Réutilisation des données de réanimation : État de lieux des bases existantes et mise en place d'un entrepôt de données de réanimation au CHU de Rouen

Julien Kallout

UFR DE SANTÉ DE ROUEN NORMANDIE

Année 2022-2023

THÈSE POUR LE DOCTORAT EN MÉDECINE

(Diplôme d'État)

Par

Julien KALLOUT

Née le 18/04/1994 à Bondy

Présentée et soutenue publiquement le 08/09/2023

Réutilisation des données de réanimation : État de lieux des bases existantes et Mise en place d'un entrepôt de données de réanimation au CHU de Rouen

PRÉSIDENT DE JURY :	Professeur Benoît VEBER
DIRECTEUR DE THÈSE :	Docteur Benjamin POPOFF
MEMBRES DU JURY :	Professeur Thomas CLAVIER
	Docteur Philippe GOUIN
	Docteur Julien GROSJEAN

ANNEE UNIVERSITAIRE 2022 - 2023

U.F.R. SANTÉ DE ROUEN

DOYEN : **Professeur Benoît VEBER**

ASSESSEURS : **Professeur Loïc FAVENNEC**

Professeur Agnès LIARD

Professeur Guillaume SAVOYE

I – MEDECINE

PROFESSEURS DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mr Frédéric **ANSELME**	HCN	Cardiologie
Mme Gisèle **APTER**	Havre	Pédopsychiatrie
Mme Isabelle **AUQUIT AUCKBUR**	HCN	Chirurgie plastique
Mr Jean-Marc **BASTE**	HCN	Chirurgie Thoracique
Mr Fabrice **BAUER**	HCN	Cardiologie
Mme Soumeya **BEKRI**	HCN	Biochimie et biologie moléculaire
Mr Ygal **BENHAMOU**	HCN	Médecine interne
Mr Jacques **BENICHOU**	HCN	Bio statistiques et informatique médicale
Mr Emmanuel **BESNIER**	HCN	Anesthésiologie - Réanimation
Mr Olivier **BOYER**	UFR	Immunologie
Mme Valérie **BRIDOUX HUYBRECHTS**	HCN	Chirurgie Digestive
Mme Sophie **CANDON**	HCN	Immunologie
Mr François **CARON**	HCN	Maladies infectieuses et tropicales
Mr Philippe **CHASSAGNE**	HCN	Médecine interne (gériatrie)
Mr Florian **CLATOT**	CB	Cancérologie - Radiothérapie
Mr Moïse **COEFFIER**	HCN	Nutrition
Mr Vincent **COMPERE**	HCN	Anesthésiologie et réanimation chirurgicale
Mr Jean-Nicolas **CORNU**	HCN	Urologie
Mr Antoine **CUVELIER**	HB	Pneumologie
Mr Jean-Nicolas **DACHER**	HCN	Radiologie et imagerie médicale

Mr Stéfan **DARMONI**	HCN	Informatique et techniques de communication
Mr Pierre **DECHELOTTE**	HCN	Nutrition
Mr Stéphane **DERREY**	HCN	Neurochirurgie
Mr Frédéric **DI FIORE**	CHB	Cancérologie
Mr Fabien **DOGUET** *(disponibilité)*	HCN	Chirurgie Cardio Vasculaire
Mr Jean **DOUCET**	SJ	Thérapeutique - Médecine interne et gériatrie
Mr Bernard **DUBRAY**	CHB	Radiothérapie
Mr Frank **DUJARDIN**	HCN	Chirurgie orthopédique - Traumatologique
Mr Fabrice **DUPARC**	HCN	Chirurgie orthopédique et traumatologique
Mr Eric **DURAND**	HCN	Cardiologie
Mme Hélène **ELTCHANINOFF**	HCN	Cardiologie
Mr Manuel **ETIENNE**	HCN	Maladies infectieuses et tropicales
Mr Jean François **GEHANNO**	HCN	Médecine et santé au travail
Mr Emmanuel **GERARDIN**	HCN	Imagerie médicale
Mme Priscille **GERARDIN**	HCN	Pédopsychiatrie
M. Guillaume **GOURCEROL**	HCN	Physiologie
Mr Dominique **GUERROT**	HCN	Néphrologie
Mme Julie **GUEUDRY**	HCN	Ophtalmologie
Mr Olivier **GUILLIN**	HCN	Psychiatrie Adultes
Mr Florian **GUISIER**	HCN	Pneumologie
Mr Claude **HOUDAYER**	HCN	Génétique
Mr Fabrice **JARDIN**	CHB	Hématologie
Mr Luc-Marie **JOLY**	HCN	Médecine d'urgence
Mr Pascal **JOLY**	HCN	Dermato - Vénéréologie
Mme Bouchra **LAMIA**	Havre	Pneumologie
Mr Vincent **LAUDENBACH**	HCN	Anesthésie et réanimation chirurgicale
Mr Hervé **LEFEBVRE**	HB	Endocrinologie et maladies métaboliques
Mr Thierry **LEQUERRE**	HCN	Rhumatologie
Mme Anne-Marie **LEROI**	HCN	Physiologie
Mr Hervé **LEVESQUE**	HCN	Médecine interne
Mme Agnès **LIARD-ZMUDA**	HCN	Chirurgie Infantile
Mr Pierre Yves **LITZLER**	HCN	Chirurgie cardiaque
M. David **MALTETE**	HCN	Neurologie
Mr Christophe **MARGUET**	HCN	Pédiatrie
Mme Isabelle **MARIE**	HCN	Médecine interne
Mr Jean-Paul **MARIE**	HCN	Oto-rhino-laryngologie

Mr Stéphane **MARRET**	HCN	Pédiatrie
Mme Véronique **MERLE** *(disponibilité)*	HCN	Epidémiologie
Mr Pierre **MICHEL**	HCN	Hépato-gastro-entérologie
M. Benoit **MISSET** *(détachement)*	HCN	Réanimation Médicale
Mr Marc **MURAINE**	HCN	Ophtalmologie
Mr Gaël **NICOLAS**	UFR	Génétique
Mr Christian **PFISTER**	HCN	Urologie
Mr Jean-Christophe **PLANTIER**	HCN	Bactériologie - Virologie
Mr Didier **PLISSONNIER**	HCN	Chirurgie vasculaire
Mr Gaëtan **PREVOST**	HCN	Endocrinologie
Mr Jean-Christophe **RICHARD** *(détachement)*	HCN	Réanimation médicale - Médecine d'urgence
Mr Vincent **RICHARD**	UFR	Pharmacologie
Mme Nathalie **RIVES**	HCN	Biologie du développement - reproduction
Mr Horace **ROMAN** *(détachement)*	HCN	Gynécologie - Obstétrique
Mr Jean-Christophe **SABOURIN**	HCN	Anatomie - Pathologie
Mr Mathieu **SALAUN**	HCN	Pneumologie
Mr Guillaume **SAVOYE**	HCN	Hépato-gastro-entérologie
Mme Céline **SAVOYE–COLLET**	HCN	Imagerie médicale
Mme Pascale **SCHNEIDER**	HCN	Pédiatrie
Mr Lilian **SCHWARZ**	HCN	Chirurgie Viscérale et Digestive
Mr Michel **SCOTTE**	HCN	Chirurgie digestive
Mme Fabienne **TAMION**	HCN	Réanimation médicale
Mr Luc **THIBERVILLE**	HCN	Pneumologie
Mr Sébastien	CB	Radiothérapie
M. Gilles **TOURNEL**	HCN	Médecine Légale
Mr Olivier **TROST**	HCN	Anatomie - Chirurgie Maxillo-Faciale
Mr Jean-Jacques **TUECH**	HCN	Chirurgie digestive
Mr Benoît **VEBER**	HCN	Anesthésiologie - Réanimation chirurgicale
Mr Pierre **VERA**	CHB	Biophysique et traitement de l'image
Mr Eric **VERIN**	Les Herbiers	Médecine Physique - Réadaptation
Mr Eric **VERSPYCK**	HCN	Gynécologie obstétrique
Mr Olivier **VITTECOQ**	HCN	Rhumatologie
Mr David **WALLON**	HCN	Neurologie
Mme Marie-Laure **WELTER**	HCN	Physiologie

MAITRES DE CONFERENCES DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mme Najate **ACHAMRAH**	HCN	Nutrition
Mme Elodie **ALESSANDRI-GRADT**	HCN	Virologie
Mr Kévin **ALEXANDRE**	HCN	Maladies Infectieuses et Tropicales
Mme Noëlle **BARBIER-FREBOURG**	HCN	Bactériologie – Virologie
Mme Carole **BRASSE LAGNEL**	HCN	Biochimie
Mr Gérard **BUCHONNET**	HCN	Hématologie
Mme Mireille **CASTANET**	HCN	Pédiatrie
Mr Damien **COSTA**	HCN	Parasitologie
Mr Pierre **DECAZES**	CB	Médecine Nucléaire
Mr Maxime **FONTANILLES**	GHH	Oncologie Médicale
M. Vianney **GILARD**	HCN	Neurochirurgie
Mr Serge **JACQUOT**	UFR	Immunologie
Mr Joël **LADNER**	HCN	Epidémiologie, économie de la santé
Mr Jean-Baptiste **LATOUCHE**	UFR	Biologie cellulaire
M. Florent **MARGUET**	HCN	Histologie
Mme Chloé **MELCHIOR**	HCN	Hépato-gastro-entérologie
M. Sébastien **MIRANDA**	HCN	Médecine Vasculaire
Mr Thomas **MOUREZ** *(détachement)*	HCN	Virologie
Mme Muriel **QUILLARD**	HCN	Biochimie et biologie moléculaire
Mme Laëtitia **ROLLIN**	HCN	Médecine du Travail
Mme Pascale **SAUGIER-VEBER**	HCN	Génétique
M. Abdellah **TEBANI**	HCN	Biochimie et Biologie Moléculaire
Mme Anne-Claire **TOBENAS-DUJARDIN**	HCN	Anatomie
Mr Julien **WILS**	HCN	Pharmacologie

PROFESSEUR AGREGE OU CERTIFIE

Mme Noémie **MARIE**	UFR	Communication
Mr Thierry **WABLE**	UFR	Communication
Mme Mélanie **AUVRAY-HAMEL**	UFR	Anglais
Mme Cécile **POTTIER-LE GUELLEC**	UFR	Anglais

II – PHARMACIE

PROFESSEURS DES UNIVERSITES

Mr Jérémy **BELLIEN** (PU-PH)	Pharmacologie
Mr Thierry **BESSON**	Chimie Thérapeutique
Mr Jean **COSTENTIN** (Professeur émérite)	Pharmacologie
Mme Isabelle **DUBUS**	Biochimie
Mr Abdelhakim **EL OMRI**	Pharmacognosie
Mr François **ESTOUR**	Chimie Organique
Mr Loïc **FAVENNEC** (PU-PH)	Parasitologie
Mr Jean Pierre **GOULLE** (Professeur émérite)	Toxicologie
Mme Christelle **MONTEIL**	Toxicologie
Mme Martine **PESTEL-CARON** (PU-PH)	Microbiologie
Mr Rémi **VARIN** (PU-PH)	Pharmacie clinique
Mr Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

MAITRES DE CONFERENCES DES UNIVERSITES

Mme Margueritta **AL ZALLOUHA**	Toxicologie
Mme Cécile **BARBOT**	Chimie Générale et Minérale
Mr Frédéric **BOUNOURE**	Pharmacie Galénique
Mr Thomas **CASTANHEIRO MATIAS**	Chimie Organique
Mr Abdeslam **CHAGRAOUI**	Physiologie
Mme Camille **CHARBONNIER (LE CLEZIO)**	Statistiques
Mme Elizabeth **CHOSSON**	Botanique
Mme Marie Catherine **CONCE-CHEMTOB**	Législation et économie de la santé
Mme Cécile **CORBIERE**	Biochimie
Mme Sandrine **DAHYOT**	Bactériologie
Mme Nathalie **DOURMAP**	Pharmacologie
Mme Isabelle **DUBUC**	Pharmacologie
Mr Gilles **GARGALA**	Parasitologie
Mme Nejla **EL GHARBI-HAMZA**	Chimie analytique

Mr Chervin **HASSEL**	Virologie
Mme Maryline **LECOINTRE**	Physiologie
Mme Hong **LU**	Biologie
Mme Marine **MALLETER**	Biologie Cellulaire
M. Jérémie **MARTINET** (MCU-PH)	Immunologie
M. Romy **RAZAKANDRAINIBÉ**	Parasitologie
Mme Tiphaine **ROGEZ-FLORENT**	Chimie analytique
Mr Mohamed **SKIBA**	Pharmacie galénique
Mme Malika **SKIBA**	Pharmacie galénique

PROFESSEURS ASSOCIES

Mme Cécile **GUERARD-DETUNCQ**	Pharmacie officinale
Mme Caroline **BERTOUX**	Pharmacie officinale
M. Damien **SALAUZE**	Pharmacie industrielle

PAU-PH

M. Mikaël **DAOUPHARS**	Pharmacie
M. Pierre **BOHN**	Radiopharmacie

PROFESSEUR CERTIFIE

Mme Mathilde **GUERIN**	Anglais

ASSISTANTS HOSPITALO-UNIVERSITAIRES

M. Eric **BARAT**	Pharmacie
M. Guillaume **FEUGRAY**	Biochimie Générale
M. Henri **GONDÉ**	Pharmacie
M. Paul **BILLOIR**	Hématologie
M. Romain **LEGUILLON**	Pharmacie
M. Thomas **DUFLOT**	Pharmacologie
Mme Alice **MOISAN**	Virologie

ATTACHES TEMPORAIRES D'ENSEIGNEMENT ET DE RECHERCHE

Mme Chaïma **EZZINE**	Pharmacologie
M. Abdelmounaim **MOUHAJIR**	Informatique Bio-informatique
M. Olivier **PERRUCHON**	Pharmacognosie
M. Maxime **GRAND**	Bactériologie

LISTE DES RESPONSABLES DES DISCIPLINES PHARMACEUTIQUES

Mme Cécile **BARBOT**	Chimie Générale et minérale
Mr Thierry **BESSON**	Chimie thérapeutique
Mr Abdeslam **CHAGRAOUI**	Physiologie
Mme Elisabeth **CHOSSON**	Botanique
Mme Marie-Catherine **CONCE-CHEMTOB**	Législation et économie de la santé
Mme Isabelle **DUBUS**	Biochimie
Mr Abdelhakim **EL OMRI**	Pharmacognosie
Mr François **ESTOUR**	Chimie organique
Mr Loïc **FAVENNEC**	Parasitologie
Mme Christelle **MONTEIL**	Toxicologie
Mme Martine **PESTEL-CARON**	Microbiologie
Mr Mohamed **SKIBA**	Pharmacie galénique
Mr Rémi **VARIN**	Pharmacie clinique
M. Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

III – MEDECINE GENERALE

PROFESSEUR MEDECINE GENERALE

Mr Matthieu **SCHUERS** (PU-MG)	UFR	Médecine générale

PROFESSEURS ASSOCIES A MI-TEMPS – MEDECINS GENERALISTE

Mr Pascal **BOULET**	UFR	Médecine générale
Mr Emmanuel **HAZARD**	UFR	Médecine Générale
Mr Emmanuel **LEFEBVRE**	UFR	Médecine Générale
Mme Elisabeth **MAUVIARD**	UFR	Médecine générale
Mme Lucille **PELLERIN**	UFR	Médecine Générale
Mme Yveline **SEVRIN**	UFR	Médecine générale

MAITRE DE CONFERENCES ASSOCIE A MI-TEMPS – MEDECINS GENERALISTES

Mr Julien **BOUDIER**	UFR	Médecine Générale
Mme Laëtitia **BOURDON**	UFR	Médecine Générale
Mme Elsa **FAGOT-GRIFFIN**	UFR	Médecine Générale
Mme Ségolène **GUILLEMETTE**	UFR	Médecine Générale
Mr Frédéric **RENOU**	UFR	Médecine Générale

ENSEIGNANTS MONO-APPARTENANTS

PROFESSEURS

Mr Paul **MULDER** (phar)	Sciences du Médicament
Mme Su **RUAN** (med)	Génie Informatique

MAITRES DE CONFERENCES

Mr Sahil **ADRIOUCH** (med)	Biochimie et biologie moléculaire (Inserm 905)
Mr Jonathan **BRETON** (med)	Nutrition
Mme Gaëlle **BOUGEARD-DENOYELLE** (med)	Biochimie et biologie moléculaire (UMR 1079**)**
Mme Carine **CLEREN** (med)	Neurosciences (Néovasc)
M. Sylvain **FRAINEAU** (med)	Physiologie (Inserm 1096)
Mme Pascaline **GAILDRAT** (med)	Génétique moléculaire humaine (UMR 1079)
Mme Rachel **LETELLIER** (med)	Physiologie
Mr Antoine **OUVRARD-PASCAUD** (med)	Physiologie (Inserm 1076)
Mr Frédéric **PASQUET**	Sciences du langage, orthophonie
Mme Anne-Sophie **PEZZINO**	Orthophonie
Mme Christine **RONDANINO** (med)	Physiologie de la reproduction
Mr Youssan Var **TAN**	Immunologie
Mme Isabelle **TOURNIER** (med)	Biochimie (UMR 1079)

DIRECTEUR ADMINISTRATIF : M. Jean-Sébastien **VALET**

HCN - Hôpital Charles Nicolle

HB - Hôpital de BOIS GUILLAUME

CB - Centre Henri Becquerel

CHS - Centre Hospitalier Spécialisé du Rouvra

CRMPR - Centre Régional de Médecine Physique et de Réadaptation

SJ – Saint Julien Rouen

Par délibération en date du 3 mars 1967, la faculté a arrêté que les opinions émises dans les dissertations qui lui seront présentées doivent être considérées comme propres à leurs auteurs et qu'elle n'entend leur donner aucune approbation ni improbation.

Remerciements

A mon président de jury

Je souhaite tout d'abord exprimer mon plus grand respect au **Pr Benoît Veber**. Vous avoir comme président de jury est un véritable honneur. Merci pour votre implication totale dans le monde médical que ce soit au chevet des patients ou au sein de chaque faculté de médecine. Vous êtes une source d'inspiration pour chaque personne qui croise votre chemin.

A mon directeur de thèse

Je remercie le **Dr Benjamin Popoff** à qui je voue un profond respect. Toi qui manies le code informatique aussi bien que les recommandations médicales, le tout avec la plus grande des gentillesses. Merci de partager la même fascination que moi pour ce monde mystique qu'est la data. Je suis infiniment fier de notre travail et je suis honoré de l'avoir fait avec toi. Puisse ce binôme perdurer encore bien des années.

Aux membres de mon jury

Je tiens à remercier le **Pr Thomas Clavier** qui a été le catalyseur de mon parcours de recherche et qui m'a encouragé et poussé à faire les bons choix au bon moment. Merci pour ton enthousiasme dès les premiers instants à l'égard de mes projets. Je suis honoré d'avoir appris mon métier à tes cotés.

Je remercie également le **Dr Phillipe Gouin** dont la présence dans ce jury a été une évidence. Merci d'avoir tant œuvré dans l'organisation du département dans bien des aspects. Merci pour tes (très) longues visites pleines d'enseignements et de bienveillance. Travailler à tes cotés est et restera un véritable plaisir.

Je remercie enfin le **Dr Julien Grosjean** qui m'a accueilli et encadré au sein du département d'informatique médicale. Ce travail n'aurait très certainement pas eu la même consistance sans ton implication. Collaborer avec toi est un plaisir au quotidien et je me réjouis d'entretenir le lien entre nos deux départements à tes cotés.

A toutes les personnes du département d'anesthésie-réanimation

Je tiens à remercier chaque **médecin**, junior ou sénior, sans qui l'apprentissage de mon métier n'aurait pas été possible.

Merci aux **infirmiers** et **infirmières**, d'anesthésie et de réanimation, dont l'expérience m'a permis et me permettra de mieux appréhender l'exercice de la médecine. Merci aux **sage-femmes** sans qui la salle de naissance n'aurait pas la même atmosphère. Merci aux **aides-soignants**, aux **agents de services hospitaliers**, aux **brancardiers**, à tous ces travailleurs de l'ombre sans qui mon métier serait bien moins aisé.

