderate for adult EDs ($r = 0.421$ to 0.548) [21].

With the widespread adoption of Electronic Health Records (EHRs), hospitals also are producing a huge amount of data - collected during the course of clinical care - that offer a window into the medical care, status and outcomes of a varied population who is representative of the actual patients [22,23]. This huge amount of data holds the promise of supporting a wide range of medical and health care functions, including, among others, clinical decision-making support, disease surveillance or population health management [24].

Hospitals are currently deploying information technologies and tools intended to facilitate access to clinical data for secondary-use purposes. Among these technologies, clinical data warehouses (CDWs) come forth as one of the solutions to address Hospital Big Data (HBD) exploitation [25]. Different projects have developed CDWs with different architectures, tools and services dedicated to the reuse of patient data coming from EHRs [26–31]. Depending on their Extract-Transform and Load process, CDWs can collect data in real-time, such as the STRIDE CDW of Stanford University [30]. The most famous CDW technology is the Informatics for Integrating Biology & the Bedside project (i2b2), developed by Harvard Medical School, that is now used worldwide in clinical research and can be updated in real-time [32,33]. At our academic hospital in Rennes (France), we developed our own CDW technology, called eHOP (formerly named Roogle [31]). Structured (laboratory, prescriptions, ICD-10 diagnoses) and unstructured (discharge summaries, histopathology, operative reports) data can be integrated in eHOP in real time. Unlike i2b2 data models, eHOP integrates the chain of clinical events into its design and allows the direct access to EHRs. eHOP consists of a powerful search engine system that can identify patients who match specific criteria retrieved either from unstructured data, via keywords, or from structured data, by querying terminology-based codes. The eHOP CDW is used routinely for clinical research purposes, such as feasibility studies, cohort detection and pre-screening, at Rennes academic hospital. The eHOP technology is currently implemented in the other five academic hospitals of the Western region of France (Angers, Brest, Nantes, Poitiers and Tours). Its use will constitute a great source of health data that cover a large part of the population of the West of France who has access to health care facilities linked to eHOP (about 11 million inhabitants; 800,000 visits per year) [34].

We believe that CDWs can help to monitor influenza-like illness (ILI) thanks to their ability to provide data in near real-time and at a local scale. Moreover, the richness of the data produced during patient management will allow a better patient characterization.

In this paper, we present a feasibility study on the production of accurate near-real-time estimates of ILI activity based on the CDW eHOP.

2 Methods

We extracted data from the eHOP CDW of the academic hospital of Rennes, from September 1, 2010 to August 31, 2015. This corresponds to the last five winter seasons defined by the Sentinel network (beginning on the first day of September of every year and ending on 31 August of the following year). The data integration and storage method was the same during the entire study period. As a reference, we used French Sentinel network data on Brittany for the same period (https://websenti.u707.jussieu.fr/sentiweb/?page=table). Brittany is the French region from where most patients at Rennes academic hospital come. We also considered internet-based ILI estimates from GFT for Brittany, from September 1, 2010 to August 10, 2015 (date of GFT closure) as an additional source for comparison (https://www.google.org/flutrends/about/data/flu/fr/data.txt).

We tested two main approaches with the purpose of identifying patients who might have ILI, from data stored in eHOP (see S1 Table for a complete query description). The first approach was based on three different full-text queries to retrieve documents that match the following keywords and constraints:

- Flu query: documents matching the keywords “flu”, in the absence of “flu vaccination,” and “avian flu.”
- Symptoms query: documents matching the keywords “fever” or “pyrexia” and “ache” or “muscle pain.”
- Emergency query: ED discharge summaries where “flu” was the final diagnosis. Only applicable to discharged patients (i.e., documents belonging to patients who were further hospitalized were not considered).

The first two queries could retrieve any kind of document, including discharge summaries of inpatients or outpatients, emergency discharge summaries, operative reports, laboratory results, Diagnosis-Related Groups (DRGs), X-ray reports or histopathology reports. The third query was focused on retrieving documents from the ED.

The second approach involved querying CDW structured data for the following appropriate terminologies:

- ICD-10 query: DRGs having at least one code belonging to the influenza-related ICD-10 chapters: J09.x, J10.x or J11.x.
- Biology query: We relied on the local terminology used by the laboratory information system to retrieve all flu PCR test results (negative and positive). The aim was to have a signal connected with ILI symptoms and not only with flu.

Given that the study purpose was not to assess query accuracy or recall, we made the assumption that potential noise was constant over time. Hence, we did not manually validate the relevance of patients retrieved by the query and we retained the entire list of patients. We then processed the weekly incidences for each query. Our definition of ILI case covered any patient visit for which a document that matched a given query was generated. The date of the case was thus the

patient's admission date. A null incidence estimate was inputted for all weeks without cases. The entire process was performed using anonymous data from the eHOP CDW.

As additional variables, we retrieved the patients' birthdate to perform analyses based on patients' age groups at the time of the visit: 0 to 4 years, 5 to 14 years, 15 to 64 years and 65 years and more. The aim was to assess whether the epidemic severity could be extrapolated from such data. We considered that severe epidemics might affect especially younger and/or older people among all hospitalized patients compared with the population covered by the Sentinel network. We computed the distribution of age groups on a calendar year basis, following a process similar to that of the Sentinel network, with the aim of comparing both distributions.

To evaluate ILI detection by our system, we compared our weekly ILI incidence results with the weekly incidences rates from the reference Sentinel network by calculating the Pearson's correlation coefficient (PCC) for the entire study period and for each winter season. For comparison purposes, we did the same comparison between weekly GFT estimates and weekly incidence rates from the Sentinel network.

As an illustration of eHOP's ability to monitor flu epidemic data, we also replicated the Serfling periodic regression analysis that is currently used by the Sentinel network to identify influenza epidemic periods [35]. We used the Sentinel's R script, available at http://marne.u707.jussieu.fr/periodic, and the parameters currently employed in routine practice by the Sentinel network [36]: a pruning threshold corresponding to the 85th quantile, a 95th unilateral confidence interval to detect the start (when the observed data exceed this threshold for two consecutive weeks) and the end (when the observed data are below the threshold for two consecutive weeks) of ILI epidemics. We fitted the following linear regression model for the whole study period:

$$Y(t) = \mu + \alpha \cdot t + \beta_k \cdot cos(\frac{2k\pi}{T} \cdot t) + y_k \cdot sin(\frac{2k\pi}{T} \cdot t) \ + \ \varepsilon(t),$$

where μ is a constant, α a linear term, k the harmonic number, β_k and γ_k are period terms. The period T is equal to 52.18 weeks and k is equal to 2. The residual error corresponds to the $\varepsilon(t)$ term.

We assessed the periodic regression performance by calculating the shift between the dates (start and end of epidemics) identified with eHOP estimates and the dates identified from Sentinel network estimates.

All analyses were performed using the R software, version 3.2.3 [37].

This study was approved by the local Ethics Committee of Rennes Academic Hospital.

3 Results

3.1 Information retrieval results

The study period included lists of patients retrieved from eHOP queries between September 1, 2010 and August 31, 2015. For this period, 14,873,482 documents were available in the eHOP CDW, as well as 2,220,741 patient visits. Performing the five eHOP queries and then processing the data to produce weekly ILI estimates took approximately 7 minutes (6m 30s for queries on unstructured data and 30s for queries on structured data) on a standard desktop computer. The "flu query" (the keyword "flu", in the absence of "flu vaccination" and "avian flu") retrieved 19,522 documents, among which there were 4,604 emergency reports (24%), 3,773 laboratory results (19.3%), 3,344 outpatient discharge summaries (17.1%), 2,882 inpatient discharge summaries (14.8%) and 798 DRGs (4%). The "symptoms query" (association of fever or pyrexia and ache or muscular pain) retrieved 2,916 documents, among which there were 1,436 emergency room reports (49.2%), 524 outpatient discharge summaries (18%) and 482 inpatient discharge summaries (16.5%). The remaining documents were connected with unclear or missing document types. The last three queries were connected with specific types of documents, particularly with emergency reports, laboratory results or DRGs. The patients' distribution according to the different settings (outpatients, inpatients and ED) is illustrated in Fig 1.

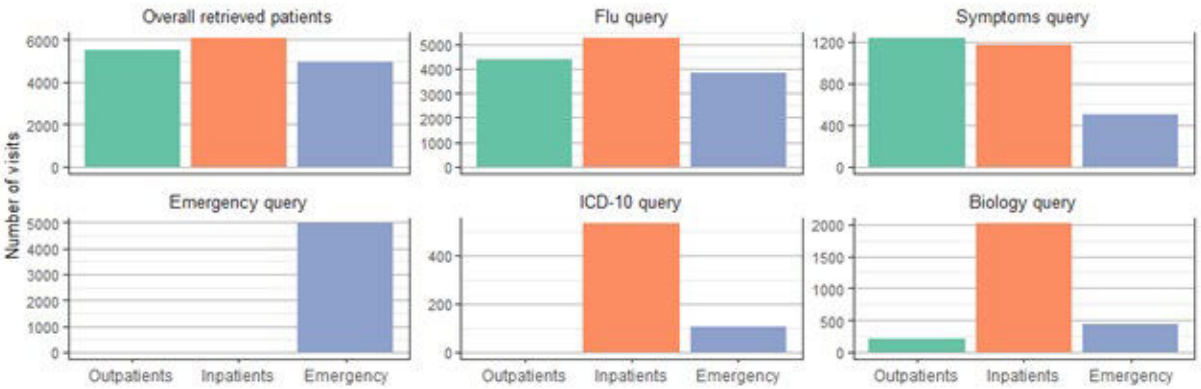


Fig 1 : Patients' settings.

Results from queries to retrieve patients with at least one document matching the following conditions: flu query = keyword "flu" in the absence of flu vaccination and avian flu; symptoms query = keywords "fever" or "pyrexia" and "ache" or "muscle pain"; emergency query = discharge summaries from the emergency department with "flu" as final diagnosis; ICD-10 query = DRGs with at least one code belonging to the ICD-10 chapters on influenza (i.e., J09.x, J10.x or J11.x.); biology query = PCR-based flu tests (negative or positive results). Emergency defined a stay in the emergency department without further hospitalization.

3.2 Overall estimates

Weekly ILI estimates computed from the eHOP query results are displayed in Fig 2. During the entire study period, the ILI estimates retrieved from the query focused on ED data were the most highly correlated with the Sentinel Network's (PCC of 0.931 compared with PCCs between 0.869 and 0.679 for other queries) (Table 1). As a comparison, the PCC for GFT with the Sentinel network was 0.925.

GFT was the data source that correlated most with the Sentinel network for the seasons 2010–11 and 2012–13 (PCC = 0.967 and 0.947, respectively). For the seasons 2011–12 and 2013–14, the eHOP query focused on EDs showed the highest correlation with the Sentinel network, but with a PCC below 0.9. For the season 2014–15, the eHOP ICD-10 query performed best, with a PCC of 0.981. The query based on symptoms was the only one with a PCC below 0.9 for this last season. For the 2013–14 flu season, both eHOP queries and GFT had PCC values below 0.9. The last complete season (2014–15) yielded the best correlations because all queries matched the Sentinel network data with PCC values up to 0.9, except for the symptoms query.

Table 1. Pearson correlation coefficients between ILI activity estimates from eHOP queries or Google Flu Trends and ILI incidence rates from the Sentinel network.

Data source /query	Entire period	Winter flu seasons (from September 1 to August 31 of the following week)				
		2010–11	2011–12	2012–13	2013–14	2014–15
GFT (up to 2015-08-10)	0.925	0.967	0.735	0.947	0.850	0.940
eHOP flu	0.869	0.871	0.862	0.911	0.818	0.939
eHOP symptoms	0.679	0.784	0.664	0.652	0.298	0.837
eHOP emergency	0.931	0.941	0.864	0.933	0.853	0.972
eHOP ICD-10	0.829	0.854	0.789	0.758	0.732	0.981
eHOP biology	0.801	0.813	0.796	0.863	0.777	0.953

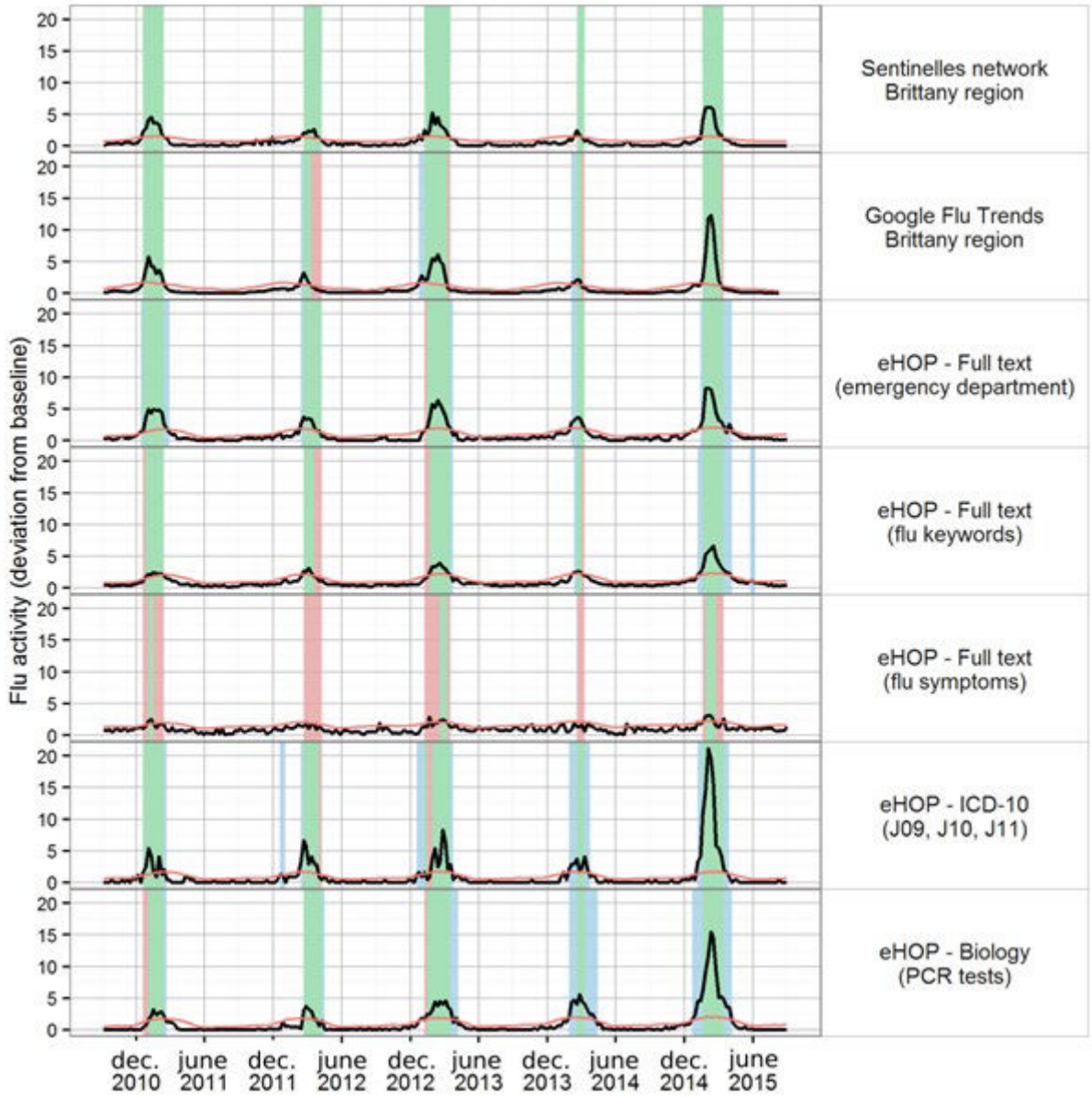


Fig 2. Weekly influenza-like illness estimates from the different data sources and periods of detected epidemics.
The reference is data from the Sentinel network for the Brittany region. Estimates from Google Flu Trends are for comparison purposes. Black curves correspond to the estimates computed from the different data sources or queries. Red curves are the upper bound of the 95% prediction interval of the periodic regression models, computed using the Serfling method to determine epidemic periods. Green areas are periods that match the Sentinel network epidemic periods. Red areas are epidemic periods not detected from data sources or queries. Blue areas are detected periods that do not match true epidemics.

3.3 Sex and age group estimates

In the Sentinel network data, the male to female ratio was 1, 0.96, 0.97, 0.93 and 1.01, respectively, for epidemics from 2010 to 2014. In comparison, the sex ratio observed in eHOP queries ranged from 0.94 to 3.2 in 2010, from 1.07 to 1.90 in 2011, from 0.94 to 1.36 in 2012,

from 1.02 to 1.78 in 2013 and from 0.92 to 1.16 in 2014. The highest sex ratio values were found in the results obtained with the biology query, indicating that PCR tests were more often performed for male patients. There was no significant difference in the age group distribution between male and female patients for the patients retrieved with this query ($p = 0.41$ using the Chi-square test).

Regarding the age group distribution, eHOP queries yielded more pediatric cases (0 to 4 years), compared with the Sentinel network data (Fig 3). The biology query retrieved more pediatric and elderly patients than the other queries.

3.4 Epidemic periods

For each GFT and eHOP query, we computed a periodic regression model (i.e., Serfling regression model) to detect epidemic periods, as done by the Sentinel network's current surveillance system (red line in Fig 2). We compared epidemic periods from GFT and eHOP with reference data from the Sentinel network for the region of Brittany (Table 2).

Table 2. Summary of epidemic detection delays using the different data sources or queries

Data source/query	No. of detected epidemics	Average delay to detect the epidemic start* (week)	Average delay to detect the epidemic end* (week)
Sentinel network	5	0	0
GFT	5	-1 ± 1	-1.4 ± 1.51
eHOP emergency	5	-0.8 ± 1.09	1 ± 1.22
eHOP flu	6	0 ± 1.58	0 ± 2.24
eHOP symptoms	3	3 ± 2.64	-2.67 ± 1.53
eHOP ICD-10	7	-0.6 ± 2.30	1 ± 1.22
eHOP biology	5	-0.8 ± 2.59	2.6 ± 1.67

* Delays are related to epidemics overlapping with the true epidemic periods from the Sentinel network

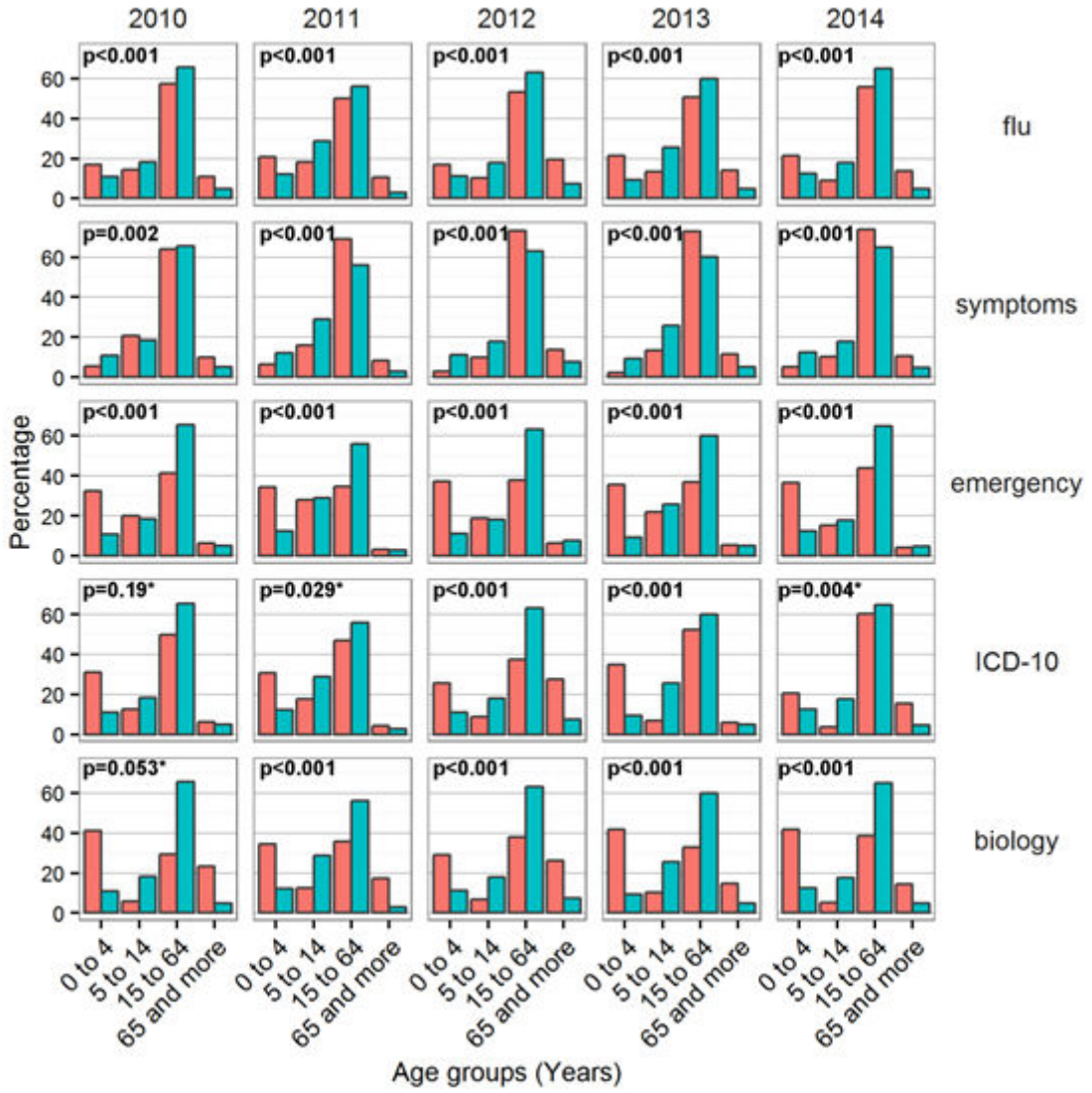


Fig 3. Age group distributions retrieved from the different eHOP queries.
Red bars show the age group distribution from the different eHOP queries. Green bars show the age group distribution from the Sentinel network.
P-values were calculated with the use of the Chi-square test or the Fisher exact test (indicated with an asterisk)
See legend to Fig 1 for a description of the eHOP queries.

GFT detected the beginning and the end of epidemics from 0 to 2 weeks before the Sentinel network. Among the different eHOP queries, the flu symptoms query yielded the worst results, particularly because it could not detect all epidemics. Laboratory and ICD-10 queries resulted in longer epidemics, particularly for the last two seasons: they anticipated the start of the two epidemics by 2 to 4 weeks and delayed the end by 2 to 5 weeks (Fig 2). The eHOP query on flu keywords and the emergency query gave the best results. Particularly, the emergency query detected the start of epidemics from 1 to 2 weeks before the Sentinel network, except in 2013, when there was a delay of one week. For the epidemic end, the emergency query tended to produce longer epidemics, ending 0 to 3 weeks after the Sentinel network's estimates (Fig 2).

4 Discussion

This study demonstrates the great potential of HBD for monitoring flu epidemics. CDWs, such as eHOP, allow researchers to leverage the richness of heterogeneous clinical data from EHRs. eHOP added value is that it provides the possibility of querying both structured and unstructured data that appear to be great candidate data sources for efficient monitoring of diseases activity. However, as it is the case with every information retrieval system, part of the results yielded by our system corresponds to noise, that is, patients who do not have ILI. The result precision depends partly on the query used. For instance, the "symptoms" query is particularly subject to noise and thus, does not seem to be specific enough for ILI monitoring. It also depends on the type of queried data. Unstructured data are, of course, more prone to produce noisy results. The main reasons are the mentioning of a personal or family history of influenza and the exclusion of influenza diagnoses in discharge summaries, although our system has several natural language processing capabilities, such as detection of negative sentences. Structured data are less susceptible to noise: laboratory results or ICD-10 codes ascertain the fact that the patient has ILI. The drawback is the lack of recall for such data sources, for instance, during epidemics the severity of which does not lead to hospitalization (i.e., without diagnosis-related groups), or with diagnoses that do not require any laboratory test. Thus, we cannot control the performance of our information retrieval system. This can be seen as a limitation of our approach: we cannot validate every potential case retrieved by the system, and we cannot ensure the retrieval of all patients with ILI. We could have investigated the system precision because eHOP provides the possibility to access the original documents to check whether the retrieved patients truly had ILI. However, the purpose of this study was not to assess the performance of our information retrieval system, but to show that it can produce ILI activity indexes in the same way as internet-based monitoring. Hence, our system is not intended to be as reliable as a traditional monitoring system, such as the Sentinel network, for producing weekly incidence rates. Nevertheless, it provides a good picture of weekly ILI activity in primary care through the ED data and in hospitalized patients.

We believe that the strength of our system is its capability to generate near real-time estimates from hospital big data. Our estimates are generated using health care activity suspected of being connected with ILI and, due to the proximity to actual ILI cases, they could be more reliable than internet-based indexes. Indeed, we can produce estimates based on data connected with patients who presented symptoms severe enough to require visiting the ED or to be transferred to hospital. On the contrary, internet-based estimates may also incorporate data from healthy internet users who can potentially be influenced by the media or are simply searching information about influenza.

The possibility to produce a fine-grained description of the diseased population is an additional strength of our system. We demonstrated this potential for simple attributes (age groups and sex ratio) that were also available in the Sentinel network annual reports, for comparison purposes. This allowed showing some differences between the population coming to hospital and the population captured by the reference system. Our system found more pediatric and geriatric cases than the Sentinel network. Particularly, the younger cases may explain the

predominance of male patients found with the PCR query because it seems than male patients are more prone to respiratory infectious diseases than female patients [38].

In addition, eHOP allows a better characterization of ILI patients by using the data available in the CDW, such as comorbidities or episode severity (e.g., requiring hospitalization or intensive care), all in near-real time.

However, one must be aware that the eHOP data loading process has various delays, depending on the data source. As a result, this process involves a high degree of heterogeneity in the availability of the data used to produce ILI estimates. For instance, discharge summaries are often generated several days after the patient's stay, which is not compatible with real-time monitoring. Conversely, ED discharge summaries are produced during the patient's visit and are made available as soon as the patient leaves the hospital or is transferred to a conventional unit. Similarly, laboratory results are produced during the patients' stay. Therefore, these two data sources are available in the CDW with a lag of one day, because they are uploaded in eHOP every night.

