

Année 2017/2018

N°

## Thèse

Pour le

### DOCTORAT EN MEDECINE

Diplôme d'État

par

**Jeremy PASCO**

Né le 11/07/1989 à LUCERNE (SUISSE)

---

### **Apport des Entrepôts de Données Cliniques en Santé : du Rêve à la Réalité**

---

Présentée et soutenue publiquement le **1 octobre 2018** devant un jury composé de :

Président du Jury : Professeur Emmanuel RUSCH, Epidémiologie, Économie de la santé et prévention, AHU, Faculté de Médecine –Tours

Membres du Jury :

Professeur Marc CUGGIA, Informatique Médicale, Faculté de Médecine – Rennes 1

Professeur Frédéric PATAT, Biophysique et médecine nucléaire, Faculté de Médecine – Tours

Docteur Leslie GUILLON-GRAMMATICO, Epidémiologie, Économie de la santé et prévention, MCU-PH, Faculté de Médecine – Tours

Docteur Solène BRUNET-HOUDARD, Epidémiologie, Économie de la santé et prévention, AHU, Faculté de Médecine –Tours

**Directeur de thèse : Docteur Solène BRUNET-HOUDARD, Epidémiologie, Économie de la santé et prévention, AHU, Faculté de Médecine –Tours**

# Apport des Entrepôts de Données Cliniques en Santé : du Rêve à la Réalité

*Académie d'Orléans – Tours*  
**UNIVERSITE François RABELAIS**  
*Faculté de Médecine de Tours*  
*Denis Diderot est le nom qu'elle s'est donné.*

**Doctorat de Médecine – DES Santé Publique et Médecine Sociale**

**Auteur : PASCO Jeremy**  
**Directeur de thèse : BRUNET-HOUDARD Solène**

**Soutenance le 1<sup>er</sup> octobre 2018**



## Résumé

**Introduction** : La réutilisation des données cliniques a longtemps été restreinte par des difficultés d'accessibilité. Le développement d'entrepôts de données cliniques au sein des établissements de santé permet théoriquement de résoudre ce problème en centralisant l'ensemble des données cliniques au sein d'une même base de données. En amont du déploiement des Centres de Données Cliniques (CDC) eHOP au sein du Ri-CDC (réseau des CDC de l'inter-région Grand Ouest), nous avons souhaité étudier l'apport réel de ces entrepôts au travers d'une revue de la littérature.

**Matériel et Méthode** : Les articles ont été obtenus à l'aide du mot clé « data warehouse » sans restriction temporelle sur le moteur de recherche Pubmed. Seuls les articles faisant usage d'un entrepôt de données cliniques centrées sur le patient étaient retenus. Les éléments suivants étaient recueillis à l'aide d'une grille de lecture : date de publication, nom de l'entrepôt, localisation de l'entrepôt, type d'étude, temporalité de l'étude, discipline d'organe concernée, effectif extrait de l'entrepôt pour l'étude.

**Résultats** : Parmi les 1004 articles présents sur Pubmed, 618 ont été inclus dans notre revue de la littérature. Depuis le 1er article en 1994, on constate une progression constante du nombre de publications par année (83 articles en 2016). Néanmoins la majorité d'entre eux aborde les aspects méthodologiques de la mise en place d'un entrepôt de données. Seuls 250 articles (40%) consistent en une véritable étude exploitant un entrepôt. Parmi celles-ci on compte 226 (90%) articles d'épidémiologie, 12 (5%) études médico-économiques et 12 (5%) évaluations des pratiques. Le mode d'exploitation des entrepôts reste très largement rétrospectif (237 articles, 95%). Rares sont les études qui proposent une analyse des données en prospectif (9 articles, 4%) voire en temps réel (4 articles, 2%).

**Discussion** : A ce jour, la réutilisation des données cliniques au travers des entrepôts reste anecdotique dans la littérature médicale. L'importante quantité d'articles méthodologiques sur le sujet souligne les nombreuses difficultés techniques liées au déploiement et à l'exploitation de ces entrepôts. De nombreux efforts seront encore nécessaires pour intégrer l'exploitation des entrepôts dans le quotidien de la recherche biomédicale.

## Abstract

**Introduction**: Clinical data reuse has been limited for long due to access limitations. The development of data warehouses among health institutions would theoretically solve this issue though the centralization of all the clinical data inside a unique database. Before the deployment of the eHOP Clinical Data Centers (CDC) within the Ri-CDC (CDC's network of the western part of France), we wanted to assess the real contribution of these data warehouses through a literature review.

**Material and Methods**: The articles were obtained with the « data warehouse » keyword on the Pubmed database with no restriction on time. Articles using a clinical data warehouse patient-centered only were included. The following items were retrieved thanks to a reading grid: publication date, warehouse name, warehouse location, study type, temporal pattern of warehouse exploitation, concerned organ discipline and warehouse effective used.

**Results**: Among the 1004 articles found on Pubmed, 618 were included in our literature review. Since the first article in 1994, a constant progression of publications per year is noted (83 articles in 2016). However, most of them addresses methodological aspects about warehouse setting up. Only 250 articles (40%) consisted in studies that operate data warehouses. Among them, 226 (90%) articles were epidemiological studies, 12 (5%) were medico-economic studies and 12 (5%) were practice evaluations. The operating mode remains largely retrospective (237 articles, 95%). Few studies make use of prospective (9 articles, 4%) or real-time analysis (4 articles, 2%).

**Discussion**: Nowadays, clinical data reuse through data warehouses remains anecdotic in medical literature. The large amount of methodological articles on the subject highlights the many technical difficulties related to warehouse deployment and exploitation. Many efforts will still be needed to integrate warehouse operations into the daily life of biomedical research.

## Mots clés

Entrepôt de données, big data, réutilisation des données, revue de la littérature

## Keywords

Data warehouse, bug data, data reuse, literature review



UNIVERSITE DE TOURS  
**FACULTE DE MEDECINE DE TOURS**

**DOYEN**

**Pr Patrice DIOT**

**VICE-DOYEN**

Pr Henri MARRET

**ASSESEURS**

Pr Denis ANGOULVANT, *P dagogie*  
Pr Mathias BUCHLER, *Relations internationales*  
Pr Hubert LARDY, *Moyens – relations avec l'Universit *  
Pr Anne-Marie LEHR-DRYLEWICZ, *M decine g n rale*  
Pr Fran ois MAILLOT, *Formation M dicale Continue*  
Pr Patrick VOURC'H, *Recherche*

**RESPONSABLE ADMINISTRATIVE**

Mme Fanny BOBLETER

\*\*\*\*\*

**DOYENS HONORAIRES**

Pr Emile ARON (†) – 1962-1966  
*Directeur de l'Ecole de M decine - 1947-1962*  
Pr Georges DESBUQUOIS (†) – 1966-1972  
Pr Andr  GOUAZE – 1972-1994  
Pr Jean-Claude ROLLAND – 1994-2004  
Pr Dominique PERROTIN – 2004-2014

**PROFESSEURS EMERITES**

Pr Daniel ALISON  
Pr Philippe ARBEILLE  
Pr Catherine BARTHELEMY  
Pr Christian BONNARD  
Pr Philippe BOUGNOUX  
Pr Alain CHANTEPIE  
Pr Pierre COSNAY  
Pr Etienne DANQUECHIN-DORVAL  
Pr Lo c DE LA LANDE DE CALAN  
Pr Alain GOUDEAU  
Pr No l HUTEN  
Pr Olivier LE FLOCH  
Pr Yvon LEBRANCHU  
Pr Elisabeth LECA  
Pr Anne-Marie LEHR-DRYLEWICZ  
Pr G rard LORETTE  
Pr Roland QUENTIN  
Pr Alain ROBIER  
Pr Elie SALIBA

**PROFESSEURS HONORAIRES**

P. ANTHONIOZ – A. AUDURIER – A. AUTRET – P. BAGROS – P. BARDOS – J.L. BAULIEU – C. BERGER – JC. BESNARD – P. BEUTTER – P. BONNET – M. BROCHIER – P. BURDIN – L. CASTELLANI – B. CHARBONNIER – P. CHOUTET – T. CONSTANS – C. COUET – J.P. FAUCHIER – F. FETISSOF – J. FUSCIARDI – P. GAILLARD – G. GINIES – A. GOUAZE – J.L. GUILMOT – M. JAN – J.P. LAMAGNERE – F. LAMISSE – Y. LANSON – J. LAUGIER – P. LECOMTE – E. LEMARIE – G. LEROY – Y. LHUINTRE – M. MARCHAND – C. MAURAGE – C. MERCIER – J. MOLINE – C. MORAIN – J.P. MUH – J. MURAT – H. NIVET – L. POURCELOT – P. RAYNAUD – D. RICHARD-LENOBLE – J.C. ROLLAND – D. ROYERE – A. SAINDELLE – J.J. SANTINI – D. SAUVAGE – D. SIRINELLI – B. TOUMIEUX – J. WEILL

## PROFESSEURS DES UNIVERSITES - PRATICIENS HOSPITALIERS

ANDRES Christian.....	Biochimie et biologie moléculaire
ANGOULVANT Denis .....	Cardiologie
AUPART Michel.....	Chirurgie thoracique et cardiovasculaire
BABUTY Dominique .....	Cardiologie
BALLON Nicolas.....	Psychiatrie ; addictologie
BARILLOT Isabelle.....	Cancérologie ; radiothérapie
BARON Christophe .....	Immunologie
BEJAN-ANGOULVANT Théodora.....	Pharmacologie clinique
BERNARD Anne .....	Cardiologie
BERNARD Louis .....	Maladies infectieuses et maladies tropicales
BLANCHARD-LAUMONNIER Emmanuelle .....	Biologie cellulaire
BLASCO Hélène.....	Biochimie et biologie moléculaire
BODY Gilles .....	Gynécologie et obstétrique
BONNET-BRILHAULT Frédérique .....	Physiologie
BRILHAULT Jean.....	Chirurgie orthopédique et traumatologique
BRUNEREAU Laurent .....	Radiologie et imagerie médicale
BRUYERE Franck.....	Urologie
BUCHLER Matthias.....	Néphrologie
CALAIS Gilles.....	Cancérologie, radiothérapie
CAMUS Vincent.....	Psychiatrie d'adultes
CHANDENIER Jacques.....	Parasitologie, mycologie
COLOMBAT Philippe.....	Hématologie, transfusion
CORCIA Philippe.....	Neurologie
COTTIER Jean-Philippe .....	Radiologie et imagerie médicale
DE TOFFOL Bertrand .....	Neurologie
DEQUIN Pierre-François.....	Thérapeutique
DESOUBEAUX Guillaume.....	Parasitologie et mycologie
DESTRIEUX Christophe .....	Anatomie
DIOT Patrice.....	Pneumologie
DU BOUEXIC de PINIEUX Gonzague .....	Anatomie & cytologie pathologiques
DUCLUZEAU Pierre-Henri.....	Endocrinologie, diabétologie, et nutrition
DUMONT Pascal.....	Chirurgie thoracique et cardiovasculaire
EL HAGE Wissam.....	Psychiatrie adultes
EHRMANN Stephan .....	Réanimation
FAUCHIER Laurent .....	Cardiologie
FAVARD Luc.....	Chirurgie orthopédique et traumatologique
FOUGERE Bertrand .....	Gériatrie
FOUQUET Bernard.....	Médecine physique et de réadaptation
FRANCOIS Patrick.....	Neurochirurgie
FROMONT-HANKARD Gaëlle .....	Anatomie & cytologie pathologiques
GAUDY-GRAFFIN Catherine.....	Bactériologie-virologie, hygiène hospitalière
GOGA Dominique.....	Chirurgie maxillo-faciale et stomatologie
GOUPILLE Philippe .....	Rhumatologie
GRUEL Yves.....	Hématologie, transfusion
GUERIF Fabrice.....	Biologie et médecine du développement et de la reproduction
GUYETANT Serge .....	Anatomie et cytologie pathologiques
GYAN Emmanuel.....	Hématologie, transfusion
HAILLOT Olivier.....	Urologie
HALIMI Jean-Michel.....	Thérapeutique
HANKARD Régis.....	Pédiatrie
HERAULT Olivier .....	Hématologie, transfusion
HERBRETEAU Denis .....	Radiologie et imagerie médicale
HOURIOUX Christophe.....	Biologie cellulaire
LABARTHE François .....	Pédiatrie
LAFFON Marc .....	Anesthésiologie et réanimation chirurgicale, médecine d'urgence
LARDY Hubert.....	Chirurgie infantile
LARIBI Saïd.....	Médecine d'urgence
LARTIGUE Marie-Frédérique .....	Bactériologie-virologie
LAURE Boris.....	Chirurgie maxillo-faciale et stomatologie
LECOMTE Thierry.....	Gastroentérologie, hépatologie
LESCANNE Emmanuel.....	Oto-rhino-laryngologie
LINASSIER Claude .....	Cancérologie, radiothérapie
MACHET Laurent .....	Dermato-vénérologie
MAILLOT François .....	Médecine interne
MARCHAND-ADAM Sylvain .....	Pneumologie