A toutes les personnes du département d'informatique médicale

Je remercie le **Pr Stefan Darmoni** de m'avoir accueilli au sein du département d'informatique médicale. Merci pour votre bienveillance à mon égard dès les premiers instants et pour votre indéfectible bonne humeur au quotidien. La façon dont vous fédérez votre équipe est un exemple et je suis très honoré d'en faire partie aujourd'hui.

Je tiens à remercier de la plus belle des manières **Mr Badisse Dahamna** sans qui le projet Icca2Omop n'aurait pas été possible. Merci pour ta réactivité et ton efficacité. Merci pour ta patience et ta bienveillance malgré le néant qui régnait parfois dans mon regard.

Je souhaite également remercier le **Dr Romain Leguillon** d'avoir partagé mes moments de solitude dans notre bureau. Je suis heureux de t'avoir rencontré et je me réjouis d'avance de collaborer avec toi dans ce département.

Enfin merci à **Elie Lacroix, Émeline Lejeune, Francesco Monti**, **Gaétan Kerdelhué, Jean-Philippe Leroy, Hélène Cieslik**, **Ivan Kergourlay** pour votre gentillesse durant mon passage chez vous.

A toutes les personnes rencontrées durant mon année de recherche

Je remercie le **Pr Bruno Falissard** qui est et restera une source d'inspiration tout au long de ma carrière. Merci pour votre pédagogie et votre capacité à transformer l'abstrait en réel.

Je tiens également à remercier toute l'équipe **de l'unité INSERM 1018** pour leur accueil et l'ensemble **du groupe Traumabase®** pour leur formidable initiative.

A ma famille

Merci à **mes parents** pour leur amour et leur sacrifice. **Baba**, **Mama**, je vous dois tout dans ma réussite. Je suis fier d'être l'étendard de votre éducation. Je vous aime.

Merci à **mes petits frères Remi et Rayan** pour tous nos moments de complicité. Vous êtes une partie de moi et je ne cesserai jamais de vous aimer et de vous protéger.

Merci à **mes grands-parents**, **mes oncles**, **mes tantes**, **mes cousins** pour tous les bons moments passés, présents et futurs.

A la famille Renault

Merci à toi **Amandine** pour toutes ces années à tes cotés. A tous nos moments d'amour et de complicité, de joie et de sacrifice. A toi ma binôme de toujours, une partie de mon cœur te restera à jamais dédiée.

Merci à **Evelyne** et **Phillipe** pour votre générosité, à **Caroline** et **Alexandre** pour votre bienveillance, à **Tiphanie** et **Benoît** pour votre gentillesse, à **Nicolas** pour ta bonne humeur, à **Annie** pour ton énergie, à **Sylvain** pour ta sagesse, à **Arnaud** et **Marie** pour votre folie, à **Annie** et **Claude** pour votre hospitalité, sans oublier les beaux gosses **Adam** et **Martin**. Vous resterez dans mon cœur comme une seconde famille.

A mes amis du lycée

Merci à **Orane** pour ta simple présence depuis le lycée jusqu'au CHU de Rouen en passant par les bancs de la fac. Nos insupportables bavardages en classe ne nous auront finalement pas empêchés de devenir anesthésiste-réanimateur.

Merci à **Leïla** et **Flore** pour votre indéfectible amitié. A tous ces moments de joie et de rire mais aussi de doute et de solitude. Puisse votre folie ne jamais s'épuiser.

Merci au club des Cincos (+ Julien). Merci à **Alexis**, ma blonde de toujours dont la distance n'épuisera jamais notre amitié. Merci à **Vincent** pour ta classe, à **Paul** pour ta joie de vivre, à **Maximilien** pour ta gentillesse et à **Charles** pour ta sensibilité.

Merci à **Adrien** pour ta sincérité. A toi qui sais tant me faire rire à chaque fois que l'on se voit.

Merci à **Pénélope** et **Coline** pour votre mignonnerie. Merci à **Pauline** pour ta simplicité. Merci à **Zoé** pour ton appétit. Merci à **Ingrid** pour notre détestable amitié. Merci à **Ellia** pour le diable qui danse dans ta tête.

Merci à tout le groupe du Mistral (ex-Panachés). Merci à **Alexandre C**, **Alexandre G**, **Antoine**, **Arthur**, **Clément**, **Florentin** sans oublier le plus dingo **Zaïm**.

Merci à toutes les pièces rapportées que je suis heureux d'avoir rencontrées. Merci **Erwin** la personne la plus gentille du monde. Merci à **Pauline D** pour ton féminisme. Merci à **Pauline G** pour ton élégance à toute épreuve.

A mes amis de la fac

Merci à **Amre** mon amour plutonique, à tous nos fou-rires communicatifs. Merci à **Eliza** pour ta veine du seum pleine d'amour, sans oublier **Axel** ton beau mari. Merci à **Romain** pour ton rire qui fait trembler les montagnes. Merci à **Sofiane** pour ta blédardise. Mais surtout, merci à **Laura** pour ta libanese-touch, le ciment et la maman du groupe.

Merci à **Alexandre Lem** mon coup de cœur de tous les temps. A nos discussions mystiques et perchées dans l'espace. A toi que je ne vois pas assez à mon gout.

Merci à **Etienne** pour ton authenticité. Merci pour ton courage et ta combativité.

Merci à **Laura A** pour ta capacité à me donner mal au ventre de rire.

Merci à **Parna** pour ta gentille gentillesse. Merci pour ta curiosité et ton intérêt envers chacun.

A mes amis de l'internat

Merci à **Alexandre Lav** pour ta folie chronique qui me rappelle à quel point la vie mérite d'être vécue. Merci d'avoir la tête en l'air encore plus haut que la mienne. Mais surtout merci pour ton sens de l'écoute et ton sérieux en cas de besoin.

Merci à **Benjamin H** pour cette formidable amitié qui dure depuis le premier jour d'internat. Merci pour ta sincérité plus que frontale mais aussi pour ta capacité à mettre en avant le bon au sein de chaque personne que tu croises.

Merci à **Samy** pour ta générosité en toute circonstance. A ton calme qui m'apaise à chaque fois que je te vois. Merci pour tes questions existentielles qui me retournent le cerveau.

Merci à **Hana** pour ta douceur mais aussi pour ton champ lexical. Merci à **Jeanne** pour ton humour et ta personnalité d'amuseuse. Merci à **Lisa** pour ta folie. Merci à **Margaux** pour ta délicatesse et ton petit rire de bijou. Merci à **Marc** pour ton sens de l'hospitalité. Merci à **Naomie** pour ta loyauté de bezo. Merci à **Yasaman** pour ton caractère bien trempé au safran.

Merci à toutes ces personnes que j'ai rencontrées dans cette belle ville de Rouen : **Alice E, Baptiste G, Boris R, Claire M, Claire V, Cloé G, Clotilde L, Ela D, Fanny B, Guillaume L, Gwenaël G, Hasan A, Hélène F, Hugo M, Ines T, Jules Fou, Julie B, Julien B, Laura M, Leila T, Marwah H, Myriam S, Nicolas D, Pauline J, Raphaël JL Sami H, Teddy C.**

Merci aux plus dingos : **Camille E, Célia V, Jasmine F, Maxime M, Thomas M, Véro S**.

Merci à la brigade des utérus, puisse vos chaussettes ne jamais rester blanches : **Amina B, Evelyne M, Marie O, Pauline L**.

Merci à tous mes co-internes sans qui mon quotidien serait bien plus monotone : **Alexis D**, **Antoine H**, **Aurore G, Clément D, Corentin M, Capucine D, Delphine S, François S, Frédéric C, Géraldine M, Grégoire J, Gwenaëlle M, Jonathan N, Jules F, Louis S, Marie G, Maxime D, Marion F, Nicolas F, Océane G, Pierre K, Rémy R, Robin T, Simon H, Stan P, Stellina BC, Vanessa K**, **Wilfried F**.

Mention spéciale pour la team réachir avec qui un semestre passe beaucoup trop vite : **Bastien, Carméliane, Charles, David, Estelle, Havana, Laure, Léa, Maxime, Quentin, Wafa**, dirigée par le plus tatillon et gentil des chefs **Yvon le citron**.

Enfin, merci à la plus belle des **promo DESAR 2018**, avec qui tout a commencé, ainsi la boucle est bouclée : **Ali, Alice, Aurélie, Bastien, Camille, Charles, Djouher, Émeline, Jordan, Léa, Maxime, Pierre, Sarah, Steffi, Thibault**.

A vous tous qui supportez mes Kallouteries depuis tant d'années. Merci de m'aimer et de m'accompagner, de me féliciter et de me sermonner, de m'encourager et de me critiquer. Merci de me faire grandir sans jamais me lâcher malgré ma tendance à toujours siphonner votre océan de patience. En espérant ne jamais l'assécher.

A la mémoire du Pr Sophie-Rym Hamada

16/10/1978 – 16/08/2022

Chère Sophie,

Je souhaite vous remercier de m'avoir accueilli au sein de votre équipe et de m'avoir encadré durant mon année de recherche. Merci pour votre bienveillance, votre disponibilité et votre réactivité à toute épreuve. Merci pour les moments privilégiés que vous m'avez accordés malgré les charges qui vous incombaient. Je vous exprime mon plus grand respect pour l'énergie, la passion et la curiosité dont vous avez fait part au quotidien et que vous m'avez transmises tout au long de l'année. Vous resterez à mes yeux un exemple à suivre tant sur le plan humain que professionnel. Vous côtoyer au quotidien aura été un véritable honneur, sans nul doute l'année la plus enrichissante de mon internat.

Malgré le temps qui passe ma tristesse reste infinie et mon chagrin inconsolable mais je me sens chanceux de vous avoir connue et c'est avec une immense fierté que je dirai avoir travaillé avec le Pr Sophie-Rym Hamada. Je tâcherai de préserver en moi cette flamme qui vous animait chaque jour afin de la transmettre à mon tour.

Le bleu ciel était votre couleur préférée et l'utiliser dans nos graphiques était une prérogative. A travers ce bleu qui vous entoure aujourd'hui je vous dédie ce travail.

Reposez en paix.

Table des matières

Listes des abréviations

AJRCCM : American Journal of Respiratory and Critical Care Medicine
AmsterdamUMCdb : Amsterdam University Medical Center data base
APACHE : Acute Physiology and Chronic Health Evaluation
APS : Acute Physiology Score
AUROC : Area Under the Curve

BIDMC : Beth Israel Deaconess Medical Center

CCU : Continuing Care Unit
CépiDC : Centre d'épidémiologie sur les causes médicales de DéCès
CIM : Classification Internationales des maladies
CNIL : Commission Nationale de l'Informatique et des Libertés
CPT : Current Procedural Terminology

DIO : Digital Object Identifier
DMI : Dossier Médical Informatisé
DMP : Dossier Médical Partagé
DRG : Diagnosis related group

ECR : Essai Contrôlé Randomisé
EDSaN : Entrepôt de Données de Santé Normand
EDSH : Entrepôts de Données de Santé Hospitaliers
EHR : Electronic Health Record
eICU-CRD : eICU Collaborative Research Database
ETL : Extraction, Transformation, and Loading

GIP : Groupement d'Intérêt Public

HAS : Haute Autorité de Santé
HDH : Health Data Hu
HDU : Hemodialysis Unit
HiRID : High time-Resolution Intensive care unit Dataset
HR : Hazard-Ratio

ICCA® : IntelliSpace Critical Care and Anesthesia©
ICU : Intensive Care Unit

JCR : Journal Citation Reports

MEWS : Modified Early Warning Score
MIT : Massachusetts Institute of Technology
MIMIC : Medical Information Mart for Intensive Care
MR : Méthodologie de Référence

OASIS : Oxford Acute Severity of Illness
OHDSI : Observational Health Data Sciences and Informatics
OMOP-CMD : Observational Medical Outcomes Partnership Common Data Model
OR : Odds-Ratio

PMSI : Programme de Médicalisation des Systèmes d'Information
POCU : Post Operative Care Unit
PRISMA-ScR : Preferred Reporting Items for Systematic Reviews and Meta-analyses extension for Scoping Reviews

RDBMS : Relational Database Management System
RMSE : Root Mean Square Error
RUH : Rouen University Hospital
RWU : Respiratory Weaning Unit

SAPS : Simplified Acute Physiology Score
SNDS : Système National des Données de Santé
SNIIRAM : Système National d'Information Inter-régimes de l'Assurance Maladie
SOFA : Sepsis-related Organ Failure Assessment
SQL : Structured Query Language

Listes des figures et tableaux

Partie I : Introduction générale et objectifs de la thèse

Partie II : Contribution of open access databases to intensive care medicine research: A scoping review

Partie III : Development of an Intensive Care Data Warehouse at Rouen University Hospital: Transformation of the ICCA Database in the OMOP Common Data Model

Partie I : Introduction générale et objectifs de la thèse

Depuis James Lind et son premier essai clinique sur le scorbut en 1753 **[1]**, les Hommes ont constamment aspiré à comprendre et améliorer la santé en se basant sur les faits. Forcé de constater que la médecine factuelle ne peut reposer seulement sur la recherche expérimentale, il était impératif d'explorer d'autres moyens de recherche pour approfondir notre compréhension de la santé.

En raison des progrès technologiques croissants, la recherche sur les données de santé occupe aujourd'hui une place centrale dans la recherche médicale. Notre capacité grandissante à collecter, stocker et réutiliser les données a permis le développement d'entrepôts de données de santé ouvrant la possibilité d'analyser un nombre toujours plus grand de patients **[2]**.

La sévérité des patients de réanimation impose une surveillance continue de leurs paramètres cliniques et paracliniques. Cela génère une grande quantité de données, généralement recueillies dans des dossiers médicaux informatiques (DMI). Au cours des dernières décennies, des initiatives de collaboration ont permis l'émergence de grandes bases de données de réanimation en libre accès **[3]**. S'appuyant sur l'opportunité de la transformation numérique, elles facilitent le partage de données à grande échelle et permettent de créer des connaissances de manière plus efficace.

1. Données de santé

Selon la Commission nationale de l'informatique et des libertés (CNIL) les données de santé sont définies par "des données à caractère personnel relatives à la santé physique ou mentale, passée, présente ou future, d'une personne physique qui révèlent des informations sur l'état de santé de cette personne" **[4]**.

En 2013, selon un rapport de l'Assemblée nationale sur les données de santé, 153 exaoctets (10^{18} octets) de données de santé ont été produits dans le monde. En moins de 10 ans, le volume des données de santé aurait été multiplié par dix avec 2 314 exaoctets produits **[5]**. Le rythme de ces changements va probablement accélérer l'accès croissant à différentes sources de données.

2. Sources de données en recherche médicale

2.1. Données issues de la recherche expérimentale

L'essai contrôlé randomisé (ECR) d'une intervention contre placebo ou contre une intervention de référence, constitue aujourd'hui la référence en termes de recherche expérimentale **[6]**. Correctement réalisé, l'ECR est le seul à pouvoir affirmer la relation causale entre l'intervention et le résultat obtenu. La stricte comparabilité des groupes obtenue au départ grâce à la randomisation, maintenue secondairement tout au long de l'essai via le double insu et assurée en fin d'essai avec l'analyse en intention de traiter permet d'inférer les différences observées à l'intervention réalisée. De ce fait, la méthodologie éprouvée des ECR est à la base de la médecine factuelle (ou *evidence based medicine*) **[7]**. Apportant le meilleur niveau de preuve, elle est considérée comme le gold standard lors de l'autorisation de mise sur le marché d'un médicament.

Néanmoins, ce qui fait la force de l'ECR constitue aussi une limite de par la rigueur de son schéma expérimental. Les patients inclus dans les essais et soigneusement sélectionnés, constituent un effectif limité et parfois peu représentatif de la majorité des patients concernés en vie réelle. Le respect scrupuleux du protocole peut entraîner des conditions parfois éloignées de la réalité et remettre en cause la généralisation des résultats en pratique courante. Par ailleurs, la durée limitée des ECR est également un facteur limitant **[8]**. Les ECR ont souvent tendance à sous-estimer les effets secondaires à long terme des intervention évaluées et ne permettent pas d'étudier leur efficacité par rapport à d'autres stratégies sur une

période de temps plus longue. Enfin, cette méthodologie est difficilement applicable avec des pathologies rares **[9]**.

Les limites des ECR, si elles n'enlèvent rien au niveau de preuve apporté par les ECR, justifient que d'autres types d'études soient réalisées. Les anglo-saxons ont d'ailleurs deux termes pour distinguer l'efficacité d'une intervention : l'*efficacy*, à savoir l'efficacité dans des conditions optimales (évaluée par les ECR) et l'*effectiveness*, l'efficacité dans les conditions habituelles d'utilisation en pratique courante **[10]**.

2.2. Données issues de la vie réelle

On désigne sous le terme "données de vie réelle" les données collectées en dehors d'un cadre expérimental et générées à l'occasion de soins réalisés en routine pour un patient. Elles reflètent donc à priori la pratique courante **[11]**.

2.2.1. Dossiers médicaux patients

Le dossier médical patient constitue le support de l'ensemble des informations recueillies concernant sa prise en charge. Il comprend les informations cliniques, diagnostiques, thérapeutique, les soins infirmiers ainsi que les informations administratives.

Le dossier médical informatisé (DMI) a permis le partage des données médicales entre tous les acteurs du soin avec la notion de dossier médical partagé (DMP) **[12]**.

2.2.2. Enquêtes observationnelles

Ces études sont réalisées à partir d'échantillons sélectionnés de manière à répondre à une ou plusieurs questions précises. Pouvant être à visée descriptive ou analytique, il en existe trois types : les études de cohorte, les études cas-témoins et les études transversales.

L'existence de nombreux biais, tels que les biais de sélection, de confusion et de classement, rend illusoire la possibilité d'établir de manière définitive une relation causale. Les critères de Bradford Hill permettent néanmoins d'approcher cette causalité sans jamais pouvoir l'attendre avec une certitude absolue **[13]**.

Malgré ces limites, les études observationnelles jouent un rôle crucial dans la recherche médicale en particulier lorsqu'il n'est pas éthique ou possible de mener des ECR. Ces études fournissent des informations précieuses pour explorer des associations, générer des hypothèses, guider la planification d'études plus rigoureuses et vérifier les conclusions des ECR en pratique courante sur le long terme.

2.2.3. Systèmes de collecte permanents

Ces systèmes enregistrent des événements (une pathologie, un décès, une situation médicale), à l'occasion de leur diagnostic ou de leur survenue afin de fournir des données pour les étudier. Pour répondre à ces objectifs, le système doit être pérenne et stable, donc simple et acceptable.

Registres de maladies

Un registre de maladie est un recueil continu et exhaustif de données nominatives intéressant un ou plusieurs événements de santé dans une population géographiquement définie, à des fins de recherche et de santé publique, par une équipe ayant les compétences appropriées **[14]**.

Les registres de morbidité concernent des maladies dont le diagnostic peut-être aisément établi et qui justifient du point de vue des enjeux de santé publique la mise en place de ce type de système lourd et complexe. Les registres sont un outil essentiel pour la surveillance épidémiologique, la recherche médicale et l'amélioration des soins. Ils sont généralement mis en place dans le but d'étudier l'évolution de la maladie, évaluer l'efficacité des traitements, identifier les facteurs de risque, faciliter la recherche épidémiologique et générer des données pour des études observationnelles.

A titre d'exemple, le registre *Traumabase*® est un observatoire français de traumatologie lourde créé en 2012. Son objectif est de recueillir les données des patients traumatisés graves dans un but à la fois sanitaire et scientifique. Actuellement ce groupe est constitué d'un réseau de 25 centres français de traumatologie lourde et recense les données de plus de 40 000 admissions en réanimation pour traumatisme grave, de la prise en charge hospitalière jusqu'à la sortie de réanimation **[15]**.

Système national des données de santé (SNDS)

Le Système national des données de santé (SNDS) est une infrastructure française qui vise à collecter, stocker, et exploiter les données de santé à des fins de recherche, de surveillance épidémiologique, d'évaluation des politiques de santé, et de prise de décisions en matière de santé publique **[16]**. Cela contribue à l'amélioration des connaissances en matière de santé, à la gestion des risques sanitaires et à l'élaboration de politiques de santé fondées sur des données probantes.