Another of the system's limitations is that we currently only have access to hospital data. This is the main cause of the differences in ILI activity compared with the Sentinel network. From the perspective of our hospital physicians working on infectious diseases this is not really a drawback, because the differences in duration and magnitude may reflect the severity of epidemics that cause more hospitalizations during a longer period. The higher estimates resulting from ICD-10 and laboratory queries also seem to be connected with more severe epidemics, as was the case in 2014–15. Moreover, local ILI activity estimates could be compared with other local indexes, such as the global hospital activity, bed occupation rates or average hospitalization length, to produce more appropriate estimates of the overcrowding risk. This is a key point for hospitals, as estimates from traditional surveillance systems do not allow them to anticipate overcrowding during severe epidemics, resulting in higher rates of hospitalization. However, we also produced estimates comparable to those of the Sentinel network, when using appropriate queries from the ED (PCC of 0.931) that correlated more closely with the Sentinel network estimates than any of the Google Correlate internet-based queries (the Google query most correlated with ILI activity from the Sentinel network for the region of Brittany and for our study period was "Tamiflu", with a PCC of 0.9265).

In our study, we were limited to the population of Rennes academic hospital that, in addition, does not entirely cover the geographical territory of Brittany. As mentioned in the Introduction, the eHOP technology is going to be deployed in all academic hospitals of the West of France. By extending the study reach, we could obtain a complete view of influenza dynamics and activity at a larger scale. We also believe that our approach is transposable to other CDW technologies, such as the i2b2 standard [32], with appropriate real-time data integration. This could allow aggregating estimates from different institutions, using a SHRINE data sharing network at different scales [39]. Indeed, the SHRINE technology allows building a multi-node, peer-to-peer infrastructure for connecting i2b2 CDWs to research networks. We are also exploring this approach by feeding an i2b2 instance with limited sets (i.e., only patients retrieved through our queries) of structured data from eHOP. Another approach could be based

on the OHDSI initiative that proposes a common data model for observational studies employing other standardization procedures [40]. However, we have not yet investigated this approach.

Finally, this study only gives the proof of concept concerning the HBD potential for ILI monitoring. The next step will be to assess eHOP prediction capabilities with appropriate statistical models, using such data to predict the data generated by the Sentinel network. Several models have been explored in previous studies with promising results. Recently, Harvard University proposed an alternative model to GFT also based on Google users' activity [11]. Briefly, for each weekly ILI activity to be predicted, a model is built using predictors consisting of the 2-year history of the CDC ILI activity, submitted to an autoregressive process of order 52, and the 100 Google queries most highly correlated with the CDC ILI activity for the same period. The model uses a LASSO method to perform variable selections to only keep the most informative predictors. This kind of model could easily use our eHOP query results as covariates instead of internet-based data. Another interesting approach could be to build models that combine internet-based data and hospital data. Besides predicting ILI activity at a population level, we also want to assess whether our data can be used for predicting ED activity that might help to better manage issues connected with overcrowding. Our results also suggests that this approach could be used for monitoring the activity of other diseases that are emerging or that require precise follow-up, especially when the population is not yet worried about them.

5 Conclusions

Our study shows that HBD are a valuable data source for ILI activity monitoring. Specific data sources, such as laboratory results or DRGs, and the patient characteristics that are available in CDWs allow a fine description of epidemics. However, further investigation is necessary to assess the near real-time prediction capabilities of models that use such data sources, and to demonstrate its extensibility to other diseases.

6 Acknowledgments

We would like to thank the French National Research Agency (ANR), for funding this work inside the INSHARE (INtegrating and Sharing Health dAta for Research) project (grant no. ANR-15-CE19-0024).

We thank our colleagues Eric Matzner-Lober from the University of Rennes 2, Jean-Marc Chapplain from the CHU of Rennes and the COREB from the French Infectious Diseases Society who provided insight and expertise that greatly assisted the research.

We also thank the French Sentinel network for making their data publicly available.

References

[1] L. Brammer, A. Budd, N. Cox, Seasonal and pandemic influenza surveillance considerations for constructing multicomponent systems, Influenza Other Respir. Viruses. 3 (2009) 51–58. doi:10.1111/j.1750-2659.2009.00077.x.

[2] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature. 457 (2009) 1012–1014. doi:10.1038/nature07634.

[3] O.M. Araz, D. Bentley, R.L. Muelleman, Using Google Flu Trends data in forecasting influenza-like–illness related ED visits in Omaha, Nebraska, Am. J. Emerg. Med. 32 (2014) 1016–1023. doi:10.1016/j.ajem.2014.05.052.

[4] J.-P. Chretien, D. George, J. Shaman, R.A. Chitale, F.E. McKenzie, Influenza Forecasting in Human Populations: A Scoping Review, PLOS ONE. 9 (2014) e94130. doi:10.1371/journal.pone.0094130.

[5] W.W. Thompson, L. Comanor, D.K. Shay, Epidemiology of Seasonal Influenza: Use of Surveillance Data and Statistical Models to Estimate the Burden of Disease, J. Infect. Dis. 194 (2006) S82–S91. doi:10.1086/507558.

[6] A.J. Valleron, E. Bouvet, P. Garnerin, J. Ménarès, I. Heard, S. Letrait, J. Lefaucheux, A computer network for the surveillance of communicable diseases: the French experiment, Am. J. Public Health. 76 (1986) 1289–1292.

[7] P.M. Polgreen, Y. Chen, D.M. Pennock, F.D. Nelson, R.A. Weinstein, Using Internet Searches for Influenza Surveillance, Clin. Infect. Dis. 47 (2008) 1443–1448. doi:10.1086/593098.

[8] R. Chunara, S. Aman, M. Smolinski, J.S. Brownstein, Flu Near You: An Online Self-reported Influenza Surveillance System in the USA, Online J. Public Health Inform. 5 (2013). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3692780/ (accessed August 1, 2016).

[9] L. Josseran, J. Nicolau, N. Caillère, P. Astagneau, G. Brücker, Syndromic surveillance based on emergency department activity and crude mortality: two examples, Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull. 11 (2006) 225–229.

[10] D.A. Broniatowski, M.J. Paul, M. Dredze, National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic, PLOS ONE. 8 (2013) e83672. doi:10.1371/journal.pone.0083672.

[11] S. Yang, M. Santillana, S.C. Kou, Accurate estimation of influenza epidemics using Google search data via ARGO, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 14473–14478. doi:10.1073/pnas.1515373112.

[12] A.F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, R.E. Rothman, Influenza Forecasting with Google Flu Trends, PLOS ONE. 8 (2013) e56176. doi:10.1371/journal.pone.0056176.

[13] D.R. Olson, K.J. Konty, M. Paladini, C. Viboud, L. Simonsen, Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales, PLOS Comput Biol. 9 (2013) e1003256. doi:10.1371/journal.pcbi.1003256.

[14] M.J. Paul, M. Dredze, D. Broniatowski, Twitter Improves Influenza Forecasting, PLoS Curr. 6 (2014). doi:10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

[15] D.A. Broniatowski, M.J. Paul, M. Dredze, National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic, PLOS ONE. 8 (2013) e83672. doi:10.1371/journal.pone.0083672.

[16] K.S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J.M. Hyman, A. Deshpande, S.Y.D. Valle, Forecasting the 2013–2014 Influenza Season Using Wikipedia, PLOS Comput Biol. 11 (2015) e1004239. doi:10.1371/journal.pcbi.1004239.

[17] N. Generous, G. Fairchild, A. Deshpande, S.Y.D. Valle, R. Priedhorsky, Global Disease Monitoring and Forecasting with Wikipedia, PLOS Comput Biol. 10 (2014) e1003892. doi:10.1371/journal.pcbi.1003892.

[18] D.J. McIver, J.S. Brownstein, Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time, PLOS Comput Biol. 10 (2014) e1003581. doi:10.1371/journal.pcbi.1003581.

[19] D. Lazer, R. Kennedy, G. King, A. Vespignani, The Parable of Google Flu: Traps in Big Data Analysis, Science. 343 (2014) 1203–1205. doi:10.1126/science.1248506.

[20] D. Butler, When Google got flu wrong, Nature. 494 (2013) 155–156. doi:10.1038/494155a.

[21] A.F. Dugas, Y.-H. Hsieh, S.R. Levin, J.M. Pines, D.P. Mareiniss, A. Mohareb, C.A. Gaydos, T.M. Perl, R.E. Rothman, Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics, Clin. Infect. Dis. 54 (2012) 463–469. doi:10.1093/cid/cir883.

[22] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (2013) 144–151. doi:10.1136/amiajnl-2011-000681.

[23] C. Saez, M. Robles, J.M. Garcia-Gomez, Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances, Stat. Methods Med. Res. (2014). doi:10.1177/0962280214545122.

[24] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Health Inf. Sci. Syst. 2 (2014) 3.

[25] S.-Y. Shin, W.S. Kim, J.-H. Lee, Characteristics Desired in Clinical Data Warehouse for Biomedical Research, Healthc. Inform. Res. 20 (2014) 109–116. doi:10.4258/hir.2014.20.2.109.

[26] J.-M. Pinon, S. Calabretto, L. Poullet, Document Semantic Model: an experiment with patient medical records., in: ELPUB, 1997. http://elpub.scix.net/data/works/att/97124.content.pdf (accessed April 21, 2015).

[27] D.A. Hanauer, EMERSE: The Electronic Medical Record Search Engine, AMIA. Annu. Symp. Proc. 2006 (2006) 941.

[28] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, H.C. Chueh, Integration of Clinical and Genetic Data in the i2b2 Architecture, AMIA. Annu. Symp. Proc. 2006 (2006) 1040.

[29] J. Rogers, C. Puleston, A. Rector, The CLEF Chronicle: Patient Histories Derived from Electronic Health Records, in: 22nd Int. Conf. Data Eng. Workshop 2006 Proc., 2006: pp. x109–x109. doi:10.1109/ICDEW.2006.144.

[30] H.J. Lowe, T.A. Ferris, P.M. Hernandez, S.C. Weber, STRIDE – An Integrated Standards-Based Translational Research Informatics Platform, AMIA. Annu. Symp. Proc. 2009 (2009) 391–395.

[31] M. Cuggia, N. Garcelon, B. Campillo-Gimenez, T. Bernicot, J.-F. Laurent, E. Garin, A. Happe, R. Duvauferrier, Roogle: an information retrieval engine for clinical data warehouse, Stud. Health Technol. Inform. 169 (2011) 584–588.

[32] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inform. Assoc. 17 (2010) 124–130. doi:10.1136/jamia.2009.000893.

[33] R.W. Majeed, R. Röhrig, Automated realtime data import for the i2b2 clinical data warehouse: introducing the HL7 ETL cell, Stud. Health Technol. Inform. 180 (2012) 270–274.

[34] C. Jaglin-Grimonprez, Organiser, moderniser, innover : quelles avancées pour les patients, (2015). http://social-sante.gouv.fr/IMG/pdf/tr2_colloque-5_jaglin_20151016.pdf (accessed May 18, 2016).

[35] R.E. Serfling, Methods for current statistical analysis of excess pneumonia-influenza deaths, Public Health Rep. 78 (1963) 494–506.

[36] C. Pelat, P.-Y. Boëlle, B.J. Cowling, F. Carrat, A. Flahault, S. Ansart, A.-J. Valleron, Online detection and quantification of epidemics, BMC Med. Inform. Decis. Mak. 7 (2007) 29. doi:10.1186/1472-6947-7-29.

[37] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015. https://www.R-project.org.

[38] M. Muenchhoff, P.J.R. Goulder, Sex Differences in Pediatric Infectious Diseases, J. Infect. Dis. 209 (2014) S120–S126. doi:10.1093/infdis/jiu232.

[39] G.M. Weber, S.N. Murphy, A.J. McMurry, D. MacFadden, D.J. Nigrin, S. Churchill, I.S. Kohane, The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories, J. Am. Med. Inform. Assoc. 16 (2009) 624–630. doi:10.1197/jamia.M3191.

[40] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers, Stud. Health Technol. Inform. 216 (2015) 574–578.

Supporting Information

S1 Table. Oracle text SQL queries used for unstructured data in the eHOP clinical data warehouse.

eHOP flu	gripp% NOT(near((vaccin%,gripp%),5,TRUE) OR aviaire
eHOP symptoms	(fuzzy(pyrexie) OR fuzzy(fievre)) AND (myalgie OR near((douleur%,musculaire%),0,TRUE) OR courbature%)
eHOP emergency	gripp% AND RETOUR DOMICILE

Article 5 : Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Study

L'article précédent a permis de montrer que les entrepôts de données biomédicales comportaient des données fortement associées aux syndromes grippaux. Ce second article avait pour objectif d'évaluer si l'utilisation de modèles d'apprentissage automatique permettait de tirer parti de l'intégration des données de l'entrepôt eHOP en quasi-temps réel pour anticiper le délai de production des indicateurs de surveillance des épidémies grippales par le réseau Sentinelles. L'évaluation a d'autre part comparé les performances de notre méthode avec celles employées pour exploiter les données du moteur de recherche Google, à la fois à une échelle nationale et régionale.

Ma contribution dans ce travail a été de réaliser l'extraction des données depuis l'entrepôt de donnée eHOP du CHU de rennes et d'assurer la supervision de la stagiaire de Master 2 ayant développé les modèles d'apprentissage automatique.

Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Study

Canelle Poirier[1,2], MSc; Audrey Lavenu[3], PhD; Valérie Bertaud[1,2,4], DMD, PhD; Boris Campillo-Gimenez[2,5], MD, MSc; Emmanuel Chazard[6,7], MD, PhD; Marc Cuggia[1,2,4], MD, PhD; Guillaume Bouzillé[1,2,4], MD, MSc

[1] Laboratoire Traitement du Signal et de l'Image, Université de Rennes 1, Rennes, France

[2] INSERM, U1099, Rennes, France

[3] Centre d'Investigation Clinique de Rennes, Université de Rennes 1, Rennes, France

[4] Centre Hospitalier Universitaire de Rennes, Centre de Données Cliniques, Rennes, France

[5] Comprehensive Cancer Regional Center, Eugene Marquis, Rennes, France

[6] Centre d'Etudes et de Recherche en Informatique Médicale EA2694, Université de Lille, Lille, France

[7] Public Health Department, Centre Hospitalier Régional Universitaire de Lille, Lille, France

Corresponding Author:

Canelle Poirier, MSc

Laboratoire Traitement du Signal et de l'Image Université de Rennes 1

2 rue Henri Le Guilloux Rennes, 35033

France

Email: canelle.poirier@outlook.fr

Abstract

Background: Traditional surveillance systems produce estimates of influenza-like illness (ILI) incidence rates, but with 1- to 3-week delay. Accurate real-time monitoring systems for influenza outbreaks could be useful for making public health decisions. Several studies have investigated the possibility of using internet users' activity data and different statistical models to predict influenza epidemics in near real time. However, very few studies have investigated hospital big data.

Objective: Here, we compared internet and electronic health records (EHRs) data and different statistical models to identify the best approach (data type and statistical model) for ILI estimates in real time.

Methods: We used Google data for internet data and the clinical data warehouse eHOP, which included all EHRs from Rennes University Hospital (France), for hospital data. We compared 3 statistical models—random forest, elastic net, and support vector machine (SVM).

Results: For national ILI incidence rate, the best correlation was 0.98 and the mean squared error (MSE) was 866 obtained with hospital data and the SVM model. For the Brittany region, the best correlation was 0.923 and MSE was 2364 obtained with hospital data and the SVM model.

Conclusions: We found that EHR data together with historical epidemiological information (French Sentinelles network) allowed for accurately predicting ILI incidence rates for the entire France as well as for the Brittany region and outperformed the internet data whatever was the statistical model used. Moreover, the performance of the two statistical models, elastic net and SVM, was comparable.

(JMIR Public Health Surveill 2018;4(4):e11361) doi:10.2196/11361

KEYWORDS : electronic health records; hospital big data; internet data; influenza; machine learning; Sentinelles network

Introduction

Background

Influenza is a major public health problem. Outbreaks cause up to 5 million severe cases and 500,000 deaths per year worldwide [1-5]. During influenza peaks, large increase in visits to general practitioners and emergency departments causes health care system disruption.

To reduce its impact and help organize adapted sanitary responses, it is necessary to monitor influenza-like illness (ILI; any acute respiratory infection with fever ≥ 38°C, cough, and onset within the last 10 days) activity. Some countries rely on clinical surveillance schemes based on reports by sentinel physicians [6], where volunteer outpatient health care providers report all ILI cases seen during consultation each week. In France, ILI incidence rate is then computed at the national or regional scale by taking into account the number of sentinel physicians and medical density of the area of interest. ILI surveillance networks produce estimates of ILI incidence rates, but with a 1- to 3-week delay due to the time needed for data processing and aggregation. This time lag is an issue for public health decision making [2,7]. Therefore, there is a growing interest in finding ways to avoid this information gap. Nsoesie et al [8] reviewed methods for influenza forecasting, including temporal series and compartmental methods. The authors showed that these models have limitations. For instance, influenza activity is not consistent from season to season, which is a problem for temporal series. Alternative strategies have been proposed, including using different data sources, such as meteorological or demographic data, combined with ILI surveillance network data [9-11] or big data, particularly Web data [12]. With over 3.2 billion Web users, data flows from the internet are huge and of all types; they can be from social networks (eg, Facebook and Twitter), viewing sites, (eg, YouTube and Netflix), shopping sites, (eg, Amazon and Cdiscount), but also from sales or rentals website between particulars (eg, Craigslist and Airbnb). In the case of influenza, some studies used data from Google [2,4,9,13-16], Twitter [17,18], or Wikipedia [19-21]. The biggest advantage of Web data is that they are produced in real time. One of the first and most famous studies on the use of internet data for detecting influenza epidemics is Google Flu Trends [13,22], a Web service operated by Google. They showed that internet users' searches are strongly correlated with influenza epidemics. However, for the influenza season 2012-2013, Google Flu Trends clearly overestimated the flu epidemic due to the announcement of a pandemic that increased the internet users' search frequency, whereas the pandemic finally did not appear. The lack of robustness, due to the sensitivity to the internet users' behavioral changes and the modifications of the search engine performance led to stop the Google Flu Trends algorithm [2,23,24].

Some authors updated the Google Flu Trends algorithm by including data from other sources, such as historical flu information for instance or temperature [2,13-16]. Yang et al [2] proposed an approach that relies on Web-based data (Centers for Diseases Control ILI activity and Google data) and on a dynamic statistical model based on a least absolute shrinkage and selection operator (LASSO) regression that allows overcoming the aforementioned issues. At the national scale, the correlation between predictions and incidence rates was 0.98.

The internet is not the only data source that can be used to produce information in real time. With the widespread adoption of electronic health records (EHRs), hospitals also produce a huge amount of data that are collected during hospitalization. Moreover, many hospitals are

implementing information technology tools to facilitate the access to clinical data for secondary-use purposes. Among these technologies, clinical data warehouses (CDWs) are one of the solutions for hospital big data (HBD) exploitation [25-28]. The most famous is the Informatics for Integrating Biology & the Bedside (i2b2) project, developed by the Harvard Medical School, which is now used worldwide for clinical research [29,30]. In addition, it has been shown that influenza activity changes detected retrospectively with EHR-based ILI indicators are highly correlated with the influenza surveillance data [31,32]. However, few HBD-based models have been developed to monitor influenza [7,33]. Santillana et al proposed a model using HBD and a machine learning algorithm (support vector machine [SVM]) with a good performance at the regional scale [7]. The correlation between estimates and ILI incidence rates ranges from 0.90 to 0.99, depending on the region and season.

Objectives

It would be interesting to determine whether HBD gives similar, better, or lower results than internet data with these statistical models (machine learning and regression). To this aim, we first evaluated HBD capacity to estimate influenza incidence rates compared with internet data (Google data). Then, we aim to find the best statistical model to estimate influenza incidence rates at the national and regional scales by using HBD or internet data. As these models have been described in the literature, we focused on two machine learning algorithms, random forest (RF) and SVM, and a linear regression model, elastic net.

Methods

Data Sources

Clinical Data Warehouse eHOP

At Rennes University Hospital (France), we developed our own CDW technology called eHOP. eHOP integrates structured (laboratory test results, prescriptions, and International Classification of Diseases 10th Revision, ICD-10, diagnoses) and unstructured (discharge letter, pathology reports, and operative reports) patients data. It includes data from 1.2 million in- and outpatients and 45 million documents that correspond to 510 million structured elements. eHOP consists of a powerful search engine system that can identify patients with specific criteria by querying unstructured data with keywords, or structured data with querying codes based on terminologies. eHOP is routinely used for clinical research. The first approach to obtain eHOP data connected with ILI was to perform different full-text queries to retrieve patients who had, at least, one document in their EHR that matched the following search criteria:

1. Queries directly connected with flu or ILI were as follows:
   "flu"
   "flu" or "ILI"
   "flu" or "ILI", in the absence of "flu vaccination"
   "flu vaccination"
   "flu" or "ILI", only in emergency department reports
2. Queries connected with flu symptoms were as follows:
   "fever" or "pyrexia"
   "body aches" or "muscular pain"

"fever or pyrexia" or "body aches or muscular pain"
"flu vaccination"
"fever or pyrexia" and "body aches or muscular pain"

3. Drug query was as follows:
"Tamiflu"

The second approach was to leverage structured data with the support of appropriate terminologies:

1. ICD-10 queries were as follows: J09.x, J10.x, or J11.x (chapters corresponding to influenza in ICD-10). We retained all diagnosis-related groups with these codes.
2. Laboratory queries were as follows: influenza testing by reverse transcription polymerase chain reaction; we retained test reports with positive or negative results because the aim was to evaluate more generally ILI symptom fluctuations and not specifically influenza.

In total, we did 34 queries. For each query, the eHOP search engine returned all documents containing the chosen keywords (often, several documents for 1 patient and 1 stay). For query aggregation, we kept the oldest document for 1 patient and 1 stay and then calculated, for each week, the number of stays with, at least, one document mentioning the keyword contained in the query. In this way, we obtained 34 variables from the CDW eHOP. Multimedia Appendix 1 shows the queries and the number of concerned stays. We retrieved retrospective data for the period going from December 14, 2003 to October 24, 2016. This study was approved by the local Ethics Committee of Rennes Academic Hospital (approval number 16.69).

Google Data

For comparison with internet data, we obtained the frequency per week of the 100 most correlated internet queries (Multimedia Appendices 2 and 3) by French users from Google Correlate [34], and we used this information to retrieve Google Trends data. Unlike Google Correlate, Google Trends data [35] are available in real time, but we had to use Google Correlate to identify the most correlated queries to a signal. The times series passed into Google Correlate are the national flu time series and the regional flu time series (Brittany region) obtained from the French Sentinelles network (see below). The time period used to calculate the correlation is from January 2004 to October 2016. We used the R package gtrendsR to obtain automatically Google Trends data from January 4, 2004 to October 24, 2016 [36,37].

Sentinelles Network Data

We obtained the national (Metropolitan France) and regional (Brittany region, because Rennes University Hospital, from which EHR data were obtained, is situated in this region) ILI incidence rates (per 100,000 inhabitants) from the French Sentinelles network [38-40] from December 28, 2002 to October 24, 2016. We considered these data as the gold standard and used them as independent historical variables for our models.

Data Preparation

Based on previous studies that included datasets with very different numbers of explanatory variables according to the used statistical model [2,7], we built two datasets (one with a large number of variables and another with a reduced number of selected variables) from eHOP and Google data, for both the national and regional analyses (Figure 1).

Each one of these four datasets was completed with historical Sentinelles data. Therefore, for this study, we used the following:

1. eHOP Complete: this eHOP dataset included all variables from eHOP and the historical data from the Sentinelles network with the ILI estimates for the 52 weeks that preceded the week under study (thus, from t-1 to t-52).
2. eHOP Custom: this eHOP dataset included the 3 most correlated variables between January 2004 and October 2016 from eHOP for the ILI signal for week t, −1 (t-1), and −2 (t-2), and historical information from the Sentinelles network with ILI estimates for t-1 and t-2.
3. Google Complete: this Google dataset included the 100 most ILI activity-correlated queries from Google Trends and historical information from the Sentinelles network with ILI estimates for t-1 to t-52.
4. Google Custom: this Google dataset included the 3 most ILI activity-correlated queries between January 2004 and October 2016 from Google Trends for t, (t-1), and (t-2) and historical data from the Sentinelles network with ILI estimates for (t-1) and (t-2).

Statistical Models

Our test period started on December 28, 2009 and finished on October 24, 2016. We fitted our models using a training dataset that corresponded to the data for the previous 6 years. Each model was dynamically recalibrated every week to incorporate new information. For instance, to estimate the ILI activity fluctuations for the week starting on December 28, 2009, the training data consisted of data from December 21, 2003 to December 21, 2009.

Elastic Net

Elastic net is a regularized regression method that takes into account the correlation between explanatory variables and also a large number of predictors [41]. It combines the penalties of the LASSO and Ridge methods, thus allowing keeping the advantages of both methods and overcoming their limitations [42,43]. With datasets that may have up to 152 potentially correlated variables, we performed the elastic net regression analysis using the R package glmnet and the associated functions [36,44]. We fixed a coefficient alpha equal to.5 to give the same importance to the LASSO and Ridge constraints. We optimized the shrinkage parameter lambda via a 10-fold cross validation.

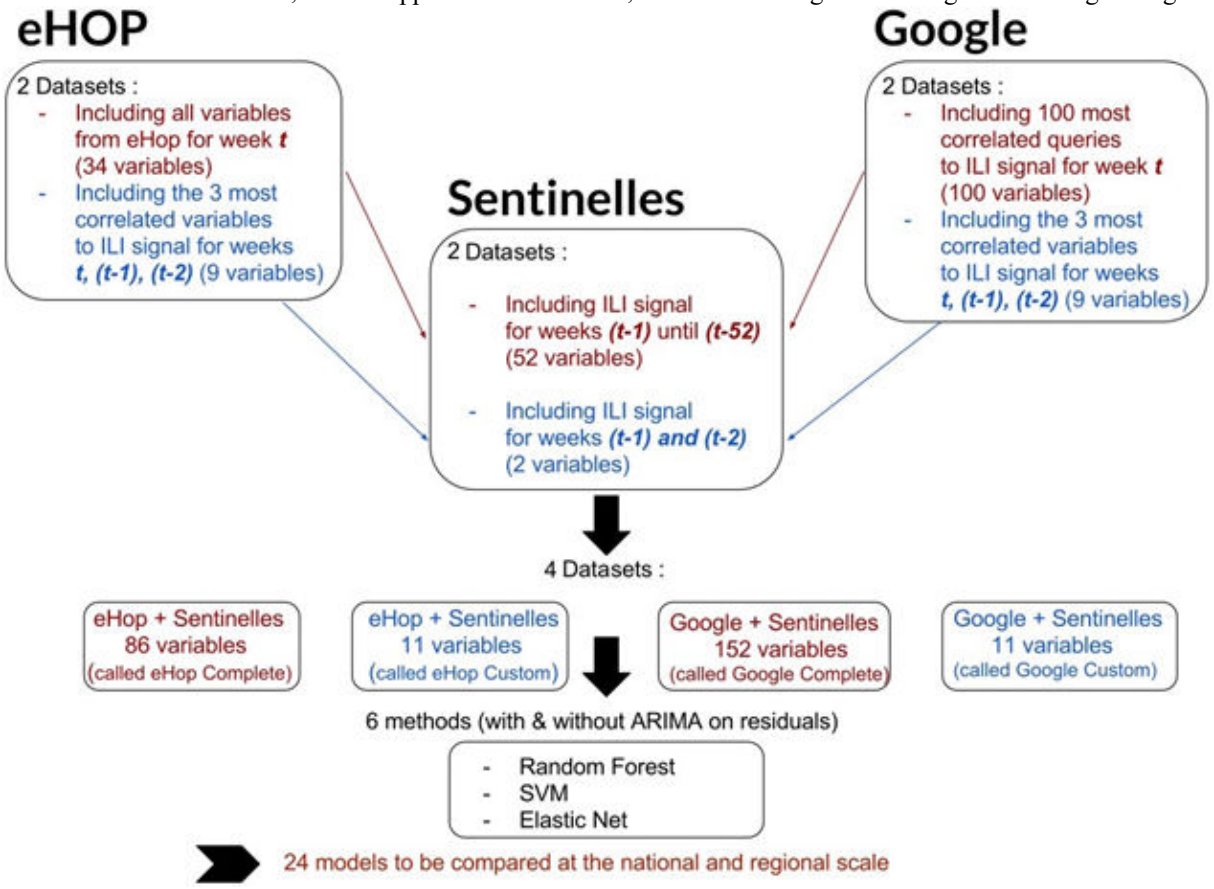
Random Forest

RF model combines decision trees constructed at training time using the general bootstrap aggregating technique (known as bagging) [45]. We used the R package randomForest to create RF models with a number of decision trees equal to 1500 [36,46].