MARRET Henri .....	Gynécologie-obstétrique
MARUANI Annabel .....	Dermatologie-vénérologie
MEREGHETTI Laurent.....	Bactériologie-virologie ; hygiène hospitalière
MORINIERE Sylvain.....	Oto-rhino-laryngologie
MOUSSATA Driffa .....	Gastro-entérologie
MULLEMAN Denis.....	Rhumatologie
ODENT Thierry.....	Chirurgie infantile
OUAISSI Mehdi .....	Chirurgie digestive
OULDAMER Lobna.....	Gynécologie-obstétrique
PAGES Jean-Christophe .....	Biochimie et biologie moléculaire
PAINTAUD Gilles .....	Pharmacologie fondamentale, pharmacologie clinique
PATAT Frédéric .....	Biophysique et médecine nucléaire
PERROTIN Dominique.....	Réanimation médicale, médecine d'urgence
PERROTIN Franck .....	Gynécologie-obstétrique
PISELLA Pierre-Jean.....	Ophtalmologie
PLANTIER Laurent.....	Physiologie
QUENTIN Roland.....	Bactériologie-virologie, hygiène hospitalière
REMERAND Francis .....	Anesthésiologie et réanimation, médecine d'urgence
ROINGEARD Philippe.....	Biologie cellulaire
ROSSET Philippe.....	Chirurgie orthopédique et traumatologique
RUSCH Emmanuel.....	Epidémiologie, économie de la santé et prévention
SAINT-MARTIN Pauline.....	Médecine légale et droit de la santé
SALAME Ephrem.....	Chirurgie digestive
SAMIMI Mahtab .....	Dermatologie-vénérologie
SANTIAGO-RIBEIRO Maria .....	Biophysique et médecine nucléaire
THOMAS-CASTELNAU Pierre .....	Pédiatrie
TOUTAIN Annick.....	Génétique
VAILLANT Loïc.....	Dermato-vénérologie
VELUT Stéphane.....	Anatomie
VOURC'H Patrick.....	Biochimie et biologie moléculaire
WATIER Hervé .....	Immunologie

## **PROFESSEUR DES UNIVERSITES DE MEDECINE GENERALE**

---

LEBEAU Jean-Pierre

## **PROFESSEURS ASSOCIES**

---

MALLET Donatien .....	Soins palliatifs
POTIER Alain .....	Médecine Générale
ROBERT Jean.....	Médecine Générale

## **MAITRES DE CONFERENCES DES UNIVERSITES - PRATICIENS HOSPITALIERS**

---

BAKHOS David.....	Physiologie
BARBIER Louise.....	Chirurgie digestive
BERHOUEZ Julien .....	Chirurgie orthopédique et traumatologique
BERTRAND Philippe.....	Biostat., informatique médical et technologies de communication
BRUNAUT Paul .....	Psychiatrie d'adultes, addictologie
CAILLE Agnès .....	Biostat., informatique médical et technologies de communication
CLEMENTY Nicolas .....	Cardiologie
DOMELIER Anne-Sophie .....	Bactériologie-virologie, hygiène hospitalière
DUFOUR Diane .....	Biophysique et médecine nucléaire
FAVRAIS Géraldine .....	Pédiatrie
FOUQUET-BERGEMER Anne-Marie.....	Anatomie et cytologie pathologiques
GATAULT Philippe.....	Néphrologie
GOUILLEUX Valérie.....	Immunologie
GUILLOIN Antoine.....	Réanimation
GUILLOIN-GRAMMATICO Leslie.....	Epidémiologie, économie de la santé et prévention
HOARAU Cyrille .....	Immunologie
IVANES Fabrice .....	Physiologie
LE GUELLEC Chantal.....	Pharmacologie fondamentale, pharmacologie clinique
MACHET Marie-Christine .....	Anatomie et cytologie pathologiques
MOREL Baptiste .....	Radiologie pédiatrique



PIVER Éric.....Biochimie et biologie moléculaire  
REROLLE Camille.....Médecine légale  
ROUMY Jérôme.....Biophysique et médecine nucléaire  
SAUTENET Bénédicte .....Néphrologie  
TERNANT David.....Pharmacologie fondamentale, pharmacologie clinique  
ZEMMOURA Ilyess .....Neurochirurgie

## **MAITRES DE CONFERENCES DES UNIVERSITES**

---

AGUILLON-HERNANDEZ Nadia.....Neurosciences  
BOREL Stéphanie.....Orthophonie  
DIBAO-DINA Clarisse .....Médecine Générale  
MONJAUZE Cécile .....Sciences du langage - orthophonie  
PATIENT Romuald.....Biologie cellulaire  
RENOUX-JACQUET Cécile .....Médecine Générale

## **CHERCHEURS INSERM - CNRS - INRA**

---

BOUAKAZ Ayache .....Directeur de Recherche INSERM – UMR INSERM 1253  
CHALON Sylvie .....Directeur de Recherche INSERM – UMR INSERM 1253  
COURTY Yves .....Chargé de Recherche CNRS – UMR INSERM 1100  
DE ROCQUIGNY Hugues .....Chargé de Recherche INSERM – UMR INSERM 1259  
ESCOFFRE Jean-Michel.....Chargé de Recherche INSERM – UMR INSERM 1253  
GILOT Philippe.....Chargé de Recherche INRA – UMR INRA 1282  
GOUILLEUX Fabrice .....Directeur de Recherche CNRS – UMR CNRS 7001  
GOMOT Marie.....Chargée de Recherche INSERM – UMR INSERM 1253  
HEUZE-VOURCH Nathalie.....Chargée de Recherche INSERM – UMR INSERM 1100  
KORKMAZ Brice.....Chargé de Recherche INSERM – UMR INSERM 1100  
LAUMONNIER Frédéric .....Chargé de Recherche INSERM - UMR INSERM 1253  
LE PAPE Alain.....Directeur de Recherche CNRS – UMR INSERM 1100  
MAZURIER Frédéric.....Directeur de Recherche INSERM – UMR CNRS 7001  
MEUNIER Jean-Christophe .....Chargé de Recherche INSERM – UMR INSERM 1259  
PAGET Christophe.....Chargé de Recherche INSERM – UMR INSERM 1100  
RAOUL William .....Chargé de Recherche INSERM – UMR CNRS 7001  
SI TAHAR Mustapha.....Directeur de Recherche INSERM – UMR INSERM 1100  
WARDAK Claire.....Chargée de Recherche INSERM – UMR INSERM 1253

## **CHARGES D'ENSEIGNEMENT**

---

### ***Pour l'Ecole d'Orthophonie***

DELORE Claire .....Orthophoniste  
GOUIN Jean-Marie.....Praticien Hospitalier  
PERRIER Danièle .....Orthophoniste

### ***Pour l'Ecole d'Orthoptie***

LALA Emmanuelle.....Praticien Hospitalier  
MAJZOUB Samuel.....Praticien Hospitalier

### ***Pour l'Ethique Médicale***

BIRMELE Béatrice.....Praticien Hospitalier

# SERMENT D'HIPPOCRATE

En présence des Maîtres de cette Faculté,  
de mes chers condisciples  
et selon la tradition d'Hippocrate,  
je promets et je jure d'être fidèle aux lois de l'honneur  
et de la probité dans l'exercice de la Médecine.

Je donnerai mes soins gratuits à l'indigent,  
et n'exigerai jamais un salaire au-dessus de mon travail.

Admis dans l'intérieur des maisons, mes yeux  
ne verront pas ce qui s'y passe, ma langue taira  
les secrets qui me seront confiés et mon état ne servira pas  
à corrompre les mœurs ni à favoriser le crime.

Respectueux et reconnaissant envers mes Maîtres,  
je rendrai à leurs enfants  
l'instruction que j'ai reçue de leurs pères.

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



# Remerciements

## A ma famille

A Nathalie, merci pour ton écoute et ta patience infinies, merci pour cet incroyable quotidien partagé. Et merci de tout ce que tu fais pour notre petit Alexandre.

A Alexandre, petit mignon de tous les jours qui me motive à travailler efficacement pour vite rentrer te retrouver le soir. Merci pour tes beaux sourires et ces 9 mois merveilleux en ta compagnie.

A mes parents, merci pour votre soutien durant toutes ces longues années d'études. Ça se termine !

A Céline et Amaury, merci pour ces quelques jours de repos en votre compagnie avant le grand jour, on a enfin découvert le soleil de Marseille !

A Chloé, plein de courage pour cette année difficile #tatabestacestlameilleure

A Caroline et Jean-Marc, Chloé et Sébastien, Stéphanie et Charles, Alfred et Constance, vous êtes une famille au top, c'est toujours un plaisir de vous retrouver.

A ma famille, merci d'être venus si nombreux me soutenir en ce jour important.

## A mes amis

A Kanav et Claire, merci pour votre énergie infaillible, ces soirées jeux pleines de rebondissements et ces beaux articles en soumission.

A Charly, merci pour ces belles aventures partagées pendant l'externat. Reviens nous vite ! #sangliercorse #lezard

Aux Tourangeaux, merci pour toutes ces soirées, et pour ce beau projet de constitution d'équipe de foot junior. #onvarepeuplerlatourraine #finilesdesertsmedicaux

A Romain et Angéline, merci pour ces années folles d'Elus UFR #revolution #courgette #coeurcoeurpaillette

Au Tutorat de Tours (Poupou, Adèle, Patate et Marie), merci pour ces premiers pas associatifs en votre compagnie, et vive le Tut ! C'est avec vous que j'ai réalisé mes débuts en informatique médicale.