Le SNDS permet de chaîner différentes sources de données, agrégées de manière anonyme et sécurisée avec notamment :

- Les données de l'Assurance Maladie (base du Système national d'information inter-régimes de l'Assurance maladie, SNIIRAM) : informations médicales et administratives liées aux remboursements des soins de santé, aux médicaments délivrés, aux hospitalisations
- Les données des hôpitaux (base du Programme de médicalisation des systèmes d'information, PMSI) : informations issues des hôpitaux et cliniques, telles que les diagnostics, les actes médicaux, les séjours hospitaliers,
- Les causes médicales de décès (base du Centre d'épidémiologie sur les causes médicales de décès de l'Inserme, CépiDC) : informations sur les décès et causes de décès délivrées par les certificats de décès émis par les médecins

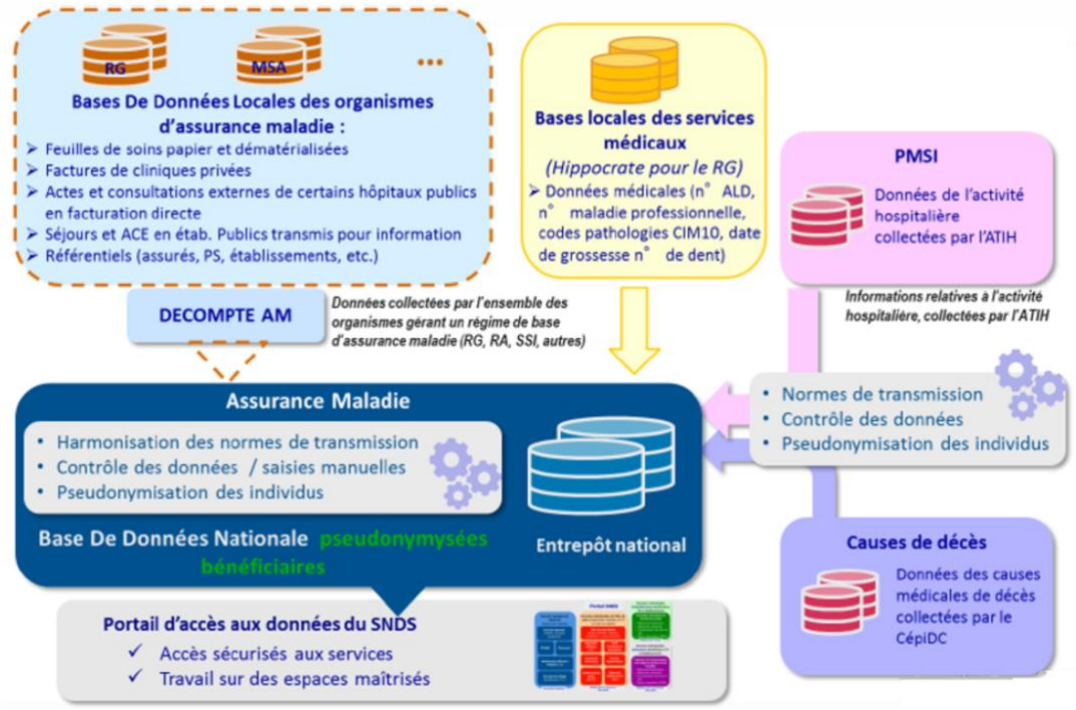


Figure 1 – Structure du SNDS [17]

3. Réutilisation des données en santé

La réutilisation des données (ou *data reuse*) fait référence à la pratique qui consiste à utiliser des données existantes à des fins multiples ou dans des contextes différents. Au lieu de collecter de nouvelles données, les organisations ou les individus exploitent des données déjà générées auparavant. Elle repose sur le fait que les données peuvent être utilisées au-delà de leur objectif initial **[18]**.

Cette approche offre plusieurs avantages, notamment des économies de coûts, de temps et de ressources humaines ainsi que des gains d'efficacité. En réutilisant les données, les acteurs de la santé peuvent éviter les dépenses liées à la collecte de nouvelles données, telles que les outils de collecte de données, l'infrastructure et le personnel. Cela peut être particulièrement avantageux lorsqu'il s'agit de travailler avec des ensembles de données volumineux ou complexes. De plus, la réutilisation des données favorise l'efficacité en maximisant la valeur tirée des données existantes. Au lieu de laisser les données inutilisées après leur objectif initial, les organisations peuvent trouver de nouvelles applications ou des questions de recherche pouvant être abordées à l'aide des mêmes données. De ce fait, cette vision encourage une gestion des données plus durable et responsable.

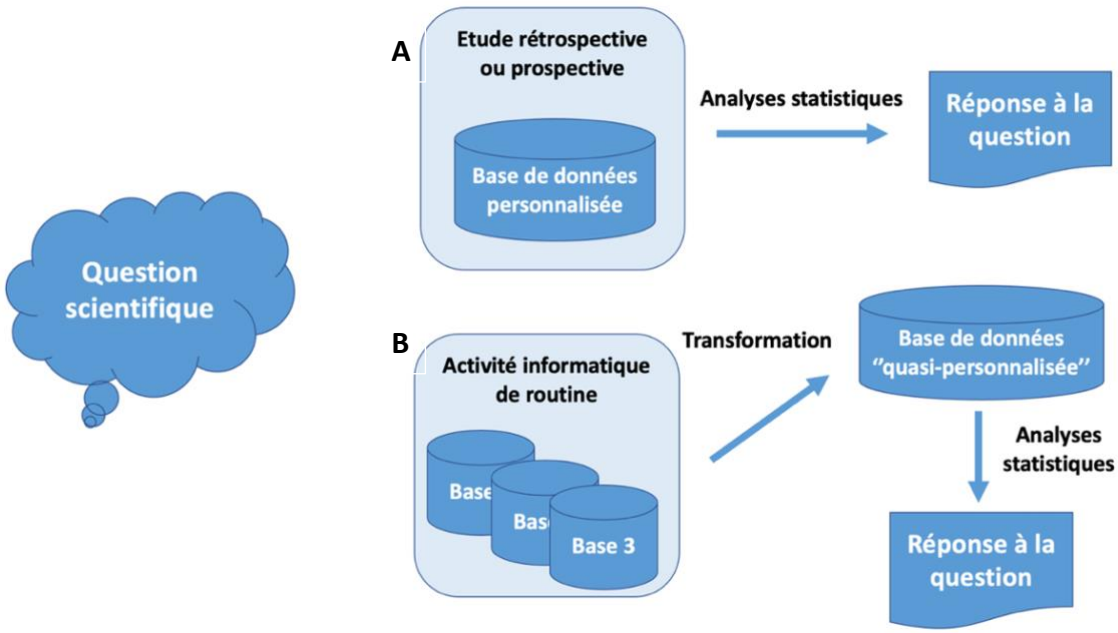


Figure 2 – Approche traditionnelle (A), Réutilisation des données (B)

4. Entrepôts de données de santés hospitaliers

Les entrepôts de données de santé hospitaliers (EDSH) sont des systèmes de stockage, de gestion et de réutilisation de données utilisés par les hôpitaux et les établissements de santé pour collecter, organiser et analyser une grande quantité de données de santé **[2]**.

Ces bases de données sont constituées pour un volume de données important et s'étendent sur une longue durée. Elles peuvent être alimentées par de multiples sources. Les EDSH présentent une organisation commune autour des données administratives, des textes cliniques, des résultats biologiques, des comptes rendus d'imagerie, du circuit du médicament et un socle variable de données dans leur format ou leur construction. Les données sont collectées depuis les différentes sources du système d'information hospitalier (SIH). Secondairement elles sont harmonisées puis agrégées afin de constituer un entrepôt exploitable. Chaque EDSH, et chaque projet sur EDSH, doit être en conformité avec les exigences de la CNIL et les prérogatives des comités scientifiques et éthiques.

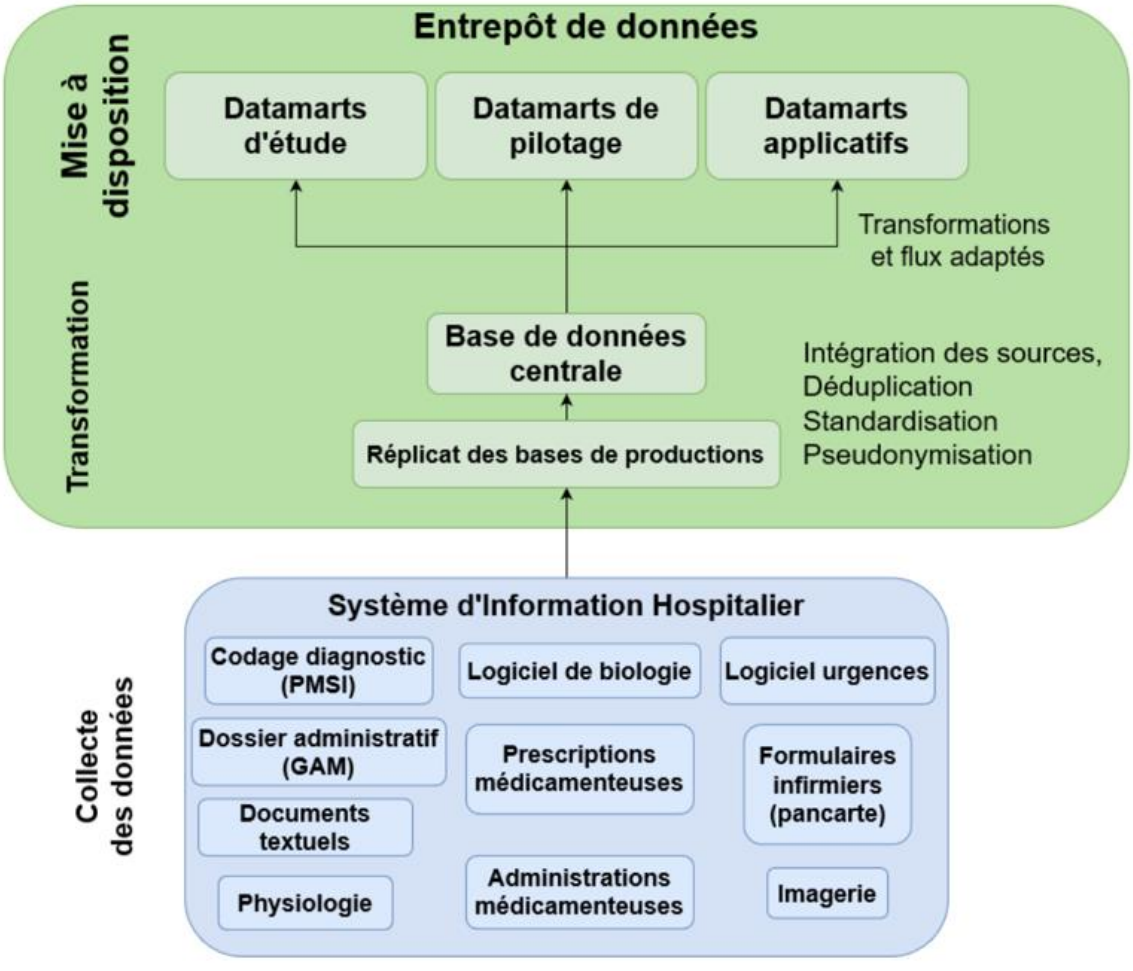


Figure 3 – Structures des EDS [19]

La mise en place des EDSH en France remonte à la fin des années 2000 et s'est renforcée à la fin des années 2010. On observe actuellement une accélération dans le déploiement de l'écosystème des EDSH, notamment grâce à des financements nationaux, la multiplication d'acteurs industriels spécialisés en données de santé et le début d'une réflexion internationale à propos de l'espace européen de données de santé.

Dans son dernier rapport de 2022, la Haute autorité de santé (HAS) a recensé 22 EDSH, dont 17 constitués au sein d'un centre hospitalo-universitaire et 5 au sein d'un autre type d'établissement de santé **[19]**.

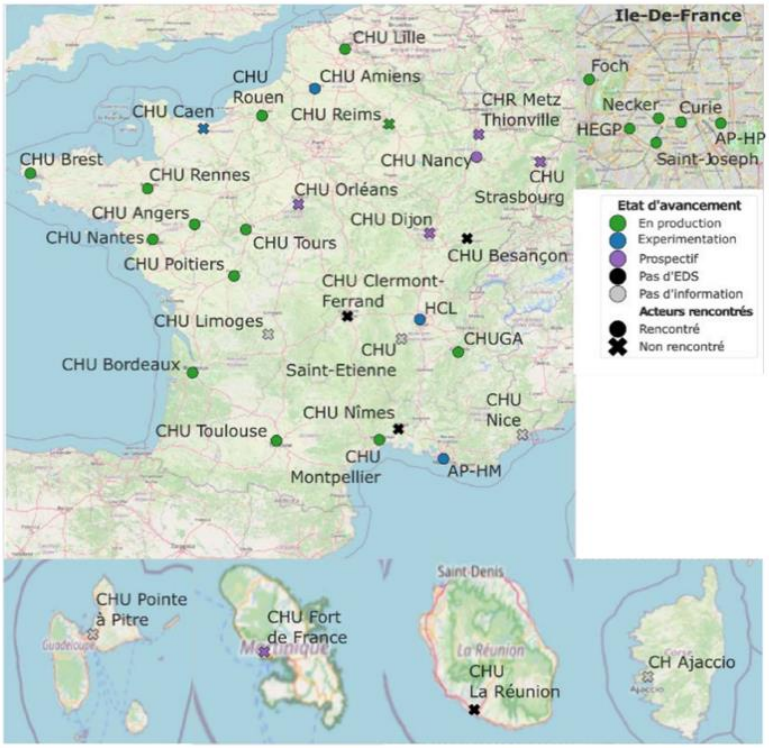


Figure 4 – Répartition des EDS sur le territoire français en 2022 [19]

Par ailleurs, les EDSH se développent également dans une perspective de structuration à l'échelle régionale et nationale. Créé en 2019, le Groupement d'Intérêt Public (GIP) Health Data Hub (HDH) est une plateforme collaborative qui vise à croiser l'ensemble des données de santé dont nous disposons sur le territoire. Le Health data hub a pour objectif de faciliter l'accès et l'utilisation des données de santé par les différents acteurs de recherche et de développement afin de promouvoir l'innovation en matière de santé numérique. Cette plateforme implique une collaboration entre plusieurs acteurs, dont l'assurance maladie, le ministère de la santé, les établissements de santé et les chercheurs **[20]**.

5. Base de données de réanimation en libre accès

5.1. Medical Information Mart for Intensive Care (MIMIC) database

La base de données MIMIC est la plus ancienne et la plus documentée des bases de données en accès libre en soins intensifs. Issue d'un partenariat de recherche parrainé par le National Institutes of Health (NIH), elle combine les ressources d'une équipe interdisciplinaire composée du *Massachusetts Institute of Technology* (MIT), *Philips Healthcare*, et le *Beth Israel Deaconess Medical Center* (BIDMC).

Développée en 2000, la première version de la base MIMIC contenait essentiellement des séries chronologiques de signes vitaux capturés à partir de moniteurs de 121 patients admis en soins intensifs. Entre 2001 et 2019, l'implémentation progressive des versions II, III et IV a permis de regrouper plusieurs sources de données comprenant deux systèmes d'information utilisées en soins intensifs au cours de la période de collecte de données (*Philips CareVue* et *iMDsoft MetaVision*) couplés à des informations issues de dossiers médicaux des hôpitaux et des laboratoires :

- Informations démographiques
- Mortalités en réanimation et à l'hôpital
- Mesures physiologiques horodatées en réanimation
- Résultats des tests de laboratoire en réanimation et à l'hôpital
- Prescriptions médicamenteuses en réanimation
- Rapports des examens de radiologie en réanimation et à l'hôpital
- Notes cliniques d'évolution documentées par les acteurs du soin
- Informations relatives à la facturation telles que les codes de la classification internationale des maladies (CIM-9), les codes de groupe de diagnostic (DRG) et les codes de procédure (CPT).

Actuellement, la version MIMIC-IV compte 50 048 patients admis en soins intensifs entres 2008 et 2019 **[21]**.

5.2. eICU Collaborative Research Database (eICU CRD)

La base de données eICU CRD est une vaste base de données multicentrique de soins intensifs mise à disposition par *Philips Healthcare* et développée en partenariat avec le laboratoire d'informatique du MIT **[22]**.

Cette base de données agrège uniquement les données de télésurveillance enregistrées en soins intensifs avec notamment : les informations démographiques, la mortalité en réanimation, les mesures physiologiques, les notes d'évolution, les prescriptions médicamenteuses, les résultats de test de laboratoires et les antécédant médicaux via les codes de la classification internationale des maladies (CIM-9).

Les données de la base eICU CRD portent sur 139 367 patients uniques admis en soins intensifs dans 208 hôpitaux à travers les Etats-Unis entre 2014 er 2015.

5.3. Amsterdam University Medical Center data base (AmsterdamUMCdb)

AmsterdamUMCdb est la première base de données de soins intensifs librement accessible en Europe. Elle est le produit de la collaboration entre la société néerlandaise de médecine de soins intensifs (NVIC), en particulier de son réseau Research Collaboration Critical Care (RCCnet) et de l'hôpital universitaire d'Amsterdam **[23]**.

Elle regroupe les données de 20 109 patient admis en soins intensifs entre 2013 et 2016. Outre son caractère monocentrique, l'absence d'information sur les antécédents médicaux est sa principale faiblesse. Néanmoins, de nombreux centres aux Pays-Bas ont publiquement exprimé leur intention future de partager les données relatives aux patients traités dans leurs services de soins intensifs.

5.4. High time-resolution intensive care unit dataset (HiRID)

HiRID est le jeu de données le plus récent et donc le moins exploré. Cet ensemble de données a été développé en coopération avec l'institut fédéral suisse de technologie (ETH) de Zurich **[24]**.

Il recense les données de 33 905 admis en unité de soins intensifs à l'hôpital universitaire de Berne en Suisse entre 2005 et 2016. Comme la base AmsterdamUMCdb, HiRID ne contient pas d'information sur les antécédents médicaux. Cependant, les données physiologiques sont stockées avec une résolution temporelle exceptionnellement élevée d'une entrée toutes les deux minutes (contre une entré toutes les trente minutes pour MIMIC-IV à titre de comparaison).

Table 1 – Caractéristiques des bases de données étudiées [3]

	MIMIC-IV	**eICU CRD**	**AmsterdamUMCdb**	**HiRID**
Centres	Boston, Etats-Unis	Multicentrique, Etats-Unis	Amsterdam, Pays-Bas	Berne, Suisse
Nombre de centres	1	208	1	1
Période d'inclusion	2008-2019	2014-2015	2013-2016	2005-2016
Nombre de patients	50 048	139 367	20 109	33 905
Information démographique	Oui	Oui	Oui	Oui
Mortalité en soins intensifs	Oui	Oui	Oui	Oui
Mortalité à l'hôpital	Oui	**Non**	**Non**	**Non**
Mesures physiologiques	Oui	Oui	Oui	Oui
Résultats de laboratoires	Oui	Oui	Oui	Oui
Prescriptions médicamenteuses	Oui	Oui	Oui	Oui
Images/Rapports de radiologie	Oui	**Non**	**Non**	**Non**
Antécédant médicaux	Oui	Oui	**Non**	**Non**
Notes cliniques d'évolution	Oui	Oui	**Non**	**Non**

6. Objectifs de la thèse

L'objectif de ce travail est d'étudier la contribution de ces bases de données dans la recherche médicale en réanimation. Secondairement et à partir de ces observations, développer notre propre entrepôt de données de réanimation à l'échelle locale au CHU de Rouen.

Partie II : Contribution of open access databases to intensive care medicine research: A scoping review

Background: Intensive care unit (ICU) handles the most critical patients with a high risk of mortality. Due to those conditions, close monitoring is necessary and therefore a large volume of data is collected. Collaborative ventures have enabled the emergence of large open access databases, leading to numerous publications in the field.

Objective: The aim of this scoping review is to identify the characteristics of studies using open access intensive care databases and to describe the contribution of these studies to intensive care research.