Support Vector Machine

SVM is a supervised machine learning algorithm that can be used for classification or regression analyses [47]. Unlike multivariate regression models, SVM can learn nonlinear functions with the kernel trick that maps independent variables in a higher dimensional feature space. As Santillana et al [7], we used the linear kernel and optimized the cost parameter via a 10-fold cross validation with the R package e1071 [36,48].

Figure 1. Schematic representation of the study design, including the data preparation and data modeling steps. ILI: influenza-like illness; SVM: support vector machine; ARIMA: autoregressive integrated moving average.



Validity

Elastic net is a model that fulfills some assumptions on residuals. Means and variances must be constant, and residuals must be not correlated. Thus, residuals are called white noise. To test the stationarity and whiteness, we used Dickey Fuller's and Box-Pierce's tests available from the R packages tseries and stats [36,49]. When assumptions were not respected, we fitted residuals with a model of temporal series, called autoregressive integrated moving average (ARIMA) model. For RF and SVM, assumptions on residuals are not required. However, for comparison purpose, we tested them with the ARIMA model on residuals (Multimedia Appendices 4 and 5). We also assessed the calibration of the models by plotting the estimates against the real observations and by adding the regression line [50] (Multimedia Appendices 6 and 7).

Evaluation

We compared our ILI estimates with ILI incidence rates from the Sentinelles network by calculating different indicators. The mean squared error (MSE); Pearson correlation coefficient (PCC); variation in the height of the epidemic peak (ΔH), which corresponds to the difference between the height of the ILI incidence rate peak during the epidemic period estimated by the models and the height estimated by the Sentinelles network; and prediction lag (ΔL), which corresponds to the time difference between the ILI incidence rate peak estimated by the models and the peak estimated by the Sentinelles network, were calcuated. For the global comparison (ie, the entire study period), we calculated only the MSE and PCC. We calculated the four metrics only for the epidemic periods (plus 2 weeks before the start and after the end of the

epidemic). The start and end date of epidemics were obtained from the Sentinelles network [39]. Indeed, clinicians want to know when an epidemic starts and finishes, as well as its amplitude and severity. Therefore, interepidemic periods are less important. We also calculated the mean of each indicator for each influenza season to assess the model robustness. We also added two indicators to the mean of (ΔH) and (ΔL): the mean of |ΔH| and |ΔL|. We used the mean of (ΔH) to assess whether the models tended to underestimate or overestimate the peak calculated by the Sentinelles network, and the mean of (ΔL) to determine whether the predictions made by our models were too late or too in advance relative to the Sentinelles data. The mean of |ΔH| and |ΔL| allowed us to assess the estimate variability.

Results

Principal Results

Here, we show the results we obtained with the four datasets and three models—RF, SVM, and elastic net+residuals fitted by ARIMA (ElasticNet+ARIMA). The model on residuals was required to fulfill the assumptions for elastic net but not for the RF and SVM models. All results are presented in Multimedia Appendices 4 and 5. Moreover, we present two influenza outbreaks, including the 2010-2011 season (flu outbreak period for which the best estimates were obtained with all models) and the 2013-2014 season (flu outbreak period for which the worst estimates were obtained with all models; Multimedia Appendix 8). The calibration plots are in presented in Multimedia Appendices 7 and 9.

National Analysis

Dataset Comparison

PCC ranged from 0.947 to 0.980 when using the eHOP datasets (Multimedia Appendix 8) and from 0.937 to 0.978 with the Google datasets. MSE ranged from 2292 to 866 for the eHOP and from 2607 to 968 for the Google datasets. The mean PCC values during epidemic periods varied from 0.90 to 0.96 for the eHOP and from 0.87 to 0.96 for the Google datasets. The mean MSE values ranged from 7597 to 2664 for the eHOP and from 9139 to 2805 for the Google datasets.

Model Comparison

The eHOP Custom dataset gave the best results with the SVM model and ElasticNet+ARIMA (Multimedia Appendix 8). The SVM model and ElasticNet+ARIMA showed similar performance concerning the global activity (PCC=0.98; MSE<900) and also during epidemic periods (mean values), although PCC decreased (0.96) and the MSE increased (> 2500). Both models tended to overestimate the height of the epidemic peaks (ΔH=6 with SVM; ΔH=26 with ElasticNet+ARIMA), but the SVM model was slightly more accurate (|ΔH|=19 for SVM; |ΔH|=30 for the ElasticNet+ARIMA model). Conversely, the SVM model showed a larger prediction lag (ΔL=+0.83). Figure 2 illustrates the estimates obtained with the best models (SVM and ElasticNet+ARIMA with the dataset eHop Custom).

The same figure with the dataset Google Custom is presented in Multimedia Appendix 10. In the same way, there is a figure with eHOP Custom and Google Custom datasets with the model ElasticNet+ARIMA presented in Multimedia Appendix 11.

For the outbreak of 2010-2011, eHOP Custom using ElasticNet+ARIMA gave the best PCC (0.98) and the best MSE (1222). With this model, there was a slight overestimation of the height of the epidemic peak (ΔH=23) and a prediction lag of 1 week. For the 2013-2014 outbreak,

eHOP Custom using SVM gave the best PCC (0.95) and MSE (996), as well as the best ΔH (19) and prediction lag (1 week; Multimedia Appendix 8).

Regional Analysis

Figure 3 shows that ILI incidence rate variations were more important at the regional than the national level. For this reason, PCC decreased and MSE increased by the order of magnitude. The same figure with the dataset Google Custom is presented in Multimedia Appendix 12.

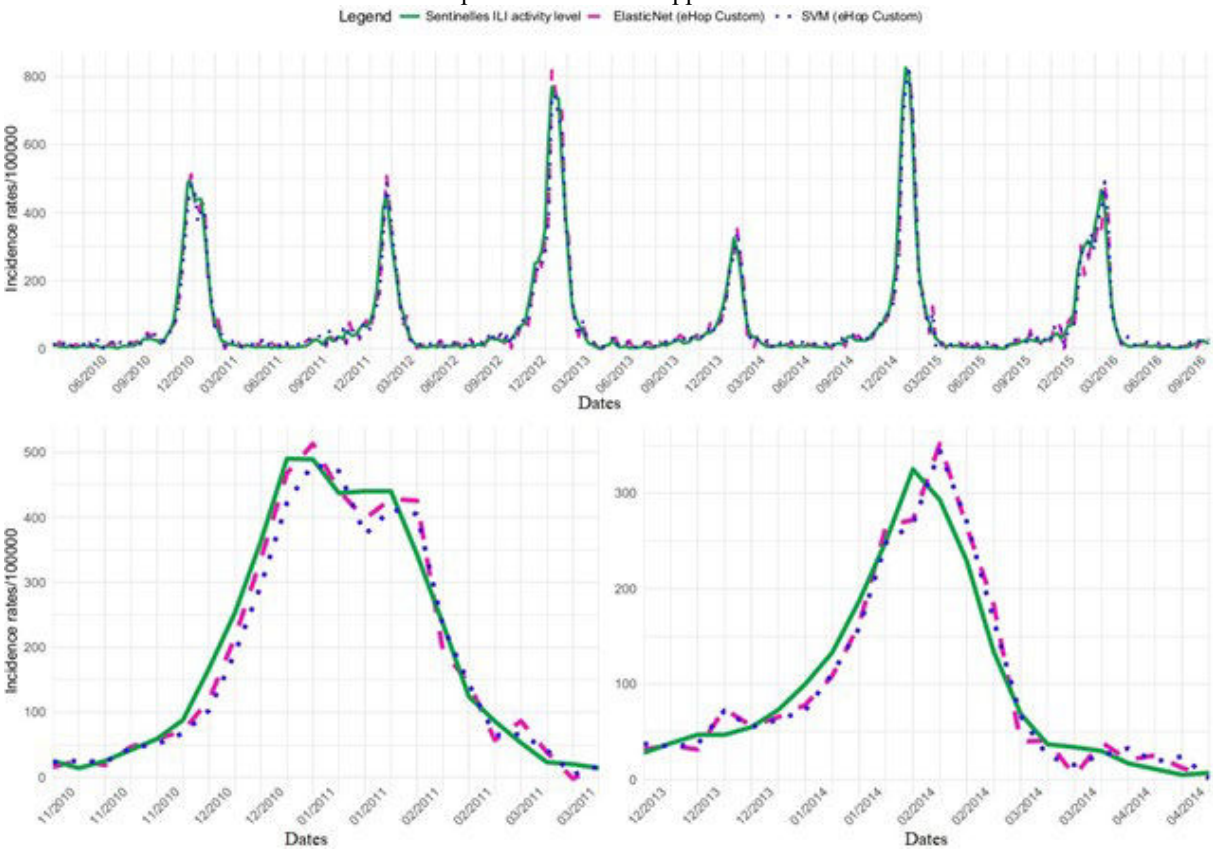
Dataset Comparison

PCC ranged from 0.911 to 0.923 (Multimedia Appendix 8) with the eHOP and from 0.890 to 0.912 with the Google datasets. MSE varied from 2906 to 2364 and from 3348 to 2736 for the eHOP and Google datasets, respectively. During epidemic periods, the mean PCC value ranged from 0.83 to 0.86 and from 0.70 to 0.83 for the eHOP and Google datasets, respectively. The mean MSE values ranged from 7423 to 5893 for the eHOP and from 9598 to 7122 for the Google datasets.

Model Comparison

Like at the national scale, eHOP Custom allowed obtaining the best PCC and MSE, and the SVM (PCC=0.923; MSE=2364) and ElasticNet+ARIMA (PCC=0.918; MSE=2451) models showed similar performances (Multimedia Appendix 8). Similar results were obtained also for the mean values during epidemic periods. Nevertheless, the PCC decreased (0.86 for SVM and 0.84 for ElasticNet+ARIMA), and the MSE increased (6050 for SVM and 5999 for ElasticNet+ARIMA). Both models tended to underestimate the height of the epidemic peaks (ΔH=−60 with SVM; ΔH=−32 with ElasticNet+ARIMA). The SVM model gave better PCC and MSE than the ElasticNet+ARIMA model, but ElasticNet+ARIMA was slightly more accurate for the epidemic peak height (|ΔH|=60 for SVM; |ΔH|=38 for the ElasticNet+ARIMA model). Although both models had a prediction lag (ΔL=+0.3), the ElasticNet+ARIMA model absolute lag value was smaller than that of SVM (|ΔL|=0.7; |ΔL|=1). For the 2010-2011 outbreak, eHOP Complete using the RF model gave the best PCC (0.92) and MSE (4263); with this model, there was a slight peak underestimation (ΔH=−40) but no prediction lag. For the 2013-2014 epidemic, the best PCC (0.78) and MSE (2113) were obtained with the Google Complete dataset and the ElasticNet+ARIMA model; there was a slight epidemic peak height underestimation (ΔH=−26) and 1 week of prediction lag.

Figure 2. National influenza-like illness (ILI) activity retrospective estimates obtained using the eHOP Custom dataset and the elastic net model with residuals fitted or the support vector machine model compared with the ILI activity levels from the French national Sentinelles networks. Global signal and 2010-2011 and 2013-2014 outbreaks are presented. SVM: support vector machine.



Discussion

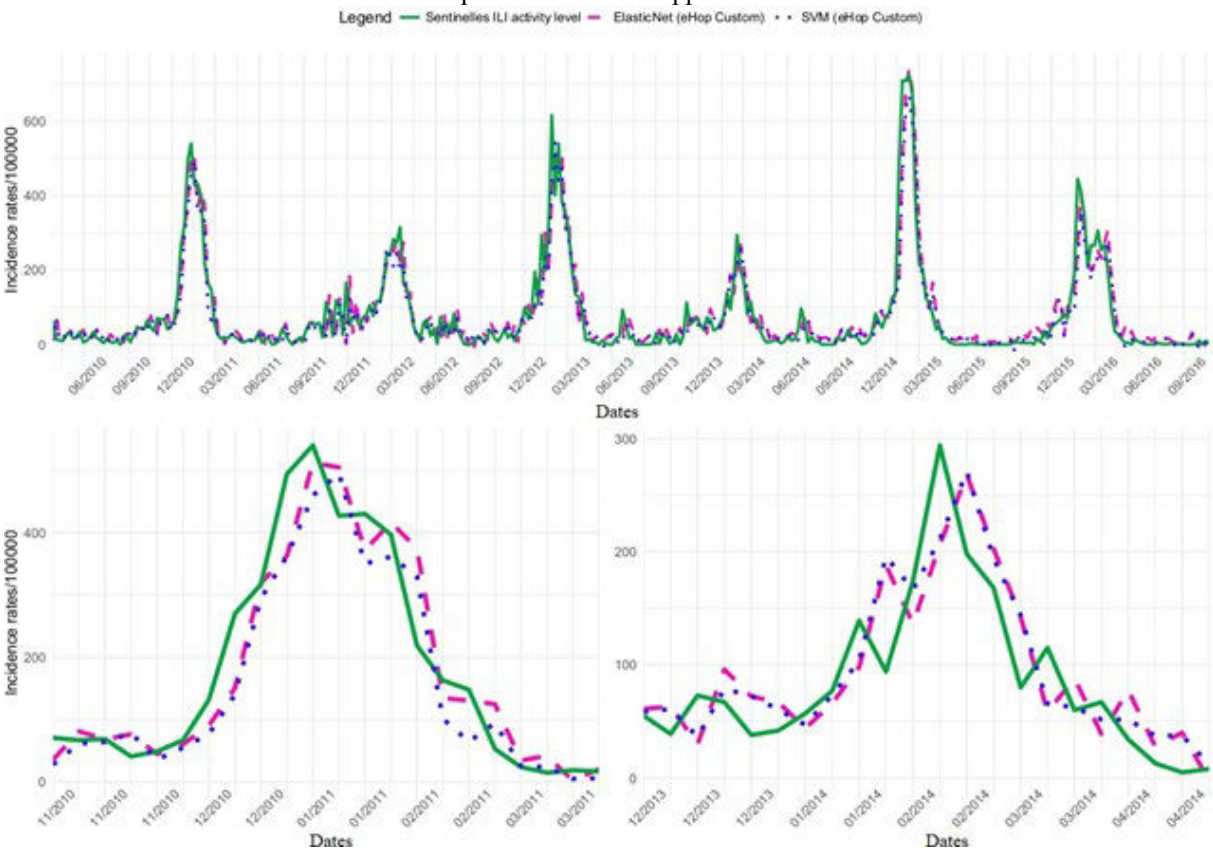
Data

Here, we show that HBD in combination with flu activity-level data from a national surveillance network allows accurately predicting ILI incidence rate at the national and regional scale and outperform Google data in most cases. The correlation coefficients obtained for the French data are comparable to those reported by studies on US data [2,7]. At the national and regional level, the best PCC and the best MSE during the entire study period or during epidemics were obtained using the eHOP Custom dataset. Moreover, the PCC and MSE values obtained with the eHOP datasets were better than those obtained with the Google datasets, particularly at the regional level (PCC 0.911-0.923 vs 0.890-0.912; MSE 2906-2364 vs 3348-2736,
respectively; Multimedia Appendix 8). However, the national signal is smoother and less noisy than the regional signal; the contribution of other data sources, such as hospital data or Web data, in addition to historical influenza data is more important at the regional level (Multimedia Appendices 4 and 5). The contribution of these external sources being less important at the national level, the differences observed between hospital data and Web data at this scale could be more significant.

Like internet data, some HBD can be obtained in near real time, especially records from emergency departments that are available on the same day or the day after. This is the most important data source for our models using eHOP datasets. Some other data, such as laboratory

results, are available only on a weekly basis; however, they are not the most important data source for our models.

Figure 3. Regional influenza-like illness (ILI) activity retrospective estimates obtained using the eHOP Custom dataset and the elastic net model with residuals fitted or the support vector machine model compared with the ILI activity levels from the French regional Sentinelles networks. Global signal and 2010-2011 and 2013-2014 outbreaks are presented. SVM: support vector machine.



Moreover, in comparison to internet data, HBD have some additional advantages. First, data extracted from CDWs are real health data can give information that cannot be extracted from internet data, particularly information about patients (sex, age, and comorbidities) [51]. In addition, an important clinical aspect is to determine the epidemic severity. With HBD, it is possible to gauge this parameter by taking into account the number of patients who were admitted in intensive care or died as the result of flu. Second, some CDW data (particularly emergency department discharge summaries and laboratory test results) can confirm that people were really affected by influenza or ILI symptoms. On the other hand, people can make internet queries not because they are ill, but for other people, for prevention purposes or just because it is a topical subject. Third, HBD could also be used to estimate the incidence rates of diseases that do not generate internet activity (eg, diseases without or with little media coverage or that are not considered interesting by the general population). Fourth, there is a spatial decorrelation between internet data and the regional estimates that were not observed with the eHOP data. It is quite reasonable that hospital-based data give a better estimate of regional epidemics, although currently, we have only data from Rennes University Hospital that might not be representative of the entire Brittany region.

A major HBD limitation is that, generally, clinical data are not publicly available. In our case, we could only access the Rennes University Hospital HBD. However, the epidemic peak in Brittany could have occurred earlier or later relative to the national peak, and this could have

introduced a bias in our estimation. We can hypothesize that ILI estimates, particularly nationwide, might be improved if we could extract information from HBD in other regions. In the United States, a patient research system allows aggregating patient observations from a large number of hospitals in a uniform way [52]. In France, several initiatives have been developed to create search systems. For instance, an ongoing project (Réseau interrégional des Centres de Données Cliniques) [53] in the Northwest area of France associates six University Hospital centers (Angers, Brest, Nantes, Poitiers, and Rennes et Tours) and Orleans Regional Hospital Centre, thus collecting data on patients in the Bretagne, Centre-Val de Loire, and Pays de la Loire regions. This corresponds to 15.5% of Metropolitan France and 14.4% of the entire French population. Another way to aggregate patient data could be a cloud-based platform, and we are currently setting up this kind of architecture; this platform will integrate two University Hospital centers, Brest and Rennes, the French health reimbursement database (Système national d'information interrégimes de l'Assurance Maladie) and registries, such as the birth defect registry or cancer registry.

Statistical Models

Regarding the statistical models, we show that SVM and elastic net with ARIMA model are fairly comparable with PCC ranging from 0.970 to 0.980 at the national scale and from 0.890 to 0.923 at the regional scale. The SVM and elastic net models in combination with the eHOP custom dataset were the most robust models, although they did not always give the best results. Indeed, they showed the best performance in term of PCC and MSE for the global signal and also for the mean values. Nevertheless, these models have some limits. The main limitation of the SVM model is the very slow parameter optimization when there are many variables. With the SVM model, it can be important to preselect the important variables to reduce the dataset size to improve the optimization speed. For this, one needs a good knowledge of the available data, which may be difficult when using big data. On the other hand, elastic net shows good performance with many variables, which is an advantage when the most relevant variables to estimate ILI incidence rates are not known in advance. The elastic net model is a parametric model that fulfills certain assumptions on residuals, differently from the SVM model. With elastic net, residuals must be fitted to have a statistically valid model. Nevertheless, if we had to choose a model, we would prefer SVM with the eHOP Custom dataset because it has a better PCC than elastic net at the regional scale.

Another limitation is that indicators are better for the global period than for epidemic periods. This implies that models are less efficient during flu outbreaks, while clinical concerns are higher during epidemics when good estimates of the outbreak starting date, amplitude, and end are needed. Finally, the results of our models with Web data may have been overestimated due to the way we obtained data from Google Correlate. Indeed, Google Correlate used information that we did not have at the beginning of our test period. The time period for our time series passed into Google Correlate is from January 2004 to October 2016. But, the beginning of our test period for our models is January 2010. To be more precise, we should recalculate the correlation coefficients for each week to predict with the data available at that time.

In the same way, to custom datasets, we calculated the 3 most correlated variables on a time period including our test period. To compare the results, we built another dataset from eHOP, including the 3 most correlated variables to ILI regional signal between December 2003 and December 2009 (before our test period), and we applied an ElasticNet+ARIMA model. In this way, we kept 2 variables on the 3 present in the eHOP custom dataset. The difference does not seem significant (Multimedia Appendix 6), but it would be interesting to test this hypothesis with all models at the national and regional scale with Google and eHOP custom datasets.

Perspectives

Future research could address clinical issues not only nationally or regionally but also at finer spatial resolutions such as a city like Lu et al did [54], a health care institution or in subpopulations. Indeed, by predicting epidemics, it will be possible to organize hospitals during epidemics (eg, bed planning and anticipating overcrowding). Moreover, in this study, we compared internet and HBD data; however, hybrid systems could be developed to take advantage of multiple sources [55,56]. For instance, internet data might avoid the limit of the local source linked to the choice or availability of HBD. Data collected by volunteers who self-report symptoms in near real time could be exploited [57]. Similarly, by combining models, we could retain the benefits of each of them and improve the estimates of ILI incidence rates. For example, we could use another algorithm, such as stacking [58], to concomitantly use the SVM and elastic net models. We could also test other kernels than the linear kernel for SVM models. Finally, we carried out a retrospective study using various models with clinical data in combination with the flu activity from the Sentinelles network to estimate ILI incidence rates in real time. Our models need now to be tested to determine whether they can anticipate and predict ILI incidence rates.

Conclusions

Here, we showed that HBD is a data source that allows predicting the ILI activity as well or even better than internet data. This can be done using two types of models with similar performance—SVM (a machine learning model) and elastic net (a model of regularized regression). This is a promising way for monitoring ILI incidence rates at the national and local levels. HBD presents several advantages compared with internet data. First, they are real health data and can give information about patients (sex, age, and comorbidities). This could allow for making predictions on ILI activity targeted to a specific group of people. Second, hospital data can be used to determine the epidemic severity by taking into account the number of patients who were admitted in intensive care or died as a result of flu. Third, hospital data (particularly the emergency department discharge summaries and laboratory test results) can confirm that people were really affected by influenza. Finally, HBD could also be used to estimate the incidence rates of diseases that do not generate internet activity. Although massive data cannot take the place of traditional influenza surveillance methods at this time, they could be used to complete them. For instance, real-time forecasting is necessary for decision making. It can also be used to manage the patients' flow in general practitioners' offices and hospitals, particularly emergency departments.

Acknowledgments

We would like to thank the French National Research Agency for funding this work inside the Integrating and Sharing Health Data for Research Project (Grant No. ANR-15-CE19-0024). We thank Magalie Fromont Renoir and Ronan Le Gue´vel from the University of Rennes 2 who provided insight and expertise that greatly assisted the research. We also thank the French Sentinelles network for making their data publicly available.

Authors' Contributions

CP, GB, AL, and BCG conceived the experiments; CP conducted the experiments and analyzed the results.

Conflicts of Interest

None declared.