A l'AIT, parce que vider un Leclerc drive de tout son stock de Canadou ça n'a pas de prix, merci. #pizzananas

Au Dissident Crew, merci pour les paillettes !

A Benjamin, merci pour ces échanges riches d'apprentissage tout au long de nos rencontres CLISP. J'espère que nous retrouverons des projets communs malgré la distance.

### **A mes collègues et jury**

A l'équipe du CIC-IT, merci de m'avoir accueilli dès mon externat et de m'avoir ouvert si tôt la voie de la recherche clinique et du traitement du signal. Je vous dois ma reconversion vers la Santé Publique :) Merci à Frédéric Patat d'accepter d'évaluer mon travail.

A l'équipe du CIC-P, merci pour vos conseils et votre bonne humeur. PS : JFK reviendra chasser le schtroumpf.

A l'équipe Health Big Data de Rennes, merci pour votre accueil chaleureux et votre disponibilité infallible durant mon trop court séjour parmi vous. A très vite pour de nouvelles aventures. #reuniontowerfall. Merci à Marc Cuggia de m'avoir soumis cette idée de thèse et d'accepter d'en évaluer le travail.

Leslie Guillon et Emmanuel Rusch, merci de m'avoir impliqué dans ce beau projet qu'est le Centre de Données Cliniques, et d'accepter d'évaluer mon travail en ce jour.

Au SIMEES, merci pour votre convivialité et la place que vous m'offrez parmi vous.

A mes co-internes, merci pour vos horizons si différents qui font la richesse de nos échanges au quotidien.

A Solène, merci d'avoir accepté de m'encadrer sur un travail aussi vaste pour lequel un regard extérieur était toujours le bienvenu.

## Table des matières

Introduction.....	15
Matériel et Méthodes .....	17
Stratégie de recherche.....	17
Critères d'inclusion et d'exclusion.....	17
Données recueillies.....	18
Consultation des articles .....	18
Outils.....	19
Résultats.....	21
Articles sélectionnés.....	21
Figure 1 – Flow Chart.....	21
Figure 2 – Motif d'exclusion des articles (n=386).....	22
Evolution des publications dans le temps .....	23
Figure 3 – Evolution des publications par année (n=618).....	23
Répartition géographique et par EDC des publications.....	24
Figure 4 – Cartographie des publications.....	25
Figure 5 – Entrepôts les plus représentés dans la littérature mondiale (origine – 06/2017).....	26
Sujets d'étude et disciplines médicales .....	27
Figure 6 – Disciplines concernées par l'exploitation des EDC .....	27
Types d'études et schéma d'exploitation des données .....	28
Figure 7 – Typologie des articles retrouvés dans la revue de la littérature.....	29
Effectifs .....	30
Figure 8 – Evolution des quantités de données structurées exploitées par étude.....	30
Discussion.....	31
Une exploitation des EDC encore anecdotique.....	31
Des cas d'usage peu variés.....	31
Une exploitation en avance aux Etats Unis .....	31
Des difficultés liées à l'interopérabilité .....	32
Limites de notre étude.....	32
Conclusion .....	33
Bibliographie.....	35
Matériel supplémentaire .....	39



## Introduction

La rencontre du système de santé par les patients entraîne une collecte d'informations (administratives, cliniques, biologiques...) nécessaire à leur prise en charge par les professionnels de santé. Ces informations, à l'origine stockées dans des dossiers médicaux papiers sont désormais de plus en plus renseignées dans des bases de données informatisées.

La réutilisation de ces données, à des fins de recherche par exemple, a longtemps été restreinte par des difficultés d'accessibilité. En effet, avant l'adoption des dossiers médicaux informatisés (EHR : Electronic Health Records), les initiatives visant à réutiliser les données cliniques nécessitaient de renseigner manuellement les informations notées sur papier issues du dossier médical du patient au sein d'une base de données dédiée au projet de recherche (Ex: ARAMIS (1)).

Les premières données cliniques informatisées ont été les codes d'actes et de diagnostics dans le cadre du remboursement des soins, notamment aux Etats-Unis dès 1983. Puis le concept fut importé et mis en application en France à partir de 1996, au sein du Programme de Médicalisation des Systèmes d'Information.

En 2009 aux Etats Unis, la "Health Information Technology for Economic and Clinical Health Act" (HITECH Act) a permis l'adoption des EHR dans les établissements de santé, grâce à des incitations financières pour les premières adoptions. Les établissements retardataires dans l'adoption des EHR ont, eux, été soumis à des pénalités (2). En France, cette intégration a été promue par le programme « Hopital numérique » de la Direction Générale de l'Offre de Soins pour la période 2012-2017(3). Ces initiatives ont permis un fort enrichissement des EHR, tant en volumétrie qu'en variété des données. En conséquence sont attendus les challenges propres au domaine des Big Data tel que défini par Gartner et al (4) : « high-volume, -velocity, and -variety information assets ».

L'une des problématiques majeures de la constitution d'un entrepôt de données est notamment la mise en commun des différentes sources de données cliniques (biologie, imagerie, texte libre, etc) au sein d'un établissement de santé, puis entre établissements. Afin de pallier aux problèmes d'interopérabilité des différents logiciels métiers, des établissements de santé ont déployé des entrepôts de données cliniques (EDC), tels que définis par Wade (5) : « un dépôt qui ajoute des niveaux d'intégration et de qualité aux données primaires (de recherche ou cliniques) d'une seule institution, afin de prendre en charge des requêtes flexibles à des fins multiples, et dont le domaine d'application est plus large que le registre ». A ce titre sont exclues les bases de données médico-administratives telles que les composantes du Système National des Données de Santé français (6). Ces bases de données n'ont en effet pas pour vocation première la clinique ou la recherche. De plus, faisant l'objet d'une constitution homogène au niveau national, elles sont exemptes des problématiques d'interopérabilité des EDC.



Plusieurs auteurs ont prédit les apports théoriques du déploiement des entrepôts de données, et plus généralement du « data reuse ». Chatellier et al (7) ont identifié les apports théoriques suivants : identifier les patients pour des cohortes, pré-screening ou screening pour des essais cliniques, études cas-témoins, comparaisons de populations, jointures avec des biobanques...

Au sein de l'interrégion Grand Ouest, les CHU et l'Institut de Cancérologie de l'Ouest déploient le CDRN RiCDC (Réseau interrégional des Centres de Données Cliniques) s'appuyant sur des entrepôts locaux nommés eHOP. Ce réseau sera effectif lorsque tous les établissements pourront déverser dans leur entrepôt les données de leurs différentes applications métiers.

L'objectif de cette étude était de réaliser un état des lieux de l'utilisation réelle des EDC au niveau mondial, depuis leurs origines, à partir d'une revue de la littérature.

## **Matériel et Méthodes**

### **Stratégie de recherche**

Afin de limiter notre recherche aux entrepôts de données cliniques, la revue de littérature a été réalisée à partir de la base Pubmed. La requête a été effectuée le 1er juin 2017 avec le terme “data warehouse” seul pour une sensibilité maximale, et sans restriction de temps.

### **Critères d’inclusion et d’exclusion**

Pour être inclus dans notre revue de la littérature, les articles devaient porter sur une étude réalisée à partir de données issues d’un EDC ou aborder des aspects méthodologiques relatifs au développement, au déploiement ou à l’exploitation des EDC tels que définis par Wade (5). Les EDC devaient être centrés sur le patient. Ces entrepôts pouvaient être mono ou multicentriques, appartenant ou non à un établissement de santé.

Les bases de données conçues dans le cadre d’une unique étude étaient exclues. Les EDC dont le contenu n’était pas conforme, même s’ils intégraient des données cliniques (ex : entrepôts centrés sur les molécules, les protéines, les bactéries, etc), les bases de données exclusivement médico-administratives et les EDC cités mais non impliqués dans l’étude, étaient exclus. Les EDC virtuels, qui ne centralisent pas physiquement les données en une même base, et les articles usant du terme « data warehouse » pour décrire un outil de visualisation de données ont été exclus. Enfin, les cohortes et datamarts nommés « data warehouse », ainsi que les publications d’erratum, de doublons et les annonces d’intentions d’études non encore réalisées étaient exclus.

## **Données recueillies**

Parmi les articles retenus, nous avons relevé les informations suivantes : date de la publication, nom de l'EDC (lorsque l'entrepôt en possédait un), localisation géographique de l'EDC, sujet(s) de l'étude, discipline(s) médicales ou chirurgicales de l'étude, type d'étude, schéma temporel d'exploitation de l'entrepôt (rétrospectif, prospectif ou temps réel), effectifs et unités des données extraites de l'EDC (séjours, patients, examens, prescriptions, analyses, etc).

La classification suivante du type d'étude a été arrêtée après lecture de 100 articles sélectionnés de façon aléatoire :

- 1) Article méthodologique
  - a) Conception d'EDC
  - b) Déploiement d'EDC
  - c) Développement d'outils pour les EDC
  - d) Suivi de l'évolution d'EDC
  - e) Autres
- 2) Etude exploitant un EDC
  - a) Etude épidémiologique
    - i) Descriptive
    - ii) Analytique
    - iii) Evaluative
  - b) Etude médico-économique
  - c) Evaluation des pratiques

L'extraction des données pour chaque article a été réalisée par le même auteur.

## **Consultation des articles**

La vérification des critères d'inclusion et la classification du type d'étude étaient réalisées dans un premier temps à la lecture du titre et du résumé de chaque article.

Pour les études exploitant des données d'EDC, le texte intégral était systématiquement consulté pour confirmer le respect des critères d'inclusion et d'exclusion, notamment vis-à-vis de la nature de l'EDC, ainsi que pour compléter le recueil de données. Lorsque la lecture du texte intégral ne permettait pas de connaître avec précision la nature de l'EDC, des informations supplémentaires sur celui-ci étaient récupérées par recherche sur Google ainsi que via la lecture des autres articles mentionnant le même EDC. En dernier lieu, si la nature de l'EDC ne pouvait être précisée, l'article était exclu.

La lecture de chaque article a été réalisée par le même auteur.

## **Outils**

La classification des articles a été réalisée au travers d'un outil web développé pour l'étude, implémentant la grille de lecture et permettant la mise en surbrillance de certains éléments clés à l'aide d'expressions régulières.

Les analyses et graphiques ont été réalisés à l'aide de R version 3.4.4 avec les paquets ggplot2, maps et Cairo.

Le diagramme Sunburst a été réalisé à l'aide de la bibliothèque D3.js (8) et de Photoshop®.



## Résultats

### Articles sélectionnés

La requête initiale via le moteur de recherche Pubmed a relevé 1004 articles. Après lecture de leurs résumés, 300 articles, ne répondant pas aux critères d'inclusion et d'exclusion, ont été exclus. A la consultation des textes intégraux, 86 autres articles ont été exclus (Figure 1).

Les causes d'exclusion sont décrites en Figure 2. Au final, 618 articles ont été inclus dans cette revue de littérature. Tous les articles faisant mention d'étude exploitant un EDC ont fait l'objet d'une consultation du texte intégral.