Methods: The research was carried out from 3 databases (PubMed – Medline, Embase, Web of science) from the creation of the database to August 1st, 2022. Were included, the original articles based on 4 open databases of adult patients admitted to intensive care units (Amsterdam University Medical Centers Database (AmsterdamUMC), Collaborative Research Database (eICUCRD), High time resolution ICU dataset (HiRID), Medical Information Mart for Intensive Care (MIMIC)). Characteristics relating to publication review, study design and statistical analyses were extracted and analyzed.

Results: We observed a consistent increase in the number of publications from these databases since 2016. MIMIC databases were the most frequently used, while the countries contributing the most were China and the United States with 683 (52.8%) and 367 (28.4%) publications respectively. The median impact factor of publications, since the creation of open databases in intensive care, is 3.8 [2.8 - 5.8]. Cardiovascular and infectious topics were the most represented with 333 (25.7%) and 319 (24.7%) articles. Regarding statistical methods, logistic regression was the most commonly employed model for both inference and prediction questions with 383 (55.5%) and 276 (47.3%) studies. A majority of the inference studies presented statistically significant results (84.1%). In prediction studies, the most recurrent performance measure was the AUC, with a median value of 0.840 [0.780 – 0.890].

Conclusions: The abundance of scientific outputs resulting from these databases and the diversity of topics addressed highlight the importance of these databases as valuable resources for clinical research and suggest their potential impact on clinical practice in intensive care. However, the quality of studies and their clinical relevance remain highly heterogeneous, with the majority of articles published in low-impact journals.

1. Introduction

Intensive care units (ICUs) provide care for critical patients at high risk of morbidity and mortality. These patients, due to their severity, require continuous monitoring and surveillance of clinical, biological and imaging parameters. This generates a large amount of data, usually collected in electronic health records (EHRs). The exploitation of this data has become a key issue in recent years **[25]**. In addition to traditional epidemiological and clinical research, the use of databases allows the emergence of new research themes such as the construction of diagnostic tools, decision support models, or predictive models of therapeutic response **[26,27]**. However, due to the sensitive nature of health data and the technical, legal and ethical challenges, medical data are still difficult to access **[28]**.

During the last decades, collaborative ventures have enabled the emergence of large open access databases **[23]**. Building on the opportunity of digital transformation, they facilitate data sharing on a large scale and enable knowledge creation more efficiently. Moreover, they are part of an ecosystem in which science is more transparent, reproducible and an effective lever for scientific integrity **[29]**. Since the 2000s, we have seen the release of several open access ICU databases **[3]**. The best-known example is the Medical Information Mart for Intensive Care (MIMIC) database, which integrates anonymized, comprehensive clinical data from more than 50,000 intensive care admissions from Beth Israel Deaconess Medical Center in Boston, Massachusetts **[30,31]**. These open databases have led to the production of numerous research works.

Currently, no effort has been made to analyze the medical literature generated from these open intensive care databases. Previous systematic reviews were interested in describing these databases to determine their exploitation potential **[3]** or were interested in machine learning techniques used to train models from these databases **[32]**. However, none of them were interested in the research themes and their potential impact on clinical practice.

We propose this scoping review including all publications of clinical research based on open intensive care databases. A scoping review will be conducted as this is the most suitable research method to map a research area **[33]**. We will first observe the bibliometric characteristics and the authors of these publications. Then, in a second step, we will investigate the research themes and the methodologies used. The objective is to glimpse the contribution of these open databases in intensive care clinical research.

2. Methods

The protocol adheres to the reporting guidance provided in the Preferred Reporting Items for Systematic Reviews and Meta-analyses extension for Scoping Reviews (PRISMA-ScR) **[34]** (see checklist, **Supplementary Materials, Table S1**). The design of this study is conceived according to the Joanna Briggs Institute Reviewers' manual for evidence synthesis **[35]**.

2.2. Eligibility criteria

We selected studies based on the following criteria: 1/ they were original articles based on open databases of adult patients admitted to intensive care units, covering all types of interventions and outcome measures; 2/ they were studies with a clinical aim (diagnostic, therapeutic, prognostic), where "clinical" pertains observations made on actual patients as opposed to theoretical, laboratory, or computer-based studies.

Exclusion criteria were: 1/ studies on patients admitted outside of ICUs; 2/ studies involving a pediatric population; 3/ studies where open databases were solely used for external validation; 4/ studies not written in English; 4/ systematic reviews and meta-analyses.

2.3. Search strategy

First step, the literature was searched for open access ICU databases. A recent review systematically identified publicly available adult clinical ICU databases **[3]**: Amsterdam University Medical Centers Database (AmsterdamUMCdb) **[23]**, eICU Collaborative Research Database (eICU-CRD) **[22]**, High time-resolution ICU dataset (HiRID) [24], Medical Information Mart for Intensive Care (MIMIC) clinical databases version II, III and IV **[21,30,36]**.

We then used PubMed-Medline (US National Library of Medicine), Embase (Elsevier), and Web of science (Clarivate Analytics) to search articles published from the creation of the database to August 1st, 2022. The search terms of the different databases are included in the **Supplementary Materials** (**Table S2**). A senior librarian developed and validated search queries. Lastly, the references of the included articles were examined for potentially relevant medical articles that may have escaped the literature search.

2.4. Selection

Two independent reviewers (B. Popoff, J. Kallout) conducted the search strategy and retrieved the references. Any disagreements were resolved through discussion between the two reviewers, or if necessary, by a third reviewer (A. Lamer). An initial selection was made after examining the titles and abstracts of the medical articles resulting from the search strategy. A more meticulous selection followed after evaluating the full text of potentially eligible articles. Disagreements were reconciled by consensus or, when necessary, through arbitration. The selection process was facilitated using the Rayyan software®.

2.5. Data Extraction, Collection and Analysis

Medical articles were referenced by their Digital Object Identifier (DOI), the name of the first author, and the date of publication. Two reviewers (J. Kallout, B. Popoff) independently extracted data using a predesigned electronic form, which was pilot tested with a random sample of 20 articles. Upon achieving consistent data abstraction ($\kappa \geq 0.8$) **[37]**, reviewers proceeded with data extraction for the entirety of included articles using the Goupile® software **[38]**. Collected data details are outlined in **Supplementary Materials** (**Table S3**). Results were aggregated into a table, organized with one column per database and one row per variable.

A descriptive analysis was then be carried out on the gathered data utilizing R® 4.1.2 software. Categorical variables were summarized as counts and a frequencies per category, while quantitative variables were expressed as medians and interquartile ranges. Bar charts were employed to visualize the distribution of categorical variables and histograms and density curves for the quantitative variables. Additionally, the temporal evolution of the studied variables was analyzed.

3. Results

3.1. Study selection

From the inception of these databases up to August 2022, a total of 4466 publications were identified. After excluding duplicates, 2063 (46.2%) titles and abstracts were retained. Among these articles, 1347 (65.3%) met the inclusion criteria. Ultimately, 1294 (96.1%) articles were selected for the final analysis after thorough reading. The selection flowchart is shown in **Figure 1**, with the characteristics of the included studies detailed in **Table 1**.

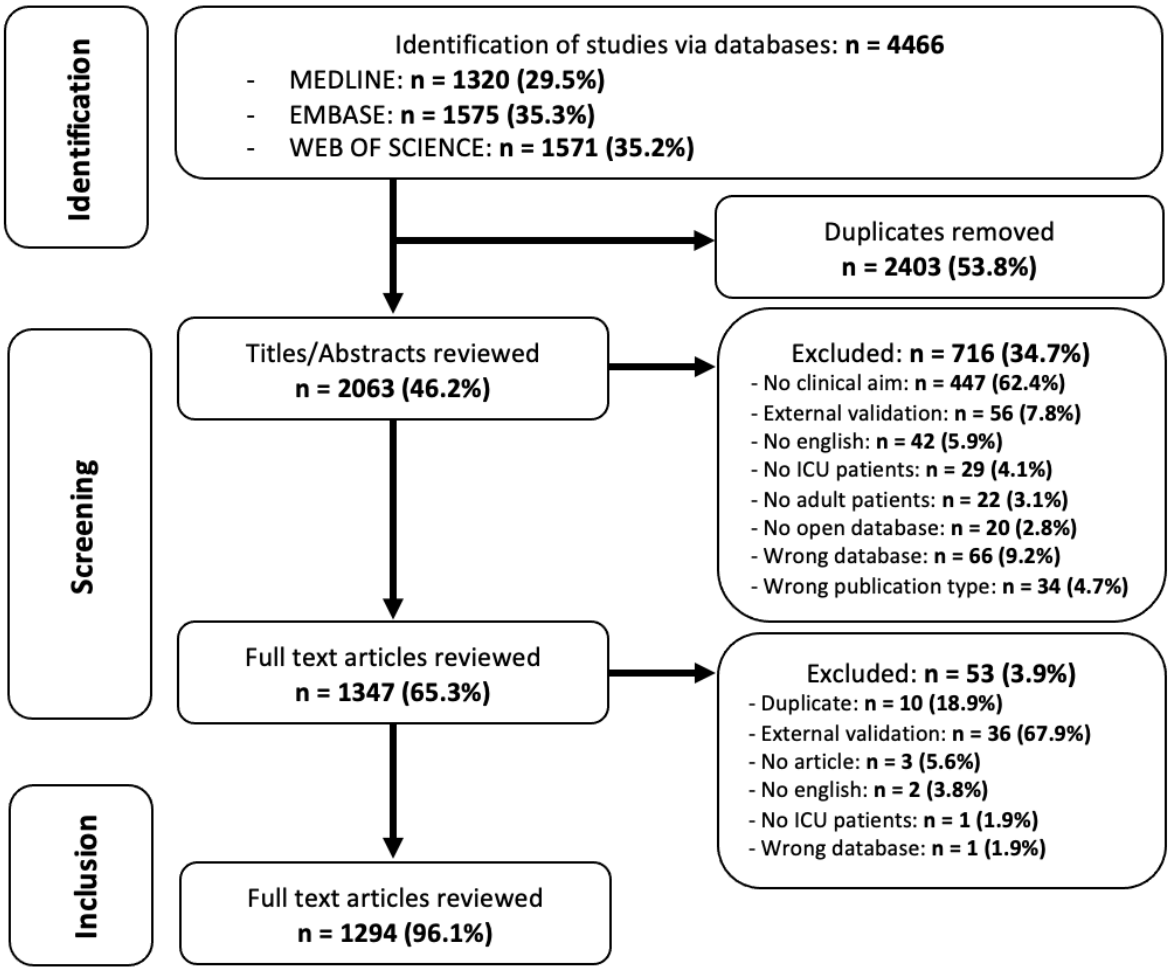


Figure 1 – Flow chart of scoping review
Abbreviations: ICU, Intensive Care Unit.

Table 1 – Characteristics of studies

Abbreviations: AJRCCM, American Journal of Respiratory and Critical Care Medicine; AmsterdamUMCdb, Amsterdam University Medical Centers Database; eICU-CRD, eICU Collaborative Research Database; HiRID, High time-resolution ICU dataset; MIMIC, Medical Information Mart for Intensive Care.

	Study (n = 1,294)
Article information	
Date of publication	
- 2000 – 2010	12 (0.9%)
- 2011 – 2015	38 (2.9%)
- 2016 – 2020	293 (22.7%)
- Post-2020	951 (73.5%)
Database used	
- AmsterdamUMCdb	6 (0.5%)
- eICU-CRD	195 (15.1%)
- HiRID	0
- MIMIC II Clinical database	105 (8.1%)
- MIMIC III Clinical database	895 (69.2%)
- MIMIC IV Clinical database	197 (15.2%)
Female corresponding author	346 (26.7%)
Country of corresponding author (top 10)	
- China	683 (52.8%)
- United States	367 (28.4%)
- United Kingdom	39 (3.0%)
- Canada	17 (1.3%)
- Germany	16 (1.2%)
- Japan	13 (1.0%)
- Australia	10 (0.8%)
- India	10 (0.8%)
- Singapore	10 (0.8%)
- Spain	10 (0.8%)
- Other	107 (8.2%)
Journal information	
Journal name (top 10)	
- Frontiers in Medicine	52 (4.0%)
- Critical Care Medicine	41 (3.2%)
- AJRCCM	40 (3.1%)
- Frontiers in Cardiovascular Medicine	37 (2.9%)
- International Journal of General Medicine	36 (2.8%)
- Intensive Care Medicine Experimental	34 (2.6%)
- Scientific Reports	31 (2.4%)
- Critical Care	26 (2.0%)
- Plos One	25 (1.9%)
- Chest	24 (1.9%)
Impact factor	3.8 [2.8 – 5.8]
Study information	
Inclusion period (years)	11 [11 – 11]
Number of participants	4,323 [1,478 – 12,956]

3.2. Article information

Since 2016, there was a consistent rise in the number of publications utilizing these databases. Most studies were published after the year 2020, with 951 (73.5%) publications. The MIMIC databases were the most frequently used. No publications employing the HiRID database were identified (**Table 1**, **Figure 2**). China and the United States were the leading countries in terms of publications, with 683 (52.8%) and 367 (28.4%) articles respectively. European countries contributed to 109 (8.4%) publications, led by the United Kingdom leading with 39 (3.0%) publications (**Table 1**, **Figure 3**).

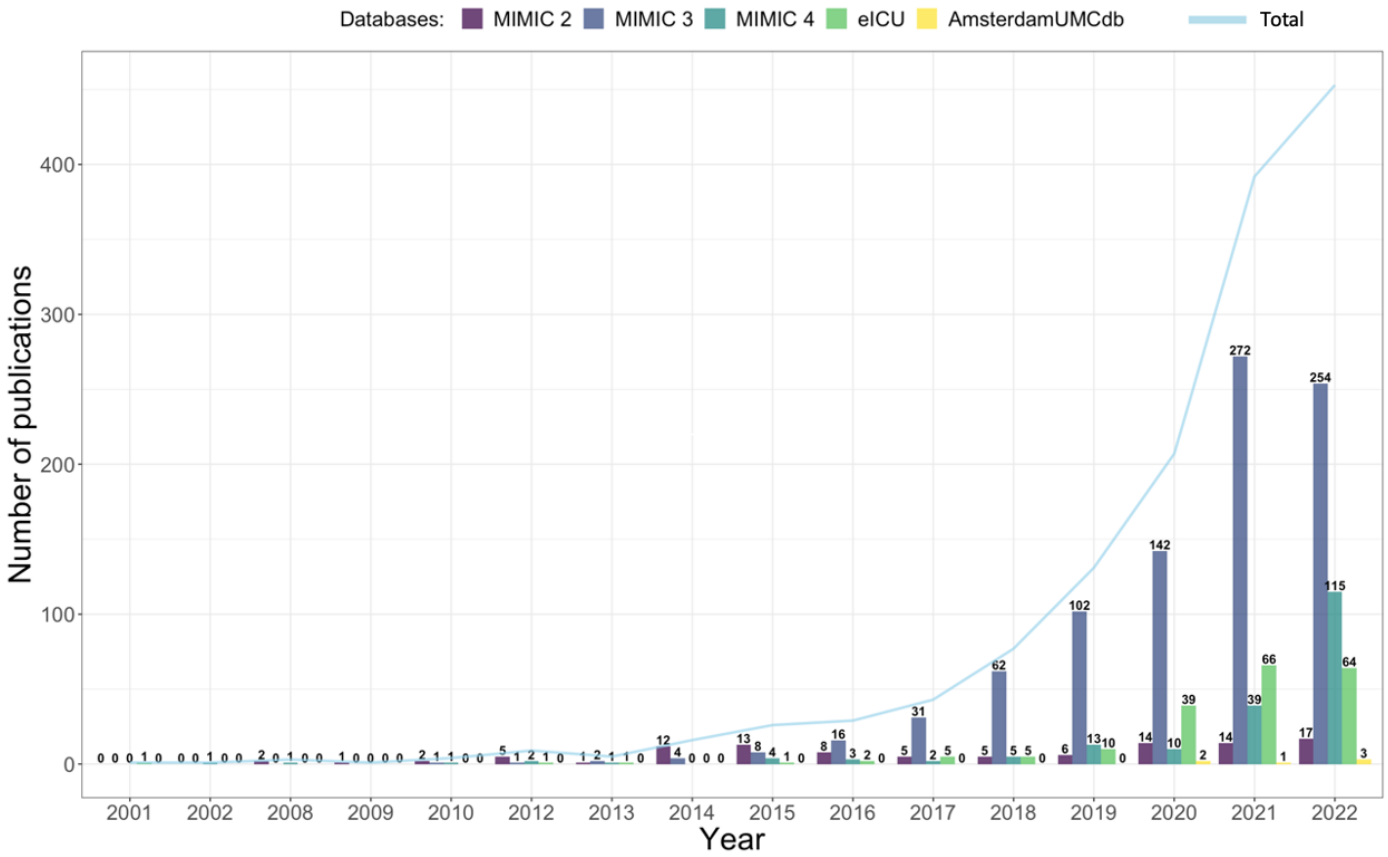


Figure 2 – Evolution of the number of publications over the years
Abbreviations: AmsterdamUMCdb, Amsterdam University Medical Centers Database; eICU-CRD, eICU Collaborative Research Database; MIMIC, Medical Information Mart for Intensive Care.