Multimedia Appendix 1

eHOP queries (with the number of concerned hospital stays from 2003 to 2016).
[PDF File (Adobe PDF File), 21KB - publichealth_v4i4e11361_app1.pdf]

Multimedia Appendix 2

The 100 most correlated Google queries at national level.
[PDF File (Adobe PDF File), 15KB - publichealth_v4i4e11361_app2.pdf]

Multimedia Appendix 3

The 100 most correlated Google queries at regional level.
[PDF File (Adobe PDF File), 14KB - publichealth_v4i4e11361_app3.pdf]

Multimedia Appendix 4

Accuracy metrics for all seasons obtained with all models for the national scale.
[PDF File (Adobe PDF File), 72KB - publichealth_v4i4e11361_app4.pdf]

Multimedia Appendix 5

Accuracy metrics for all seasons obtained with all models for the regional scale.
[PDF File (Adobe PDF File), 80KB - publichealth_v4i4e11361_app5.pdf]

Multimedia Appendix 6

Comparison between two datasets with ElasticNet + ARIMA model: Dataset 1 corresponds to the dataset called eHOP Custom used in the paper and including the 3 most correlated variables to ILI signal between December 2009 to October 2016 (our test period). Dataset 2 includes the 3 most correlated variables to ILI signal between December 2003 to December 2009 (before our test period).
[PDF File (Adobe PDF File), 25KB - publichealth_v4i4e11361_app6.pdf]

Multimedia Appendix 7

National calibration.
[PNG File, 185KB - publichealth_v4i4e11361_app7.png]

Multimedia Appendix 8

Accuracy metrics for the 2010-2011 (flu outbreak period for which the best estimates were obtained with all models) and 2013-2014 (flu outbreak period for which the worst estimates were obtained with all models) seasons. PCC and MSE for the global period (Global) and mean values (Means) of all indicators for each model during the epidemic periods. In bold, the best results for each dataset. a. Data for the whole France. b. Data for the Brittany region.
[PDF File (Adobe PDF File), 52KB - publichealth_v4i4e11361_app8.pdf]

Multimedia Appendix 9

Regional calibration.
[PNG File, 202KB - publichealth_v4i4e11361_app9.png]

Multimedia Appendix 10

National ILI activity retrospective estimates obtained using the Google Custom dataset and the Elastic Net model with residuals fitted (pink dashed line) or the SVM model (blue dotted line) compared with the ILI activity levels from the French national Sentinelles networks (green solid line). a. Global signal. b. 2010-2011 and c. 2013-2014 outbreaks.
[PNG File, 142KB - publichealth_v4i4e11361_app10.png]

Multimedia Appendix 11

National ILI activity retrospective estimates obtained using the Google Custom dataset and the Elastic Net model (blue dotted line) or eHOP Custom dataset and the Elastic Net model (pink dashed line) compared with the ILI activity levels from the French national Sentinelles networks (green solid line). a. Global signal. b. 2010-2011 and c. 2013-2014 outbreaks.
[PNG File, 142KB - publichealth_v4i4e11361_app11.png]

Multimedia Appendix 12

Regional ILI activity retrospective estimates obtained using the Google Custom dataset and the Elastic Net model with residuals fitted (pink dashed line) or the SVM model (blue dotted line) compared with the ILI activity levels from the French regional Sentinelles networks (green solid line). a. Global signal. b. 2010-2011 and c. 2013-2014 outbreaks.
[PNG File, 159KB - publichealth_v4i4e11361_app12.png]

References

1. Ferguson NM, Cummings DAT, Fraser C, Cajka JC, Cooley PC, Burke DS. Strategies for mitigating an influenza pandemic. Nature 2006 Jul 27;442(7101):448-452. [doi: 10.1038/nature04795] [Medline: 16642006]
2. Yang S, Santillana M, Kou S. Accurate estimation of influenza epidemics using Google search data via ARGO. Proceedings of the National Academy of Sciences 2015 Nov 24:14473. [doi: 10.1038/srep25732]
3. Si-Tahar M, Touqui L, Chignard M. Innate immunity and inflammation--two facets of the same anti-infectious reaction. Clin Exp Immunol 2009 May;156(2):194-198 [FREE Full text] [doi: 10.1111/j.1365-2249.2009.03893.x] [Medline: 19302246]
4. Yang W, Lipsitch M, Shaman J. Inference of seasonal and pandemic influenza transmission dynamics. Proc Natl Acad Sci USA 2015 Feb 17;112(9):2723-2728. [doi: 10.1073/pnas.1415012112] [Medline: 25730851]
5. Nichol KL. Cost-benefit analysis of a strategy to vaccinate healthy working adults against influenza. Arch Intern Med 2001 Mar 12;161(5):749-759. [Medline: 11231710]
6. Fleming DM, Van Der Velden J, Paget WJ. M. Fleming WJP J van der Velden. The evolution of influenza surveillance in Europe and prospects for the next 10 years. Vaccine ? 2003;21:1753.
7. Santillana M, Nguyen AT, Louie T, Zink A, Gray J, Sung I, et al. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. Sci Rep 2016 Dec 11;6:25732 [FREE Full text] [doi: 10.1038/srep25732] [Medline: 27165494]
8. Nsoesie E, Brownstein J, Ramakrishnan N. A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza and Other Respiratory Viruses ? 2014;8:316.
9. Chretien J, George D, Shaman J, Chitale RA, McKenzie FE. Influenza forecasting in human populations: a scoping review. PLoS One 2014;9(4):e94130 [FREE Full text] [doi: 10.1371/journal.pone.0094130] [Medline: 24714027]
10. Soebiyanto RP, Adimi F, Kiang RK. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. PLoS One 2010 Mar 01;5(3):e9450 [FREE Full text] [doi: 10.1371/journal.pone.0009450] [Medline: 20209164]
11. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012-2013 season. Nat Commun 2013;4:2837 [FREE Full text] [doi: 10.1038/ncomms3837] [Medline: 24302074]
12. Milinovich GJ, Williams GM, Clements ACA, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 2014 Feb;14(2):160-168. [doi: 10.1016/S1473-3099(13)70244-5] [Medline: 24290841]

13. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
14. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. Proceedings of the National Academy of Sciences 2012 Nov 26;109(50):20425-20430. [doi: 10.1073/pnas.1208772109] [Medline: 23184969]
15. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol 2013;9(10):e1003256 [FREE Full text] [doi: 10.1371/journal.pcbi.1003256] [Medline: 24146603]
16. Zhang Y, Bambrick H, Mengersen K, Tong S, Hu W. Using Google Trends and ambient temperature to predict seasonal influenza outbreaks. Environment International 2018;117:91.
17. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 2013;8(12):e83672 [FREE Full text] [doi: 10.1371/journal.pone.0083672] [Medline: 24349542]
18. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr 2014 Oct 28;6 [FREE Full text] [doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117] [Medline: 25642377]
19. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013-2014 influenza season using Wikipedia. PLoS Comput Biol 2015 May;11(5):e1004239 [FREE Full text] [doi: 10.1371/journal.pcbi.1004239] [Medline: 25974758]
20. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol 2014 Nov;10(11):e1003892 [FREE Full text] [doi: 10.1371/journal.pcbi.1003892] [Medline: 25392913]
21. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol 2014 Apr;10(4):e1003581 [FREE Full text] [doi: 10.1371/journal.pcbi.1003581] [Medline: 24743682]
22. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 2009 Nov 15;49(10):1557-1564. [doi: 10.1086/630200] [Medline: 19845471]
23. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014 Mar 14;343(6176):1203-1205. [doi: 10.1126/science.1248506] [Medline: 24626916]
24. Butler D. When Google got flu wrong. Nature 2013 Feb 14;494(7436):155-156. [doi: 10.1038/494155a] [Medline: 23407515]
25. Hanauer DA. EMERSE: The Electronic Medical Record Search Engine. 2006 Presented at: AMIA Annual Symposium Proceedings; 2006/11/11; Washington p. 941.
26. Murphy SN, Mendis ME, Berkowitz DA. Integration of Clinical and Genetic Data in the i2b2 Architecture. 2006 Presented at: AMIA Annual Symposium Proceedings; 2006; Washington p. 1040.
27. Lowe HJ, Ferris TA, Hernandez PM. STRIDE ? An Integrated Standards-Based Translational Research Informatics Platform. 2009 Presented at: AMIA Annual Symposium Proceedings; 2009; San Francisco p. 391.
28. Cuggia M, Garcelon N, Campillo-Gimenez B. Roogle: an information retrieval engine for clinical data. Studies in Health Technology and Informatics 2011;169:8. [doi: 10.3233/978-1-60750-806-9-584]
29. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
30. Murphy S, Wilcox A. Mission and Sustainability of Informatics for Integrating Biology and the Bedside (i2b2). EGEMS (Wash DC) 2014;2(2):1074 [FREE Full text] [doi: 10.13063/2327-9214.1074] [Medline: 25848608]
31. Viboud C, Charu V, Olson D, Ballesteros S, Gog J, Khan F, et al. Demonstrating the use of high-volume electronic medical claims data to monitor local and regional influenza activity in the US. PLoS One 2014;9(7):e102429 [FREE Full text] [doi: 10.1371/journal.pone.0102429] [Medline: 25072598]
32. Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert ML, Chabot M, Chazard E, et al. Leveraging hospital big data to monitor flu epidemics. Computer Methods and Programs in Biomedicine 2018:160.
33. Santillana M, Nsoesie EO, Mekaru SR, Scales D, Brownstein JS. Using clinicians' search query data to monitor influenza epidemics. Clin Infect Dis 2014 Nov 15;59(10):1446-1450 [FREE Full text] [doi: 10.1093/cid/ciu647] [Medline: 25115873]
34. Google Correlate. URL: https://www.google.com/trends/correlate [accessed 2018-06-19] [WebCite Cache ID 70IClAsSD]

35. Google Trends. URL: https://trends.google.fr/trends/?geo=FR [accessed 2018-06-20] [WebCite Cache ID 70JjgMxmh]
36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing 2015 [FREE Full text]
37. Massicotte P, Eddelbuettel D. gtrendsR: Perform and Display Google Trends Queries. https://github.com/PMassicotte/gtrendsR 2017 [FREE Full text]
38. Valleron AJ, Bouvet E, Garnerin P. A computer network for the surveillance of communicable diseases: the French experiment. American Journal of Public Health 1986;76:92.
39. Flahault A, Blanchon T, Dorléans Y, Toubiana L, Vibert JF, Valleron AJ. Virtual surveillance of communicable diseases: a 20-year experience in France. Stat Methods Med Res 2006 Oct;15(5):413-421. [doi: 10.1177/0962280206071639] [Medline: 17089946]
40. Réseau Sentinelles. URL: https://websenti.u707.jussieu.fr/sentiweb [accessed 2018-06-19] [WebCite Cache ID 70IEHtetc]
41. Zou H, Hastie T. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society 2005;67:320.
42. Kennard EH. Ridge regression: biaised estimation for nonorthogonal problems. Technometrics 1970;1.
43. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society 1996;58:267-288.
44. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 2010;33:1-22.
45. Breiman L. Random Forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
46. Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2:18-22.
47. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(3):273-297. [doi: 10.1007/BF00994018]
48. Meyer D, Dimitriadou E, Hornik K. e1071: Misc Functions of the Department of Statistics. Probability Theory Group (Formerly: E1071) https://CRAN.R-project.org/package=e1071 2015.
49. Trapletti A, Hornik K. tseries: Time Series Analysis and Computational Finance. http://CRAN.R-project.org/package=tseries 2015.
50. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 2010:128-138.
51. Olson D, Heffernan R, Paladini M, Konty K, Weiss D, Mostashari F. Monitoring the Impact of Influenza by Agemergency Department Fever and Respiratory Complaint Surveillance in New York City. PLOS Medicine 2007;4(8).
52. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One 2013;8(3):e55811 [FREE Full text] [doi: 10.1371/journal.pone.0055811] [Medline: 23533569]
53. Bouzillé G, Westerlynck R, Defossez G. Sharing health big data for research - A design by use cases: the INSHARE platform approach. Studies in Health Technology and Informatics 2017.
54. Lu F, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. JMIR Public Health Surveillance 2018;4(1).
55. Groupment Interrégional de Recherche Clinique et d'Innovation Grand Ouest. URL: https://www.girci-go.org/ [accessed 2018-06-20] [WebCite Cache ID 70JklABe6]
56. Simonsen L, Gog JR, Olson D, Viboud C. Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems. J Infect Dis 2016 Dec 01;214:S380-S385 [FREE Full text] [doi: 10.1093/infdis/jiw376] [Medline: 28830112]
57. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. Big Data for Infectious Disease Surveillance and Modeling. J Infect Dis 2016 Dec 01;214:S375-S379 [FREE Full text] [doi: 10.1093/infdis/jiw400] [Medline: 28830113]
58. Wolpert DH. Stacked generalization. Neural Networks 1992.

II. Recherche clinique et épidémiologie

En recherche clinique, plusieurs apports de la réutilisation des données peuvent être envisagés. Sur le plan expérimental, il peut s'agir de réexploiter les données d'essais cliniques afin de répondre à de nouvelles questions de recherche tout en bénéficiant de données de bonne qualité ou alors d'exploiter des données de soins pour générer de nouvelles hypothèses en simulant un cadre expérimental classique en recherche clinique.

Par ailleurs, un grand axe dans l'exploitation des données pour la recherche clinique est de faciliter le déroulement des essais cliniques ou l'épidémiologie, et ce à toutes les étapes : études de faisabilité, préscreening, aide à la collecte de données voire alimentation automatique de bases de données. L'identification exhaustive des individus correspondant aux critères d'inclusion des cohortes épidémiologiques ou d'essais cliniques est indispensable pour produire des analyses sans biais de sélection. Dans le cadre d'études cas-témoins, il est également indispensable d'identifier les témoins les plus adaptés dans la population source.

Ces critères d'éligibilité sont le plus souvent extrêmement précis, les identifier à partir des données est un processus complexe pouvant faire appel à différentes méthodes telles que la recherche d'information ou le traitement automatique du langage.

Article 6 : Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation

L'article ci-après présente un travail en traitement automatique du langage qui avait pour objectif d'évaluer la capacité de ces méthodes à extraire les critères d'éligibilité de type numérique depuis des documents textuels pour faciliter l'inclusion de patients dans les essais cliniques.

Ma contribution à ce travail a été de définir la méthodologie d'annotation des documents afin que celle-ci soit pertinente pour le cas d'usage et de réaliser l'annotation du corpus.

Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation

Vincent Claveau[1], Lucas Emanuel Silva Oliveira[2], Guillaume Bouzillé[3], Marc Cuggia[3], Claudia Maria Cabral Moro[2], Natalia Grabar[4]

1 IRISA - CNRS, Rennes, France
2 PUCPR - Pontifíicia Universidade Católica do Paraná, Curitiba, Brazil
3 INSERM/LTSI, HBD; CHU de Rennes; Université Rennes 1
4CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
vincent.claveau@irisa.fr, lucas.oliveira@pucpr.br, guillaume.bouzille@chu-rennes.fr, marc.cuggia@chu-rennes.fr, c.moro@pucpr.br, natalia.grabar@univ-lille3.fr

Abstract. Clinical trials are fundamental for evaluating therapies and diagnosis techniques. Yet, recruitment of patients remains a real challenge. Eligibility criteria are related to terms but also to patient laboratory results usually expressed with numerical values. Both types of information are important for patient selection. We propose to address the processing of numerical values. A set of sentences extracted from clinical trials are manually annotated by four annotators. Four categories are distinguished: *C* (concept), *V* (numerical value), *U* (unit), *O* (out position). According to the pairs of annotators, the inter-annotator agreement on the whole annotation sequence *CV U* goes up to 0.78 and
0.83. Then, an automatic method using CFRs is exploited for creating a supervised model for the recognition of these categories. The obtained F-measure is 0.60 for *C*, 0.82 for *V* , and 0.76 for *U* .

Keywords: Natural Language Processing, Supervised Learning, Clinical Trials, Patient Eligibility, Numerical Criteria

1 Introduction

In the clinical research process, recruitment of patients for clinical trials (CTs) remain an unprecedented challenge, while they are fundamental for evaluating therapies or diagnosis techniques. They are the most common research design to test the safety and efficiency of interventions on humans. CTs are based on statistical inference and require appropriate sample sizes from well identified population. The challenge is to enroll a sufficient number of participants with suitable characteristics to ensure that the results demonstrate the desired effect with a limited error rate. Hence, CTs must define a precise set of inclusion and exclusion criteria (eg, age, gender, medical history, treatment, biomarkers). With paper files and EHRs as the main sources of information, only human operators are capable to efficiently detect the eligible patients [3]. This is a laborious and costly task, and it is common that CTs fail because of the difficulty to meet the necessary recruitment target in an acceptable time [6]: almost half of all

trial delays are caused by participant recruitment problems. Only 18% in Europe, 17% in Asia-Pacific, 15% in Latin America, and 7% in the USA complete enrollment on time [4]. The existing enrollment systems are facing the gap between the free text representation of clinical information and eligibility criteria [11, 17]. Most of them propose to fill in this gap manually, while automatic NLP methods may help to overcome this issue.

The traditional NLP work is dedicated to the recognition and extraction of terms. Yet, there is an emerging work on detection of temporality, certainty, and numerical values. Such information has the purpose to complete, enrich and more generally make more precise the terminological information. In the general language, framework for automated extraction and approximation of numerical values, such as height and weight, has been proposed [5]. It uses relation patterns and WordNet and shows the average precision up to 0.84 with exact matching and 0.72 with inexact matching. Another work proposes two extraction systems: rule based extractor and probabilistic graphical model [9] for extraction of life expectancy, inflation, electricity production, etc. It reaches 0.56 and 0.64 average F-measure for the rule-based and probabilistic systems, respectively. On the basis of a small set of clinical annotated data in French, a CRF model is trained for the recognition of concepts, values, and units. Then, a rule-based system is designed for computing the semantic relations between these entities [2]. The results obtained show average F-measure 0.86 (0.79 for concepts, 0.90 for values and 0.76 for units). On English data, extraction of numerical attributes and values from clinical texts is proposed [13]: after the extraction of numerical attributes and values with CRFs, relations for associating these attributes to values are computed with SVMs. The system shows 0.95 accuracy with entities and 0.87 with relations. Yet another work is done on cardiology radiological reports in English [10] and achieves 93% F1-measure. In contrast with these studies, here we focus on clinical trial protocols written in English.

2 Material

Clinical Trials. In December 2016, we downloaded protocols of the whole set of CTs from *www.clinicaltrials.com*. The corpus counts 211,438 CTs. We focus on inclusion and exclusion criteria (more than 2M sentences).

Reference Annotations. 1,500 randomly selected sentences are annotated by 3 annotators with different backgrounds (medical doctor and computer scientists). Each sentence is annotated by at least two of them. On such typical sentences:

- *Absolute neutrophil count ≥ 1,000 cells/µl.*
- *Exclude if T3 uptake is less than 19%; T4 less than 2.9 ((g/dL); free T4 index is less than 0.8.*

the annotators have to mark up three categories of entities: *C* (concepts *Absolute neutrophil count, T3 uptake, T4, free T4 index*), *V* (numerical values *≥ 1,000, less than 19, less than 2.9, less than 0.8*), *U* (units *cells/µl, %, g/dL*).

3 Methods

The main objective of the methods is to create an automatic model for the detection of numerical values (concept, value and unit).

Inter-annotator agreement. In order to assess the inter-annotator agreement, we compute Cohen's κ [1] between each pair of annotators. The final version is obtained after a consensus is reached among the annotators.

Automatic annotation. Conditional Random Fields (CRFs) [7] are undirected graphical models that represent the probability distribution of annotation y on observations x. They are widely used in NLP thanks to their ability to take into account the sequential aspect and rich descriptions of text sequences. CRFs have been successfully used in many tasks casted as annotation problems: information extraction, named entity recognition, tagging, etc. [18, 12, 14]. From training data, CRF models learn to assign a label to each word of a sentence such that the tag sequence is the most probable given the sentence given as input. We want the CRFs to learn to label words denoting a concept with the tag *C*, values with *V*, units with *U*, while every other words will receive a void label noted *O*. In order to handle multi-word concepts, values and units, we adopt the so-called BIO scheme: the first word of a multi-word concept is labeled as *B C* (*B* stands for *beginning*), next words are labeled as *I C* (*I* stands for *inside*), the same for values and units. To find the most probable label of a word at position *i* in a sentence, CRFs exploit features describing the word (for example, Part-of- Speech tags, lemmas [15], graphemic clues) and its context (words and features at positions i 1, $i + 1$, $i-2$) up to 4 words. The CRF implementation used for our experiments is Wapiti [8], which is known for its efficiency.

Evaluation of automatic annotation. The evaluation is performed against the reference data and is measured with token errors (percentage of words wrongly labeled with respect to the human annotation) and F-measure [16].

4 Results and Discussion

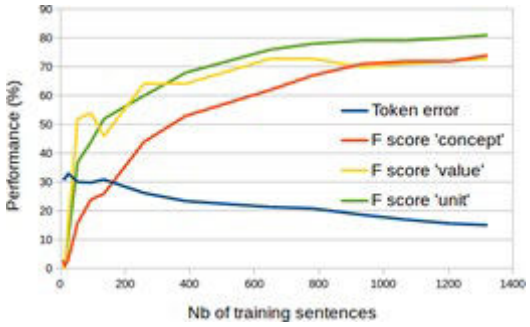
Inter-annotator agreement. InTable1,we indicate the inter-annotator agreement for each pair of annotators and taking into account two tagsets: whole set of tags and the tagset without concepts. The figures indicate that A1 and A2 show the highest agreement: both have important experience in medical area. When concepts are not taken into account the agreement is even better: manual annotation of concepts is more complicated than annotation of the two other categories. With annotations from A1 and A2, the consensual annotation is built. This version of data is used for training and evaluation of the supervised model. **Automatic annotation.** In Figure 1, we present the evaluation of automatic annotation in terms of token errors and F-score for each category. In order to estimate the ideal amount of the required training data, we also display the evolution of the performance according to the size of the training data used. First, the global error rate tends to decrease. Since its decrease continues, more training data would help reaching better results. Among the categories aimed,

Table 1. Inter-annotator agreement (Cohen's κ) on the whole and reduced tag sets

	A1 vs. A2	A1 vs. A3	A2 vs. A3

κ whole tagset	**0.78**	**0.51**	**0.47**
κ without'concept'	**0.83**	**0.60**	**0.64**

Fig. 1. CRF annotation performance (globally in terms of token errors, and by category in terms of F-score) according to the amount of training data (number of sentences)



the best performance is obtained with units, while the concept category is the most difficult to detect. For these two categories, the performance continues to grow up: a larger set of annotated data would be helpful. As for the value category, its evolution is less linear and finally it seems to find a "plateau" with no more apparent evolution. Otherwise, the detection efficiency of this category is in between the two other categories. The obtained F-measure is 0.60 for *C*, 0.82 for *V* , and 0.76 for *U*.

5 Conclusion and Future Work

Recruitment of patients for CTs is a difficult task. We proposed a contribution to this task. We generate an automatic model for the detection of numerical values, composed of three items (concept *C*, value *V* and unit *U*), in narrative text in English. These results are evaluated against reference data and show F-measure 0.60 for *C*, 0.82 for *V* , and 0.76 for *U*. We have several directions for future work: to normalize the units; to build resources and rules for their standardization (*cell/mm3* instead of *cell/cm3*); to prepare a larger set of reference annotations; to complete these annotations with temporal information; to apply the models for enrollment of patients in French and Brazilian hospitals.

Acknowledgements. This work was partly funded by CNRS-CONFAP project FIGTEM for Franco-Brazilian collaborations and a French government support granted to the CominLabs LabEx managed by the ANR in Investing for the Future program under reference ANR-10-LABX-07-01.

References

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics 34(4) (2008)
2. Bigeard, E., Jouhet, V., Mougin, F., Thiessard, F., Grabar, N.: Automatic extraction of numerical values from unstructured data in EHRs. In: MIE (Medical Informatics in Europe) 2015. Madrid, Spain (2015)
3. Campillo-Gimenez, B., Buscail, C., Zekri, O., Laguerre, B., Le Prisé, E., De Crevoisier, R., Cuggia, M.: Improving the pre-screening of eligible patients in order to increase enrollment in cancer clinical trials. Trials 16(1), 1–15 (2015)
4. Center Watch: State of the clinical trials industry: A sourcebook of charts and statistics. Tech. rep., Center Watch (2013)

5. Davidov, D., Rappaport, A.: Extraction and approximation of numerical attributes from the web. In: 48th Annual Meeting of the Association for Computational Linguistics. pp. 1308–1317 (2010)
6. Fletcher, B., Gheorghe, A., Moore, D., Wilson, S., Damery, S.: Improving the recruitment activity of clinicians in randomised controlled trials: A systematic review. BMJ Open 2(1), 1–14 (2012)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (ICML) (2001)
8. Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 504–513. Association for Computational Linguistics (July 2010), http://www. aclweb.org/anthology/P10-1052
9. Madaan, A., Mitta, A., Mausam, Ramakrishnan, G., Sarawagi, S.: Numerical relation extraction with minimal supervision. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
10. Nath, C., Albaghdadi, M., Jonnalagadda, S.: A natural language processing tool for large-scale data extraction from echocardiography reports. PLoS One 11(4), 153749–64 (2016)
11. Olasov, B., Sim, I.: Ruleed, a web-based semantic network interface for constructing and revising computable eligibility rules. In: AMIA Symposium. p. 1051 (2006)
12. Pranjal, A., Delip, R., Balaraman, R.: Part Of speech Tagging and Chunking with HMM and CRF. In: Proceedings of NLP Association of India (NLPAI) Machine Learning Contest (2006)
13. R., S.P., Mandhan, S., Niwa, Y.: Numerical attribute extraction from clinical texts. CoRR 1602.00269 (2016), http://arxiv.org/abs/1602.00269
14. Raymond, C., Fayolle, J.: Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In: Actes de la conférence Traitement Automatique des Langues Naturelles. Montréal, Canada (2010)
15. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proc of International Conference on New Methods in Language Processing. pp. 44–49 (1994)
16. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
17. Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P., Elhadad, N., Johnson, S., Lai, A.: A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 21(2), 221–30 (2014)
18. Wang, T., Li, J., Diao, Q., Wei Hu, Y.Z., Dulong, C.: Semantic event detection using conditional random fields. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06) (2006)

III. Pharmacovigilance

La pharmacovigilance a pour objet la mesure des événements en lien avec l'usage des produits de santé en vie réelle, notamment les effets indésirables. La réutilisation des données permet de réaliser cette surveillance à une large échelle ou à l'inverse dans des sous-populations.

Trois grands objectifs peuvent être identifiés et être mis en relation avec les différentes approches méthodologiques :

1. Surveiller la survenue d'effets indésirables ou d'interactions, connus en population, c'est-à-dire faciliter la notification d'effets indésirables, en lien avec la pharmacovigilance : méthodes de recherche d'information.
2. Identifier des effets indésirables ou des interactions qui n'avaient pas été détectés lors des phases d'évaluation du médicament en recherche clinique afin d'améliorer la connaissance sur le médicament : fouille de données.
3. Prédire la survenue d'événements liés aux médicaments (effet indésirable, observance, etc.) : apprentissage automatique.

Article 7 : Drug safety and big clinical data: detection of drug-induced anaphylactic shock events

L'article ci-après présente une approche de recherche d'information pour l'identification des chocs anaphylactiques médicamenteux à partir de l'entrepôt de données biomédicales du CHU de Rennes. L'objectif était d'évaluer si un entrepôt de données biomédicales apportait une information supplémentaire par rapport aux méthodes traditionnelles de surveillance du médicament.

Ma contribution a porté sur la réalisation des recherches sur l'entrepôt de données du CHU de Rennes afin d'identifier la requête la plus performante pour extraire un signal d'intérêt pour la pharmacovigilance et de réaliser les analyses permettant de mesurer les performances de cette approche.

Drug safety and big clinical data: detection of drug-induced anaphylactic shock events

Guillaume Bouzillé, MD MSc[1234*], Marie-Noëlle Osmont, PharmD[5], Louise Triquet, PharmD[5], Natalia Grabar, PhD[67], Cécile Rochefort-Morel, MD[8], Emmanuel Chazard MD[9] PhD, Elisabeth Polard, PharmD[5] and Marc Cuggia MD PhD[1234]

[1]INSERM, U1099, Rennes, F-35000, France
[2]Université de Rennes 1, LTSI, Rennes, F-35000, France
[3]CHU Rennes, CIC Inserm 1414, Rennes, F-35000, France
[4]CHU Rennes, Centre de Données Cliniques, Rennes, F-35000, France
[5]CHU Rennes, Centre Régional de Pharmacovigilance, Rennes, F-35000, France
[6]CNRS, UMR 8163, F-59000 Lille, France
[7]Université de Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
[8]CHU Rennes, Service de pneumologie et d'allergologie, Rennes, F-35000, France
[9]Département de Santé Publique, Université de Lille EA 2694, CHU Lille, F-59000 Lille, France

Corresponding author
E-mail: guillaume.bouzille@univ-rennes1.fr

Abstract

Rationale, aims and objectives: The spontaneous reporting system currently used in pharmacovigilance is not sufficiently exhaustive to detect all adverse drug reactions (ADRs). With the widespread use of electronic health records (EHRs), biomedical data collected during the clinical care process can be reused and analyzed to better detect ADRs. The aim of this study was to assess whether querying a Clinical Data Warehouse (CDW) could increase the detection of drug-induced anaphylaxis.
Methods: All known cases of drug-induced anaphylaxis that occurred or required hospitalization at Rennes Academic Hospital in 2011 (n=19) were retrieved from the French pharmacovigilance database, which contains all reported ADR events. Then, from the Rennes Academic Hospital CDW, a training set (all patients hospitalized in 2011) and a test set (all patients hospitalized in 2012) were extracted. The training set was used to define an optimized query, by building a set of keywords (based on the known cases) and exclusion criteria to search structured and unstructured data within the CDW in order to identify at least all known cases of drug-induced anaphylaxis for 2011. Then, the real performance of the optimized query was tested in the test set.
Results: Using the optimized query, 59 cases of drug-induced anaphylaxis were identified among the 253 patient records extracted from the test set as possible anaphylaxis cases. Specifically, the optimal query identified 41 drug-induced anaphylaxis cases that were not detected by searching the French pharmacovigilance database, but missed seven cases detected only by spontaneous reporting.
Discussion: We proposed an information retrieval-based method for detecting drug-induced anaphylaxis, by querying structured and unstructured data in a CDW. CDW queries are less specific than spontaneous reporting and DRG queries, although their sensitivity is much higher. CDW queries can facilitate monitoring by pharmacovigilance experts. Our method could be easily incorporated in the routine practice.