Figure 1 – Flow Chart

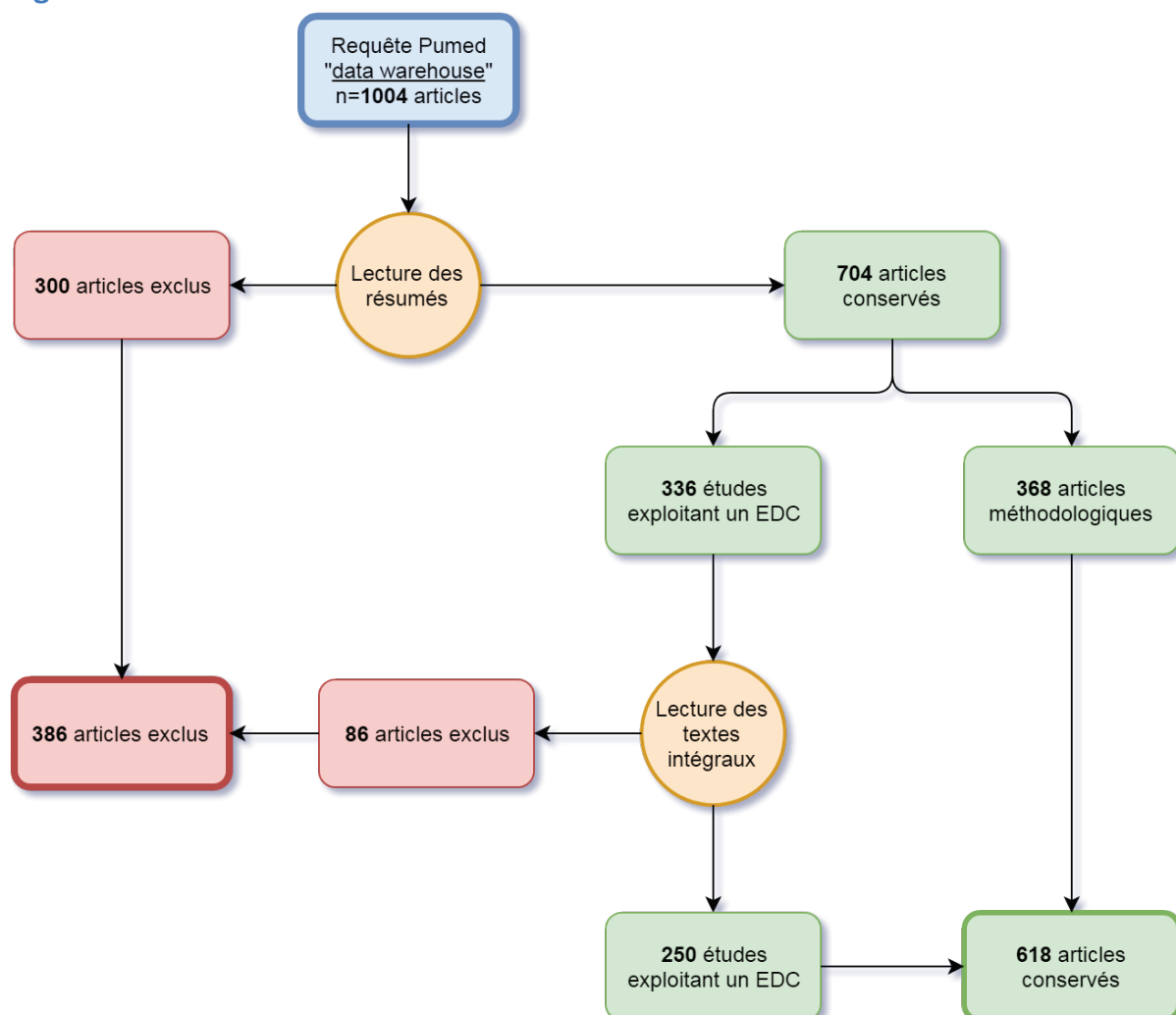
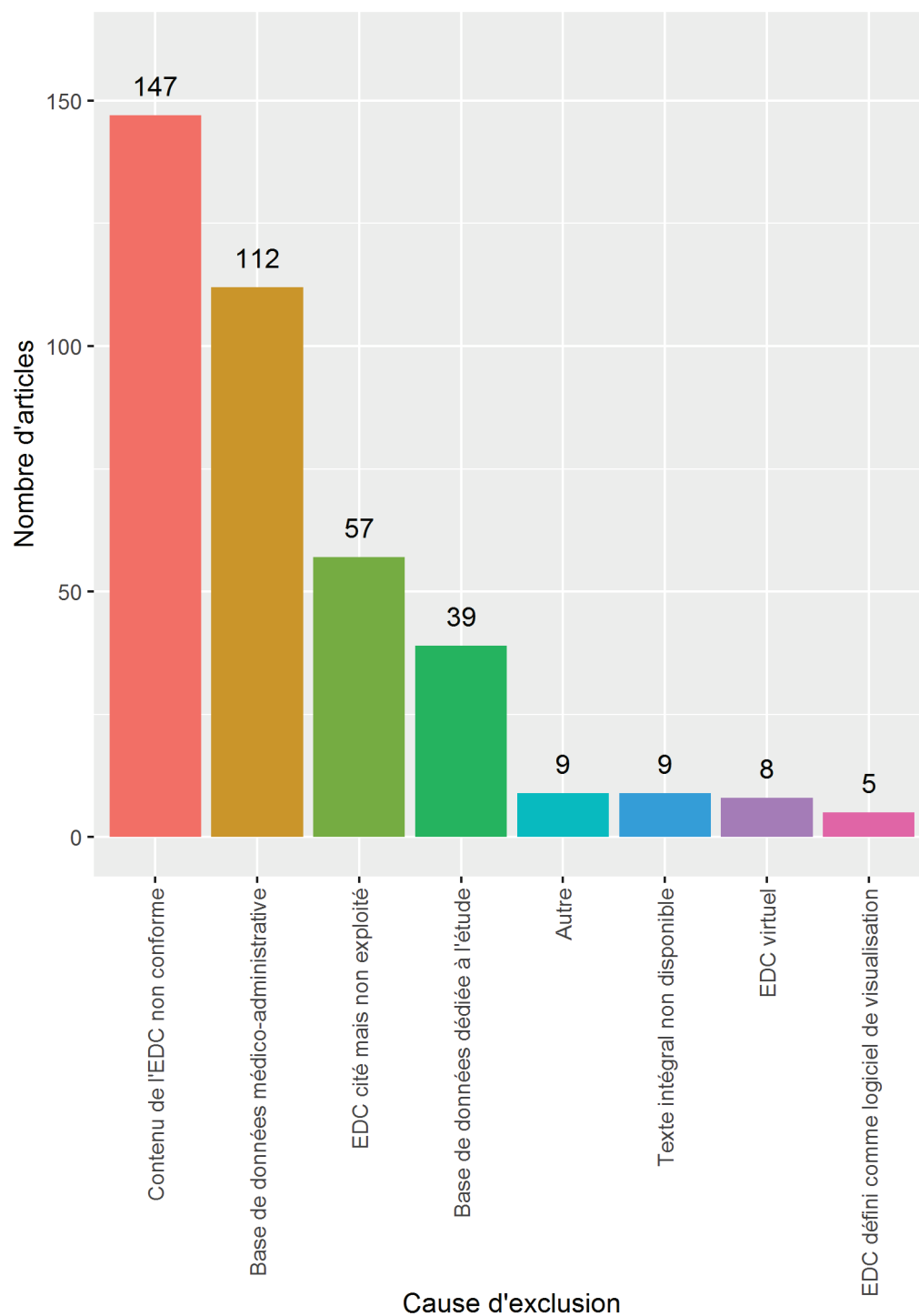


Figure 2 – Motif d'exclusion des articles (n=386)



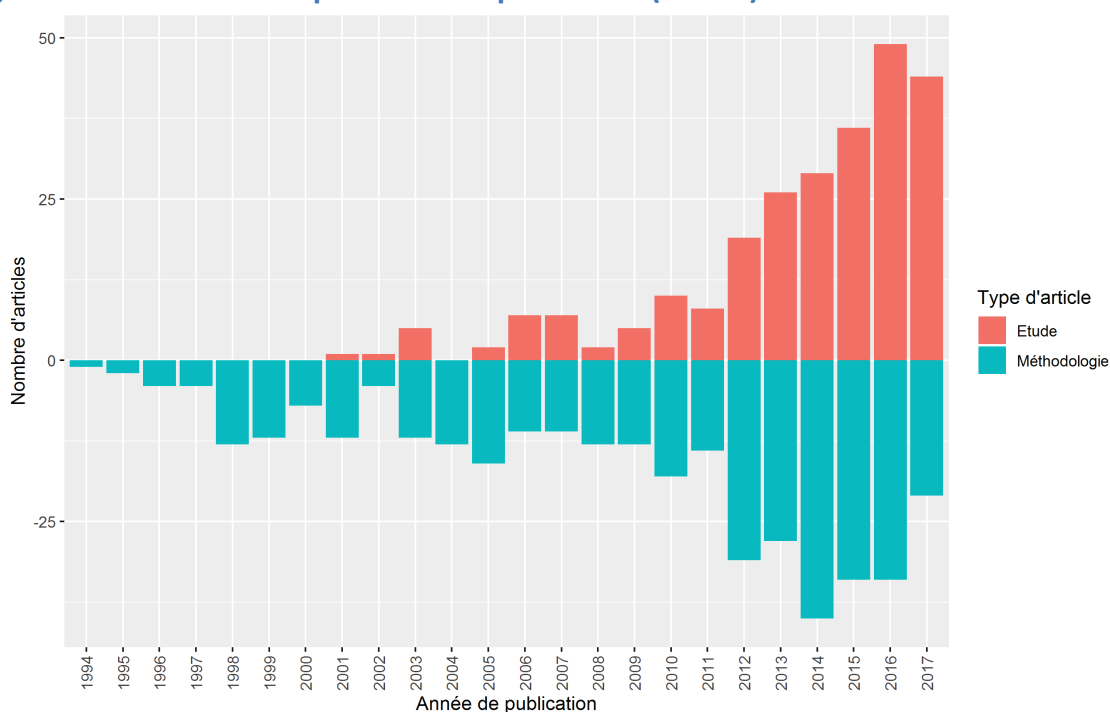
### Evolution des publications dans le temps

La première mention de « data warehouse » dans la littérature médicale est retrouvée en 1994 dans un article méthodologique (9) vantant l'intérêt futur des EDC. La première étude exploitant un EDC date de 2001 (10) : l'entrepôt utilisé comprenait alors pour chaque chirurgie la liste des complications, des temps chirurgicaux et du personnel présent en plus des codes d'actes et de diagnostics.

Le nombre de publications d'articles méthodologiques et d'études exploitant des EDC est en croissance constante depuis 2011, avec jusqu'à 83 articles en 2016 (34 exploitations d'EDC et 49 articles méthodologiques). La baisse de publications en 2017 s'explique par un recueil réalisé au cours de la même année.

L'évolution des publications est représentée en Figure 3.