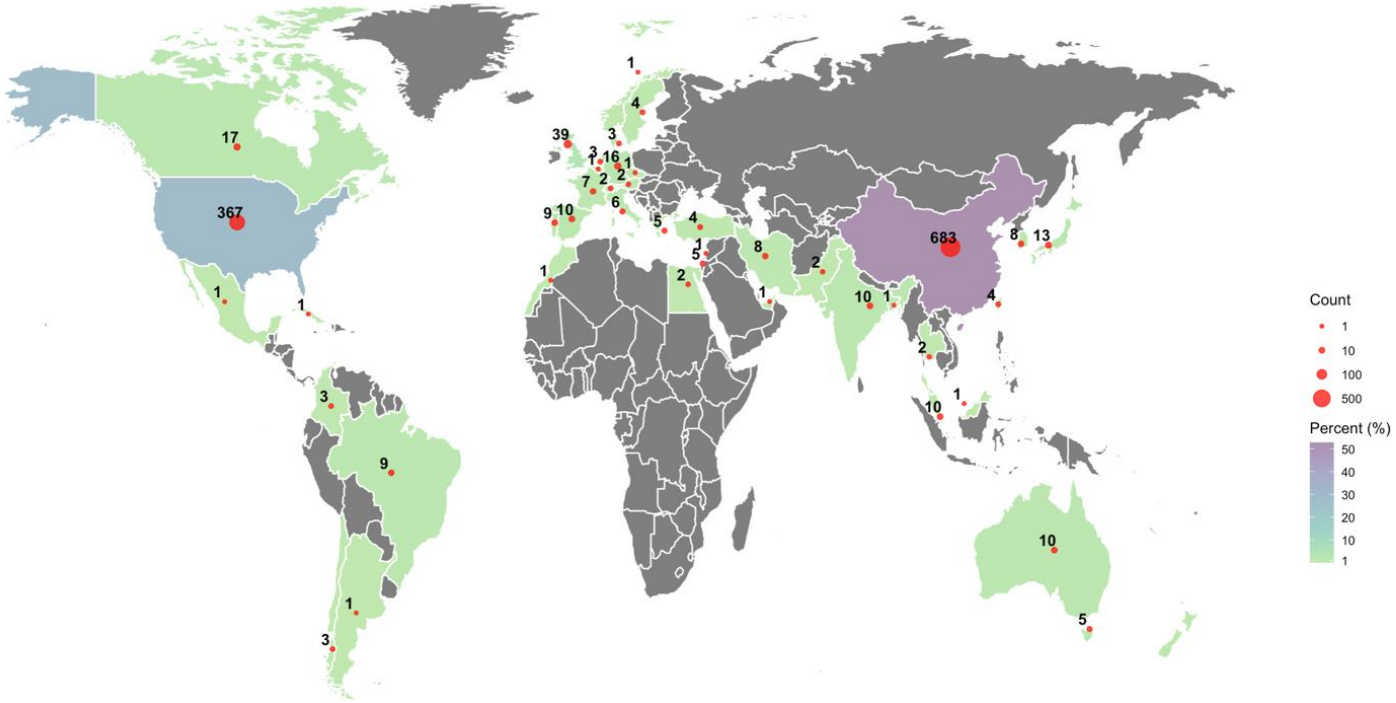


Figure 3 – Worldwide publications number distribution

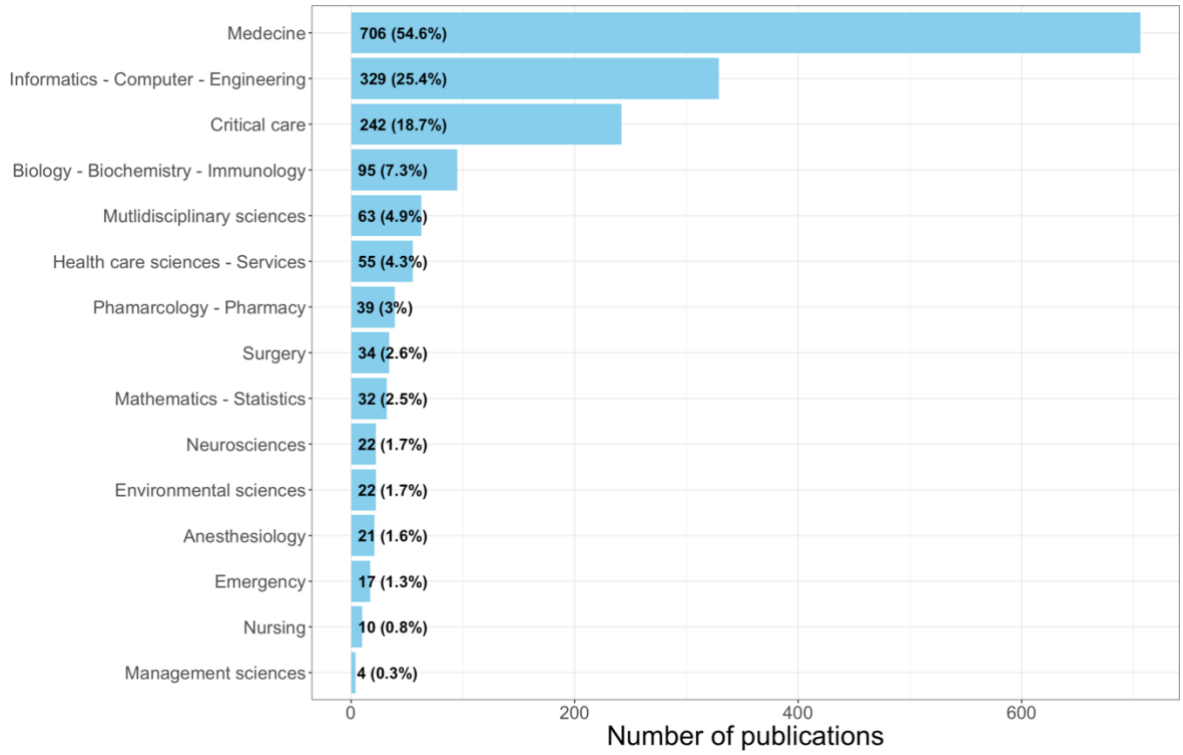


Figure 4 – Field of the journal

3.3. Journal information

The median impact factor of publications since the creation of open databases in intensive care is 3.8 [2.8 - 5.8] (**Table 1**). As for the journal field, the most commonly represented journals are those in medicine, computer science, and intensive care, with 706 (54.6.%), 329 (25.4%) and 242 (18.7%) publications respectively (**Figure 4**).

3.4. Study information

Studies included a median of 4,323 patients [1,478 – 12,956] with a maximum sample size reaching up to 219,306 by combining patients from several databases. The median inclusion period was 11 years [11 – 11] (**Table 1**).

Research predominantly explored the cardiovascular system, particularly hemodynamics, and infectious diseases, specifically sepsis, with 333 (25.7%) and 319 (24.7%) publications respectively. Renal failure and metabolic disorders were also heavily studied, with 231 (17.9%) publications, along with respiratory failure and mechanical ventilation, which had 201 (15.5%) publications. Notably, many studies investigated overall mortality among critically ill patients admitted to the ICU, with 178 (13.8%) publications (**Figure 5**).

Furthermore, 422 (32.6%) studies investigated the entire population admitted to the ICU. A majority of studies focused on patients experiencing specific organ failures, represented in 812 (62.8%) publications (**Figure 6**). Various exposure or prediction factors were evaluated. Biological data and vital signs were the most commonly analyzed, featuring in 664 (51.3%) and 421 (32.5%) studies, respectively. Demographic information, medication prescriptions, and comorbidities were also widely included, with 329 (25.4%), 278 (21.5%) and 270 (20.9%) publications, respectively (**Figure 7**). The primary outcome most frequently assessed was mortality, whether it occurred in the ICU, the hospital, or after a certain period of time, with 871 (67.3%) studies (**Figure 8**).

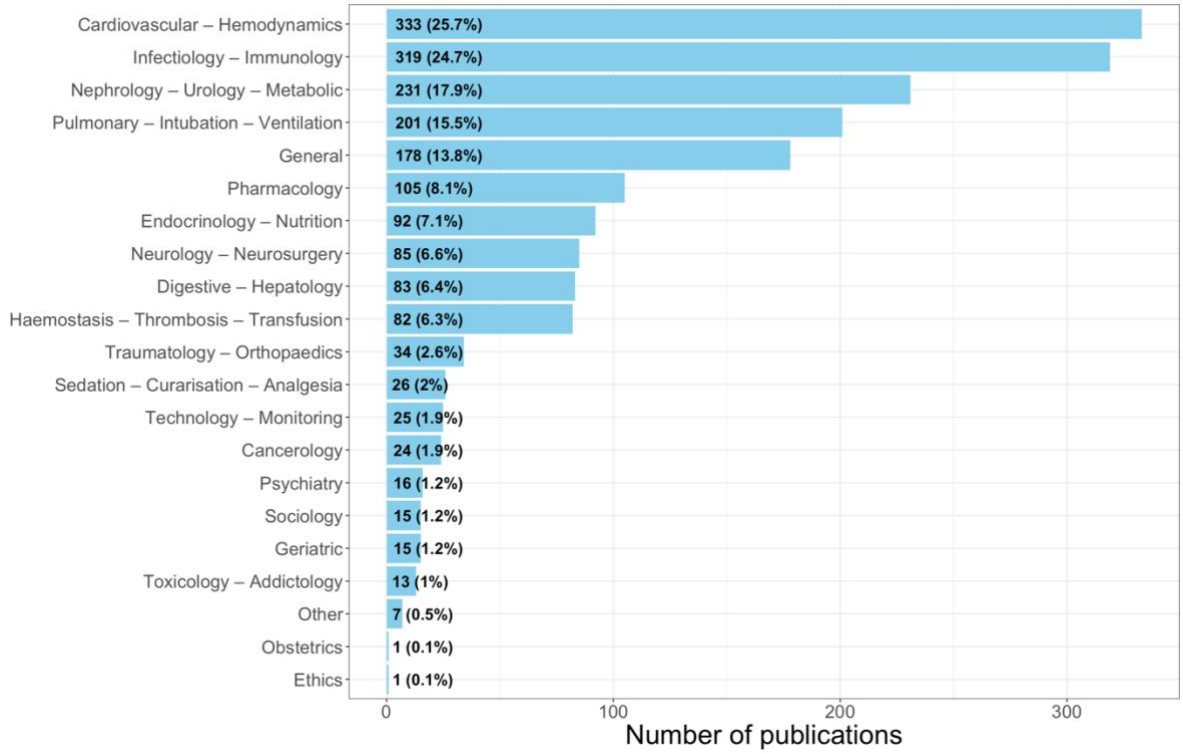


Figure 5 – Research topics

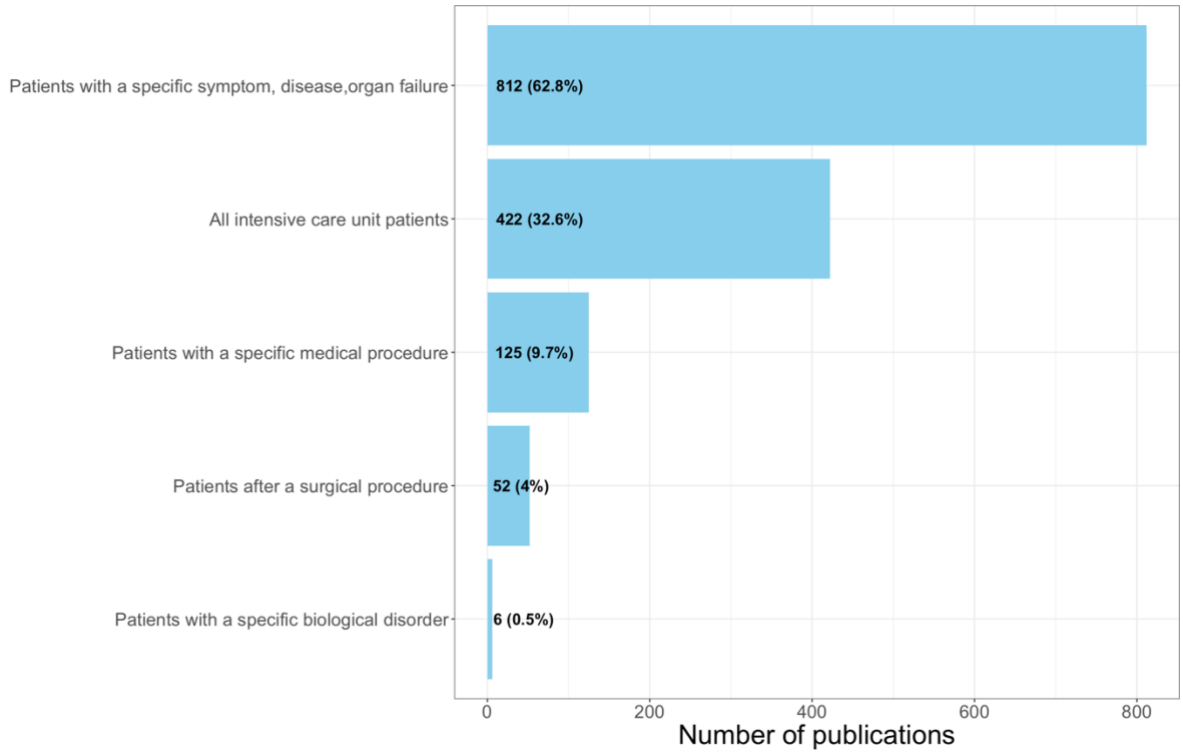


Figure 6 – Analyzed population

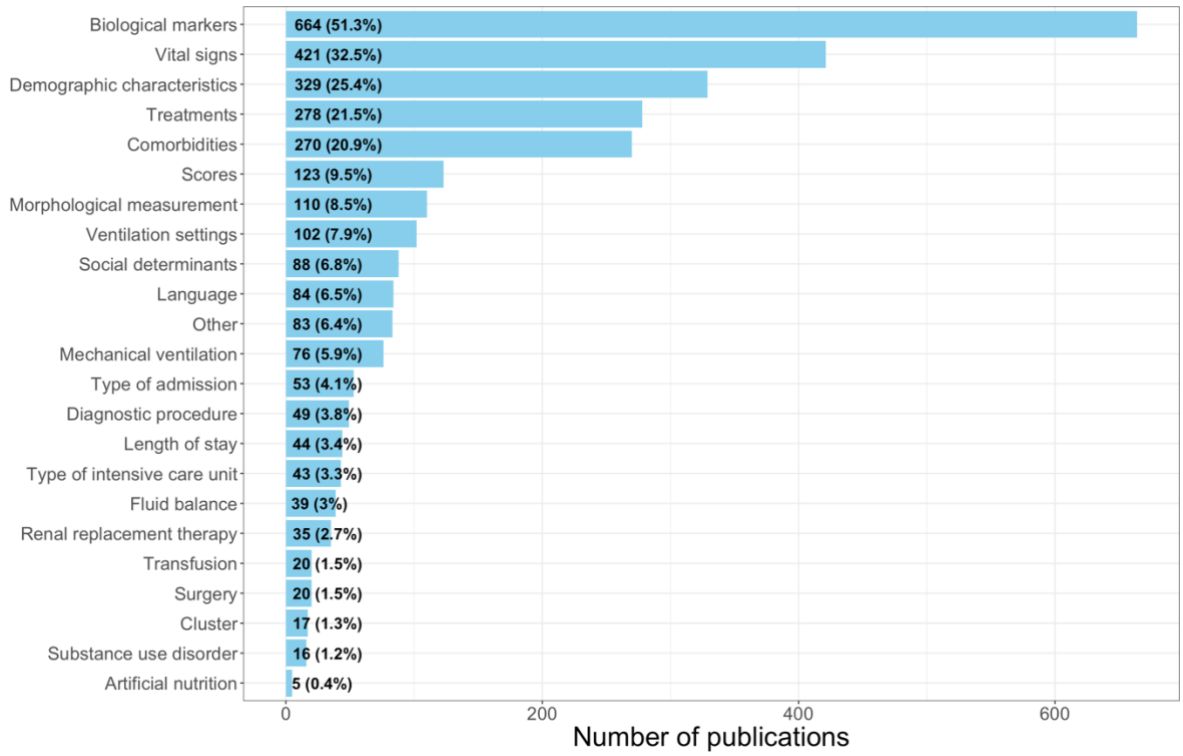


Figure 7 – Analyzed exposures

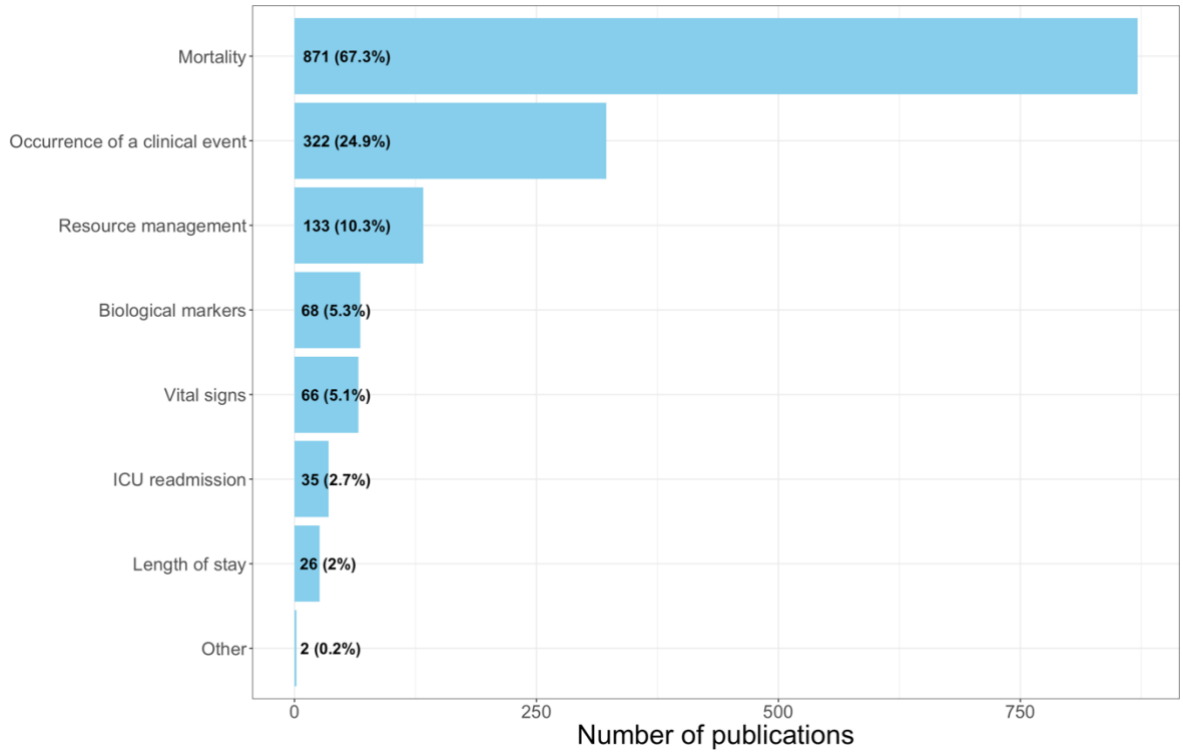


Figure 8 – Analyzed outcomes

3.5. Statistical methods used and results

Regarding the study objectives, 708 (54.7%) studies addressed an inference question, while 586 (45.2%) studies targeted a prediction question (**Table 2, Figure 9**). Supervised learning and deep learning methods were the most commonly applied machine learning techniques, with 424 (32.8%) and 244 (18.9%) publications respectively (**Figure 9**).

When examining specific models, logistic regression was the predominant choice for both inference questions with 383 (55.5%) publications (**Figure 10A**). Regarding prediction questions, logistic regression was the most widely used classical machine learning method, surpassing survival models and generalized models with 276 (47.3%), 29 (5.0%) and 11 (1.9%) publications respectively. Neural networks were the most frequently applied advanced machine learning methods outpacing tree-based methods and boosting methods, with 234 (40.1%), 184 (31.5%) and 167 (28.6%) publications respectively **(Figure 10B**).

Inference studies reported statistically significant results in 84.1% of cases. Furthermore, the prevailing performance measure in prediction studies was the area under the curve (AUC) with 424 (88.3%) studies. The median value of the AUC was 0.840 [0.780 – 0.890] (**Table 2**).

Pre-existing clinical scores were applied for comparison with the performance of prediction models implemented in the studies. The most commonly used scores were SAPS (Simplified Acute Physiology Score), SOFA (Sepsis-related Organ Failure Assessment), and APACHE (Acute Physiology and Chronic Health Evaluation) scores, with 55 (43.0%), 24 (18.8%) and 14 (10.9%) publications, respectively. The median AUC for these scores was 0.732 [0.677 – 0.782] (**Table 2**).

Table 2 – Statistical analyses

Abbreviations: AUROC, Area Under the Curve; RMSE, Root Mean Square Error; SAPS, Simplified Acute Physiology Score; SOFA, Sepsis-related Organ Failure Assessment; APACHE, Acute Physiology and Chronic Health Evaluation; APS, Acute Physiology Score; MEWS, Modified Early Warning Score; OASIS, Oxford Acute Severity of Illness; OR, Odds-Ratio; HR, Hazard-Ratio.