Keywords: Adverse Drug Reaction Reporting Systems, Drug-Related Side Effects and Adverse Reactions, Electronic Health Records, Information Storage and Retrieval

1. Introduction

Pharmacovigilance is an essential component of drug safety. The main goals of pharmacovigilance are the post-marketing drug surveillance and reducing the risk of adverse drug reactions (ADR) associated with drug use. The French pharmacovigilance system is based on a network of 31 regional pharmacovigilance centers. Spontaneous reporting (i.e., the unrequested notification of ADR events) by healthcare professionals or consumers to pharmacovigilance centers is the cornerstone of this surveillance system. Moreover, pharmaco-epidemiological studies assess the benefit/risk ratio when a drug is prescribed in "real life" to a large population of patients. Safety data are then analyzed by the competent authorities at different levels (national, European, global) that can then call attention to pharmacovigilance signals. Validated signals allow the authorities to develop regulatory measures to better control drug-related risks.

Spontaneous reporting has the advantage of covering a large number of patients (ideally, the entire population) and a wide range of drugs. It is a cost-effective method for monitoring drug safety [1], but has some limitations, mainly the under-reporting of ADR events, particularly those occurring in hospitals or leading to hospitalization. Indeed, it is estimated that more than 90% of ADR events are currently not reported to the authorities [1].

Several studies have assessed whether medico-economic databases could be used to improve ADR detection [2–4]. These authors detected more ADRs using the billing codes of the international classification of diseases, 10th edition (ICD-10), compared to the number of ADRs identified via spontaneous reporting for the same period. Moreover, they observed little or no overlap between the numbers of ADR events detected using the two methods.

Similarly, the French Medical Information System Program (PMSI) database generates Diagnosis-related Groups (DRGs) that contain administrative and medical data, including diagnoses that are classified according to the ICD-10 codes. By querying this database using selected ICD-10 billing codes, it is possible to identify serious ADR events. This system is now routinely used by the Regional Pharmacovigilance Center of Rennes[5] to detect serious ADRs, such as generalized rash, anaphylaxis, nephropathy, liver damage, polyneuropathy, neuroleptic malignant syndrome and interstitial lung disease (see Appendix 1 for a list of ICD-10 billing codes used by the Regional Pharmacovigilance Center of Rennes). All hospital stays involving a hospitalization summary that includes at least one of the selected ICD-10 billing codes are investigated to validate whether the extracted ADR is relevant to pharmacovigilance.

Big data analysis also could contribute to improving the relevance of the detected signals. In France, this could be done with the support of institutions that run large databases and/or manage several health professionals' databases, such as the French National Agency for Medicines and Health Products Safety (ANSM), the French National Health Insurance Agency (CNAM), the French Public Regional Health Agencies (ARS), the Hospital Information Technology Agency (ATIH), or the French national inter-scheme information system of health insurance (SNIIRAM, Système national d'information inter-régimes de l'Assurance maladie), which combines the French healthcare reimbursement system and PMSI databases. Nevertheless, they should be considered as extra sources of information rather than as replacements for spontaneous notifications that remain the most effective tool to date [6]. As these

methods are not exhaustive, the use of other data sources and methods becomes necessary to effectively identify ADR events in hospitals [7]. Due to the widespread use of electronic health records (EHRs), biomedical data collected during the clinical care process could be reused and analyzed to improve ADR detection[8–10]. Technologies, such as Clinical Data Warehouses (CDW), have these abilities. They are currently deployed in many hospitals, thus making possible the efficient exploitation of clinical data [11–13]. We believe that such technologies can also be used to address some of the issues currently encountered in pharmacovigilance.

We are interested in improving the detection of anaphylactic shock events. We previously carried out a study to evaluate the performance of a data-gathering method using a CDW for cases of anaphylactic shocks (IgE-mediated) that specifically occurred during anesthesia at the Academic Hospital of Rennes (CHU-RENNES) [14]. In that study, anaphylactic shock during anesthesia proved to be an easily identifiable clinical entity. However, we think that the identification of anaphylactic shock events due to other causes also could be improved by using the same kind of information retrieval method. Therefore, the objective of the current study was to assess whether clinical data from narrative EHRs can help pharmacovigilance centers to detect drug-induced anaphylaxis. More specifically, we propose an optimal method, based on a CDW, to detect drug-induced anaphylaxis events.

2. Material and methods

2.1. Data sources and studied population

At the CHU-RENNES, we developed our own CDW solution, called eHOP (formerly Roogle) [15,16]. Briefly, this CDW integrates all types of documents produced by the hospital information system and connected with healthcare:

- structured data using reference terminologies (e.g., ICD-10 diagnoses from DRGs, local terminology codes for laboratory tests, Association for the Development of Informatics in Cytology and Pathology codes for pathology diagnoses, Anatomical Therapeutic Chemical terminology corresponding to drug prescriptions and administration);
- unstructured data, such as clinical narrative notes, surgical protocols, X-rays or pathology reports.

Hence, a unique attribute of eHOP is that it allows users to search for information from both structured and unstructured data. Additionally, two different ways of querying data can be combined. Users can build queries based on reference terminologies (e.g., ICD-10), or simply submit keywords to retrieve both structured (e.g., terminology labels) or unstructured documents that contain these terms or keywords. Users can then access documents via a dedicated interface that incorporates functionalities to allow navigating through the entire patient EHR. eHOP is routinely used at the CHU-RENNES to support clinical research in feasibility studies, or for screening patients for eligibility criteria.

The eHOP CDW currently provides the possibility to search among 25 million unstructured data and 170 million structured elements. Some unstructured data, such as laboratory results or

diagnoses, are also recorded in a structured form thanks to the corresponding terminological codes. All these data are collected from EHRs and cover more than 1.2 million patients.

For this study, we defined two datasets from eHOP: i) a training set that contained information on patients who stayed at the hospital between January 1 and December 31, 2011; and ii) a test set that contained data on patients who stayed at the hospital between January 1 and December 31, 2012.

The inclusion period of one year for the two sets was extended by three months to ensure that the data obtained would be complete. This was done to take into account possible delays in the production of documents for patients with a suspected anaphylaxis event who, at the end of the studied period, had allergy-related consultations. All documents related to these patients and produced during the studied period (including the three additional months) were used as part of the information retrieval process.

2.2. Definition of anaphylaxis

This study focused on patients who had drug-induced anaphylaxis. Anaphylaxis is an acute, potentially life-threatening hypersensitivity reaction that involves the release of mediators from mast cells, basophils and recruited inflammatory cells. Anaphylaxis is defined by a number of signs and symptoms, alone or in combination, that occur within minutes, or up to a few hours after exposure to a causative agent[17]. Concerning the underlying mechanisms, the anaphylactic reaction can be IgE-mediated or non-IgE-mediated. Based on the reaction severity, anaphylaxis can be classified in four levels:[15]

- level I: presence of cutaneous signs;
- level II: presence of measurable, but not life-threatening symptoms, including cutaneous effects, arterial hypotension, cough or ventilation difficulties;
- level III: presence of life-threatening symptoms (e.g., cardiovascular collapse, tachycardia or bradycardia, arrhythmias, severe bronchospasms);
- level IV: circulatory failure, cardiac and/or respiratory arrest.

All anaphylaxis episodes that occurred at CHU-RENNES or led to hospitalization or required allergy investigations at the hospital were considered relevant for our study. As in France, there are 31 regional pharmacovigilance centers, we included only the anaphylaxis cases that occurred in the geographic area covered by the Regional Pharmacovigilance Center of Rennes.

Moreover, a given patient was included in the study and considered as having had a new anaphylaxis episode each time that she/he was hospitalized or required allergy investigations at the hospital for drug-induced anaphylaxis during the study period.

2.3. Extraction of the known drug-related anaphylaxis cases

The known cases of drug-related anaphylaxis (reference drug-related anaphylaxis cases throughout the text) that occurred within or required care at CHU-RENNES during the training

period (2011) were retrieved from the French pharmacovigilance database, which contains all reported ADRs, spontaneously reported and identified using DRGs (based on two ICD-10 billing codes: T88.2 and T88.6) (see Figure 1 for the study design). The *Medical Dictionary for Regulatory Activities* (MedDRA®) term "Anaphylactic responses" and the high-level term belonging to the System Organ Classes "Immune system disorders" were used to query the database. Criteria on the active substance responsible for the anaphylaxis were not included in the query.

The same procedure was used to extract the drug-related anaphylaxis cases from the French pharmacovigilance database for 2012 for the comparison with the results obtained with the test set (2012).

2.4. Identification of keywords related to the reference drug-related anaphylaxis cases

Two different methods were used to identify potentially relevant keywords:

a) Experts from the Regional Pharmacovigilance Center of Rennes reviewed the discharge summaries of the reference drug-related anaphylaxis cases extracted from the French pharmacovigilance database, in order to identify terms used by clinicians in connection with such episodes.

b) Analysis of the keywords used in the previous study to detect cases of anaphylactic shock occurring during anesthesia [14]. These keywords were "anaphylaxis" and "anesthesia". Only the cases in which "anaphylaxis" OR "anaphylactic" OR "anaphylactoid" was not mentioned in the discharge summaries could not be identified in this study. Moreover, these previous results suggested that keywords connected with symptoms of anaphylaxis were not sufficiently specific.

2.5. Definition of the optimized query

To define the optimized query, the set of keywords identified in 2.4 were used to query structured and particularly unstructured data to find, at least, the reference drug-related anaphylaxis cases in the training set (2011) from the eHOP CDW. Moreover, the performance of each keyword was independently tested in terms of retrieval of the reference drug-related anaphylaxis cases, by using the following standard evaluation measures:

- Precision: fraction of drug-related anaphylaxis cases among all retrieved patients
- Recall: fraction of retrieved drug-related anaphylaxis cases among all anaphylaxis cases
- Synthetic F-measure:

A given keyword was considered to be relevant if its recall was at least 0.5. The recall measure was prioritized because it is essential for pharmacovigilance experts not to miss any relevant case. The optimized query was built by combining the retained keywords (recall ≥ 0.5) with the help of the inclusive disjunction ("OR"), so that the query would retrieve all documents matching at least one of the relevant keywords.

2.6. Evaluation of the optimized query in the test set

The real performance (precision, recall and F-measure) of the optimized query and of each keyword forming this query (i.e., all keywords with a recall ≥0.5 in the training set) was assessed in the test set.

Two pharmacovigilance experts from the Regional Pharmacovigilance Center of Rennes reviewed all potential cases returned by the query in an independent and blind validation process to confirm that they were real drug-related anaphylaxis cases. Any disagreements between experts required a consensus session before reaching the final decisions on the patient's status and relative to the definition of anaphylactic shock (and to the inclusion criteria) mentioned above. Validated cases from the test set were then compared with the drug-related anaphylaxis cases (spontaneous and DRG-based reporting) identified in the French pharmacovigilance database for 2012. Duplicated cases present in different data sources were identified using administrative data (initials, date of birth, etc.), the characteristics of the side effects, suspected drugs and date of occurrence. This step was necessary because a given patient could present with several episodes of anaphylaxis during the study period.

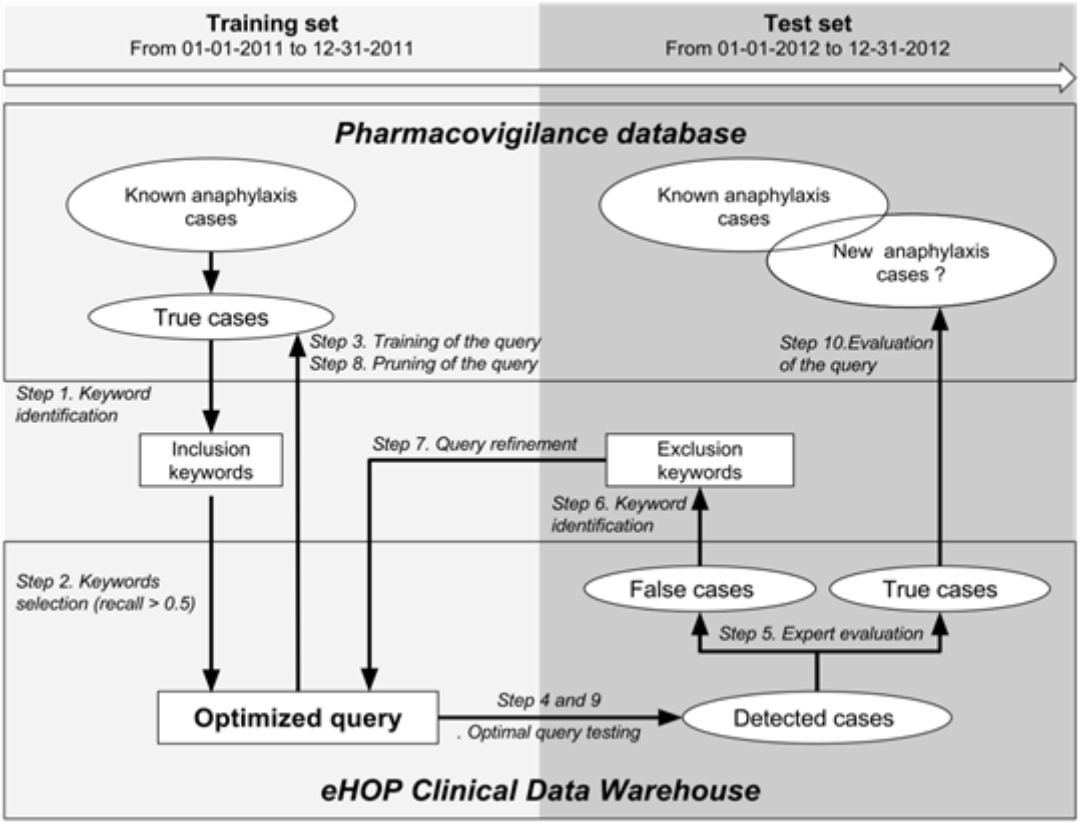


Figure 1. Study design to build the optimized query used to identify drug-related anaphylaxis cases

2.7. Final tuning of the query

During the final step, exclusion criteria were added to the optimized query, consisting of keywords connected with the causes of anaphylaxis, identified by experts, in patients who had not suffered drug-induced anaphylaxis. This pruning phase was validated by testing the

optimized, pruned query using the training set to check whether all reference drug-induced anaphylaxis cases were still detected.

The study design is summarized in Figure 1.

2.8. Ethics statement and funding sources

This study was approved by the Ethics Committee of the academic hospital of Rennes and performed in accordance with the Declaration of Helsinki guidelines.

The funding source, the French National Agency for Medicines and Health Products Safety (ANSM), had no role in the study.

3. Results

3.1. Extraction of reference drug-induced anaphylaxis cases from the French pharmacovigilance database

In total, 30 anaphylaxis cases were identified in the French pharmacovigilance database for the area covered by the Regional Pharmacovigilance Center of Rennes for the year 2011. Among them, 19 cases were selected as confirmed cases that occurred or required care at the CHU-RENNES. For the remaining 11 cases, 9 did not occur at the CHU-RENNES, one had a final diagnosis of mastocytosis and one included an anaphylactoid reaction.

3.2. Keyword identification and evaluation in the training set

On the basis of the review of the 19 reference drug-related anaphylaxis cases by the experts at the Regional Pharmacovigilance Center of Rennes and the previous study on anaphylaxis during anesthesia, the following keywords were selected:

- Anaphylaxis query: words prefixed by “anaphy.”
- Tryptase query: “level” AND “tryptase” OR “increase” AND “tryptase,” with a maximum of two words between them.
- Allergo-anesthesia query: words prefixed by “allergo-anesth.”
- Allergen-specific IgE query: “allergen-specific IgE.”
- Prick testing query: “prick-test.”
- Intradermal allergy testing query: “intradermal” OR “intra-dermal.”
- Histamine query: words “histamine” AND “increased” with a maximum of two words between them.
- Contraindication query: “strictly contraindicated.”
- Immunoallergic query: words prefixed by “immunoaller.”
- RAST inhibition query: “RAST inhibition.”

The next step was to evaluate the capacity of different queries in which the term “shock” or “collapse” was combined with one of these keywords (all Oracle Text SQL queries are in Appendix 2) to retrieve the reference drug-induced anaphylaxis cases from the training set (n=178,676 patients with at least one hospital visit in 2011). Among the tested queries, “anaphylaxis” showed the highest recall, because it retrieved 17 of the 19 reference drug-

induced anaphylaxis (0.89). It was also the keyword with the lowest precision (0.08) (Table 1). The queries "allergo-anesthesia," "tryptase," and "allergen-specific IgE" retrieved at least half of the reference cases (i.e., with a recall higher than 0.5).

On the basis of these results, an optimized query that included all individual queries with recall ≥ 0.5 was built by using the inclusive disjunction ("OR"), so that any fulfilled condition was sufficient to retrieve a relevant case. This optimized query had a recall value of 1 and a precision value of 0.07 (all 19 reference cases among the 270 patients retrieved by the query) (Table 1).

Table 1. Query results using the training set (2011)

Id	Query	No. of retrieved cases out of 19	No. of returned patients out of 178,676	Precision	Recall	F-measure
1	anaphylaxis	17	226	0.08	0.89	0.15
2	allergo-anesthesia	13	44	0.30	0.68	0.42
3	tryptase	10	38	0.26	0.53	0.22
4	allergen-specific IgE	10	87	0.11	0.53	0.18
5	prick-testing	7	28	0.25	0.37	0.30
6	intradermal allergy testing	6	100	0.06	0.32	0.10
7	histamine	4	9	0.44	0.21	0.28
8	contraindication	4	6	0.67	0.21	0.32
9	immunoallergic	2	13	0.15	0.11	0.13
10	RAST-inhibition	1	2	0.50	0.05	0.09
	Optimized query (1 OR 2 OR 3 OR 4)	19	270	0.07	1.00	0.13

3.3. Optimized query performance in the test set

The test set included 182,127 patients with at least one hospital visit in 2012. The optimized query identified 253 patients and 452 matching documents: 159 outpatient discharge summaries (35% of all documents), 110 biology results (24%), 73 inpatient discharge summaries (16%), 42 DRGs (9%), 25 discharge summaries from the emergency department (6%) and miscellaneous narrative documents (10%). The "anaphylaxis" and "allergo-anesthesia" queries (which correspond to signs and symptoms) mainly matched outpatient discharge summaries, whereas the "tryptase" and "allergen-specific IgE" queries (which correspond to biological results) matched laboratory results (Table 2).

Table 2. Summary of the types of retrieved documents (test set, 2012)

	Overall		Outpatient discharge summary		Inpatient discharge summary		DRGs		Discharge summary from emergency department		Laboratory results		Other	
Query	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
anaphylaxis	343	100	**149**	**43.4**	73	21.3	42	12.3	25	7.3	11	3.2	42	12.5
allergo-anesthesia	38	100	**36**	**94.8**	1	2.6	0	0	0	0	0	0	1	2.6
tryptase	65	100	19	29.2	2	3.1	0	0	0	0	**42**	**64.6**	2	3.1
allergen-specific IgE	127	100	33	25.9	3	2.4	0	0	2	1.6	**88**	**69.3**	1	0.8
Optimized query	452	100	**159**	**35.2**	73	16.2	42	9.3	25	5.5	110	24.3	43	9.5

Bold data correspond to keywords relevant for the optimized query.

3.4. Expert evaluation

The two experts identified 59 cases of drug-induced anaphylaxis among the 253 patients retrieved from the test set using the optimized query: 38 cases of drug-induced anaphylaxis that occurred or required care at the CHU-RENNES, and 21 cases that required allergy investigations at the hospital (Table 3). The detailed results on each keyword performance can be found in Table 4.

Table 3. Comparison of data sources for the identification of drug-induced anaphylaxis cases (2012)

Source	Number of cases that occurred at CHU-RENNES	Number of cases that required allergy investigation at CHU-RENNES	Total number of cases
eHOP CDW (test set)			
total	38	21	59
only detected in eHOP	**26**	**15**	**41**
not detected in eHOP	3	4	7
DRGs			
total	10	n/a	10
only detected with DRGs	0	n/a	0
not detected with DRGs	31	n/a	31
Spontaneous reports			
total	7	10	17
only detected by spontaneous reports	3	4	7
not detected by spontaneous reports	34	15	59
Total*	41	25	66

*Total number of individual cases of drug-related anaphylaxis identified using the three sources.

Specifically, among the 253 patients, one expert identified 60 and the other one 56 cases of drug-related anaphylaxis. Ten disagreements between experts were resolved during consensus sessions. These disagreements included three cases that were validated by one of the experts, although the date of occurrence of the shock was not included in the study period; one case that was wrongly excluded because the date of occurrence was inaccurate, but within the study

period; three cases that were included by one expert, although the name of the drug was not reported (only its pharmacological class); and three other cases that were not identified by one of the two experts when reading the records.

Table 4. Optimized query evaluation using the test set (2012)

Id	Query	No. of retrieved cases out of 59	No. of returned patients out of 182,127	Precision	Recall	F-measure
1	anaphylaxis	44	190	0.23	0.74	0.15
2	allergo-anesthesia	11	35	0.31	0.19	0.42
3	tryptase	22	56	0.39	0.37	0.22
4	allergen-specific IgE	42	114	0.37	0.71	0.18
	Optimized query (1 OR 2 OR 3 OR 4)	59	253	0.23	1	0.13

3.5. Comparison of data sources for the identification of anaphylaxis cases

Analysis of the in French pharmacovigilance database for the year 2012 highlighted the presence of 17 drug-induced anaphylaxis cases at the CHU-RENNES and 10 cases that required care at the hospital (Table 3). Comparison of the data concerning the drug-induced anaphylaxis cases identified in the test set (eHOP) and in the French pharmacovigilance database showed that among the 41 cases that occurred at the hospital (test set + French pharmacovigilance database cases, see Table 3), 17 were already recorded in the French pharmacovigilance database (10 cases identified from DRGs and 7 spontaneously reported by physicians). Only two cases were shared by these two last data sources. Among these 17 cases, only three cases, which were spontaneously reported by physicians, were not found by querying the test set. One had a missing discharge summary, one was described as an "allergic reaction", and the last one was only described using symptoms related to anaphylaxis. Twenty six cases were only found by querying the test set from eHOP.

Among the 25 cases that required allergy investigations at the hospital (Table 3), 10 were spontaneously reported by physicians. Four of these spontaneously reported cases were not detected via eHOP. These four cases occurred in the area covered by the Regional Pharmacovigilance Center of Rennes, but allergy investigations were not conducted at the CHU-RENNES during the studied period. Fifteen cases were detected only in the test set.

3.6. Final tuning

Besides drug-induced anaphylaxis, the main causes of anaphylaxis identified by experts were food and insect stings. Therefore, the following keywords were added to the optimized query as exclusion criteria: "flour," "rye," "wheat," "venom," "wasp," "hymenopteran," "stings," "bee," "bumble bee" and "insect". The other source of false positive cases was the mention of "shock" or "collapse" in the documents describing the patients' medical history. Consequently, another exclusion criterion was added to exclude patients for whom only the keywords "shock" or "collapse" appeared in their medical history (see Appendix 3, containing the complete Oracle text SQL query). The new optimized query that included also the exclusion criteria retrieved

200 potential cases from the test set (2012), compared to the 253 cases obtained by using the first version of the optimized query. This reduced group still contained all 59 drug-related anaphylaxis cases, thus resulting in a final precision of 0.29. The same process using the training set (2011) still retrieved the 19 reference drug-related anaphylaxis cases among 202 potential cases, compared with 270 cases retrieved by the original query.

Finally, based on the dates on the retrieved documents, it was calculated that the median number of potential cases per week was three (IQR: [2-5]), with a median number of validated drug-induce anaphylaxis cases of one every two weeks (IQR: [0-3]).

4. Discussion

4.1. Comparison of data sources for the identification of anaphylaxis cases

Our study demonstrates the added value of using information technologies, such as CDWs to improve the current practice, specifically in terms of identifying ADRs. The methods currently used, such as spontaneous reporting and DRG queries, do not detect all relevant ADRs and thus lead to underestimating drug safety issues. As suggested by some authors, it is extremely valuable to use additional data sources that allow better ADR detection and contribute to improving drug safety for patients [7].

We do not intend to offer a method that exhaustively detects every case of drug-induced anaphylaxis. Producing reliable estimates for missing cases is undoubtedly a complex and time-consuming task because it would require reviewing all hospital visits during a given period to confirm both the occurrence of the anaphylaxis episode and the causal link with the administration of a drug. Yet, we do believe that spontaneous reporting, DRG queries and CDW (eHOP in our study) queries are complementary and can be integrated into a global strategy for the systematic detection of ADRs.

4.2. Evaluation of the query on the test set

If we compare the performance of three ADR detection methods on the test set, all cases spontaneously reported by physicians were valid drug-related anaphylaxis cases. In addition, DRG reporting allowed the validation of 10 of the 11 (90.9%) potential cases of anaphylactic shock (ICD-10 billing codes T88.2 and T88.6). In contrast, only 59 cases were validated among the 200 anaphylaxis cases (29.5%) identified in the test set (from eHOP) using the improved query with exclusion criteria (38 cases that occurred or required care at the CHU-RENNES [19%], and 21 cases that required allergy investigations at the hospital [10.5%]). The eHOP query is less specific than spontaneous reporting and DRG queries, although its sensitivity is far higher, thus allowing the detection of a larger number of relevant cases (41 of the 59 cases identified via the eHOP query were new). Concerning the drug-related anaphylaxis cases that occurred or required care at the CHU-RENNES, the proposed eHOP query detected all the cases identified through the DRG query, four of the seven spontaneously reported cases, and 26 not previously known cases. The eHOP query also detected several cases that did not occur at the CHU-RENNES, but that only required allergy investigations (including outpatients). This kind of case cannot be identified via DRG queries because DRGs are only produced for inpatients.