Figure 3 – Evolution des publications par année (n=618)





### **Répartition géographique et par EDC des publications**

Les EDC relevés étaient répartis dans 15 pays distincts. 58 d'entre eux possédaient une dénomination précise. Leur exploitation était majoritairement réalisée par les Etats Unis avec 188 études (75%).

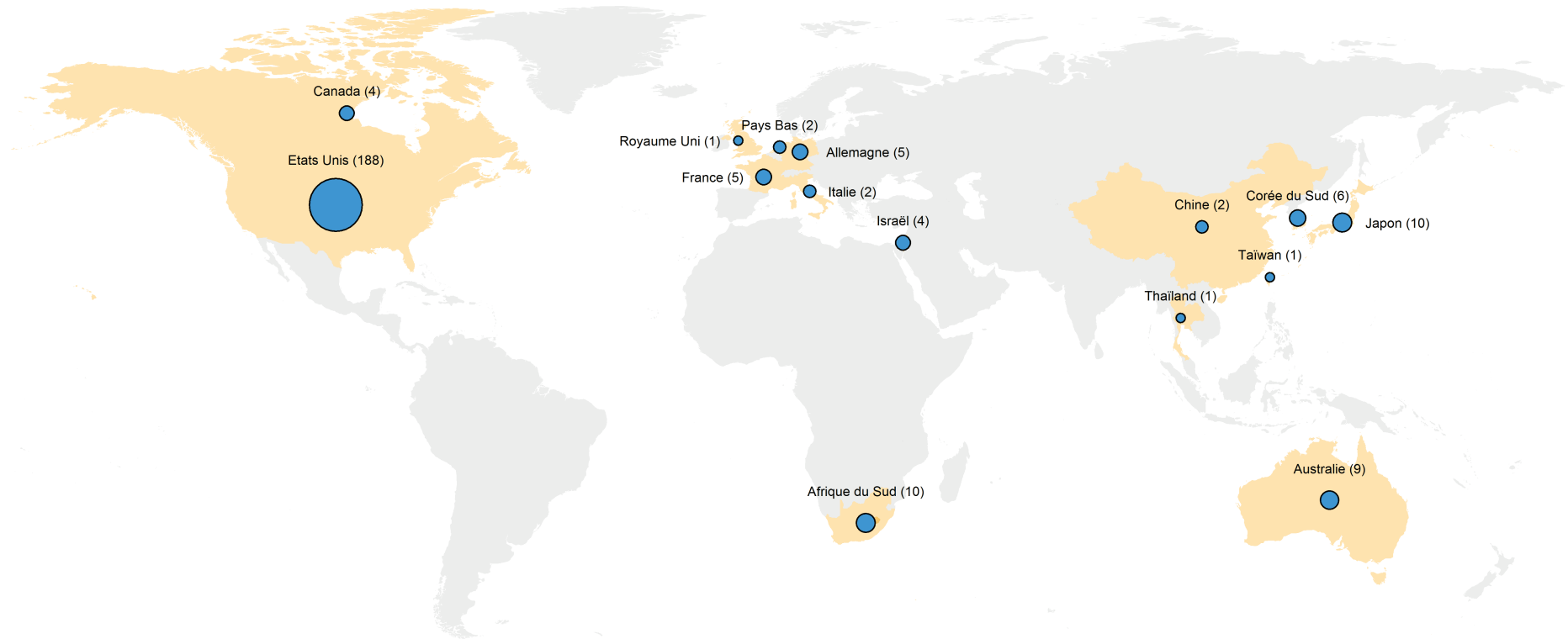
Sur l'ensemble des publications sélectionnées, les 3 EDC les plus fréquemment exploités étaient américains : le Veterans Health Administration Corporate Data Warehouse (57 études, 23%), le Pediatric BabySteps Clinical Data Warehouse (16 études, 6%) et l'Intermountain Healthcare Enterprise Data Warehouse (15 études, 6%).

La France était en 6<sup>e</sup> position en termes d'exploitation d'EDC avec 5 études publiées (2%) dont 4 réalisées sur l'EDC I2b2 de l'Hôpital Européen Georges-Pompidou, Paris et 1 sur l'EDC eHOP du Centre Hospitalier Universitaire de Rennes.

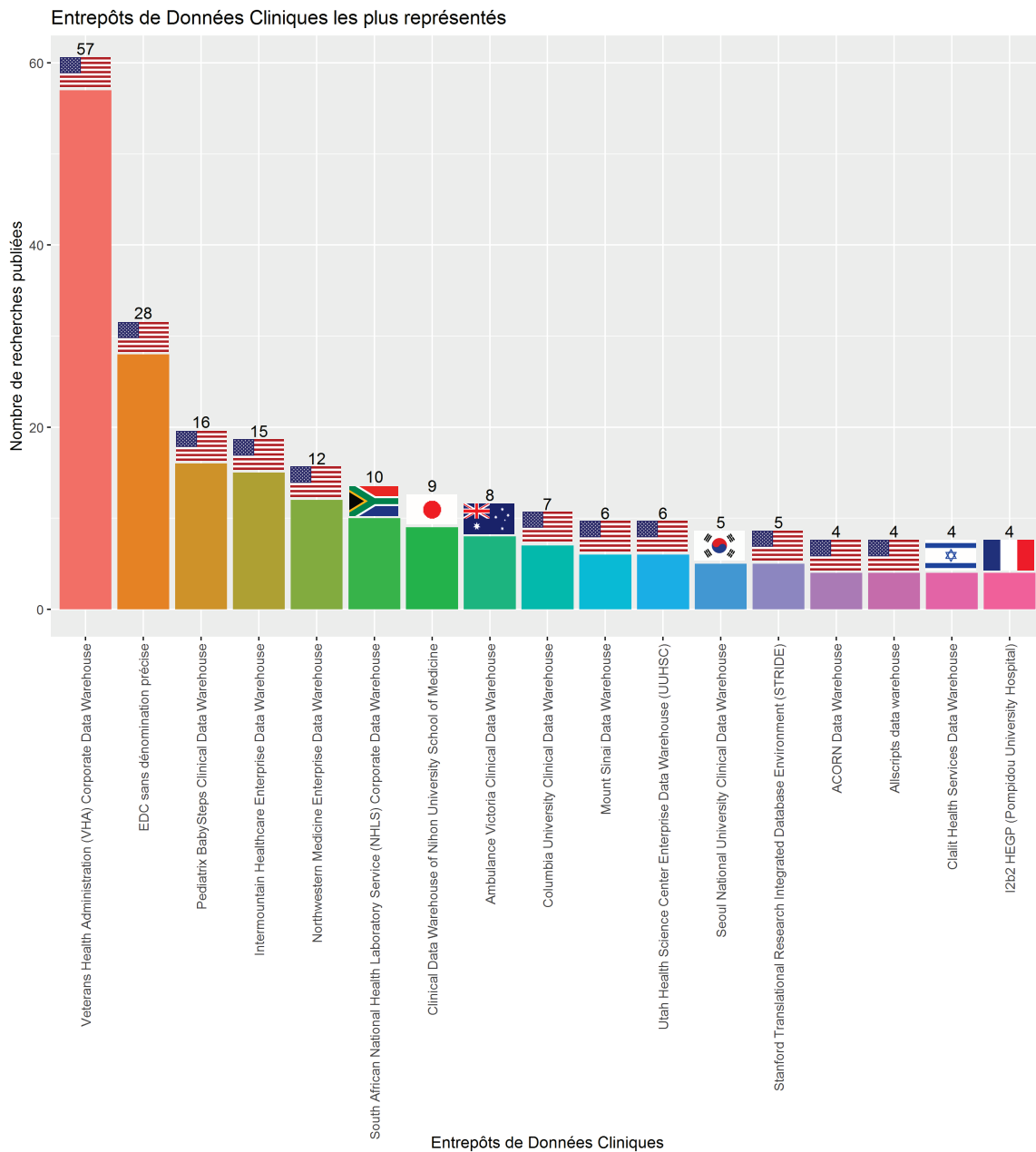
La répartition géographique détaillée est disponible sur la Figure 4, ainsi que le classement des EDC les plus exploités (au moins 4 études publiées) en Figure 5.

Figure 4 – Cartographie des publications

Nombre de publications d'études exploitant des EDC par pays (origine – 06/2017).



**Figure 5 – Entrepôts les plus représentés dans la littérature mondiale (origine – 06/2017)**

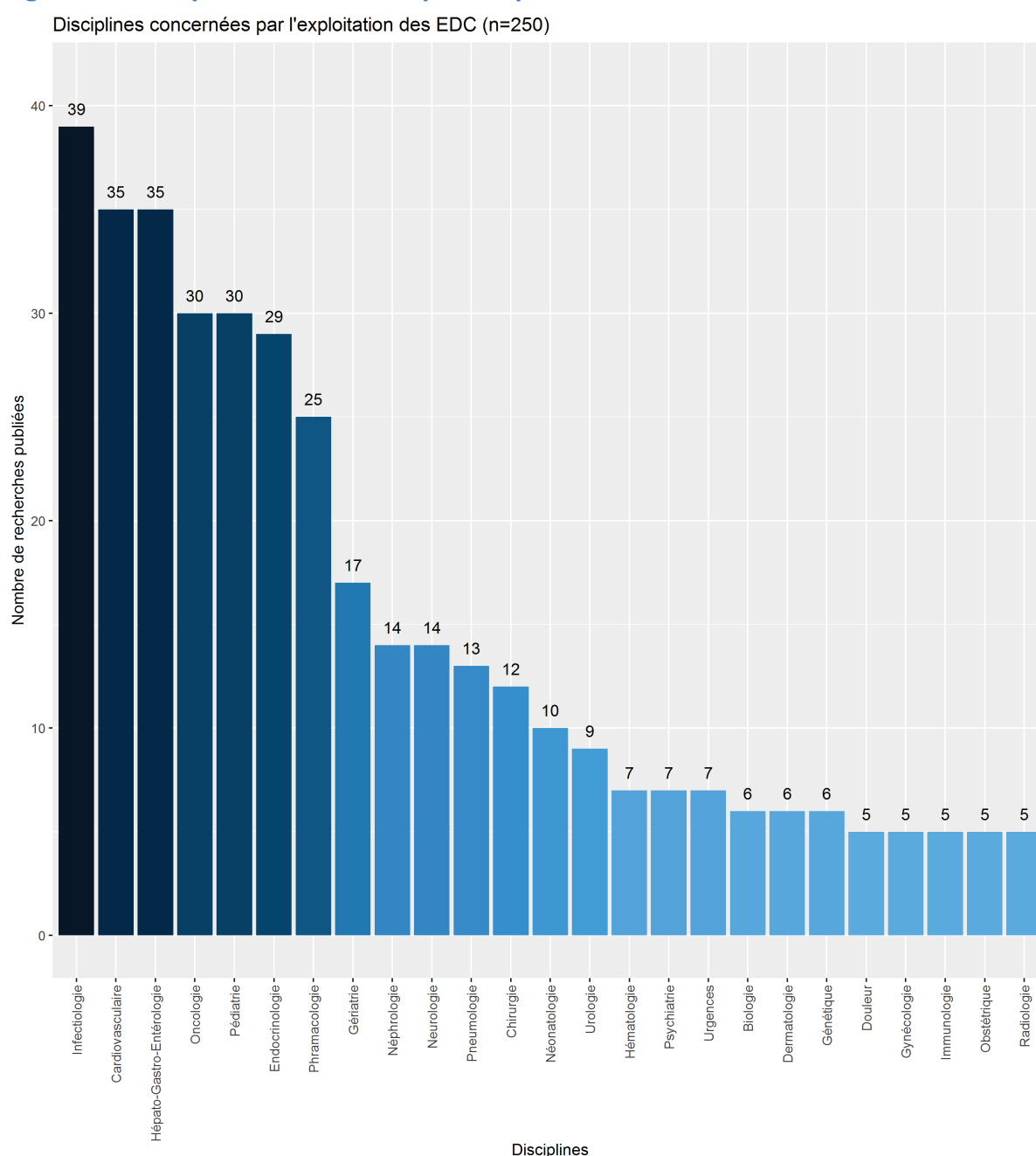


## Sujets d'étude et disciplines médicales

Les 3 disciplines les plus représentées dans l'exploitation des EDC sont l'infectiologie (39 articles, 16%), le cardiovasculaire (35 articles, 14%) et l'hépto-gastro-entérologie (35 articles, 14%). Les disciplines ayant au moins 4 publications sont représentées en Figure 6.