	Study (n = 1,294)
Statistical methods used	
Aim of the study	
- Inference	708 (54.7%)
- Prediction	586 (45.2%)
Effect-size measure (if inference)	
- Odds-Ratio	379 (59.1%)
- Hazard-Ratio	248 (38.7%)
- Coefficient	14 (2.2%)
Prediction performance measure (if prediction)	
- AUROC	424 (88.3%)
- Sensibility/Specificity	143 (29.8%)
- Accuracy	139 (28.8%)
- F1-Score	95 (19.8%)
- C-Index	74 (15.4%)
- Recall	42 (8.8%)
- RMSE	8 (1.7%)
Known prediction scores used for comparison (if prediction)	
- SAPS	55 (43.0%)
- SOFA	24 (18.8%)
- APACHE	14 (10.9%)
- APS	9 (7.0%)
- MEWS	4 (3.1%)
- OASIS	3 (2.3%)
- Other (< 3 publications)	19 (14.8%)
Key Findings Obtained	
Effect-size value (OR/HR)	
- Protective effect	0.65 [0.50 – 0.79]
- Adverse effect	1.64 [1.28 – 2.35]
P-value	
- > 0.05	97 (15.9%)
- 0.01 – 0.05	151 (24.8%)
- 0.001 – 0.01	94 (15.4%)
- < 0.001	267 (43.8%)
Performance of models used (AUC)	0.840 [0.780 – 0.890]
Performance of known scores used (AUC)	0.732 [0.677 – 0.782]

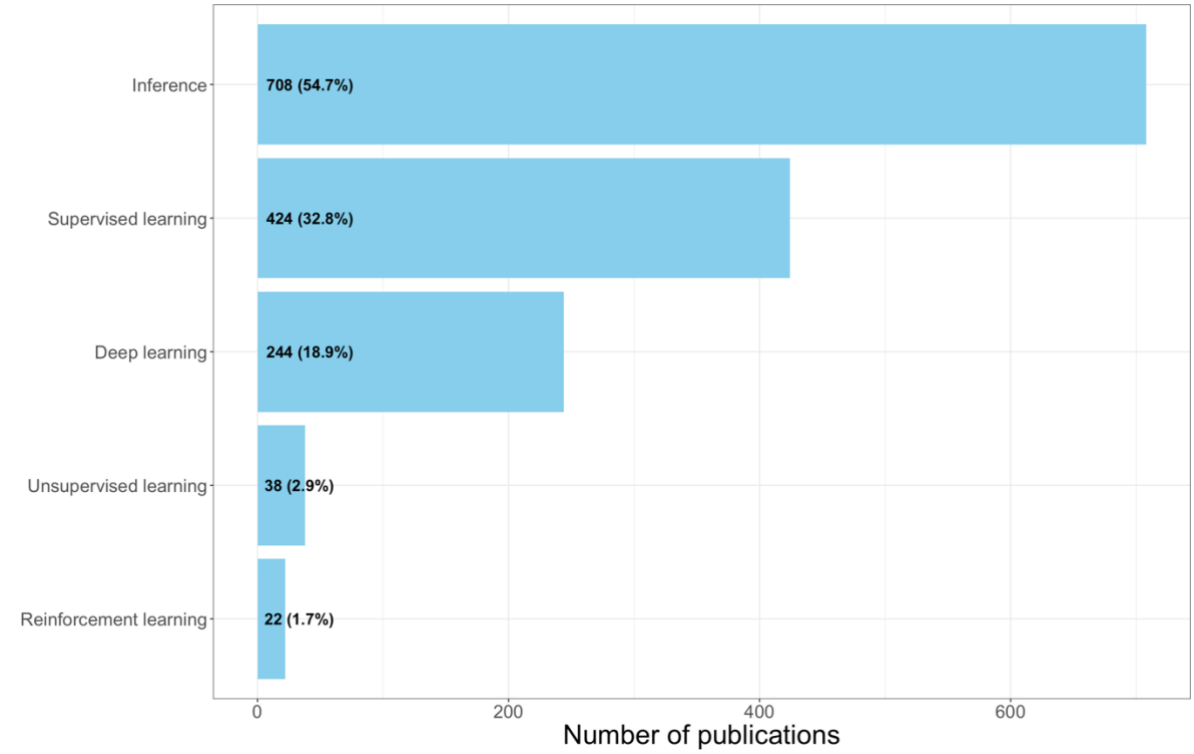


Figure 9 – Algorithm used

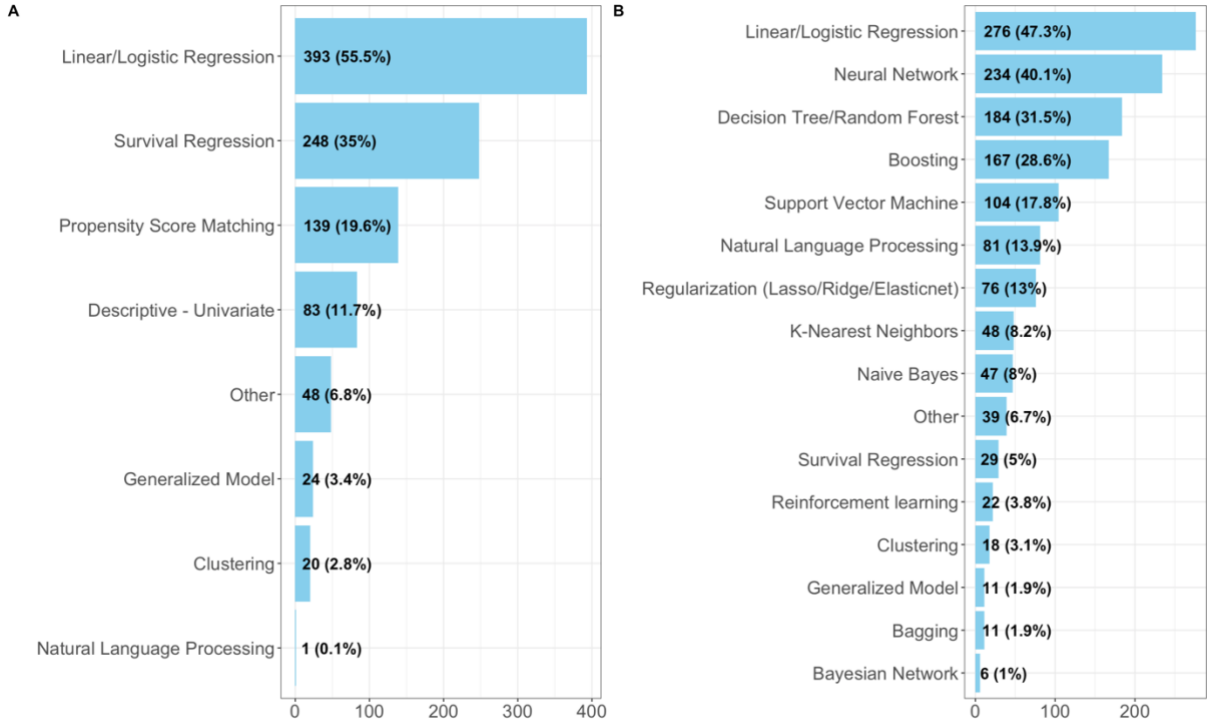


Figure 10 – Specific model used, inference (A) and prediction (B)

4. Discussion

4.1. Main findings

This review examined clinical publications from open databases in the field of intensive care. We observed a consistent increase in the number of publications from these databases since 2016, with the majority of articles being published after 2020. MIMIC databases were the most frequently used, while the countries contributing the most were China and the United States. Most studies were published in journals with a low impact factor. Cardiovascular and infectious topics, particularly those related hemodynamics and sepsis, were the most represented. Other significant subjects included renal failure coupled with metabolic disorders and respiratory failure with mechanical ventilation. The most studied outcome measure was mortality in the ICU. Biological data, vital signs, demographic details, medication prescriptions, and comorbidities were extensively used as exposure or predictor factors.

Regarding statistical methods, logistic regression was the most commonly employed model for both inference and prediction questions. Neural networks were the most frequently used advanced machine learning methods surpassing supervised and reinforcement methods. A majority of the inference studies presented statistically significant results. In prediction studies, the most recurrent performance measure was the AUC, with a median value of 0.840.

4.2. Results in context with literature

Scientific contribution

The rising number of publications using open databases is steadily increasing, highlighting the potential impact of these resources on clinical research within intensive care. Collaborative initiatives have led to the emergence of large open-access databases, enhancing data availability in this field. Researchers have now access to data from numerous centers, enabling the production of robust results via larger sample sizes. Furthermore, this access has significantly boosted statistical power, enabling the study of specific topics and populations that were previously overlooked due to insufficient sample sizes. In our review, the median sample size in the studies was 4,323 [1,478 – 12,956], in contrast to typical sample sizes in prospective studies barely reaching a few hundred patients. In a literature review by *van de Sande and al.* studying machine learning in intensive care, the median sample size was 179 [94 – 1,411] and 142 [40 – 380] across all prospective observational and clinical studies, respectively **[39]**.

The wealth and diversity of information contained in these databases have opened up a broad spectrum of research possibilities, resulting in a wide array of topics covered in publications. Our review categorized these into over 20 different categories. The cardiovascular, infectious, respiratory, and metabolic systems emerged as the most commonly studied research domains, mirroring the reality of clinical practice in intensive care **[40]**. Mortality in ICU was the most frequently analyzed outcome measure. In a literature review by *Syed et al.*, which examined the application of machine learning using the MIMIC dataset, mortality prediction was also the most studied outcome measure, followed by sepsis prediction, cardiac events, and acute kidney injury prediction **[32]**.

Despite the significant number and diversity of publications, the clinical relevance of studies from these databases remains heterogeneous. In our review, the majority of studies were published in journals with low impact factors. The results were often statistically significant but had limited effect sizes. Nevertheless, there are few criteria to objectively evaluate the clinical relevance of a publication. Additionally, the quantitative nature of our analyses did not allow us to fully explore this aspect. A qualitative methodology specifically assessing clinical relevance would provide more insights and discussion points.

Towards precision intensive care medicine

In recent years, due to advancing technology, data availability, and the need to analyze increasingly larger databases, machine learning has emerged to develop increasingly accurate predictive models. Machine learning is a field of computer science and a part of artificial intelligence that defines both the science and engineering for which computer systems can analyze data and "learn" from the information contained within **[41]**.

In our study, 586 out of 1,294 (45.2%) publications utilized at least one machine learning model, with a significant increase in their implementation after 2015. The most commonly used machine learning method was logistic regression and neural networks. *Shillan et al.* literature review, summarizing the characteristics and results of machine learning methods used in intensive care, found similar results. Nearly half of the studies identified in that review were published after 2015, and the most commonly used methods were neural networks, support vector machines, and decision trees **[42]**.

Although the use of machine learning models in intensive care has increased in the scientific literature, their adoption in actual clinical practice remains limited or even absent at present. Machine learning models are often complex and require specialized skills to develop and implement. Their interpretation also poses a significant barrier to their use, with the phenomenon known as the "black box" issue **[43]**. Moreover, these models need to be rigorously validated to ensure their reliability and accuracy. External validation of models on

independent datasets is crucial to evaluate their performance under real-world conditions. In our review, the vast majority of publications analyzed came from databases implemented in the United States, limiting the generalizability of these models to the rest of the population. Additionally, only 92 out of 2,063 (4.5%) articles focused on external validation using an independent dataset, which is consistent with *Shillan and al.* findings **[42]**. Furthermore, prospective clinical evaluation of these models is still rare. In a recent review, only 10 out of 494 (2.0%) articles clinically evaluated artificial intelligence in real clinical settings, with 5 studies being randomized clinical trials **[39]**. Finally, the use of machine learning models raises ethical and regulatory questions, particularly concerning data privacy, automated decision-making, and liability in case of errors. Clear guidelines and regulations must be established to govern the use of these models in a clinical context.

4.3. Strengths and Limitations of the study

This study, to our knowledge, is the first to examine the contribution of open databases in clinical research in intensive care. This study follows a rigorous methodology and its research protocol was made public prior to the study **[44]**. Furthermore, this review adopts a comprehensive approach by including all clinical research publications generated from open databases in ICU. By examining a wide range of publications, it provides a detailed overview of research themes, methodologies used, and results obtained. This can assist clinicians and researchers in understanding how these databases can contribute to improving patient care in intensive care and offering a solid foundation for future research and discussions.

However, it is important to acknowledge the limitations of this literature review. Firstly, only publications written in English were included, potentially introducing a linguistic bias. Secondly, there exists and inherent risk of publication bias in literature reviews. Published and accessible studies may not represent all research carried out on open databases in intensive care. Studies with negative or non-significant results are less likely to be published, potentially resulting in an overestimation of positive results. While several open databases in intensive care were included in the study, it is possible that there exist other databases that were not taken into account. This could limit the representativeness of the sampled studies. Lastly, despite exhaustive research, no publications from the HiRID database were identified. With these limitations in mind, it is important to consider the results of this study with caution and interpret them in the appropriate context.

5. Conclusion

Open databases in intensive care have facilitated clinical research and provided new perspectives for enhancing care in intensive care. The abundance of scientific outputs resulting from these databases and the diversity of topics addressed highlight the importance of these databases as valuable resources for clinical research and suggest their potential impact on clinical practice in intensive care. However, the quality of studies and their clinical relevance remain highly heterogeneous among publications and are challenging elements to evaluate.

6. Supplementary materials

Table S1 – Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

Section and Topic	Item	PRISAM-ScR checklist item	Reported on page number
TITLE			
Title	1	Identify the report as a scoping review.	34
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	34
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	35
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements or other relevant key elements used to conceptualize the review questions and/or objectives	35
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed and if available, provide registration information, including the registration number.	36
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria and provide a rationale.	36
Information sources	7	Describe all information sources in the search, as well as the date the most recent search was executed.	36
Selection of sources of evidence	8	State the process for selecting sources of evidence included in the scoping review.	36-37
Data charting process	9	Describe the methods of charting data from the included sources of evidence and any processes for obtaining and confirming data from investigators	37
Data collection process	10	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	37
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	54-56
Critical appraisal of individual sources of evidence	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis.	37

Table S2 – Search terms used for studies selection

Database	Search terms
Pubmed	"Medical Information Mart for Intensive Care"[Text word] OR "MIMIC database"[Text word] OR "MIMIC II"[Text word] OR "MIMIC III"[Text word] OR "MIMIC IV"[Text word] OR "Amsterdam University Medical Centers Database"[Text word] OR AmsterdamUMCdb[Text word] OR "eICU-CRD"[Text word] OR "eICU Collaborative Research Database"[Text word] OR "High time resolution ICU dataset"[Text word] OR HiRID[Text word]
Embase	TS=("Medical Information Mart for Intensive Care") OR TS=("MIMIC database") OR TS=("MIMIC II") OR TS=("MIMIC III") OR TS=("MIMIC IV") OR TS=("Amsterdam University Medical Centers Database") OR TS=(AmsterdamUMCdb) OR TS=(eICU-CRD) OR TS=("eICU Collaborative Research Database") OR TS=("High time resolution ICU dataset") OR TS=(HiRID)
Web of science	'Medical Information Mart for Intensive Care' OR 'MIMIC database' OR 'MIMIC II' OR 'MIMIC III' OR 'MIMIC IV' OR 'Amsterdam University Medical Centers Database' OR AmsterdamUMCdb OR eICU-CRD OR 'eICU Collaborative Research Database' OR 'High time resolution ICU dataset' OR HiRID

Table S3 – Collected variables

Variables	Definitions
Article information	
Title Digital	
Digital Object Identifier (DOI)	
Date of publication (YYYY)	
Database used - AmsterdamUMCdb - eICU-CRD - HiRID - MIMIC II Clinical database - MIMIC III Clinical database - MIMIC IV Clinical database	
Corresponding author name	
Corresponding author e-mail	
Corresponding author gender	
Corresponding author country	
Journal information	
Journal name	
Field of the journal	Categories according to the Journal Citation Reports (JCR) **[45]**.
Impact factor	
Study information	
Inclusion period (years)	From the first inclusion of the first database to the last inclusion of the last database (if several databases).
Number of participants	Sums of patients included (from different databases if several databases).
Research topics - General - Cardiovascular – Hemodynamics - Cancerology - Digestive – Hepatology - Endocrinology – Nutrition - Ethics - Geriatric - Hematology – Haemostasis – Transfusion - Infectiology – Immunology - Nephrology – Urology – Metabolic - Neurology – Neurosurgery - Obstetrics - Pharmacology - Psychiatry - Pulmonary – Intubation – Ventilation - Sedation – Curarisation – Analgesia - Technology – Monitoring - Toxicology – Addictology - Traumatology – Orthopedic - Other	Studies without specific theme (eg, mortality in critical patients). Not belonging to the previous categories.

Population	
- All intensive care unit patients	Publication studied the whole population of ICU patients.
- Patients with a specific symptom, disease or organ failure - Patients with a specific biological disorder - Patients after a surgical procedure - Patients with a specific medical procedure	Publication studied a specific population of ICU patients.
- Other	Not belonging to the previous categories.
Exposure/Predictor	
Demographic characteristics	Age, Gender.
Social determinants	The conditions in which people are born, grow, live, work and age.
Morphological measurement	Size and shape characters are quantified and reported as lengths or indices.
Substance use disorder	The persistent use of drugs despite substantial harm and adverse consequences as a result of their use.
Comorbidities	The simultaneous presence of two or more diseases or medical conditions in a patient.
Type of admission	Admission from the emergency department or for scheduled elective or emergency surgery.
Type of intensive care unit	Admission to a medical, surgical or specialized ICU.
Length of stay	Length of stay in ICU or hospital.
Vital signs	Clinical measurements that indicate the state of a patient's essential body functions.
Fluid balance	Measurement including intravenous fluids volume or output/input difference
Biological markers	Biological measures of a biological state evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a intervention.
Diagnostic procedure	Test used to help diagnose a disease or condition other than laboratory tests.
Mechanical ventilation	Use of invasive or non-invasive mechanical ventilation.
Ventilation settings	The inputs to a mechanical ventilator that determine the mode and how much support is provided for the patient.
Renal replacement therapy	Use of a renal replacement method.
Artificial nutrition	Administration of enteral or parenteral nutrition.
Transfusion	Administration of labile blood products
Treatments	A set of measures applied to cure a disease, relieve symptoms, or prevent their onset including pharmacological and electrolytic treatments.
Surgery	Recourse to surgical intervention or type of surgical intervention.
Scores	Medical decision support tool, aggregating in a single value several clinical observations.
Language	Variables obtained from a natural language processing method.
Cluster	Groups obtained from an unsupervised learning method.
Other	Not belonging to the previous categories.

Outcome	
Mortality	Explanation or prediction of mortality.
Occurrence of a clinical event	Explanation or prediction of a symptom, disease or organ failure.
Vital signs	Explanation or prediction of a vital signs.
Biological markers	Explanation or prediction of a biological markers.
Resource management	Explanation or prediction of therapeutic or management strategy.
ICU readmission	Explanation or prediction of discharge and readmission.
Length of stay	Explanation or prediction length of stay in ICU or hospital.
Other	Not belonging to the previous categories.
Statistical methods used and results	
Aim of the study	
Inference	Process of evaluating the relationship between the explanatory and response variables.
Prediction	Process of using a model to make a prediction about something that is yet to happen.
Algorithm used	
Unsupervised learning	A machine learning method whose goal is to describe the associations and patterns among a set of input measures without outcome measure **[46]**.
Supervised learning	A machine learning method whose goal is to predict the value of an outcome measure based on input measures **[46]**.
Reinforcement learning	A machine learning method where an agent is faced an a problem and that learns behavior through trial-and-error interactions with a dynamic environment **[47]**.
Deep learning	A machine learning method usually performed by an artificial neural network composed of several layers of neurons arranged hierarchically and interacting with each other to predict the value of an outcome measure **[46]**.
Specific model used	Exhaustive list of used models.
Effect-size measure (if inference)	Type of measurement of effect size of primary outcome (if several).
Effect-size value (if inference)	Value of effect size of primary outcome (if several).
P-value (if inference)	P-value of primary outcome (if several).
Prediction performance measure (if prediction)	Type of measurement of performance of better model used (if several).
Prediction performance value (if prediction)	Value of performance of better model used (if several).

Partie III : Development of an Intensive Care Data Warehouse at Rouen University Hospital: Transformation of the ICCA Database in the OMOP Common Data Model

Background: Intensive care units (ICUs) generate a vast amount of data, collected in electronic health records (EHRs), which are essential for improving clinical practices and patient care. However, the structural complexity of ICU databases and the lack of standardized data hinder their effective analysis and utilization.

Objective: This study aimed to transform the data from the ICCA® information system, widely used in ICUs at Rouen University Hospital, into the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM). The objective was to enable the integration, standardization, and large-scale analysis of ICU data.

Methods: The transformation process involved structural and semantic mapping of selected tables from the ICCA® database to the OMOP-CDM. An extraction, transformation, and loading (ETL) process were implemented to restructure and normalize the data into the target format.

Results: The study successfully transformed and integrated demographic and administrative data of 20,235 patients admitted to the ICUs at Rouen University Hospital. The transformed data allowed for the description of patient characteristics and the flow of patients within each ICU. All the generated information was aggregated in an interactive dashboard.

Conclusions: The study demonstrated the feasibility of transforming ICCA® ICU data into the OMOP-CDM, enabling their utilization and analysis at scale. Future perspectives include expanding the transformation process to include additional clinical concepts, addressing the challenge of structuring unstructured data from clinical notes, and promoting the creation of multicentric ICU data networks. These efforts will unlock the full potential of the OMOP-CDM model for research and improvements in intensive care, fostering significant advancements in critical care medicine.