In addition, our method retrieved patients mainly from outpatient discharge summaries or laboratory results that are inaccessible to the DRG query.

Comparison of the keyword recall values in the training (2011) and test (2012) sets showed that the keywords “anaphylaxis” and “allergen-specific IgE” allowed the detection of more relevant cases. Conversely, “allergo-anesthesia” and “tryptase” displayed a lower recall value in the test set than in the training set. This could lead to false positive results when using these two keywords, which can be explained by the fact that consultations for anesthesia-related allergy problems and tryptase tests are only performed after the identification and investigation of drug-induced anaphylaxis (i.e., the type of cases that were identified in the training set). Thus, these keywords are not as pertinent for the identification of unknown cases that have not been spontaneously declared or identified via DRG. These observations highlight the complementarity of these three methods to improve the use of the available data sources for ADR detection.

4.3. Added value of the eHOP CDW

One of the strengths of the eHOP technology is the possibility to query both unstructured and structured data. For instance, 56.9% of all retrieved documents were unstructured: discharge summaries from inpatients, outpatients or emergency departments (Table 2). Structured data (DRG codes and laboratory results) constituted only 33.6% of all retrieved documents, and most information found in EHRs is recorded as free text. More specifically, the search terms “tryptase” and “allergen-specific IgE” may improve the thoroughness of anaphylaxis detection. Most institutions currently use the standard i2b2 platform as their CDW technology, although its main purpose is the integration of structured data. Therefore, when only structured data are obtainable (or full-text information retrieval systems are not available), querying tryptase laboratory tests or allergen-specific IgE assays from i2b2, or even directly from laboratory information systems, can help to improve the detection of drug-induced anaphylaxis cases.

We consider that it is crucial to have the appropriate tools to adequately leverage the richness of the different data sources. Therefore, the eHOP CDW is currently deployed in the six main academic hospitals of western France. This could lead to the implementation of our query method by other pharmacovigilance centers, thus allowing the assessment of its potential and also an easier ADR detection at a larger scale.

Another advantage of eHOP over other CDW technologies is its easier access to EHR data that could facilitate the investigation process by pharmacovigilance experts. We found that if the eHOP query was used in routine practice by pharmacovigilance experts, the average number of potential cases to be investigated would be three per week (IQR: [2-5]). This method would result in one valid case detected every two weeks (IQR: [0-3]). This additional workload would be viable and could be handled by pharmacovigilance experts as part of their practice. In addition, eHOP provides a user interface for navigating through EHRs, thus making the investigation process even easier.

4.4. Perspectives

Several options can be further explored to improve the proposed method. For instance, natural language processing (NLP) could increase the detection accuracy, as the presence of the keyword "anaphylaxis" in the patients' medical history yielded several false positive cases. Most importantly, the available amount of health data and their heterogeneity require using machine learning and text mining approaches. Our study demonstrates that simple information retrieval methods are very efficient when the concepts to be retrieved can be described with relatively specific keywords. This is particularly true in pharmacovigilance, where there is still space for significantly improving ADR detection rate and accuracy. Our method could be applied to other ADR type. However, many diseases have complex characteristics and several etiologies, besides ADRs. In such cases, machine learning approaches could help to detect hidden or latent characteristics that are specific to complex ADRs.

5. Conclusion

Pharmacovigilance is crucial for the efficient long-term management of drug safety. This requires the development of suitable tools. Here, we described an information retrieval-based method for the detection of drug-induced anaphylaxis, based on querying both structured and unstructured data from a CDW. Besides the 25 cases already known from spontaneous and DRG reporting for 2012, with this method we could identify 41 additional cases. Our method can be easily implemented in the routine practice and could be proposed to other regional pharmacovigilance centers to better identify well-defined ADRs. Additional improvements may be necessary for the detection of more complex ADRs, possibly by using NLP processing, as well as machine learning and text mining methods.

Acknowledgements

We would like to thank the French National Agency for Medicines and Health Products Safety (ANSM), for funding this work inside the Breizh project (Evaluation of under-reporting of adverse drug reactions in public hospitals in Brittany: the contribution of PMSI and a biomedical data warehouse) (grant no. AAP-2014-014).

References

1. Hazell L, Shakir SAW. Under-Reporting of Adverse Drug Reactions. *Drug Saf*. 2012;29(5):385-396.
2. Jorup-Rönström C, Keisu M, Wiholm B-E. Could Swedish 'Yellow Cards' Be Substituted by E-Coded Summaries? *Drug Saf*. 2012;5(1):72-77.
3. Wodtke JM, Generali JA. Use of medical record codes to identify adverse drug reactions. *Am J Hosp Pharm*. 1993;50(9):1915-1916.

4. Cox AR, Anton C, Goh CHF, Easter M, Langford NJ, Ferner RE. Adverse drug reactions in patients admitted to hospital identified by discharge ICD-10 codes and by spontaneous reports. *Br J Clin Pharmacol*. 2001;52(3):337-339.
5. Osmont M-N, Cuggia M, Polard E, Riou C, Balusson F, Oger E. Use of the PMSI for the detection of adverse drug reactions. *Therapie*. 2013;68(4):285-295.
6. Vial T. French pharmacovigilance: Missions, organization and perspectives. *Thérapie*. 2016;71(2):143-150.
7. Asfari H, Bousquet C, Trombert Paviot B, et al. Drug-related anaphylactic shocks: under-reporting and PMSI. *Therapie*. 2014;69(6):483-490.
8. Coloma PM, Trifirò G, Schuemie MJ, et al. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? *Pharmacoepidemiol Drug Saf*. 2012;21(6):611-621.
9. Trifirò G, Sultana J, Bate A. From Big Data to Smart Data for Pharmacovigilance: The Role of Healthcare Databases and Other Emerging Sources. *Drug Saf*. August 2017.
10. Black C, Tagiyeva-Milne N, Helms P, Moir D. Pharmacovigilance in children: detecting adverse drug reactions in routine electronic healthcare records. A systematic review. *Br J Clin Pharmacol*. 2015;80(4):844-854.
11. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – An Integrated Standards-Based Translational Research Informatics Platform. *AMIA Annu Symp Proc*. 2009;2009:391-395.
12. Cuggia M, Garcelon N, Campillo-Gimenez B, et al. Roogle: an information retrieval engine for clinical data warehouse. *Stud Health Technol Inform*. 2011;169:584-588.
13. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17(2):124-130.
14. Osmont M-N, Campillo-Gimenez B, Metayer L, et al. Peranesthesic Anaphylactic Shocks: Contribution of a Clinical Data Warehouse. *Therapie*. October 2015.
15. Delamarre D, Bouzille G, Dalleau K, Courtel D, Cuggia M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. *Stud Health Technol Inform*. 2015;210:702-706.
16. Cuggia M, Garcelon N, Campillo-Gimenez B, et al. Roogle: an information retrieval engine for clinical data warehouse. *Stud Health Technol Inform*. 2011;169:584-588.
17. Montañez MI, Mayorga C, Bogas G, et al. Epidemiology, Mechanisms, and Diagnosis of Drug-Induced Anaphylaxis. *Front Immunol*. 2017;8:614.
18. Dong SW, Mertes PM, Petitpain N, Hasdenteufel F, Malinovsky JM, GERAP. Hypersensitivity reactions during anesthesia. Results from the ninth French survey (2005-2007). *Minerva Anestesiol*. 2012;78(8):868-878.

Appendices

Appendix 1: List of ICD-10 codes used by the pharmacovigilance center of Rennes

ICD-10 billing code	Description
G21.0	Malignant neuroleptic syndrome

G62.0	Drug-induced polyneuropathy
J70.2	Acute drug-induced interstitial lung disorders
K71.0	Toxic liver disease with cholestasis
K71.1	Toxic liver disease with hepatic necrosis
K71.2	Toxic liver disease with acute hepatitis
K71.6	Toxic liver disease with hepatitis, not elsewhere classified
K71.8	Toxic liver disease with other disorders of liver
K71.9	Toxic liver disease, unspecified
L27.0	Generalized skin eruption due to drugs and medicaments
L51.0	Nonbullous erythema multiforme
L51.1	Bullous erythema multiforme
L51.2	Toxic epidermal necrolysis [Lyell]
N14.1	Nephropathy induced by other drugs, medicaments and biological substances
N14.2	Nephropathy induced by unspecified drug, medicament or biological substance
N17.0	Acute renal failure with tubular necrosis
N17.1	Acute renal failure with acute cortical necrosis
N17.8	Other acute renal failure
N17.9	Acute renal failure, unspecified
T88.2	Shock due to anesthesia
T88.6	Anaphylactic shock due to adverse effect of correct drug or medicament properly administered

Appendix 2: Oracle text queries based on keywords identified by experts

Id	Query name	Oracle text query expression
1	anaphylaxis	(shock or collapse) and anaphy%
2	allergo-anesthesia	(shock or collapse) and allergo-anesth%
3	tryptase	(shock or collapse) and (NEAR((level, tryptase), 2, TRUE) OR NEAR((increase%, tryptase), 2, FALSE))

4	allergen-specific IgE	(shock or collapse) and allergen-specific IgE%
5	prick-testing	(shock or collapse) and prick-test
6	Intradermal allergy testing	(shock or collapse) and intradermo or intra-dermo
7	histamine	(shock or collapse) and NEAR ((increase%, histamine), 2, FALSE)
8	contraindication	(shock or collapse) and contraindication for life
9	immunoallergic	(shock or collapse) and immuno-aller%
10	RAST inhibition	(shock or collapse) and RAST inhibition
	Optimized query	(shock or collapse) and (anaphy% OR NEAR((level, tryptase), 2, TRUE) OR NEAR((increase%, tryptase), 2, FALSE) OR allergo-anesth% OR allergen-specific IgE%)

Appendix 3: Optimized query with exclusion criteria: complete Oracle text SQL query

(shock or collapse) and (anaphy% OR NEAR((level, tryptase), 2, TRUE) OR NEAR((increase%, tryptase), 2, FALSE) OR allergo-anesth% OR allergen-specific IgE%) not (flour or rye or wheat or venom% or wasp% or hymenopter% or sting% or bee% or bumble bee% or insect% or near((medical history%, shock% or collapse), 50, true))

Article 8 : An automated detection system of drug-drug interactions from electronic patient records using big data analytics

L'article précédent a permis de montrer que les données massives hospitalières permettaient d'extraire des signaux pertinents pour la pharmacovigilance.

La richesse des données permet d'extraire des informations plus complexes et de tirer parti des associations entre les données.

L'article ci-après illustre l'emploi de technologies de bases de données « graphe » pour intégrer des bases de connaissances médicamenteuses permettant d'extraire des signaux complexes tels que les interactions entre substances actives.

Par ailleurs, l'objectif était de réaliser une preuve concept pour valider une approche d'apprentissage automatique permettant de prendre en considération les retours utilisateurs afin d'améliorer les performances de prédictions d'effets indésirables chez les patients.

Ma contribution à ce travail a été de coordonner le travail entre une interne de pharmacie apportant son expertise sur le médicament et un ingénieur en informatique et statistiques en stage de fin d'études qui a réalisé le développement des modèles prédictifs.

An automated detection system of drug-drug interactions from electronic patient records using big data analytics

Guillaume Bouzillé[a], Camille Morival[b], Richard Westerlynck[c], Pierre Lemordant[a], Emmanuel Chazard[d], Pascal Lecorre[b], Yann Busnel[c], Marc Cuggia[a]

[a] Univ Rennes, CHU Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France,

[b] Laboratoire de Pharmacie Galénique, Biopharmacie et Pharmacie Clinique, IRSET U1085, Faculté de Pharmacie, Université de Rennes 1, F-35043, Rennes Cedex, France,

[c] IMT Atlantique, F-35576, Cesson-Sévigné, France,

[d] *Univ Lille, CHU Lille, CERIM EA2694, F-59000 Lille, France*

Abstract

The aim of the study was to build a proof-of-concept demonstrating that big data technology could improve drug safety monitoring in a hospital and could help pharmacovigilance professionals to make data-driven targeted hypotheses on adverse drug events (ADEs) due to drug-drug interactions (DDI). We developed a DDI automatic detection system based on treatment data and laboratory tests from the electronic health records stored in the clinical data warehouse of Rennes academic hospital. We also used OrientDb, a graph database to store informations from five drug knowledge databases and Spark to perform analysis of potential interactions between drugs taken by hospitalized patients. Then, we developed a machine learning model to identify the patients in whom an ADE might have occurred because of a DDI. The DDI detection system worked efficiently and computation time was manageable. The system could be routinely employed for monitoring.

Keywords: Computing Methodologies, Drug Interaction, Machine Learning.

Introduction

Drug-drug interactions (DDIs) are a critical issue in patient care because they can lead to adverse events and ultimately increase care costs and patient mortality. Therefore, these events must be identified and prevented as early as possible [1]. However, many new drugs are released each year, and therefore, it is very difficult for healthcare professionals to be informed and to consider all DDIs. Moreover, the alarm functionalities of drug computerized physician order entry (CPOE) systems are frequently not used because they do not focus on clinically relevant DDIs and lead users to alarm fatigue. Although focused on specific interactions, studies on DDI prevalence show the existence of risks for polymedicated patients and highlight the importance of pharmacovigilance programmes [2,3].

With the unprecedented development of digital health and hospital clinical data warehouses (CDW), data produced during the healthcare process are now easily reusable [4]. Electronic health records (EHR) contain real-time information on drug prescription/regimens during hospitalization as well as all clinical information. Such data could be analysed to estimate DDI

prevalence, to facilitate health professionals' practice assessment and to detect the occurrence of DDI-linked adverse drug events (ADE). In France, pharmacovigilance currently relies mainly on the spontaneous reporting by physicians or/and detection of diagnoses that could be related to ADE from the hospital billing system (diagnosis related group, DRG, database). New data sources, such as national claim databases, are also leveraged to improve DDI and ADE detection [5,6]. EHR data-mining also could help pharmacovigilance professionals to improve drug safety assessment.

All these health-related databases fit perfectly with the big data paradigm because they contain voluminous, highly complex and heterogeneous information that is produced in real time [7]. In the last few years, many big data technologies have been developed. However, their implementation in a hospital information system for processing healthcare big data in real-world condition of use is still largely uncharted.

Here, we describe a method, which propose to use big data technology to improve drug safety monitoring in a hospital and could help pharmacovigilance professionals to make data-driven targeted hypothesis on ADEs.

Methods

Figure 1 presents the overall approach of the study and the big data technologies used in each step.

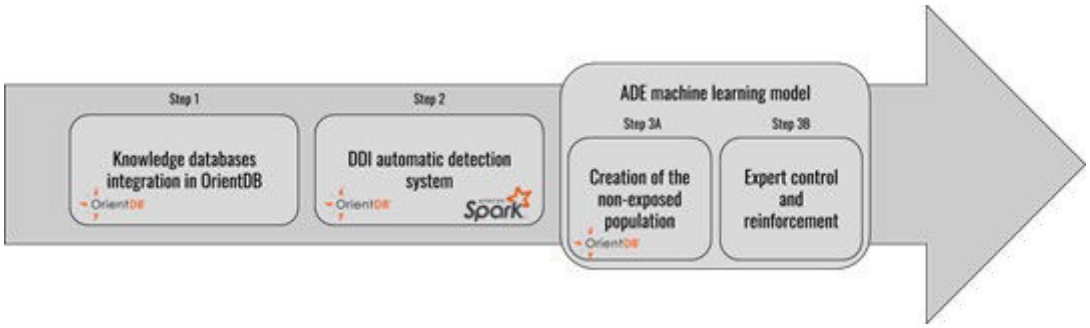


Figure 1 - Overall approach of the study

Patient data

We used the Rennes academic hospital EHRs that are stored in a CDW called eHOP (entrepot HOPital). This CDW includes both structured data (e.g., laboratory results, drug prescriptions and regimens) and unstructured data (e.g., operative reports, discharge summaries), and is dedicated to data reuse for clinical research [8]. The eHOP's star schema architecture and graphic user interface allows researchers, even without any database language knowledge, to quickly access and efficiently search information within millions of patient records.

For this study, we used information about drug administrations (used drug(s) and regimens) and laboratory results (date, nature of the test and results: normal, abnormally high, or abnormally low).

Knowledge databases integration (step 1)

To identify and collect information on potential DDIs, and also to compare information from different sources, we selected five drug knowledge databases: Thesaurus, Vidal, Theriaque, Micromedex and Drugs.com [9–13]. These databases are commonly used by health professionals, but are not specifically targeted to DDI detection. They are available via a web application programming interface (API) that requires a specific procedure because each database stores data with its own structure. To avoid this, we extracted the relevant information from these databases and stored it in OrientDB, a graph-oriented model database [14] that fits well with our objective because a DDI can be modelled as an edge between two drugs. Thus, once the information is stored in a single OrientDB database, no more computation is required to access such information.

DDI automatic detection system from patient records (step 2)

For DDI identification, we collected drug data from the patient EHRs stored in eHOP and computed the active interval (i.e., the period during which a drug was effective) for all drugs taken by a patient during the hospital stay. If two active intervals overlapped (fully or partially), then analysis of the data collected in the OrientDB database allowed determining whether the two drugs interacted. In this case, the potential DDI event was stored in eHOP. As these are independent processes (each drug pair is checked independently), the Spark cluster- computing framework was used to perform distributed computing [15,16]. As all the potential DDI events can be stored in eHOP, then we could compute the prevalence of a DDI for any specific drug, molecule, or population.

Creation of a machine learning model (step 3A)

The data stored in the CDW eHOP do not allow direct confirmation of whether a patient reported a DDI-linked ADE or not. Indeed, this needs to be validated by the pharmacovigilance experts who do not have the proper means to check all the patient records. Therefore, we wanted to create a system to report to drug safety professionals only the most interesting cases among all DDIs detected by the DDI automatic detection system (i.e., patients in whom an ADE might have occurred because of a DDI).

We assume that laboratory results will change if an ADE occurs. So, we can train a machine learning model with two populations: those who experienced an ADE and those who did not. Unfortunately, we cannot identify manually who experienced an ADE. For this reason, we performed one of the research design presented by Hennessy et al. [17]: we choose to compare the population exposed to a DDI with another population non-exposed to this DDI and who did not experience an ADE, by design. There are likely many patients who do not experience an ADE in the exposed population, but the model will present only the most suspected cases and this problem will be solved with the gradual feedback of drug safety professionals: the system will adjust weights of patients in the model, giving a greater weight to the well-predicted patients.

We developed an artificial neural network system that allows us to predict an output. This system has a single hidden layer and the number of perceptrons was decided during cross-validation. Our machine learning model works in two phases. First, it uses all data available for patients who experienced a specific DDI and those who did not (exposed and non-exposed

populations) to classify them as having reported an ADE or not. Then, the model is reinforced with information coming from drug safety professionals who infirm or confirm the previous classification (Fig. 2).

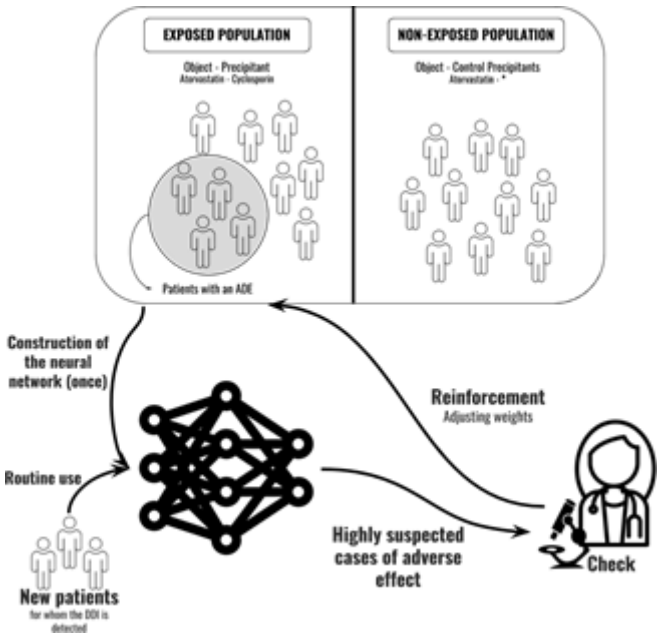


Figure 2 - Creation and use of the artifical neural network

We then had to form the non-exposed population. Within a DDI, we called “Object” the drug under study, and “Precipitant” the other drug. Moreover, we called “Control-precipitant” any drug that has the same therapeutic use as the Precipitant, but that does not interact with the Object. For a given Object, we compared the exposed population, found with the DDI automatic detection system, to the non-exposed population. The non-exposed population included all patients, who were not in the exposed population and who had an overlap (fully or partial) between the action interval of the Object and of the Control-precipitant. We created this non-exposed population using the same process as for the exposed population.

Data processing was performed with Java 8, Spark 2.10 and OrientDB 2.2.4 on Intel(R) Xeon(R) CPU E5-2609 1,90GHz computer with 32,0 Go of RAM.

Big data technologies: convenient tools for complex data processing

Here, we proposed a complete automated data treatment system, from the collection of heterogeneous data to their enhancement in a machine learning model. This system can monitor DDI prevalence and try to identify patients with a possible DDI-linked ADE, without the intervention of drug safety professionals. To achieve this, we used several convenient tools:

OrientDB is an easy-to-use tool to store pre-computed data. The OrientDB database model includes two main classes: vertices and edges that connect two vertices. In our study, the “vertex” interface represented the class “Drug” and included drug name, ID-code and half-life. The “edge” interface represented the class “Interactions” and included DDI severity level. We also specified from which drug database the information on the DDI came. Thus, via OrientDB, each drug knowledge database can be interrogated separately. The “edge” interface is also used to represent the class “Control-precipitant”.

Figure 3 presents the database model through an example: Drug1 has an interaction with Drug3 according two different databases (two edges of class “Interaction”). Let consider the Object-Precipitant couple Drug1-Drug3, then Drug4 is a control-precipitant of Drug3 (one oriented edge of class “Control-precipitant”). An example of query would be: “give all the drugs that have an interaction with Drug1 according Micromedex and where the severity level is 1”.

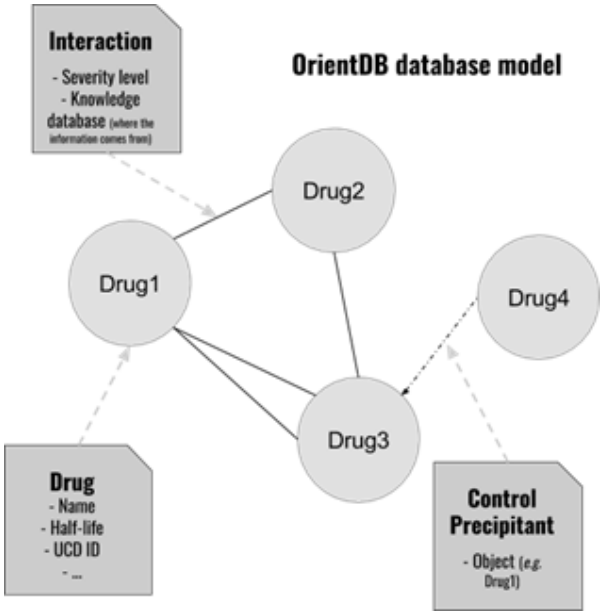


Figure 3 - OrientDb database model

The query language was very close to the structured query language (SQL) and allowed searching a vertex that walks along edges to another vertex, according to the chosen conditions. Data uploading is fast and based on a convenient Java Graph API. We manipulated a graph Java object that is automatically committed at the end of the process. Moreover, if access to a part of the graph is required (e.g., all the drugs that interact with pravastatin according to a severity level of 2), we used this object as a temporary store before processing.

Switching between different knowledge databases, stored in the same OrientDB database, involves only a variable on an edge. Ultimately, the little amount of time spent for the pre-calculation facilitates the storage and the access to multiple data sources. Only one kind of request is needed for all five databases. We could easily add information from other data sources (for example, composition of a drug and half-life of the active substances), or more precise information about DDI-linked ADEs (such as the relevant laboratory tests). Nevertheless, this task demands a manual work for each group of drugs [18].

Spark allows parallel processing easily. As many processes are independent from each other, their parallel treatment with Spark leads to a big time saving [19].

Evaluation (step 3B)

To evaluate our DDI detection system (step 2 in Fig 1), we focused on a class of drugs called statins that are prescribed (long-term treatment) to patients with cardiovascular diseases, and particularly to elderly patients who are usually polymedicated and consequently prone to DDIs. We selected the study population (i.e., all patients taking statins) from all patients included in

eHOP from January, 1 2015 to July, 8 2016. It included 10,506 hospitalized patients with a median hospitalization of 7 days, and a median age of 72 years (range: 19 to 98 years).

We defined statins as the "Object" and all the drugs that interact with them were considered as candidate "Precipitants". We selected as Control-precipitants (symbolized by * in fig 3) all the drugs that are in the same fifth level (i.e., chemical substance) as the Precipitant in the Anatomical Therapeutic Chemical (ATC) classification [20], but do not interact with the Object. Thus, Control-precipitants have the same (or a similar) therapeutic usage as the Precipitant. We stored all these data in OrientDB because each DDI is a link (i.e., edge) between drugs (i.e., two vertices).

Concerning the action intervals, we chose a period of seven half-lives for each statin molecule and arbitrarily selected one day for the Precipitant, because this information could not always be extracted automatically from the five drug knowledge databases.

To determine how well the machine learning model can identify patients who may have a DDI-linked ADE (step 3B in Fig 1), we evaluated the model prediction error using the out-of-bag (OOB) error method: several models are built with a bootstrapped dataset, the OOB error is the mean of the errors computed with non-used data in each model.

The neural network gives the probability to belong to a class. We used cross-validation resampling to optimize the threshold separating the two class. We chose to study a specific DDI in which atorvastatin was the Object and cyclosporine the Precipitant (i.e., exposed population). The non-exposed population consisted of patients who took atorvastatin and a Control-precipitant (Fig 3). The used variables were: demographic data, pathologies (ICD-10 codes) and laboratory test results. We used all the laboratory test results included between the beginning of the event and 3 days later. If a laboratory test appeared more than once, we took the mode of the results.

The reinforcement phase was not evaluated because it is currently under construction in collaboration with drug safety specialists.

Results

DDI identification with the automatic detection system was very fast due to the use of a graph-oriented model. For instance, for the simple query "is there an interaction between these two drugs?", or the more complex query "select all drugs that interact with this specific drug", the OrientDB database was always faster (less than 20ms) than the Theriaque SQL database (several seconds). Moreover, switching to another drug knowledge database was very easy with OrientDB because it only needed to change a condition in the query (which database = 'Theriaque').