Peu de thématiques récurrentes se démarquaient parmi les sujets d'étude. On retrouve en tête le diabète (18 articles, 7%), le VIH (12 articles, 5%), les parcours de soin (8 articles, 3%) et l'hypertension artérielle (8 articles, 3%)

Figure 6 – Disciplines concernées par l'exploitation des EDC



### **Types d'études et schéma d'exploitation des données**

Parmi les 618 articles inclus, 250 articles (40%) relatent une étude exploitant un EDC et 368 articles (60%) traitent des aspects méthodologiques autour de la conception et du déploiement des EDC.

Les articles méthodologiques traitent à 63% (230 articles) de points techniques de conception, à 24% (90 articles) du déploiement des EDC, à 10% (38 articles) de développement d'outils pour les EDC et à 2% (8 articles) du suivi de l'évolution d'EDC. Deux autres articles (1%) ont regroupés dans la catégorie « Autres », l'un consistait en une revue de la littérature (11), l'autre abordait l'usage des EDC lors de la phase 2 du développement de médicaments (12).

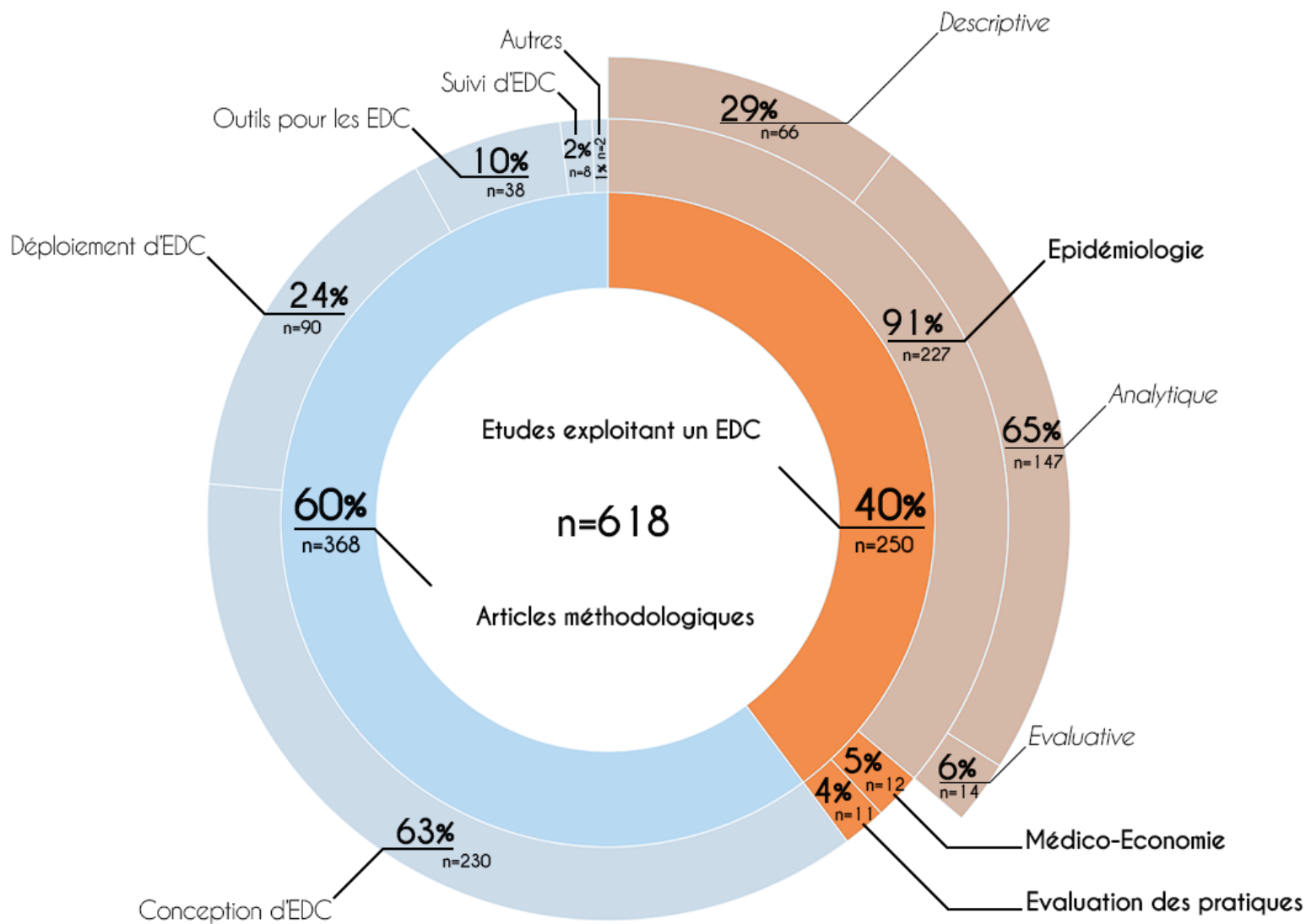
Parmi les 250 études exploitant un entrepôt on compte 227 études épidémiologiques (91%), 12 études médico-économiques (5%) et 11 évaluations des pratiques (4%). Les études épidémiologiques étaient à 29% descriptives (66 articles), 65% analytiques (147 articles) et 6% évaluatives (14 articles).

Le mode d'exploitation des entrepôts était quasi exclusivement rétrospectif (237 articles, 95%), seuls 9 articles (4%) proposaient une étude prospective et 4 articles une étude en temps réel (2%).

Les études prospectives et temps réel usaient de différentes approches : des données déjà présentes dans l'EDC pouvaient être extraites en fin d'étude (13,14) ou de façon itérative tout au long de l'étude (15-18). Des données spécifiques à l'étude pouvaient être renseignées dans les observations médicales afin de les récupérer en fin d'étude via une extraction de l'EDC (19), permettant ainsi de sursoir à l'usage d'un eCRF (electronic case report form). Le développement de modèles prédictifs pouvait être réalisé sur données rétrospectives de l'EDC puis mis en pratique au fil de l'eau (20-24). Enfin, la recherche du statut vital à distance d'une étude de survie pouvait faire l'objet de requêtes itératives sur un EDC (25).

La répartition complète des types d'articles est décrite en Figure 7.

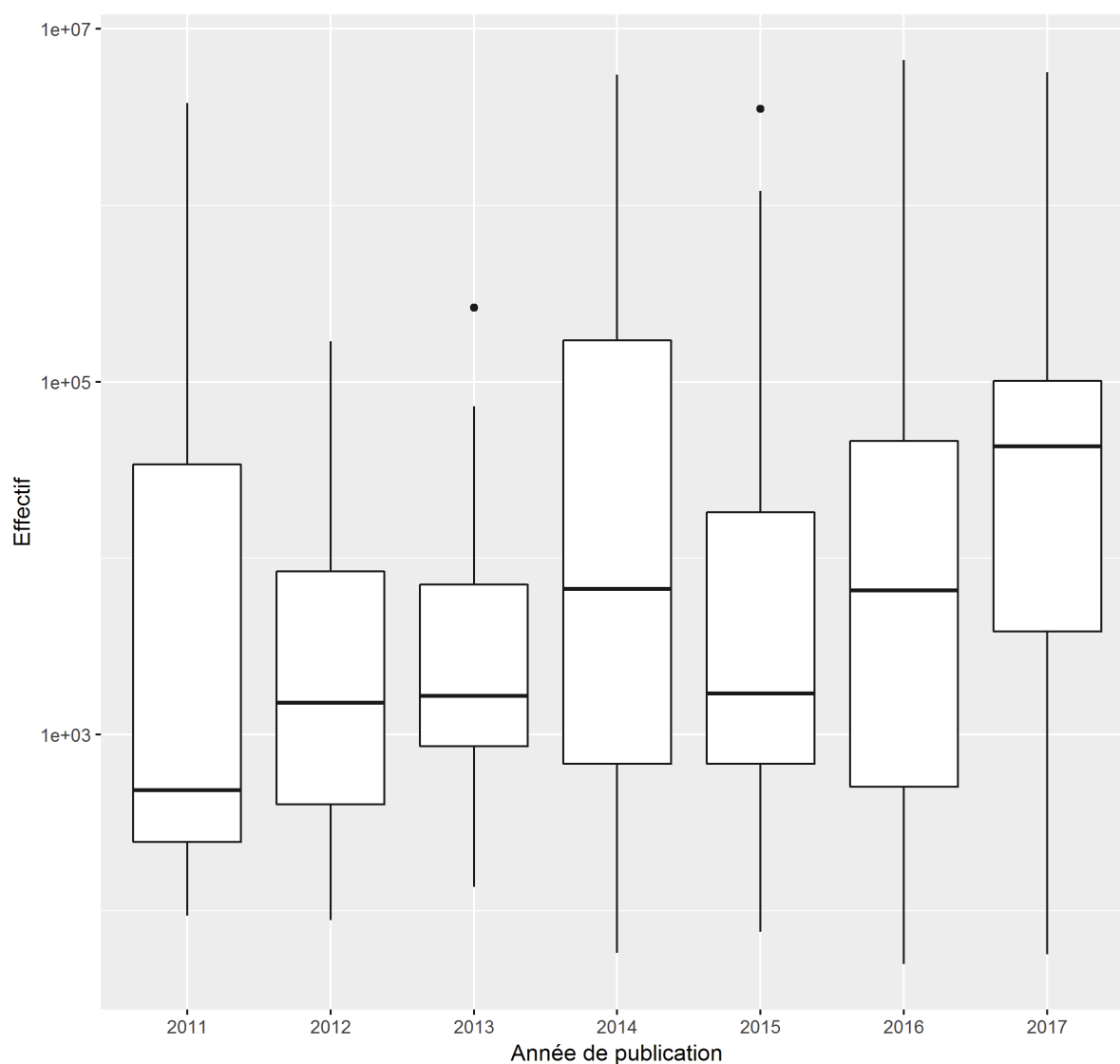
Figure 7 – Typologie des articles retrouvés dans la revue de la littérature



## Effectifs

A partir de 2011, on constate une augmentation moyenne de la quantité de données structurées issues d'EDC exploitées par étude. En 2011 la médiane était de 486 données structurées (min=94, max=3,8 millions) par étude pour 12 572 en 2017 (min=57, max=7,4 millions), soit un effectif médian multiplié par 26 en 6 ans. L'unité d'étude était le plus souvent le patient, le reste du temps il s'agissait principalement d'hospitalisations, de résultats d'analyses ou de prescriptions. L'évolution des effectifs est décrite en Figure 8.