1. Introduction

Intensive care units (ICUs) play a crucial role within healthcare providing specialized care to critically ill patients in need of immediate and specialized intervention. These patients experience one or more acute organ failures, in many cases leading to death. in the ICU, continuous monitoring of both clinical and paraclinical parameters is needed, leading to the accumulation of a substantial data collected in Electronic Health Records (EHRs). This data, pivotal for the amelioration of clinical practices and patient care quality, requires scrupulous collection, analysis, and interpretation **[48]**. Several information systems are available in the ICU settings and has proven to be an invaluable tool in managing these data in the ICU **[49]**. One example is ICCA® (IntelliSpace Critical Care and Anesthesia©, Koninklijke Philips N.V., The Netherlands), widely deployed across France, encompassing over 65 hospitals, including the Rouen University Hospital (RUH). It comprehensively records data related to ICU patients, including demographic, clinical, physiological, biological parameters, medical directives, and longitudinal clinical trajectories **[50]**.

Nevertheless, the structural complexity of the databases housing this renders them resistant to comprehensive analysis or utilization. Additionally, the disparities in data entry and storage methodologies across different hospitals hinder multi-multicentric studies. To capitalize on these extensive data repositories, the adoption of methods for data structuring and normalization, facilitating large-scale analysis, become imperative.

The Observational Health Data Sciences and Informatics (OHDSI) initiative, an open scientific community, is committed to enhancing healthcare outcomes through collaborative evidence-based innovation. To surmount the aforementioned challenges, OHDSI developed, in 2008, the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM), along with a wide range of open-source tools and methods. OMOP-CDM is designed to facilitates the structuring of EHRs and standardize vocabularies, enhancing data utilization and interoperability **[51]**. Although numerous databases have been successfully converted to the OMOP-CDM, leading to a global network of hundreds of databases across more than 20 countries and capturing aver a billion patient records **[52]**, the structuring of ICU data has not been addressed yet.

We propose the development of a methodology for the transformation of data extracted from the ICCA® information system into the OMOP-CDM framework, to enable their effective utilization. Then, we will describe the demographic profiles of patients admitted to RUH's ICUs. The objective is to foster a robust health data repository tailored for ICU data.

2. Methods

2.1. Study Data

ICCA® Intensive Care Data

Since April 2016, the ICUs at Rouen University Hospital have been using the ICCA® information software. The generated data is stored and managed by a commercial relational database management system (RDBMS), specifically Microsoft® SQL Server 2008 R2 (SP3) 10.50.600.34.

A relational database (RDB) organizes data within tables and establishes relationships between these tables. With an RDB, data is organized into multiple tables (**Figure 1**), with rows and columns representing records (tuples), and characteristics (attributes). Keys establish relationships between tables, with primary uniquely identifying each record and foreign keys referencing primary keys in different tables. The relational model enables logical relationships and join operations, thereby structuring for efficient querying, manipulation, and scrutiny. The RDBMS employs Structured Query Language (SQL) to query the databases.

Among the 61 tables that constitute the ICCA® database, 7 were selected to fulfill the study's objectives. Medical stay and demographic information were found in the D_Encounter and PtDemographic tables. Movements within care units and beds were found in D_ClinicalUnit, PtCensus, D_Bed and PtBedStay tables. Weight and height measurements were recorded in the PtAssessement table.

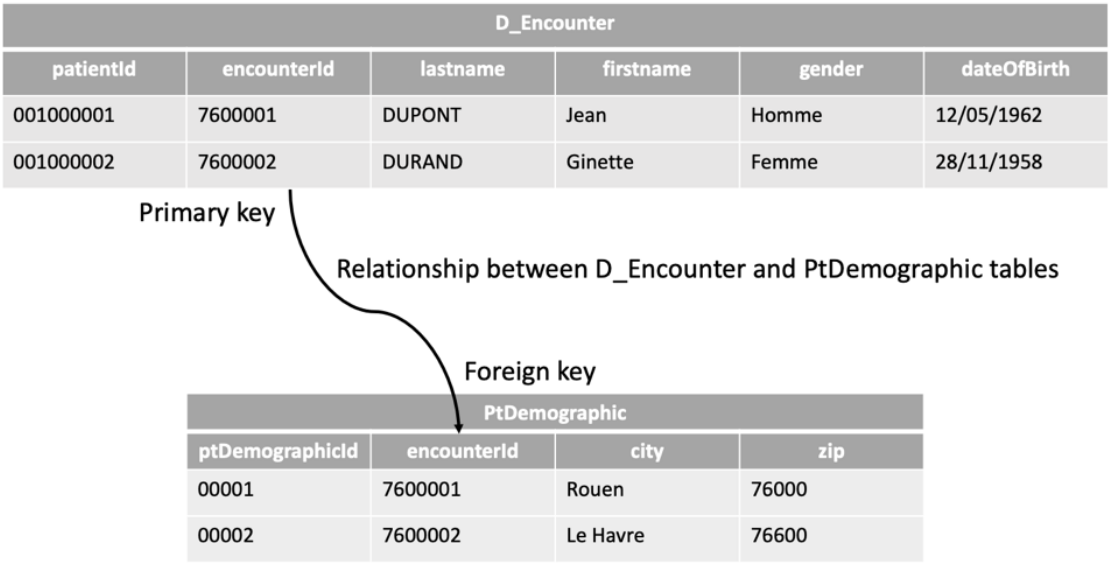


Figure 1 – Simplified Representation of an RDB Involving D_Encounter and PtDemographic tables

OMOP Common Data Model

The OMOP-CDM, developed by OHDSI, was implemented in the Normandy Health Data Warehouse (EDSaN) through scripts provided by the OHDSI community. The OMOP-CDM version 5.4 consists of 15 clinical data tables centered around the PERSON table, 3 tables for healthcare system data, 3 tables for health economics data, and 12 for standardized vocabulary (**Figure 2**).

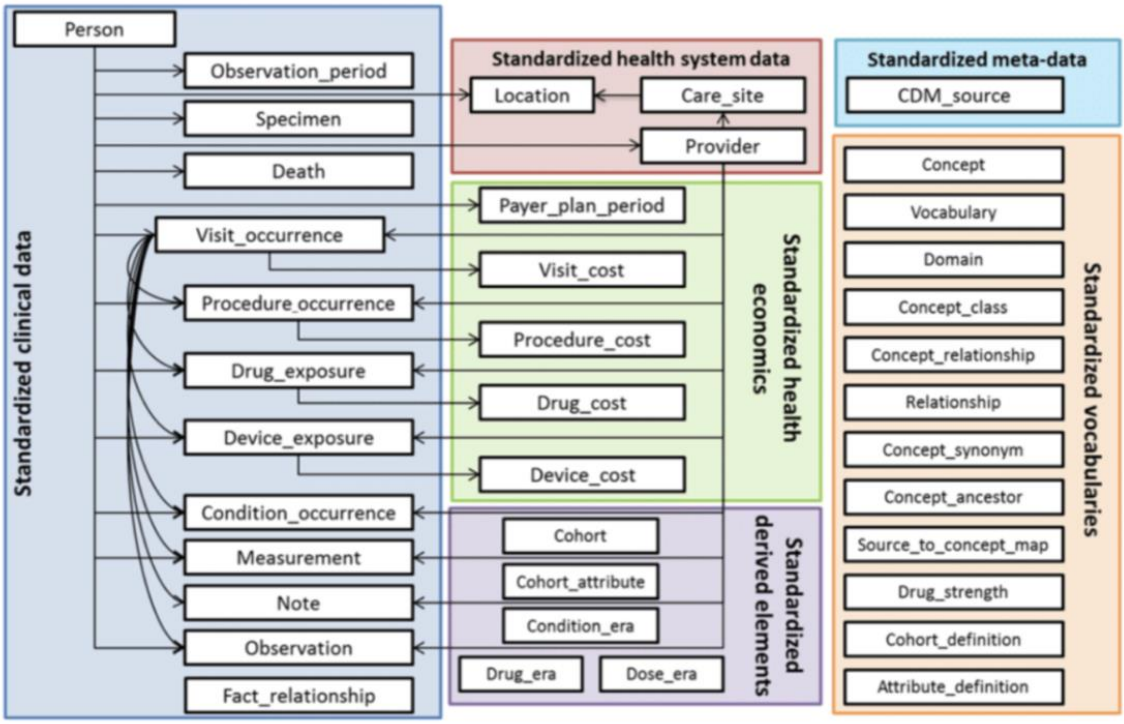


Figure 2 – Structure of the OMOP-CDM [53]

These standardized vocabularies are a fundamental component of the OHDSI research network and the OMOP-CDM, serving to homogenize methods, definitions, and outcomes. The standardized vocabularies are freely available to the community. The CONCEPT, VOCABULARY, and DOMAIN tables were loaded from files via the Athena interface developed by OHDSI (**Figure 4**).

The source code was implemented in PostgreSQL 10.4, an open-source RDBMS renowned for its performance, reliability, and extensive feature set. Moreover, PostgreSQL benefits from an active open-source community that provides support, regular updates, security patches, and detailed documentation.

2.2. Structural and Semantic Mapping

Structural mapping aimed to identify the relevant data within the ICCA® database and integrate it into the appropriate locations in the OMOP database (**Figure 3**). Semantic mapping, on the other hand, aimed to align ICCA®'s local terminologies with OMOP's standardized counterparts, ensuring consistency across different sources and enabled meaningful inquiry within the OMOP format (**Figure 4**).

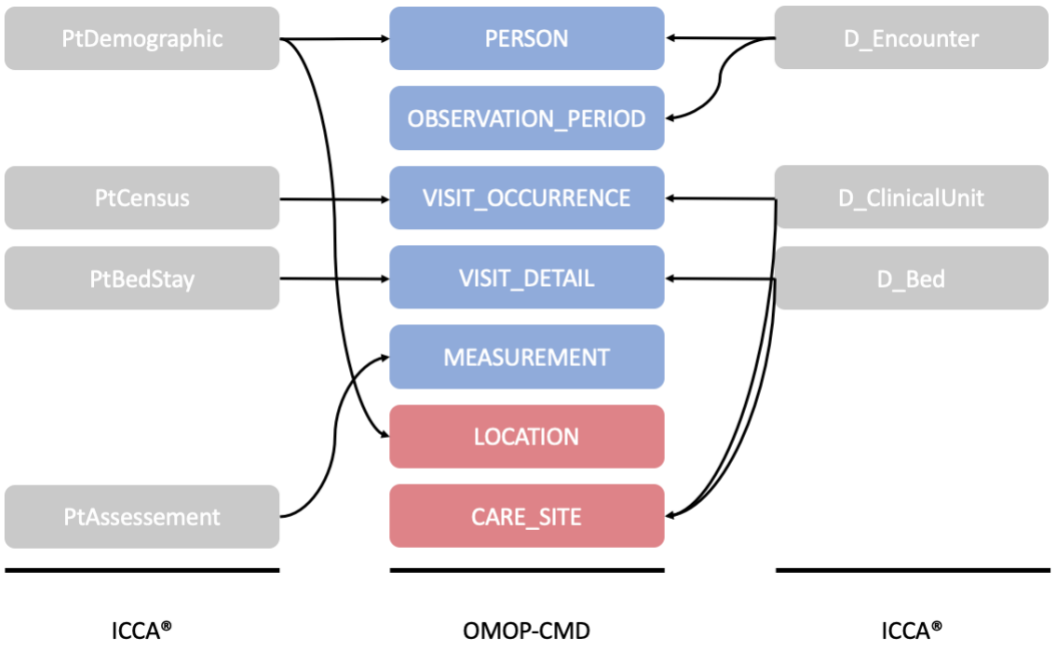
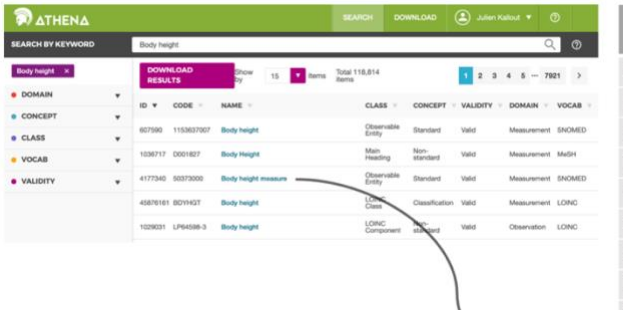


Figure 3 – Structural Mapping from ICCA® to OMOP-CDM



Concepts used	Standardized Vocabularies
Gender (Female)	8532
Gender (Male)	8507
Country (France)	4320324
Location (Hospital)	38004515
Location (Intensive care unit)	32037
Location (Medical intensive care unit)	40481392
Location (Surgical intensive care unit)	4305366
Location (Cardiac intensive care unit)	4149943
Location (Neurological intensive care unit)	4148496
Period covering healthcare encounters	44814724
EHR encounter record	32035
Measurement (Body height)	4177340
Unit (Centimeter)	8582
Measurement (Body weight)	4099154
Unit (Kilogram)	9529

Figure 4 – Semantic Mapping using OMOP Standardized Vocabulary

2.3. Extraction, Transformation, and Loading (ETL)

An ETL process was implemented to transpose the raw ICCA® source data into the OMOP-CDM. The **extraction** step involved retrieving the data from the source system. The **transformation** step, performed after extraction, modified the structure of the source data to meet the requirements of the target format. The names of the source variables were changed to match the target format variable names. Then, patient and visit identifiers were anonymized using a mapping table. Additional variables specific to the target format were added, and standardized vocabulary was incorporated during this step. Incorrect or outlier values were removed. Lastly, the **loading** step integrated the extracted and transformed data into the target database. Loading could be done incrementally (updating modified data) or in bulk (loading all data in each execution). An example of an ETL script for the "person" table is provided in the **Supplementary Materials** (**Figure S1**).

This process restructured the ICCA® data into the OMOP-CDM (structural mapping) and incorporated mappings to the standardized vocabularies (semantic mapping). It was implemented through a set of automated scripts using the Python programming language.

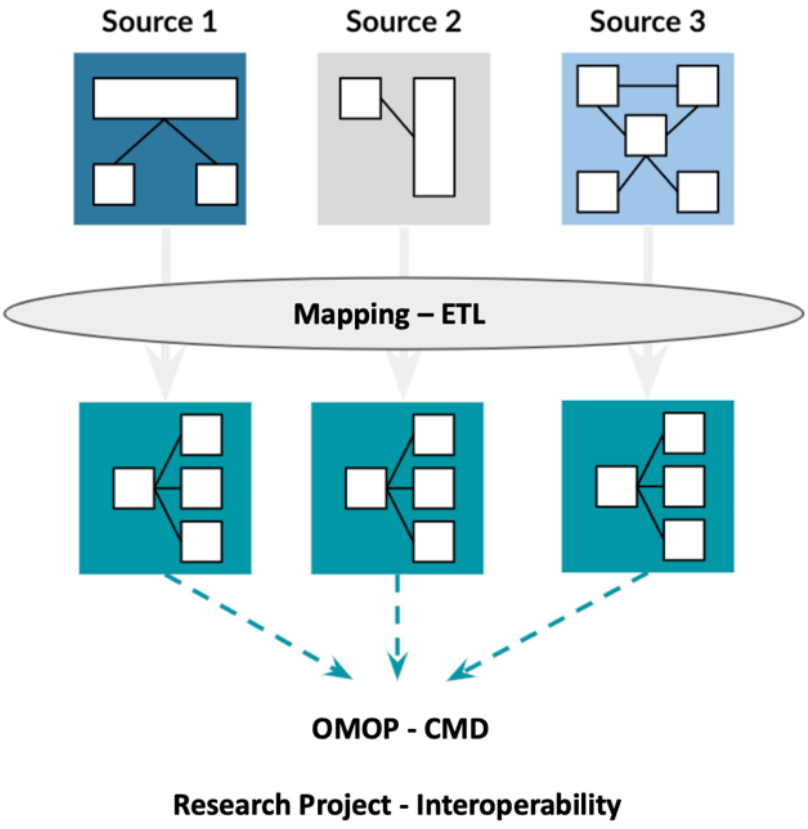


Figure 5 – Summary of the ICCA Database Conversion Process to the OMOP-CDM [53]

2.4. Ethics

All patient data underwent anonymization, by creating a new identifier, aligned with the reference methodology (MR-004) stipulated by the National Commission for Data Protection and Liberties (CNIL).

2.5. Analysis and Visualization

Data extraction was performed as a data frame using PostgreSQL scripts, followed by a descriptive via Python® 3.11.2 software.

Categorical variables were summarized as counts and frequencies, and quantitative variables were expressed as means and standard deviations. Bar charts were employed to visualize the distribution of categorical variables and histograms and density curves for the quantitative variables. The graphs were drawn using the Plotly 5.16.0 package. The interactive map was built using the Folium 0.14.0 package. The geographic coordinates of the cities data was downloaded via the Data gouv website **[54]**. All the generated information was aggregated in an interactive dashboard using the Dash 2.11.1 package.

3. Results

3.1. Patient information

From April 2016 to February 2022, a total of 20,235 unique patients were admitted to the ICUs at Rouen University Hospital, comprising 24,351 stays. The majority of admitted patients were between the ages of 61 and 70, with a mean age of 58.7 (±16.1) years. Among these patients, men were predominant, accounting for 14,878 (61.1%) admissions (**Table 1**, **Figure 6**). Geographically, the patients hailed from various regions across France, with a preponderance from the Normandie region, including 2,404 (9.8%) patients from Rouen itself (**Table 1**, **Figure 7**).

3.2. Care site information

Out of 27,631 admissions in the various units, the medical, surgical, and cardiac ICUs recorded the highest number of admissions, accounting for 6,454 (23.4%), 4,457 (16.1%), and 5,197 (18.8%) admissions, respectively (**Table 2**). There was a stability in the number of admissions over the years **(Figure 8**). The average length of stay was 5.6 (±65.4) days. The respiratory weaning unit (RWU) and the neurosurgical ICU registered the longest average lengths of stay (LOS), with 28.8 (±233.5) days and 9.1 (±62.6) days. The postoperative care unit (POCU) had an average LOS of less than 2 days (Table 2, Figure 9). A total of 209 (0.8%) admissions were recorded in the COVID unit with an average duration of 7.6 (6.8) days. However, other temporary COVID units were established without being equipped with the ICCA® software.