The time required to create these graph databases was reasonable: for instance, the information coming from the Theriaque database, which is equivalent to 18,800 vertices and 23 million edges, was integrated in one hour. Afterwards, data access was immediate.

Once the OrientDB database was ready, from the eHOP CDW, we checked the DDI occurrence for all drug couples in the study population. To this aim, we computed all the fully or partially overlapping action intervals for all drug couples involving a statin. For each patient, we visualized all the detected DDIs: between 22.5% and 52.2% (depending on the drug knowledge

database) of the 10,506 patients who were taking statins presented at least one DDI involving a statin.

Computation time was reduced with the use of the Spark framework: the processing time of 800,000 rows of patient records decreased from 60 minutes initially to only 12 minutes with Spark.

To test the ADE prediction performance of the machine learning model, we then focused only on one specific DDI (atorvastatin-cyclosporine) to create the training sample. We could identify 102 patients with atorvastatin-cyclosporine DDIs (i.e., the exposed population) and 150 patients without this DDI (i.e., the non-exposed population) (Table 1).

Table 1- Demographic data of the exposed and non-exposed population samples

	Exposed population (n=102)	**Non-exposed population (n=150)**
Age (mean ± Sd)	72.1 ± 11.6	72.9 ± 10.9
Sex (% of men)	79.8	83.5
Cardiac pathology (%)	38.2	37.8

For the optimal threshold, the neural network out-of-bag error was 17.06%, sensitivity and specificity were 90.20% and 78% respectively, and the AUC was 0.757. The processing time was short (less than 30 seconds) and could be easily performed again during the reinforcement phase.

Discussion

DDI automatic detection system: a new source of refined data for drug safety professionals

With this DDI detection system and the CDW, we can compute the overall DDI prevalence for any drug pairs, and also according to a chosen interaction severity level, or for a specific population subset. These data are useful for drug safety monitoring/research and have been already used in a study on the use of statins [21-22]. Moreover, currently, pharmacovigilance studies use different case report databases [23]. We find DDIs directly in the patient EHRs. Therefore, after DDI detection, we can link this information to other data included in the EHR (e.g., demographic data, laboratory test, etc.) to contextualize the case.

However, our DDI detection system cannot identify all DDIs. This could be due to several reasons. First, the choice of the drug knowledge database is important, and we actually observed heterogeneity between these databases that might lead to variability in DDI detection [22]. Moreover, with more information concerning the changes in the blood concentration (and half-life) of the involved drugs, we could compute more precise action intervals, thus improving the

identification of overlapping treatment periods. However, this would require extensive manual search of literature data. Finally, our system cannot detect a DDI caused by a drug prescribed/administered outside the hospital. For instance, the regular treatment is usually stopped when a patient is hospitalized in the emergency service and is recorded in the emergency report. Accessing this information requires a specific treatment of unstructured text. Another option could be to link data on the drugs prescribed in primary care settings (i.e., the national health insurance database) to the hospital data (e.g., eHOP). Despite the linkage problems and the issues due to the national health insurance database features (data only on refundable drugs and only on the drug purchase but not the regimen), the analysis of the entire patient path could bring useful information on treatment ruptures, which could suggest DDIs.

A machine learning model for search reinforcement

The automatic way used to create the non-exposed population works and selects a population similar to the exposed group in terms of demographics and pathology. If the sample is big enough, we can ask the system to select the most similar patients.

Although the study of the temporal correlations between laboratory test changes and drug administration is relevant for ADE detection [24,25], we chose a robust prediction-oriented machine learning model that can work without requiring too many adjustments. Indeed, we expect that clinical variables in the exposed population will change in the presence of a DDI. However, we do not know whether the detection of a DDI implies automatically an ADE, and accessing the information to confirm the ADE involves a considerable work for drug safety professionals that we want to avoid. Therefore, to automate the monitoring of DDI-linked ADEs, we took the data immediately available from eHOP.

As they have very similar demographic characteristics, comparing exposed and non-exposed populations seemed to be an effective way to initialize the system. An improvement would be to take into account also the information included, for example, in ADE report databases. However, this system can be easily improved even without more data. Indeed, the model predicts candidate ADE cases that are likely to have been caused by DDIs and proposes them to drug safety professionals. If these cases are confirmed by drug safety professionals, they are included in the training sample to automatically enhance the model.

On the other hand, and like for any automatic detection model, our neural network model does not allow understanding which anomaly led to the prediction of an ADE and for this the analysis of the patient record is required. A machine learning model requires a lot of work, especially the choice of the model and the features engineering. In particular, a larger sample could ollow other resampling strategies to be used, that do not require the out of bag error, which is prone to overestimation of the true prediction error [26]. These questions need a suitable study including a better evaluation with drug safety specialists.

Conclusions

This study shows how to employ healthcare data for automated DDI monitoring and ADE prediction. It involves the complete data processing chain: data collection, processing and enrichment as well as the creation of a machine learning model. The developed statistical model

is the first step for a simple and convenient use of data, and could be enriched with additional information from other databases that must be integrated (more specific drug knowledge databases, ADE report databases …).

Although no drug safety professional is required during the monitoring, their expertise is essential to properly understand the data and put them into context. Their recommendations were also important to build the monitoring system and to improve the model.

Acknowledgements

We would like to thank the French National Research Agency (ANR), for funding this work inside the INSHARE (INtegrating and Sharing Health dAta for Research) project (grant no. ANR-15-CE19-0024).

References

[1] F. Meier, R. Maas, A. Sonst, A. Patapovas, F. Müller, B. Plank-Kiegele, B. Pfistermeister, O. Schöffski, T. Bürkle, H. Dormann, Adverse drug events in patients admitted to an emergency department: an analysis of direct costs, Pharmacoepidemiol. Drug Saf. 24 (2015) 176–186.

[2] C. Bonnet, P. Boudou-Rouquette, E. Azoulay-Rutman, O. Huillard, J.-L. Golmard, E. Carton, G. Noé, M. Vidal, G. Orvoen, et al., Potential drug-drug interactions with abiraterone in metastatic castration-resistant prostate cancer patients: a prevalence study in France, Cancer Chemother. Pharmacol. 79 (2017) 1051–1055.

[3] E. Ramirez, A.J. Carcas, A.M. Borobia, S.H. Lei, E. Piñana, S. Fudio, J. Frias, A Pharmacovigilance Program From Laboratory Signals for the Detection and Reporting of Serious Adverse Drug Reactions in Hospitalized Patients, Clin. Pharmacol. Ther. 87 (2010) 74–86.

[4] J. Price, What Can Big Data Offer the Pharmacovigilance of Orphan Drugs?, Clin. Ther. 38 (2016) 2533–2545.

[5] K. Martin-Latry, B. Bégaud, Pharmacoepidemiological research using French reimbursement databases: yes we can!, Pharmacoepidemiol. Drug Saf. 19 (2010) 256–265.

[6] M.-L. Yeh, Y.-J. Chang, S.-J. Yeh, L.-J. Huang, Y.-T. Yen, P.-Y. Wang, Y.-C. Li, C.-Y. Hsu, Potential drug–drug interactions in pediatric outpatient prescriptions for newborns and infants, Comput. Methods Programs Biomed. 113 (2014) 15–22.

[7] G. Bouzillé, R. Westerlynck, G. Defossez, D. Bouslimi, S. Bayat, C. Riou, Y. Busnel, C. Le Guillou, J.-M. Cauvin, C. Jacquelinet, P. Pladys, E. Oger, E. Stindel, P. Ingrand, G. Coatrieux, M. Cuggia, Sharing health big data for research - A design by use cases: the INSHARE platform approach, in: 16th World Congr. Med. Health Inform. MedInfo2017, Hangzhou, China, 2017.

[8] G. Bouzillé, E. Sylvestre, B. Campillo-Gimenez, E. Renault, T. Ledieu, D. Delamarre, M. Cuggia, An Integrated Workflow For Secondary Use of Patient Data for Clinical Research., Stud Health Technol Inf. 216 (2015) 913.

[9] ANSM: Agence nationale de sécurité du médicament et des produits de santé. http://ansm.sante.fr (accessed August 1, 2017).

[10] VIDAL - La base de données en ligne des prescripteurs libéraux. https://www.vidal.fr (accessed August 1, 2017).

[11] Thériaque. http://www.theriaque.org (accessed August 1, 2017).

[12] Micromedex. https://www.micromedexsolutions.com (accessed August 1, 2017).

[13] Drugs.com | Prescription Drug Information, Interactions & Side Effects, Drugs.Com. https://www.drugs.com.

[14] OrientDB - Distributed Graph/Document Multi-Model Database. httpss//orientdb.com (accessed August 1, 2017).

[15] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker, I. Stoica, Spark: Cluster computing with working sets., HotCloud. 10 (2010) 95.

[16] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, I. Stoica, Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing, in: USENIX Association, 2012: pp. 2–2.

[17] S. Hennessy, C. Leonard, J. Gagne, J. Flory, X. Han, C. Brensinger, W. Bilker, Pharmacoepidemiologic Methods for Studying the Health Effects of Drug-Drug Interactions (DDIs), Clin. Pharmacol. Ther. 99 (2016) 92–100.
[18] A. Neubert, H. Dormann, H.-U. Prokosch, T. Bürkle, W. Rascher, R. Sojer, K. Brune, M. Criegee-Rieck, E-pharmacovigilance: development and implementation of a computable knowledge base to identify adverse drug reactions, Br. J. Clin. Pharmacol. 76 (2013) 69–77.
[19] A.G. Shoro, T.R. Soomro, Big data analysis: Apache spark perspective, Glob. J. Comput. Sci. Technol. 15 (2015).
[20] WHOCC - Structure and principles. https://www.whocc.no/atc/structure_and_principles/ (accessed August 1, 2017).
[21] J. Stausberg, International prevalence of adverse drug events in hospitals: an analysis of routine data from England, Germany, and the USA, BMC Health Serv. Res. 14 (2014) 125.
[22] C. Morival, R. Westerlynck, G. Bouzillé, M. Cuggia, P.L. Corre, Prevalence and nature of statin drug-drug interactions in a university hospital by electronic health record mining, Eur. J. Clin. Pharmacol. (2017) 1–10.
[23] F. Kaguelidou, F. Beau-Salinas, A.P. Jonville-Bera, E. Jacqz-Aigrain, Neonatal adverse drug reactions: an analysis of reports to the French pharmacovigilance database, Br. J. Clin. Pharmacol. 82 (2016) 1058–1068.
[24] A. Newe, S. Wimmer, A. Neubert, L. Becker, H.-U. Prokosch, M.W. Beckmann, R. Fietkau, C. Forster, M.F. Neurath, G. Schett, T. Ganslandt, Towards a Computable Data Corpus of Temporal Correlations between Drug Administration and Lab Value Changes, PLOS ONE. 10 (2015) e0136131.
[25] A. Newe, Dramatyping: a generic algorithm for detecting reasonable temporal correlations between drug administration and lab value alterations, PeerJ. 4 (2016) e1851.
[26] S. Janitza, and R. Hornung, On the overestimation of random forest's out-of-bag error, *PLoS One*. **13** (2018).

Address for correspondence

Bouzillé Guillaume: guillaume.bouzille@univ-rennes1.fr

Discussion

Le champ de la science des données est vaste, à la fois sur les méthodes, les données et les applications médicales possibles. La quantité de données aujourd'hui disponible et les performances des algorithmes laissent entrevoir des retombées multiples pour de nombreux usages. Cependant, il ne faut pas perdre de vue que le terreau de ces approches est constitué de données observationnelles, principalement rétrospectives qui, dans le cadre de leur réutilisation, ne répondent pas à des objectifs prédéfinis. Ainsi, les études menées sur ces données sont inévitablement sujettes à de potentiels biais, que ce soit sur les populations étudiées ou sur les données en elles-mêmes (32).

Une des difficultés fréquemment rencontrées est la sélection sans biais des patients d'intérêt pour une étude. Il est en effet relativement difficile de s'assurer que la sélection est effectivement représentative de la population ciblée par l'étude dès lors que l'on mobilise des données rétrospectives. Si les étapes de pré screening permettent de s'affranchir du bruit présent dans la sélection, il n'est en revanche pas possible d'évaluer le nombre de patients non identifiés qui auraient pu être sélectionnés. De plus, les travaux en sciences des données mobilisent de larges cohortes de patients dont la taille n'est pas compatible avec une revue manuelle des dossiers. Un autre exemple concret concerne le développement de systèmes d'aide à la décision s'appuyant sur des données issues de prises en charge hétérogènes parfois inappropriées, qui peuvent biaiser les propositions du modèle (62).

La qualité des données et la méthodologie employée pour les exploiter jouent donc un rôle majeur dans la robustesse des résultats issus des data sciences (63). De la forte hétérogénéité des sources de données et des aspects de qualité de données découlent les verrous majeurs de la réutilisation des données de santé. L'expert « data scientist » doit donc nécessairement avoir une connaissance fine des données à sa disposition et des processus les ayant produites pour adapter sa méthodologie. La proximité de l'expertise métier, en particulier médicale peut alors s'avérer cruciale pour appréhender correctement les données et tenir compte de leurs avantages et limites.

I. Enjeux concernant les données

Une fois les caractéristiques des données appréhendées, il peut être envisagé la mise en œuvre de méthodes permettant d'optimiser leurs qualités. Une première possibilité consiste en

l'amélioration des données à la source, par le producteur de données. Ceci n'est le plus souvent envisageable que s'il y a un intérêt pour le producteur de données à la réaliser. Par exemple, rédiger un compte rendu de façon non structurée convient parfaitement à un clinicien dans le cadre de la prise en charge des patients. La légitimité à demander un effort de structuration de la donnée dans le cadre de la réutilisation secondaire peut d'ailleurs se poser, dans le sens où il s'agit de deux processus contradictoires. En revanche, la réutilisation des données peut permettre d'initier une démarche vertueuse par le développement de méthodes qui illustrent l'apport de l'usage de la donnée et le gain attendu par une mise en qualité des données à la source. Il s'agit alors d'une démarche incitative initiée par une réponse à un besoin et qui inclut le producteur de données dans la démarche de réutilisation des données. La mise en qualité des données ne doit donc pas être présentée comme un prérequis à leur réutilisation, mais plutôt comme une retombée.

Il est également possible d'agir sur la qualité des données a posteriori. On peut distinguer deux axes complémentaires allant dans ce sens.

La structuration des données concerne principalement les sources textuelles, ce qui constitue une problématique de recherche en soi avec par exemple le champ du traitement automatique du langage. Les approches employées aujourd'hui s'appuient majoritairement sur l'apprentissage automatique, bien que le recours à des méthodes plus classiques à base de règles soit également souvent nécessaire. Les méthodes d'apprentissage automatique ont cependant une contrainte forte puisqu'elles demandent un effort d'annotation manuelle des documents. Les ressources françaises en matière de corpus de textes médicaux sont quasiment inexistantes, d'autant plus que l'annotation est le plus souvent spécifique d'un problème ou d'une discipline. Cette annotation est un processus coûteux et dont la fiabilité peut souvent être remise en question comme nous l'avons vu dans l'article portant sur l'extraction des critères d'éligibilité numériques à partir de documents (64). La reproductibilité inter-annotateurs reportée dans la littérature est souvent faible, ce qui pose la question de savoir si cette étape d'annotation experte est compatible avec la philosophie de l'approche « data driven » souvent mise en avant dans les data sciences (65). On peut alors se demander si les modèles ne prédisent pas plus les annotations qui seraient réalisées sur de nouveaux documents que la réelle information que l'on cherchera à identifier. Ceci est d'autant plus vrai que l'évaluation des modèles est le plus souvent réalisée en s'appuyant sur un corpus de documents annoté de la même façon que le corpus ayant servi à l'apprentissage. La richesse des sources de données peut parfois permettre d'obtenir une annotation indirecte des documents dont il faut pouvoir tirer parti dans les

approches de traitement automatique de la langue. Par exemple, les diagnostics du PMSI sont étroitement liés aux comptes rendus d'hospitalisations, les prescriptions électroniques de médicaments se retrouvent également dans les comptes rendus ou encore, s'agissant de la désidentification, les identités patients sont connues dans le système d'information et pourraient servir de base d'apprentissage. Dans cette approche, l'annotation est indirecte, mais le fait qu'elle ne requiert pas de procédure manuelle permet d'envisager la constitution de corpus de volume plus important, potentiellement issus de plusieurs centres.

Parmi les processus de mise en qualité a posteriori, les méthodes de phénotypage automatique à partir des données se sont développées depuis quelques années. Par ailleurs, la mise en qualité des données repose sur le principe du phénotypage à partir des données. Ce processus revient à diminuer l'hétérogénéité des données en réconciliant des sources de données plus ou moins redondantes ou complémentaires par le biais d'algorithmes à dire d'experts ou produits à partir des données pour aboutir à une information phénotypique fiable. Ces méthodes peuvent évidemment combiner sources textuelles et données structurées pour produire une information consolidée. Ces approches ont été largement décrites dans la littérature et les algorithmes de phénotypage font l'objet de propositions de standardisation telles que la base de données phénotype knowledge base (66). Ces concepts extraits à partir des données doivent évidemment être décrits selon des ressources terminologiques standards.

Cette standardisation peut s'envisager à la fois sur les données non pré-traitées si elles répondent en tant que telles à un besoin (par exemple, un résultat numérique à un dosage biologique) ou sur des données ayant déjà subi un prétraitement, que ce soit par un phénotypage ou un processus de traitement automatique du langage.

II. Enjeux concernant le partage des données

Le partage de données est aujourd'hui indispensable pour plusieurs raisons : en premier lieu, le changement d'échelle permet de lever certains verrous liés au volume de données nécessaires pour obtenir des résultats robustes et généralisables comme on peut l'entrevoir avec l'exemple précédent en traitement automatique du langage (67). Par ailleurs, l'intérêt du partage de données réside dans la complémentarité des sources aujourd'hui disponibles. L'exemple le plus évident est peut-être la complémentarité des données sur SNIIRAM avec les données hospitalières, puisque ces deux sources combinées permettent de couvrir l'ensemble de la trajectoire de soins des patients. La complémentarité provient également des sources issues de

différentes disciplines : par exemple les données cliniques et les données d'imagerie ou les données OMICs. Là encore, la proximité des utilisateurs finaux, cliniciens, chercheurs est indispensable pour cibler les besoins concrets en matière de partage de données.

La mise en place d'un projet de partage de données requiert une réflexion sur les axes technologiques, méthodologiques, de sécurité ou encore sur les aspects réglementaires et de gouvernance. Sur le plan technique, l'enjeu majeur se situe au niveau de la normalisation et la standardisation des données, ainsi que dans l'interopérabilité des systèmes produisant les données. L'implication des producteurs de données peut permettre d'envisager cet effort à la source. Cela peut s'avérer indispensable lorsqu'il s'agit de données extrêmement spécialisées que seul un expert du domaine peut appréhender. Dans le cas contraire, il faudra mettre en place des procédures permettant de réaliser des alignements terminologiques a posteriori pour réconcilier les données ou des méthodes d'appariement en ce qui concerne les identités.

Les ruptures technologiques sur les architectures de stockage des données permettent d'aborder ces problématiques de façon plus souple. Les approches « Data Lake » permettent en effet de collecter des données dans leur format brut puisqu'il n'y a pas de contrainte de modèle d'intégration à respecter. Les transformations peuvent être réalisées dans un second temps, ce qui laisse entrevoir la possibilité d'évaluer différentes approches de transformation, d'adapter ces dernières en fonction des besoins, de rejouer des procédures, voire d'employer des méthodes d'alignement automatique basées sur l'apprentissage automatique. Il s'agit donc de trouver l'équilibre entre l'approche data-driven et l'expertise médicale nécessaire pour réaliser les traitements pertinents.

III. Enjeux concernant les méthodes d'exploitation des données

Il y a aujourd'hui une attente importante concernant les projets mobilisant la réutilisation de données, pour lesquels les investigateurs espèrent souvent des résultats de niveaux de preuve similaires à ce qui pourrait être réalisé en recherche clinique via des essais thérapeutiques ou des études épidémiologiques. De la même façon, les méthodes développées en sciences des données pour la prise en charge des patients sont souvent développées pour une mise en pratique après leur phase de conception. Les attentes sont donc souvent plus grandes que les possibilités offertes actuellement par les données et les méthodes.

Le champ des data sciences est actuellement saturé par les approches dites d'intelligence artificielle censées couvrir l'ensemble des besoins. Pourtant, certains besoins sont toujours confrontés à des verrous pouvant paraître triviaux : l'identification de patients éligibles en recherche clinique, la notification des effets indésirables, l'anticipation des épidémies ou encore l'aide au diagnostic. Les différents cas d'usage présentés illustrent bien que des réponses peuvent être apportées par l'emploi de méthodes simples ou plus traditionnelles et mieux adaptées au volume ou au type de données à traiter (68). À l'inverse l'utilisation de méthodes d'apprentissage automatique est vue comme faussement simple d'utilisation, avec l'idée reçue qu'il suffit d'appliquer les modèles sur les données pour obtenir des résultats pertinents, fiables et robustes.

Cette discordance entre la réalité de terrain et les prétentions du champ des data sciences s'explique principalement par la très forte communication de la part de certains acteurs, depuis l'avènement du Big Data, du Deep Learning et de l'intelligence artificielle, sur les retombées attendues par ces méthodes très prometteuses et qui s'avèrent parfois décevantes (69). Une meilleure communication au sujet des objectifs accessibles aujourd'hui par la réutilisation secondaire des données et les data sciences semble donc nécessaire, car le risque est de provoquer de façon définitive une déception des utilisateurs par rapport aux possibilités des data sciences dans ce domaine.

IV. Enjeux concernant les usages des données

L'apport des data sciences devrait être présentée comme consistant avant tout à pouvoir faciliter les pratiques, quels que soient les domaines, avec pour conséquence de diminuer les coûts. Le développement de méthodes totalement automatisées est encore marginal et le recours à une validation experte est le plus souvent indispensable. En fait, il est même préférable de positionner et de développer les méthodes dans ce sens afin d'obtenir une meilleure acceptabilité des outils. Cette dernière est également fortement liée à l'utilisabilité des outils qui sont développés et mis à disposition des professionnels (70,71).

La science des données peut également s'envisager selon son impact direct, ou non, sur les individus. Il en découle des niveaux de risques différents selon les domaines d'application.

A. Surveillance syndromique

En ce qui concerne la surveillance syndromique, les impacts se situent en matière de santé publique, en particulier sur l'organisation des soins. Dans ce contexte, il existe déjà des outils très performants que les méthodes issues de la science des données ne cherchent pas à substituer. L'objectif est avant tout de développer des méthodes complémentaires aux outils traditionnels. Il s'agit d'anticiper la production des indicateurs de référence, mais également de produire des méthodes de détection de nouveaux signaux à surveiller. Deux leviers sont disponibles pour répondre à ces besoins.

En premier lieu, l'intégration de nouvelles sources de données parfois quasi temps réel ou permettant de tirer de l'information depuis d'autres domaines : nous avons montré que les données hospitalières en combinaison avec les données du réseau Sentinelles ont un fort potentiel pour la surveillance des épidémies grippales (72). De nombreuses autres sources ont été évaluées et ont monté leur intérêt : données d'internet (google, twitter, wikipedia), environnementale, de météorologie (59,73,74).

En second lieu, l'exploitation de ces diverses sources de données nécessite l'usage de modèles prédictifs adaptés. Sur ce plan également, de nombreux travaux ont évalué différents types de modèles, qu'ils soient issus des statistiques, de l'apprentissage automatique ou profond (75–79). Ces modèles permettent d'envisager des prévisions à des échéances de 3 ou 4 semaines, d'anticiper le début de l'épidémie ou le pic épidémique. Comme notre étude l'a montré, ces modèles sont généralement évalués sur des données rétrospectives. De nombreux facteurs n'ont donc potentiellement pas été pris en compte : surapprentissage du modèle, disponibilité des données à la fréquence attendue, existence d'une tendance dans les données liées à des changements d'activité ou présence de nouvelles sources de données. Il est donc indispensable de réaliser des études prospectives pour valider ces modèles (80).

Par ailleurs, ces modèles sont évalués à différentes échelles, le plus souvent à des niveaux local, régional ou national. Il y a en effet un fort besoin notamment au niveau des services d'urgence à pouvoir anticiper les afflux de patients. La mise en œuvre de ce type de modèle au sein des établissements pourrait, à moindre coût, permettre de suivre les pathologies prises en charge ce qui pourrait constituer des marqueurs précoces de phénomènes nouveaux ou réémergeants qui ne disposent pas de dispositifs de surveillance.

B. Recherche clinique

En recherche clinique, nous avons montré que les technologies d'entrepôt de données avec leurs fonctionnalités de recherche d'informations permettent d'optimiser la recherche de patients éligibles à des études ou de réaliser des études de faisabilité. Le décloisonnement des sources de données permet de couvrir aujourd'hui la quasi-exhaustivité des critères d'inclusion ou de non-inclusion : critères diagnostics, résultats biologiques, prise en charge thérapeutique, ou encore données de biobanques. Les méthodes de recherche d'informations ont pour avantage de tirer parti des données non structurées, qui sont une des sources les plus riches pour retrouver l'information d'intérêt.

Les limites actuelles de ces approches tiennent dans leur capacité à tenir compte de la synonymie, du contexte (notion d'antécédents, de certitude) et de la temporalité afin d'optimiser le rappel et la précision des recherches de patients. De la même façon, les critères d'éligibilité des études sont généralement exprimés de façon non structurée ce qui ne laisse pas entrevoir d'approche d'alignement terminologique sans approche de traitement automatique du langage au préalable (81).

Les enjeux en matière de recherche d'information sont donc de deux ordres, à la fois sur l'amélioration de la qualité des données par le biais de la structuration des concepts et des contextes présents dans les documents, mais également dans le développement d'outils de recherche d'informations permettant de formaliser les critères d'éligibilité des patients.

Le premier axe d'amélioration a fait l'objet de nombreux travaux de recherche notamment en traitement du langage et en phénotypage à partir des données, mais restent souvent peu généralisables. On retrouve ici les verrous connus des méthodes de traitement du langage liés à l'annotation, la reproductibilité et la grande hétérogénéité des critères d'éligibilité à couvrir.

Le second axe d'amélioration est l'amélioration de l'expressivité des outils de recherche d'information pour être en mesure de formaliser les critères d'éligibilité des patients. La plupart des outils de recherche visent à traduire directement les critères d'éligibilité par des critères sur les données, ce qui impose à la fois une expertise en recherche clinique et en science des données. Deux niveaux de logique peuvent pourtant être distingués : une logique liée aux relations entre les critères d'éligibilité et une logique liée aux critères et relations à retrouver au sein des données. Cette approche en deux étapes aurait plusieurs avantages :

- l'étape de formalisation des critères cliniques : elle permet de rendre intelligible la recherche (c'est-à-dire la requête) pour un professionnel de recherche clinique et de rendre partageable ce premier niveau de logique sur les différents sites participant à une étude.
- l'étape de formalisation des critères cliniques sur les données : cette étape revient à un processus de phénotypage de concepts plus élémentaires qui deviennent potentiellement réutilisables.