**Figure 8 – Evolution des quantités de données structurées exploitées par étude**



## **Discussion**

### **Une exploitation des EDC encore anecdotique**

Notre revue de la littérature n'a relevé que 250 études exploitant un EDC dans le domaine de la santé. Comparativement aux 368 publications traitant du développement et du déploiement des EDC cela semble faible. Ce différentiel laisse supposer que la mise en production et l'exploitation d'un EDC demeure un défi technique et humain complexe que peu d'institutions sont parvenues à maîtriser.

### **Des cas d'usage peu variés**

L'exploitation des EDC consiste aujourd'hui quasi exclusivement en la réalisation d'études rétrospectives, lesquelles sont majoritairement épidémiologiques descriptives ou analytiques. L'exploitation prospective voire en temps réel des EDC permet d'envisager des modifications rapides des prises en charge médicales, ce qui peut avoir un impact potentiellement notable sur le parcours de soins. Cette responsabilité ne peut être assumée qu'en cas de pleine maîtrise de la fiabilité des données déversées dans l'EDC. Dans le cas contraire, une erreur sur le flux de données pourrait aboutir à une perte de chances pour le patient. Au vu du peu d'exploitations publiées à ce jour, l'absence d'études prospectives semble être liée à un besoin préalable de maturation des solutions actuelles.

La quantité de données exploitée dans chaque étude, bien que croissante, ne permet pas à ce jour de parler de Big Data d'un point de vue volumétrique. L'émergence de réseaux d'entrepôts nommés CDRN (Clinical Data Research Network) (26) à un niveau supérieur devrait permettre la mise en commun d'un plus grand nombre de centres et donc de données. A titre d'exemple, le CDRN Shrine (27) permet la mise en commun des données issues d'entrepôts I2b2 (28).

A terme, une plus grande volumétrie permettra d'aborder certains axes de recherches encore peu exploités telle que la médecine personnalisée.

### **Une exploitation en avance aux Etats Unis**

On constate de grandes disparités géographiques de l'utilisation des entrepôts de données en santé, au niveau mondial. Plusieurs éléments peuvent apporter une explication à cela.

L'adoption de la Health Insurance Portability and Accountability Act (2) aux USA permet la réutilisation des données cliniques sous réserve du respect de certains critères comme la suppression de variables nominatives (Protected Health Information). A ce jour il n'existe, dans certains pays, pas d'équivalent de cette réglementation. En France la réutilisation des données cliniques nécessite l'information individuelle et le recueil de la



non opposition auprès de chaque patient (méthodologie de référence MR 004) (29) et ce pour chaque étude. Ces différences de réglementations jouent probablement un rôle majeur dans les inégalités de déploiement des EDC.

Les incitations et les moyens dédiés à l'informatisation des SIH sont très disparates. On note par exemple en France que la part du budget des établissements de santé dédié à l'informatisation n'a évolué que de 0,01 points entre 2015 et 2016 (1,70% à 1,72% du budget total) (30). En comparaison, les Etats-Unis annoncent un taux de croissance du budget alloué au Systèmes d'Information Hospitaliers (SIH) de 11,4% (31).

### **Des difficultés liées à l'interopérabilité**

La balkanisation des SIH que l'on retrouve en France pose aujourd'hui un défi d'interopérabilité du fait d'une hétérogénéité importante des applications et référentiels utilisés en santé. Il en résulte un frein majeur au développement des EDC puisque la mise en commun des données médicales ne peut s'effectuer qu'au prix de conversions complexes. A défaut, les données demeurent cloisonnées à un SIH. Cette difficulté persiste à divers degrés selon les états. A l'inverse, certains comme le Danemark ont cherché très tôt à unifier leurs SIH. Ce dernier affiche aujourd'hui l'une des meilleures interopérabilités (32), mais au prix d'un travail laborieux (33) qui illustre la grande difficulté que représente l'unification des SIH.

### **Limites de notre étude**

Notre étude était restreinte au moteur de recherche Pubmed. Bien que celui-ci soit le plus pertinent pour explorer l'usage des EDC dans le domaine de la santé, certaines études peuvent ne pas y avoir été référencées. Néanmoins concernant la littérature méthodologique autour du développement des entrepôts, il est probable que d'autres moteurs de recherche davantage centrés sur les sciences de l'informatique aient été plus pertinents.

Malgré une attente forte des EDC dans les études prospectives, ce mode d'exploitation n'a été retrouvé que de façon anecdotique. Classiquement les articles décrivent peu la procédure de pré-screening. L'usage d'un EDC dans ce cadre peut donc être passé sous silence et ne pas apparaître dans notre revue de la littérature. De façon générale, l'usage d'un EDC pour la récupération des données, en prospectif comme en rétrospectif pouvait ne pas être mentionné dans l'article.

Il était parfois difficile de vérifier que l'EDC exploité par une étude était conforme à la définition telle que proposée par Wade (5). Lorsque nous ne pouvions être certains que le terme « data warehouse » réponde à notre définition nous avons exclu l'article, potentiellement au prix d'une perte de sensibilité. A l'inverse, le terme « data warehouse » n'étant que peu connu du monde des cliniciens, il est probable que certaines

études ait fait usage de termes plus réducteurs par méconnaissance du sujet (tels que « database », « data bank » ou « registre ») et n'apparaissent pas dans notre recherche. Elargir notre requête en incluant ces termes n'était pas raisonnable du fait de leur ubiquité dans la recherche.

### **Conclusion**

Finalement beaucoup de cas d'usages théorisés n'ont pas été retrouvés dans la littérature. Malgré les limitations de notre étude, il est probable que ce soit avant tout le caractère récent des EDC qui explique cette différence. Beaucoup d'axes restent donc à explorer. Avant de pouvoir modifier la prise en charge des patients sur la base des données d'EDC, de nombreuses contraintes techniques doivent être encore levées et la fiabilité des données déversées doit être assurée.

La réalisation de cette revue de la littérature aura permis de dégager les tendances vis à vis des EDC et constitue une base de données riche d'inspiration pour les établissements qui souhaiteraient débiter le déploiement et l'exploitation de leur propre EDC. L'ensemble des articles est disponible en matériel supplémentaire.

Les auteurs publiant des articles reposant sur des données issues d'entrepôts de données cliniques (tels que défini par Wade (5)) devraient utiliser le mot clé « data warehouse » pour qualifier leur recherche.



## Bibliographie

1. Fries JF, McShane DJ. ARAMIS (the American Rheumatism Association Medical Information System). A prototypical national chronic-disease data bank. *West J Med.* déc 1986;145(6):798-804.
2. 111th Congress Public Law 5. Health Information Technology for Economic and Clinical Health Act (HITECH Act) [Internet]. Sect. TITLE XIII. Disponible sur: <https://www.gpo.gov/fdsys/pkg/PLAW-111publ5/html/PLAW-111publ5.htm>
3. Le programme hôpital numérique - Ministère des Solidarités et de la Santé [Internet]. [cité 27 août 2018]. Disponible sur: <http://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/e-sante/sih/hopital-numerique/Hopital-Numerique>
4. What Is Big Data? - Gartner IT Glossary - Big Data [Internet]. [cité 28 août 2018]. Disponible sur: <https://www.gartner.com/it-glossary/big-data>
5. Wade TD. Traits and types of health data repositories. *Health Inf Sci Syst.* 2014;2:4.
6. Accueil | SNDS [Internet]. [cité 3 sept 2018]. Disponible sur: <https://www.snds.gouv.fr/SNDS/Accueil>
7. Chatellier G, Varlet V, Blachier-Poisson C, Beslay N, Behier J-M, Braunstein D, et al. « Big data » et « open data » : quel accès pour la recherche ? *Thérapie.* févr 2016;71(1):97-105.
8. Bostock M. D3.js - Data-Driven Documents [Internet]. [cité 28 août 2018]. Disponible sur: <https://d3js.org/>
9. Ruffin M. The importance of data warehouses for physician executives. *Physician Exec.* nov 1994;20(11):45-7.
10. Stausberg J, Lang H, Obertacke U, Rauhut F. Classifications in routine use: lessons from ICD-9 and ICPM in surgical practice. *J Am Med Inform Assoc JAMIA.* févr 2001;8(1):92-100.
11. Statnikov Y, Ibrahim B, Modi N. A systematic review of administrative and clinical databases of infants admitted to neonatal units. *Arch Dis Child - Fetal Neonatal Ed.* mai 2017;102(3):F270-6.
12. Craig BM, Han G, Munkin MK, Fenstermacher D. Simulating the contribution of a biospecimen and clinical data repository in a phase II clinical trial: A value of information analysis. *Stat Methods Med Res.* août 2016;25(4):1303-12.
13. Alban RF, Nisim AA, Ho J, Nishi GK, Shabot MM. Readmission to Surgical Intensive Care Increases Severity-Adjusted Patient Mortality: *J Trauma Inj Infect Crit Care.* mai 2006;60(5):1027-31.
14. Salem RM, Pandey B, Richard E, Fung MM, Garcia EP, Brophy VH, et al. The VA Hypertension Primary Care Longitudinal Cohort: Electronic medical records in the post-genomic era. *Health Informatics J.* déc 2010;16(4):274-86.

15. Nault V, Pepin J, Beaudoin M, Perron J, Moutquin J-M, Valiquette L. Sustained impact of a computer-assisted antimicrobial stewardship intervention on antimicrobial use and length of stay. *J Antimicrob Chemother.* 20 déc 2016;dkw468.
16. Nayar V, Kennedy A, Pappas J, Atchley KD, Field C, Smathers S, et al. Improving Cardiac Surgical Site Infection Reporting and Prevention By Using Registry Data for Case Ascertainment. *Ann Thorac Surg.* janv 2016;101(1):190-9.
17. Valsangkar NP, Eppstein AC, Lawson RA, Taylor AN. Effect of Lean Processes on Surgical Wait Times and Efficiency in a Tertiary Care Veterans Affairs Medical Center. *JAMA Surg.* 1 janv 2017;152(1):42.
18. Levin JC, Hron J. Automated Reporting of Trainee Metrics Using Electronic Clinical Systems. *J Grad Med Educ.* juin 2017;9(3):361-5.
19. Oldenmenger WH, Witkamp FE, Bromberg JEC, Jongen JLM, Lieverse PJ, Huygen FJPM, et al. To be in pain (or not): a computer enables outpatients to inform their physician. *Ann Oncol.* sept 2016;27(9):1776-81.
20. Evans RS, Benuzillo J, Horne BD, Lloyd JF, Bradshaw A, Budge D, et al. Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation. *J Am Med Inform Assoc.* sept 2016;23(5):872-8.
21. Kaplan DE, Dai F, Aytaman A, Baytarian M, Fox R, Hunt K, et al. Development and Performance of an Algorithm to Estimate the Child-Turcotte-Pugh Score From a National Electronic Healthcare Database. *Clin Gastroenterol Hepatol.* déc 2015;13(13):2333-2341.e6.
22. Meyer AND, Murphy DR, Al-Mutairi A, Sittig DF, Wei L, Russo E, et al. Electronic Detection of Delayed Test Result Follow-Up in Patients with Hypothyroidism. *J Gen Intern Med.* juill 2017;32(7):753-9.
23. Murphy DR, Thomas EJ, Meyer AND, Singh H. Development and Validation of Electronic Health Record-based Triggers to Detect Delays in Follow-up of Abnormal Lung Imaging Findings. *Radiology.* oct 2015;277(1):81-7.
24. Balicer RD, Shadmi E, Lieberman N, Greenberg-Dotan S, Goldfracht M, Jana L, et al. Reducing Health Disparities: Strategy Planning and Implementation in Israel's Largest Health Care Organization: Reducing Health Disparities. *Health Serv Res.* août 2011;46(4):1281-99.
25. Sedlis SP, Hartigan PM, Teo KK, Maron DJ, Spertus JA, Mancini GBJ, et al. Effect of PCI on Long-Term Survival in Patients with Stable Ischemic Heart Disease. *N Engl J Med.* 12 nov 2015;373(20):1937-46.
26. Holmes JH, Elliott TE, Brown JS, Raebel MA, Davidson A, Nelson AF, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. *J Am Med Inform Assoc.* 1 juill 2014;21(4):730-6.

27. 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.
28. i2b2: Informatics for Integrating Biology & the Bedside [Internet]. [cité 28 août 2018]. Disponible sur: <https://www.i2b2.org/>
29. Délibération n° 2018-155 du 3 mai 2018 portant homologation de la méthodologie de référence relative aux traitements de données à caractère personnel mis en œuvre dans le cadre des recherches n'impliquant pas la personne humaine, des études et évaluations dans le domaine de la santé (MR-004) | Legifrance [Internet]. JORF. Disponible sur: [https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=4941B11FB7C531A046EC7E6199A84A60.tplgfr24s\\_1?cidTexte=JORFTEXT000037187498&dateTexte=&oldAction=rechJO&categorieLien=id&idJO=JORFCONT000037186583](https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=4941B11FB7C531A046EC7E6199A84A60.tplgfr24s_1?cidTexte=JORFTEXT000037187498&dateTexte=&oldAction=rechJO&categorieLien=id&idJO=JORFCONT000037186583)
30. ATIH. Atlas des SIH 2018 - Etat des lieux des systèmes d'information hospitaliers [Internet]. [cité 27 août 2018]. Disponible sur: [http://solidarites-sante.gouv.fr/IMG/pdf/dgos\\_atlas\\_sih\\_2018.pdf](http://solidarites-sante.gouv.fr/IMG/pdf/dgos_atlas_sih_2018.pdf)
31. HIS (Hospital Information Systems) Market Size | Industry Report, 2025 [Internet]. [cité 5 sept 2018]. Disponible sur: <https://www.grandviewresearch.com/industry-analysis/hospital-information-systems-his-market>
32. Kierkegaard P. Interoperability after deployment: persistent challenges and regional strategies in Denmark. *Int J Qual Health Care* [Internet]. 19 févr 2015 [cité 27 août 2018]; Disponible sur: <https://academic.oup.com/intqhc/article-lookup/doi/10.1093/intqhc/mzv009>
33. Kierkegaard P. eHealth in Denmark: A Case Study. *J Med Syst* [Internet]. déc 2013 [cité 27 août 2018];37(6). Disponible sur: <http://link.springer.com/10.1007/s10916-013-9991-y>



## Matériel supplémentaire

La base de données de l'ensemble des articles récupérés via la requête initiale sur Pubmed est disponible à l'adresse suivante : <https://github.com/JeremyPasco/cdw-review>

Voici le détail de la liste des champs

- **doi** : Digital Object Identifiant de l'article
- **pmid** : PubMed IDentifiant de l'article
- **pmcid** : PubMed Central IDentifiant de l'article
- **publication\_year** : année de publication de l'article
- **publication\_month** : mois de publication de l'article
- **publication\_day** : jour de publication de l'article
- **cdw\_name** : nom de l'entrepôt de données clinique (uniquement pour les études exploitant un entrepôt)
- **cdw\_country** : pays de rattachement de l'entrepôt de données cliniques (uniquement pour les études exploitant un entrepôt)
- **article\_type** : type de l'article, peut prendre une valeur parmi
  - "exclusion" : exclusion de l'article
  - "methods:cdw-conception" : article méthodologique sur la conception des entrepôts
  - "methods:cdw-deployment" : article méthodologique sur le déploiement des entrepôts
  - "methods:cdw-follow-up" : article méthodologique sur le suivi d'évolution d'entrepôts
  - "methods:cdw-tools" : articles sur la conception d'outils liés aux entrepôts
  - "methods:other" : autres articles méthodologiques
  - "study:epidemiology:descriptive" : étude épidémiologique descriptive exploitant un entrepôt
  - "study:epidemiology:analytical" : étude épidémiologique analytique exploitant un entrepôt
  - "study:epidemiology:evaluative" : étude épidémiologique évaluative exploitant un entrepôt
  - "study:medico-economy" : étude médico-économique exploitant un entrepôt
  - "study:practice-evaluation" : évaluation des pratiques exploitant un entrepôt
- **exclusion\_criteria** : critère ayant mené à l'exclusion de l'article
- **study\_disciplines** : disciplines concernées par l'étude (pour les études exploitant un entrepôt)
- **cdw\_exploitation\_mode** : mode d'exploitation de l'entrepôt, peut prendre une valeur parmi :
  - "retrospective" : exploitation des données de façon rétrospective
  - "prospective" : exploitation des données de façon prospective
  - "real time" : exploitation des données en temps réel
- **cdw\_effective\_used** : effectif extrait de l'entrepôt et exploité lors de l'étude





**Vu, le Directeur de Thèse**

**Vu, le Doyen  
De la Faculté de Médecine de Tours  
Tours, le**



## **RESUME**

**Introduction** : La réutilisation des données cliniques a longtemps été restreinte par des difficultés d'accessibilité. Le développement d'entrepôts de données cliniques au sein des établissements de santé permet théoriquement de résoudre ce problème en centralisant l'ensemble des données cliniques au sein d'une même base de données. En amont du déploiement des Centres de Données Cliniques (CDC) eHOP au sein du Ri-CDC (réseau des CDC de l'inter-région Grand Ouest), nous avons souhaité étudier l'apport réel de ces entrepôts au travers d'une revue de la littérature.

**Matériel et Méthode** : Les articles ont été obtenus à l'aide du mot clé « data warehouse » sans restriction temporelle sur le moteur de recherche Pubmed. Seuls les articles faisant usage d'un entrepôt de données cliniques centrées sur le patient étaient retenus. Les éléments suivants étaient recueillis à l'aide d'une grille de lecture : date de publication, nom de l'entrepôt, localisation de l'entrepôt, type d'étude, temporalité de l'étude, discipline d'organe concernée, effectif extrait de l'entrepôt pour l'étude.

**Résultats** : Parmi les 1004 articles présents sur Pubmed, 618 ont été inclus dans notre revue de la littérature. Depuis le 1er article en 1994, on constate une progression constante du nombre de publications par année (83 articles en 2016). Néanmoins la majorité d'entre eux aborde les aspects méthodologiques de la mise en place d'un entrepôt de données. Seuls 250 articles (40%) consistent en une véritable étude exploitant un entrepôt. Parmi celles-ci on compte 226 (90%) articles d'épidémiologie, 12 (5%) d'études médico-économiques et 12 (5%) évaluations des pratiques. Le mode d'exploitation des entrepôts reste très largement rétrospectif (237 articles, 95%). Rares sont les études qui proposent une analyse des données en prospectif (9 articles, 4%) voire en temps réel (4 articles, 2%).

**Discussion** : A ce jour, la réutilisation des données cliniques au travers des entrepôts reste anecdotique dans la littérature médicale. L'importante quantité d'articles méthodologiques sur le sujet souligne les nombreuses difficultés techniques liées au déploiement et à l'exploitation de ces entrepôts. Celle-ci nécessitera encore de nombreux efforts pour s'intégrer dans le quotidien de la recherche biomédicale.

## PASCO Jeremy

40 pages – 8 figures – 1 matériel supplémentaire

### **Résumé :**

**Introduction** : La réutilisation des données cliniques a longtemps été restreinte par des difficultés d'accessibilité. Le développement d'entrepôts de données cliniques au sein des établissements de santé permet théoriquement de résoudre ce problème en centralisant l'ensemble des données cliniques au sein d'une même base de données. En amont du déploiement des Centres de Données Cliniques (CDC) eHOP au sein du Ri-CDC (réseau des CDC de l'inter-région Grand Ouest), nous avons souhaité étudier l'apport réel de ces entrepôts au travers d'une revue de la littérature.

**Matériel et Méthode** : Les articles ont été obtenus à l'aide du mot clé « data warehouse » sans restriction temporelle sur le moteur de recherche Pubmed. Seuls les articles faisant usage d'un entrepôt de données cliniques centrées sur le patient étaient retenus. Les éléments suivants étaient recueillis à l'aide d'une grille de lecture : date de publication, nom de l'entrepôt, localisation de l'entrepôt, type d'étude, temporalité de l'étude, discipline d'organe concernée, effectif extrait de l'entrepôt pour l'étude.

**Résultats** : Parmi les 1004 articles présents sur Pubmed, 618 ont été inclus dans notre revue de la littérature. Depuis le 1er article en 1994, on constate une progression constante du nombre de publications par année (83 articles en 2016). Néanmoins la majorité d'entre eux aborde les aspects méthodologiques de la mise en place d'un entrepôt de données. Seuls 250 articles (40%) consistent en une véritable étude exploitant un entrepôt. Parmi celles-ci on compte 226 (90%) articles d'épidémiologie, 12 (5%) d'études médico-économiques et 12 (5%) évaluations des pratiques. Le mode d'exploitation des entrepôts reste très largement rétrospectif (237 articles, 95%). Rares sont les études qui proposent une analyse des données en prospectif (9 articles, 4%) voire en temps réel (4 articles, 2%).

**Discussion** : A ce jour, la réutilisation des données cliniques au travers des entrepôts reste anecdotique dans la littérature médicale. L'importante quantité d'articles méthodologiques sur le sujet souligne les nombreuses difficultés techniques liées au déploiement et à l'exploitation de ces entrepôts. De nombreux efforts seront encore nécessaires pour intégrer l'exploitation des entrepôts dans le quotidien de la recherche biomédicale.

**Mots clés** : entrepôt de données, big data, réutilisation des données, revue de la littérature

### **Jury :**

Président du Jury : Professeur Emmanuel RUSCH

Directeur de thèse : Docteur Solène BRUNET-HOUDARD

Membres du Jury : Professeur Marc CUGGIA

: Professeur Frédéric PATAT

: Docteur Leslie GUILLON-GRAMMATICO

: Docteur Solène BRUNET-HOUDARD

: Date de soutenance : 1 octobre 2018