L'objectif derrière cette approche est de tirer parti de l'usage actuel qui est fait des données en recherche clinique pour structurer indirectement l'information. Valider l'éligibilité d'un patient revient à annoter ce patient avec les critères d'éligibilité de l'essai (par exemple la présence du statut fumeur).

Ceci fait le lien avec un autre aspect important qui est la structuration des données pour l'alimentation de bases de données de recherche. L'export de données déjà structurées à la source (par exemple, un entrepôt de données) ne pose pas de problème technique majeur en dehors de problèmes de granularité ou d'alignement terminologique à réaliser. L'alimentation d'une base de recherche par des données issues de documents non structurés est beaucoup plus difficile puisque de telles bases cibles demandent une fiabilité des données recueillies. Ceci n'est pas réellement compatible avec les performances d'outils à base de règles et reste encore hors de portée des méthodes d'apprentissage automatique.

Là encore, l'apport des data sciences et plus largement de l'informatique médicale réside dans la capacité à proposer des outils intégrés à la pratique de la recherche clinique. Il s'agit de faciliter l'accès à l'information pertinente nécessaire au recueil manuel de données et profiter dans le même temps de cette ressaisie d'information pour annoter les données brutes. Ces approches permettraient donc sur le long terme d'obtenir une annotation, bien qu'indirecte, de l'information contenue dans les documents, orientée par l'usage et potentiellement plus pertinente qu'une annotation ad hoc réalisée dans le cadre de projets de recherche en traitement automatique du langage.

C. Pharmacovigilance

Une part des besoins de la recherche clinique se retrouve également en matière de pharmacovigilance, notamment en ce qui concerne l'identification de patients ayant eu des effets indésirables médicamenteux. Si certains de ces effets sont relativement aisés à identifier,

d'autres sont exprimés de façon très peu spécifiques, voire non évoqués. Ils ne pourront alors être retrouvés que par l'association de symptômes ou d'anomalies biologiques souvent également peu spécifiques. Les avancées des méthodes de traitement automatique de la langue sont très prometteuses (82). Cependant, les effets indésirables médicamenteux sont des événements rares demandant généralement de gros volumes de données pour obtenir une détection satisfaisante par les modèles d'apprentissage automatique. Le cas d'usage présenté a montré que bien que le bruit apporté par les méthodes de recherche d'information demeure important, le rappel est bien supérieur aux autres méthodes traditionnellement employées (notification spontanée et base PMSI).

Là encore, les entrepôts de données biomédicales offrent des fonctionnalités de recherche d'informations adaptées à la pharmacovigilance pour accéder rapidement aux éléments permettant d'identifier les cas réels d'effets indésirables médicamenteux. La présence de bruit dans les résultats est donc bien acceptée par le gain de temps apporté par ces outils pour valider les cas potentiels. Cette validation repose d'une part sur la présence réelle de l'effet et d'autre part à l'imputabilité de l'effet à un médicament qui repose essentiellement sur la relation temporelle entre la prise du médicament et l'apparition de l'effet. Il est donc important de pouvoir prendre en compte cette temporalité via les moteurs de recherche, ce qui se révèle complexe puisque cette notion est le plus souvent exprimée elle aussi dans les documents non structurés et parfois de façon relative. En revanche, la représentation des données à l'utilisateur sous forme de chronologie peut permettre de faciliter la reconstitution de la trajectoire de soin des patients et ainsi permettre de mieux évaluer les relations temporelles entre les événements (83).

Au-delà de l'identification d'effets indésirables déjà connus, un besoin important en pharmacovigilance est de développer des méthodes permettant de détecter de nouveaux événements indésirables. L'usage des données de vie réelle du big data prend ici tout son sens puisqu'elles permettent potentiellement de lever les limites de l'évaluation des produits de santé lors des essais de phase III : puissance suffisante pour détecter des événements rares, prise en compte des caractéristiques individuelles ou de prise en charge telles que les coprescriptions pouvant induire des interactions médicamenteuses (84). Bien que les données textuelles soient tout à fait pertinentes dans ce contexte, de nombreuses sources de données produisent des données structurées intéressantes pour répondre au besoin :

- les médicaments : données du SNIIRAM, données issues des logiciels de prescriptions électroniques.
- Les données en lien avec la présence d'un effet indésirable : données de laboratoire, diagnostics CIM-10 issus du PMSI.

Il y a donc un axe à la fois sur l'intégration des données d'intérêt, ambulatoires, hospitalières voire des données web comme pour la surveillance syndromique et un axe sur les méthodes de détection des associations médicaments-événement indésirable (85). Les méthodes de fouilles de données paraissent ici parfaitement adaptées afin d'associer la détection des associations présentes dans des données de grandes dimensions et l'expertise nécessaire à l'interprétation.

D. Médecine personnalisée

Enfin, outre le suivi et l'identification des événements indésirables, il s'agit d'exploiter ces connaissances sur les produits de santé dans le cadre de la médecine personnalisée afin par exemple d'adapter la surveillance du traitement ou de proposer une alternative thérapeutique en fonction de la probabilité de survenue d'événements indésirables. Néanmoins, la réutilisation des données de santé dans le cadre du soin s'avère nettement plus complexe que les cas d'usage présentés, du fait de la sensibilité du domaine et donc de l'impact potentiel des propositions formulées par les modèles.

L'exigence de fiabilité des modèles est donc essentielle alors que les algorithmes et les méthodes sont extrêmement sujets au surajustement, aux facteurs de confusion potentiels, aux biais de sélection ou toute autre source de biais engendré par les caractéristiques des données massives.

L'amélioration de la prise en charge des patients guidée par les données constitue la part médiatique de l'exploitation des données massives en santé. C'est dans ce contexte que se positionnent de nombreuses start-ups. Les premières avancées en matière d'exploitation du big data concernent la segmentation des images médicales notamment pour la stadification des tumeurs où les performances des modèles surpassent parfois l'expertise humaine, le radiologue gardant évidemment la conclusion finale de l'examen. L'enjeu est donc que le développement de ces systèmes d'aide à la décision à partir des données massives à l'usage du soin suive une méthodologie rigoureuse et progressive. La phase de conception et d'évaluation sur données massives rétrospectives ne constitue que la première étape et il est essentiel que ces modèles puissent être évalués dans un cadre classique de recherche clinique sur des données prospectives

et dans une démarche expérimentale classique de comparaison par rapport à une méthode de référence comme tout outil d'aide à la décision.

À la manière des médicaments qui continuent à être surveillés après leur mise sur le marché, les systèmes d'aide à la décision notamment ceux basés sur des algorithmes d'apprentissage automatique doivent être monitorés afin d'évaluer si la calibration demeure stable dans le temps. Des mesures simples peuvent être utilisées comme la variation de l'aire sous la courbe ROC au cours du temps (s'il est possible d'obtenir la réalité des faits), auxquels on peut rajouter des méthodes de génération d'alarme afin de détecter des variations anormales de la performance des modèles (86–88).

Enfin, la limite principale des méthodes d'apprentissage automatique est leur caractère boîte noire ce qui, à la différence des systèmes experts, les rend ininterprétables. Cet aspect constitue un axe de recherche en plein développement avec des méthodes qui commencent à apparaître pour les rendre interprétables et des retombées dans le domaine médical extrêmement importantes (55,89–91). Tout d'abord, l'interprétation du modèle est indispensable pour permettre de valider la pertinence de la modélisation par rapport aux connaissances métiers. En ce qui concerne les modèles mis en œuvre de façon opérationnelle, l'interprétation permet d'une part de valider la cohérence de la proposition et d'autre part d'être en mesure d'expliquer la proposition aux patients. Ces éléments peuvent donc apporter des éléments de transparence par rapport à ces nouveaux outils et accélérer leur adoption grâce à une meilleure maîtrise de leur comportement et de leurs limites.

Conclusion

Cette thèse s'appuie sur des exemples concrets d'application des data sciences pour la réutilisation des données massives en santé. Elle ne permet évidemment pas de couvrir l'ensemble des usages possibles, mais souligne certains points forts communs à l'ensemble des domaines d'application des data sciences.

Ainsi, on retrouve les constats identifiés depuis longtemps sur la qualité des données, le manque de standardisation des sources, et la faible interopérabilité des systèmes. La donnée étant le point de départ du champ de la science des données, l'enjeu est évidemment d'être en capacité d'améliorer sa qualité. Les avancées technologiques récentes en matière de stockage, de calcul ou de traitement de données permettent d'envisager de nouvelles approches plus souples et pouvant passer à l'échelle pour lever certains verrous, liés à la qualité, dans un cadre de partage des données. En parallèle, il est indispensable de promouvoir l'usage de cette masse de données hétérogène pour répondre à des besoins afin de susciter l'adhésion de telles approches.

Au-delà des aspects méthodologiques et techniques, la place des data sciences dépend de son intégration dans les pratiques actuelles des utilisateurs finaux. Elle doit pouvoir répondre avant tout à des besoins concrets, ne pas nécessairement remettre en cause des processus déjà établis et performants, ne pas se substituer à l'utilisateur. L'objectif est de pouvoir proposer des outils robustes et fiables fluidifiant les pratiques et contribuant à l'amélioration de la donnée.

Cette rupture technologique annoncée doit toutefois être appréhendée au regard de ses apports et limites. Le rôle du « data scientist » et plus largement de l'informatique médicale réside dans sa capacité à apporter une méthodologie adaptée aux données et un usage raisonné des data sciences, ce qui conditionne l'acceptabilité des outils développés et la confiance qui leur sera accordée. Ceci ne pourra se faire que par la mise en place d'organisations adaptées rassemblant les données, les besoins, les experts et les utilisateurs finaux.

Références

1. Adam NR, Wieder R, Ghosh D. Data science, learning, and applications to biomedical and health sciences. Ann N Y Acad Sci. 2017;1387(1):5-11.
2. Beyer MA, Laney D. The importance of 'big data': a definition. Stamford CT Gart. 2012;2014–2018.
3. van der Aalst W. Data Science in Action. In: van der Aalst W, éditeur. Process Mining: Data Science in Action [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016 [cité 12 avr 2019]. p. 3-23. Disponible sur: https://doi.org/10.1007/978-3-662-49851-4_1
4. Viju Raghupathi WR. An Overview of Health Analytics. J Health Med Inform. 2013;04(03).
5. Wang Y, Hajli N. Exploring the path to big data analytics success in healthcare. J Bus Res. 1 janv 2017;70:287-99.
6. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc. 1 janv 2007;14(1):1-9.
7. Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. Adv Ther. nov 2018;35(11):1763-74.
8. Institute of Medicine (US) Roundtable on Evidence-Based Medicine. The Learning Healthcare System: Workshop Summary [Internet]. Olsen L, Aisner D, McGinnis JM, éditeurs. Washington (DC): National Academies Press (US); 2007 [cité 17 avr 2019]. (The National Academies Collection: Reports funded by National Institutes of Health). Disponible sur: http://www.ncbi.nlm.nih.gov/books/NBK53494/
9. McLachlan S, Potts HWW, Dube K, Buchanan D, Lean S, Gallagher T, et al. The Heimdall Framework for Supporting Characterisation of Learning Health Systems. J Innov Health Inform. 15 juin 2018;25(2):77-87.
10. Kuo M-H, Sahama T, Kushniruk AW, Borycki EM, Grunwell DK. Health big data analytics : current perspectives, challenges and potential solutions. Int J Big Data Intell. 2014;1:114-26.
11. Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. Int J Med Inf. 1 févr 2017;98:22-32.
12. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. 5 nov 2014;14(1):1144.
13. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Rev Epidemiol Sante Publique. oct 2017;65 Suppl 4:S149-67.
14. Health Data Hub - Ministère des Solidarités et de la Santé [Internet]. [cité 2 mai 2019]. Disponible sur: https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/acces-aux-donnees-de-sante/article/health-data-hub
15. Pletcher MJ, Forrest CB, Carton TW. PCORnet's Collaborative Research Groups [Internet]. Patient Related Outcome Measures. 2018 [cité 15 janv 2019]. Disponible sur: https://www.dovepress.com/pcornets-collaborative-research-groups-peer-reviewed-article-PROM
16. Akgül CB, Rubin DL, Napel S, Beaulieu CF, Greenspan H, Acar B. Content-Based Image

Retrieval in Radiology: Current Status and Future Directions. J Digit Imaging. 1 avr 2011;24(2):208-22.
17. Girardeau Y, Doods J, Zapletal E, Chatellier G, Daniel C, Burgun A, et al. Leveraging the EHR4CR platform to support patient inclusion in academic studies: challenges and lessons learned. BMC Med Res Methodol. 28 févr 2017;17(1):36.
18. Claerhout B, Kalra D, Mueller C, Singh G, Ammour N, Meloni L, et al. Federated electronic health records research technology to support clinical trial protocol optimization: Evidence from EHR4CR and the InSite platform. J Biomed Inform. 1 févr 2019;90:103090.
19. Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. JAMA J Am Med Assoc. 2014;311(24):2479-80.
20. Martin EG, Helbig N, Birkhead GS. Opening Health Data: What Do Researchers Want? Early Experiences With New York's Open Health Data Platform. J Public Health Manag Pract JPHMP. oct 2015;21(5):E1-7.
21. Boslaugh S. Secondary Data Sources for Public Health: A Practical Guide. Cambridge University Press; 2007. 164 p.
22. Banque Nationale de Données Maladies Rares [Internet]. Banque Nationale de Données Maladies Rares. [cité 18 janv 2019]. Disponible sur: http://www.bndmr.fr/
23. Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open. 1 déc 2017;7(12):e018647.
24. DMP : Dossier Médical Partagé [Internet]. [cité 3 mai 2019]. Disponible sur: https://www.dmp.fr/
25. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. Health Inf Sci Syst. déc 2018;6(1):8.
26. Karampela M, Ouhbi S, Isomursu M. Personal health data: A systematic mapping study. Int J Med Inf. 1 oct 2018;118:86-98.
27. Orphanidou C. A review of big data applications of physiological signal data. Biophys Rev. févr 2019;11(1):83-7.
28. Yu X-T, Zeng T. Integrative Analysis of Omics Big Data. Methods Mol Biol Clifton NJ. 2018;1754:109-35.
29. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. avr 2013;309(13):1351-2.
30. Savaris A, Härder T, von Wangenheim A. DCMDSM: a DICOM decomposed storage model. J Am Med Inform Assoc. 1 sept 2014;21(5):917-24.
31. Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, et al. The HL7 Clinical Document Architecture. J Am Med Inform Assoc JAMIA. déc 2001;8(6):552-69.
32. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform. août 2017;26(1):38.
33. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. J Manag Inf Syst. 1 mars 1996;12(4):5-33.
34. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for Data Quality Assessment and Improvement. ACM Comput Surv. juill 2009;41(3):16:1–16:52.
35. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data [Internet]. 15 mars 2016 [cité 3 mai 2019];3. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/
36. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: The next

frontier for innovation, competition, and productivity. mai 2011 [cité 16 janv 2019]; Disponible sur: http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation
37. Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Res Int [Internet]. 2015;2015. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4468280/
38. Roski J, Bo-Linn GW, Andrews TA. Creating Value In Health Care Through Big Data: Opportunities And Policy Implications. Health Aff (Millwood). 1 juill 2014;33(7):1115-22.
39. Larson D, Chang V. A review and future direction of agile, business intelligence, analytics and data science. Int J Inf Manag. 1 oct 2016;36(5):700-10.
40. Phillips-Wren G, Iyer LS, Kulkarni U, Ariyachandra T. Business analytics in the context of big data: A roadmap for research. Commun Assoc Inf Syst. 2015;37:448-72.
41. Aluko V, Sakr S. Big SQL systems: an experimental evaluation. Clust Comput [Internet]. 11 févr 2019 [cité 3 mai 2019]; Disponible sur: https://doi.org/10.1007/s10586-019-02914-4
42. Hristovski D, Kastrin A, Dinevski D, Rindflesch TC. Constructing a Graph Database for Semantic Literature-Based Discovery. Stud Health Technol Inform. 2015;216:1094.
43. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2 nov 2010;153(9):600.
44. Burton PR, Banner N, Elliot MJ, Knoppers BM, Banks J. Policies and strategies to facilitate secondary use of research data in the health sciences. Int J Epidemiol. 1 déc 2017;46(6):1729-33.
45. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc JAMIA. oct 2009;16(5):624-30.
46. DRIM France IA [Internet]. [cité 3 mai 2019]. Disponible sur: http://www.sfrnet.org/sfr/professionnels/drim-france-ia/index.phtml
47. Durojaiye AB, Puett LL, Levin S, Toerper M, McGeorge NM, Webster KLW, et al. Linking Electronic Health Record and Trauma Registry Data: Assessing the Value of Probabilistic Linkage. Methods Inf Med. 2018;57(5-06):261-9.
48. Oellrich A, Collier N, Groza T, Rebholz-Schuhmann D, Shah N, Bodenreider O, et al. The digital revolution in phenotyping. Brief Bioinform. sept 2016;17(5):819-30.
49. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2018;2017:188-96.
50. Martinez D, Otegi A, Soroa A, Agirre E. Improving search over Electronic Health Records using UMLS-based query expansion through random walks. J Biomed Inform. oct 2014;51:100-6.
51. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med. 10 nov 2011;365(19):1758-9.
52. Sacchi L, Holmes JH. Progress in Biomedical Knowledge Discovery: A 25-year Retrospective. Yearb Med Inform. 20 mai 2016;(Suppl 1):S117-29.
53. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Mag. 15 mars 1996;17(3):37-37.
54. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data. 24 févr 2015;2(1):1.

55. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 27 nov 2018;19(6):1236-46.
56. Santé publique France - Accueil [Internet]. [cité 4 mai 2019]. Disponible sur: https://www.santepubliquefrance.fr/
57. Réseau Sentinelles > France > Accueil [Internet]. [cité 18 mai 2016]. Disponible sur: https://websenti.u707.jussieu.fr/sentiweb/
58. Réseau OSCOUR® / Surveillance syndromique - SurSaUD® / Veille et alerte / Dossiers thématiques / Accueil [Internet]. [cité 20 mai 2016]. Disponible sur: http://www.invs.sante.fr/Dossiers-thematiques/Veille-et-alerte/Surveillance-syndromique-SurSaUD-R/Reseau-OSCOUR-R
59. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013-2014 influenza season using Wikipedia. PLoS Comput Biol. mai 2015;11(5):e1004239.
60. Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. PLOS ONE. 9 déc 2013;8(12):e83672.
61. Araz OM, Bentley D, Muelleman RL. Using Google Flu Trends data in forecasting influenza-like–illness related ED visits in Omaha, Nebraska. Am J Emerg Med. sept 2014;32(9):1016-23.
62. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 30 avr 2018;361:k1479.
63. Safran C. Update on Data Reuse in Health Care. Yearb Med Inform. août 2017;26(1):24.
64. Claveau V, Oliveira LES, Bouzillé G, Cuggia M, Moro CMC, Grabar N. Numerical Eligibility Criteria in Clinical Protocols: Annotation, Automatic Detection and Interpretation. In: Artificial Intelligence in Medicine [Internet]. Springer, Cham; 2017 [cité 1 févr 2018]. p. 203-8. (Lecture Notes in Computer Science). Disponible sur: https://link.springer.com/chapter/10.1007/978-3-319-59758-4_22
65. Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. J Biomed Inform. 1 oct 2009;42(5):950-66.
66. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc JAMIA. 2016;23(6):1046-52.
67. Munkhdalai T, Liu F, Yu H. Clinical Relation Extraction Toward Drug Safety Surveillance Using Electronic Health Record Narratives: Classical Learning Versus Deep Learning. JMIR Public Health Surveill. 25 avr 2018;4(2):e29.
68. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform. 2017;21(1):4-21.
69. IBM pitched Watson as a revolution in cancer care. It's nowhere close [Internet]. STAT. 2017 [cité 30 janv 2019]. Disponible sur: https://www.statnews.com/2017/09/05/watson-ibm-cancer/
70. Pressler TR, Yen P-Y, Ding J, Liu J, Embi PJ, Payne PRO. Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. BMC Med Inform Decis Mak. 30 mai 2012;12:47.
71. Ledieu T, Bouzillé G, Thiessard F, Berquet K, Van Hille P, Renault E, et al. Timeline representation of clinical data: usability and added value for pharmacovigilance. BMC Med Inform Decis Mak. 19 2018;18(1):86.
72. Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert M-L, Chabot M, Chazard E, et al. Leveraging hospital big data to monitor flu epidemics. Comput Methods Programs Biomed. 1 févr 2018;154:153-60.

73. Shaman J, Kandula S, Yang W, Karspeck A. The use of ambient humidity conditions to improve influenza forecast. PLoS Comput Biol. nov 2017;13(11):e1005844.
74. Moss R, Zarebski A, Dawson P, McCaw JM. Forecasting influenza outbreak dynamics in Melbourne from Internet search query surveillance data. Influenza Other Respir Viruses. 2016;10(4):314-23.
75. Volkova S, Ayton E, Porterfield K, Corley CD. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. PloS One. 2017;12(12):e0188941.
76. Zhang J, Nawata K. Multi-step prediction for influenza outbreak by an adjusted long short-term memory. Epidemiol Infect. 2018;146(7):809-16.
77. Ertem Z, Raymond D, Meyers LA. Optimal multi-source forecasting of seasonal influenza. PLoS Comput Biol. 2018;14(9):e1006236.
78. Osthus D, Hickmann KS, Caragea PC, Higdon D, Del Valle SY. Forecasting seasonal influenza with a state-space SIR model. Ann Appl Stat. mars 2017;11(1):202-24.
79. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV. A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza Other Respir Viruses. 1 mai 2014;8(3):309-16.
80. Kakarmath S, Golas S, Felsted J, Kvedar J, Jethwani K, Agboola S. Validating a Machine Learning Algorithm to Predict 30-Day Re-Admissions in Patients With Heart Failure: Protocol for a Prospective Cohort Study. JMIR Res Protoc. 4 sept 2018;7(9):e176.
81. Si Y, Weng C. An OMOP CDM-Based Relational Database of Clinical Research Eligibility Criteria. Stud Health Technol Inform. 2017;245:950-4.
82. Li F, Liu W, Yu H. Extraction of Information Related to Adverse Drug Events from Electronic Health Record Notes: Design of an End-to-End Model Based on Deep Learning. JMIR Med Inform. 26 nov 2018;6(4):e12159.
83. Ledieu T, Bouzillé G, Polard E, Plaisant C, Thiessard F, Cuggia M. Clinical Data Analytics With Time-Related Graphical User Interfaces: Application to Pharmacovigilance. Front Pharmacol [Internet]. 30 août 2018 [cité 18 oct 2018];9. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6127627/
84. Liu F, Jagannatha A, Yu H. Towards Drug Safety Surveillance and Pharmacovigilance: Current Progress in Detecting Medication and Adverse Drug Events from Electronic Health Records. Drug Saf. 1 janv 2019;42(1):95-7.
85. Harpaz R, DuMouchel W, Schuemie M, Bodenreider O, Friedman C, Horvitz E, et al. Toward multimodal signal detection of adverse drug reactions. J Biomed Inform. déc 2017;76:41-9.
86. Davis SE, Lasko TA, Chen G, Matheny ME. Calibration Drift Among Regression and Machine Learning Models for Hospital Mortality. AMIA Annu Symp Proc. 16 avr 2018;2017:625-34.
87. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc JAMIA. 1 nov 2017;24(6):1052-61.
88. Neuburger J, Walker K, Sherlaw-Johnson C, van der Meulen J, Cromwell DA. Comparison of control charts for monitoring clinical performance using binary data. BMJ Qual Saf. nov 2017;26(11):919-28.
89. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable Deep Models for ICU Outcome Prediction. AMIA Annu Symp Proc AMIA Symp. 2016;2016:371-80.
90. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc JAMIA. 8 juin 2018;25(10):1419-28.
91. Arras L, Horn F, Montavon G, Müller K-R, Samek W. « What is relevant in a text

document? »: An interpretable machine learning approach. PloS One. 2017;12(8):e0181142.

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE RENNES 1

Titre : Enjeux et place des data sciences dans le champ de la réutilisation secondaire des données massives cliniques. Une approche basée sur des cas d'usage

Mots clés : Réutilisation secondaire des données, Données massives en santé, Sciences des données, Surveillance syndromique, Recherche clinique, pharmacovigilance

Résumé : La dématérialisation des données de santé a permis depuis plusieurs années de constituer un véritable gisement de données provenant de tous les domaines de la santé.
Ces données ont pour caractéristiques d'être très hétérogènes et d'être produites à différentes échelles et dans différents domaines. Leur réutilisation dans le cadre de la recherche clinique, de la santé publique ou encore de la prise en charge des patients implique de développer des approches adaptées reposant sur les méthodes issues de la science des données. L'objectif de cette thèse est d'évaluer au travers de trois cas d'usage, quels sont les enjeux actuels ainsi que la place des data sciences pour l'exploitation des données massives en santé.

La démarche utilisée pour répondre à cet objectif consiste dans une première partie à exposer les caractéristiques des données massives en santé et les aspects techniques liés à leur réutilisation. La seconde partie expose les aspects organisationnels permettant l'exploitation et le partage des données massives en santé. La troisième partie décrit les grandes approches méthodologiques en science des données appliquées actuellement au domaine de la santé. Enfin, la quatrième partie illustre au travers de trois exemples l'apport de ces méthodes dans les champs suivant : la surveillance syndromique, la pharmacovigilance et la recherche clinique. Nous discutons enfin les limites et enjeux de la science des données dans le cadre de la réutilisation des données massives en santé.

Title : Issues and place of the data sciences for reusing clinical big data: a case-based study

Keywords : Data reuse, Health big data, Data sciences, Syndromic surveillance, Clinical research, Drug safety

Abstract : The dematerialization of health data, which started several years ago, now generates na huge amount of data produced by all actors of health.
These data have the characteristics of being very heterogeneous and of being produced at different scales and in different domains. Their reuse in the context of clinical research, public health or patient care involves developing appropriate approaches based on methods from data science. The aim of this thesis is to evaluate, through three use cases, what are the current issues as well as the place of data sciences regarding the reuse of massive health data.

To meet this objective, the first section exposes the characteristics of health big data and the technical aspects related to their reuse. The second section presents the organizational aspects for the exploitation and sharing of health big data. The third section describes the main methodological approaches in data sciences currently applied in the field of health. Finally, the fourth section illustrates, through three use cases, the contribution of these methods in the following fields: syndromic surveillance, pharmacovigilance and clinical research. Finally, we discuss the limits and challenges of data science in the context of health big data.