Table 1 – Patient characteristics at admission

	Stays (n = 24,351)
Unique patient	20,325
Demographic information	
Age (year)	58.7 (16.1)
Gender (Female)	9,473 (38.9%)
Patients' hometown (major cities)	
- Rouen	2404 (9.8%)
- Le Havre	1363 (5.6%)
- Evreux	456 (1.9%)
- Dieppe	267 (1.1%)
- Elbeuf	195 (0.8%)
Morphological measurement	
Height at admission (kg)	79.5 (19.2)
Weight (cm)	169.7 (9.5)

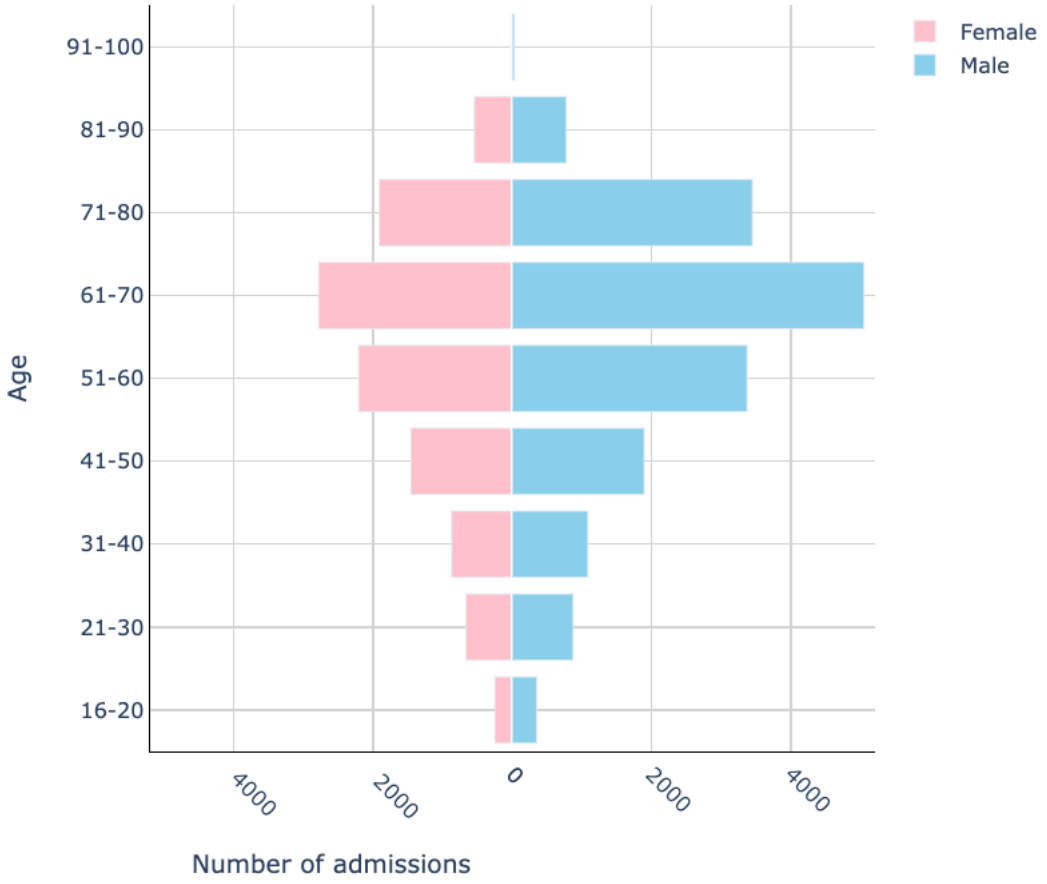


Figure 6 – Pyramid of ages of patients at admission

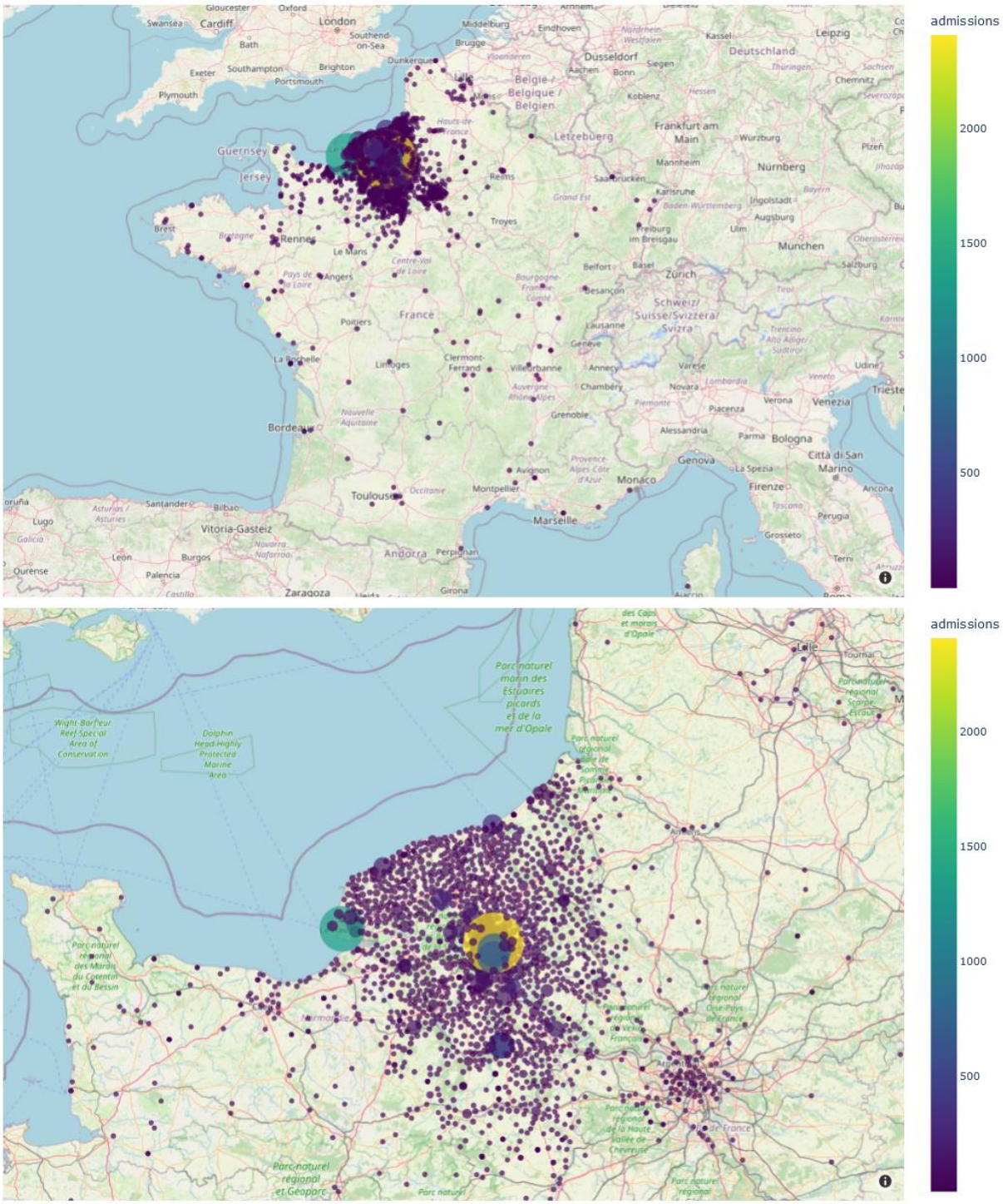


Figure 7 – City of origin of patients admitted to the intensive care units of Rouen University Hospital

Table 2 – Care site characteristics

Abbreviations: CCU, Continuing Care Unit; HDU, Hemodialysis Unit; ICU, Intensive Care Unit; POCU, Post Operative Care Unit; RWU, Respiratory Weaning Unit.

	Admissions
Intensive care stays	24,351
Clinical unit movements	
- Total	27,631
- Cardiac ICU	4,457 (16.1%)
- Surgical ICU	5,197 (18.8%)
- Medical ICU	6,454 (23.4%)
- Neurological ICU	2,166 (7.8%)
- Neurological CCU	1,460 (5.3%)
- Neurological POCU	2,987 (10.8%)
- Cardiac POCU	387 (1.4%)
- Surgical POCU	2,924 (10.6%)
- Medical RWU	911 (3.3%)
- Medical HDU	479 (1.7%)
- Covid unit	209 (0.8%)
Length of stay (days)	
- Total	5.6 (65.4)
- Cardiac ICU	4.1 (5.8)
- Surgical ICU	7.1 (70.0)
- Medical ICU	5.7 (51.6)
- Neurological ICU	9.1 (62.6)
- Neurological CCU	5.9 (106.3)
- Neurological POCU	0.9 (1.0)
- Cardiac POCU	1.0 (1.3)
- Surgical POCU	1.5 (1.8)
- Medical RWU	28.8 (233.5)
- Covid unit	7.6 (6.8)
Bed movements	28,610

Figure 8 – Number of admissions per unit and per year
Abbreviations: CCU, Continuing Care Unit, ICU, Intensive Care Unit; POCU, Post Operative Care Unit; RWU, Respiratory Weaning Unit

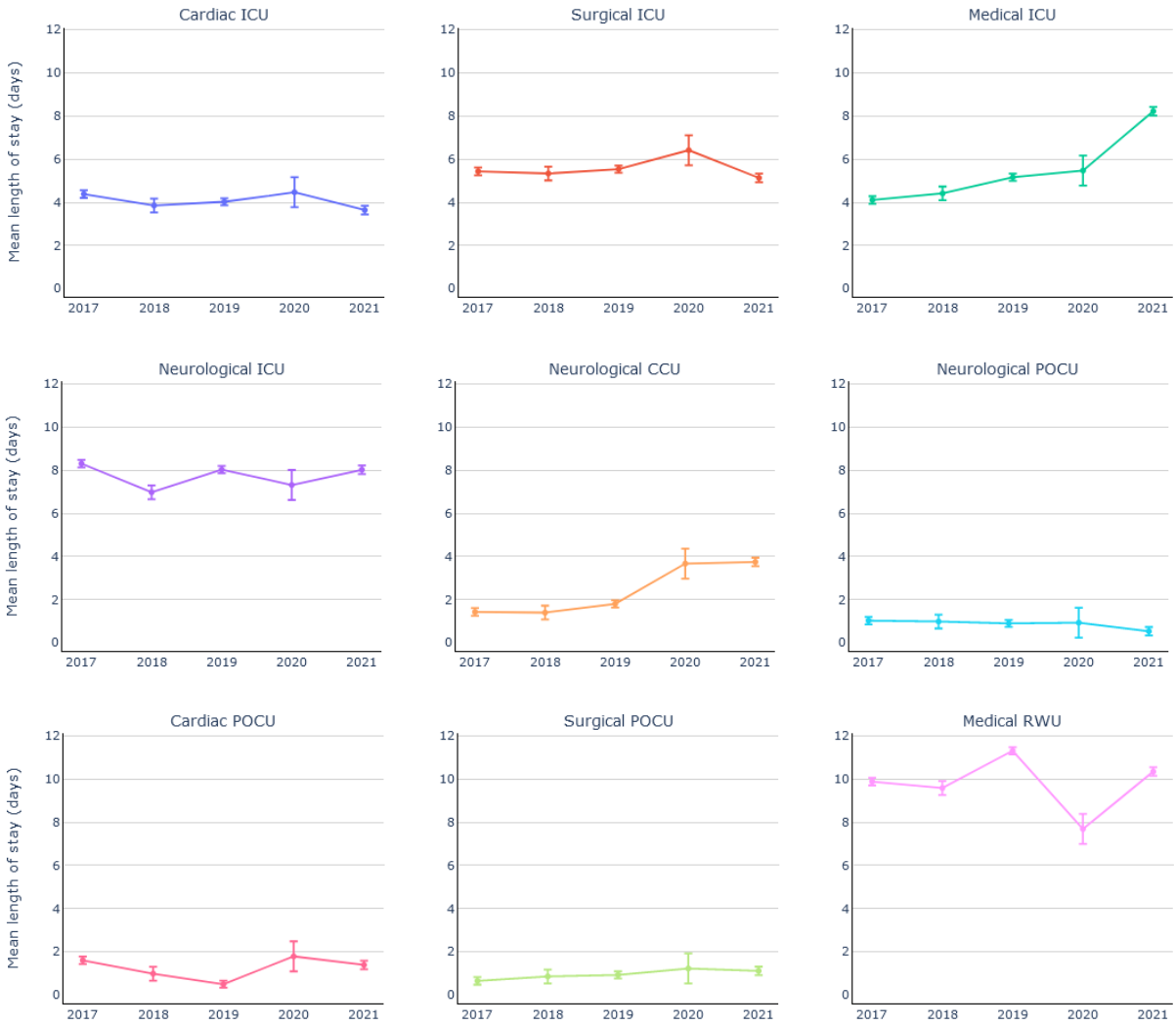


Figure 9 – Mean length of stay per unit and per year
Abbreviations: CCU, Continuing Care Unit, ICU, Intensive Care Unit; POCU, Post Operative Care Unit; RWU, Respiratory Weaning Unit.

4. Discussion

4.1. Main findings

The present study, successfully structured and normalized ICU data at Rouen University Hospital, collected through the ICCA® solution, enabling their integration into the OMOP-CDM standardized format. This transformation not only facilitated the data analysis but also allowed the description of the demographic characteristics of patients admitted to the ICUs as well as the patient flow within each unit.

4.2. Comparison with previous works

To our knowledge, no study has been published on the transformation of ICU data from the ICCA® solution to the OMOP-CDM format. However, prior works in other clinical contexts have explored the transformation of medical databases to the OMOP model. *Lamer et al.* described a methodology for transforming anesthesia data from the DIANE Anesthésie® anesthesia information management system to the OMOP-CDM **[52]**, emphasizing concepts such as patient history, measurement units, medications, different stages of intervention, and anesthesia procedures. While our study focused primarily on ICU patient demographic and administrative data, we recognize the potential for expanding this process to include additional concepts, mirroring studies like *Paris and al.* on the open-access MIMIC database **[55]**.

Furthermore, other studies are aiming to standardize French administrative healthcare data (Système national des données de santé, SNDS) to the OMOP-CDM format could intersect our findings **[56]**, enhancing the integration and analysis of combined SNDS and ICCA® data.

4.3. Limitations and future perspectives

Nevertheless, our study has some limitations. The transformation we effected focused exclusively on specific tables and variables relating to the demographics of patients and admissions within ICUs at Rouen University Hospital. Consequently, vital intensive care parameters such as clinical observations, medication prescriptions, and specific interventions, remained outside this initial transformation's scope. Future work should consider broadening the transformation to include a more extensive set of ICU data, establishing a more comprehensive and versatile data repository.

Additionally, the current OMOP methodology's inability to structure unstructured clinical notes (medical history, clinical evolution) constituting nearly 80% of EHRs medical information **[57]**, also emerge as a significant constraint. Ongoing research targets advanced natural language processing (NLP) and machine learning methods for this purpose. Initiatives such as the OHDSI NLP Working Group are striving to address this challenge by creating tools and methodologies for extracting and normalizing clinical information from unstructured text **[58]**. The successful integration of unstructured data into the OMOP model would significantly enhance research and analysis capabilities, allowing for more in-depth exploration of risk factors, epidemiological trends, and clinical outcomes.

Finally, this study is limited to Rouen University Hospital's ICCA® data and additional efforts should be made to promote reproducibility and interoperability of the process to enable the adoption of the OMOP-CDM model. Establishing a network of ICUs using a standardized data model would create a valuable resource for collaborative and multicenter research, while enabling the comparison of clinical outcomes on a national scale and enhancing the collective understanding of intensive care practices. This network of structured databases represents a significant opportunity for the application of federated learning in the field of critical care medicine, enabling the training of machine learning models while keeping data locally at different sites, without centralizing their sharing **[59]**.

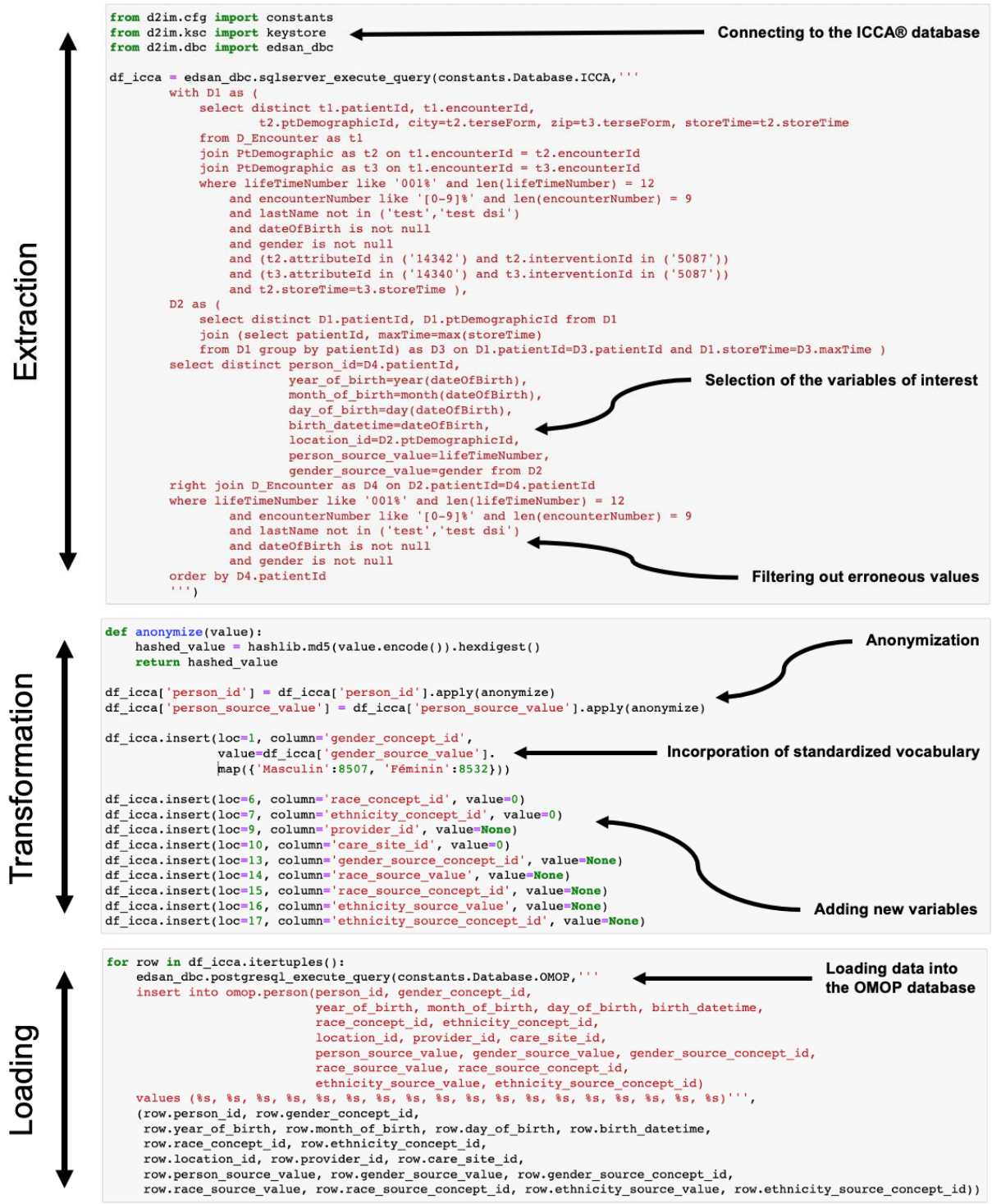
5. Conclusion

The present study demonstrated the feasibility of transforming ICU data in a tertiary university hospital collected using the ICCA® solution into the OMOP-CDM format. By structuring and normalizing this diverse dataset, we achieved integration into a standardized format, enabling easy utilization and large-scale analysis.

Looking ahead, further expanding the transformation process to include additional concepts, enhancing the structuring of textual data, and promoting the establishment of multicentric networks for ICU data will enable us to fully leverage the potential of the OMOP-CDM model for research and advancements in intensive care. This will undoubtedly lead to significant progress in the field of critical care medicine, ultimately improving patient outcomes and healthcare practices.

6. Supplementary materials

Figure S1 – ETL script for the "person" table




```
from d2im.cfg import constants
from d2im.ksc import keystore
from d2im.dbc import edsan_dbc

df_icca = edsan_dbc.sqlserver_execute_query(constants.Database.ICCA,'''
        with D1 as (
            select distinct t1.patientId, t1.encounterId,
                    t2.ptDemographicId, city=t2.terseForm, zip=t3.terseForm, storeTime=t2.storeTime
            from D_Encounter as t1
            join PtDemographic as t2 on t1.encounterId = t2.encounterId
            join PtDemographic as t3 on t1.encounterId = t3.encounterId
            where lifeTimeNumber like '001%' and len(lifeTimeNumber) = 12
                and encounterNumber like '[0-9]%' and len(encounterNumber) = 9
                and lastName not in ('test','test dsi')
                and dateOfBirth is not null
                and gender is not null
                and (t2.attributeId in ('14342') and t2.interventionId in ('5087'))
                and (t3.attributeId in ('14340') and t3.interventionId in ('5087'))
                and t2.storeTime=t3.storeTime ),
        D2 as (
            select distinct D1.patientId, D1.ptDemographicId from D1
            join (select patientId, maxTime=max(storeTime)
            from D1 group by patientId) as D3 on D1.patientId=D3.patientId and D1.storeTime=D3.maxTime )
        select distinct person_id=D4.patientId,
                        year_of_birth=year(dateOfBirth),
                        month_of_birth=month(dateOfBirth),
                        day_of_birth=day(dateOfBirth),
                        birth_datetime=dateOfBirth,
                        location_id=D2.ptDemographicId,
                        person_source_value=lifeTimeNumber,
                        gender_source_value=gender from D2
        right join D_Encounter as D4 on D2.patientId=D4.patientId
        where lifeTimeNumber like '001%' and len(lifeTimeNumber) = 12
                and encounterNumber like '[0-9]%' and len(encounterNumber) = 9
                and lastName not in ('test','test dsi')
                and dateOfBirth is not null
                and gender is not null
        order by D4.patientId
        ''')
```

```
def anonymize(value):
    hashed_value = hashlib.md5(value.encode()).hexdigest()
    return hashed_value

df_icca['person_id'] = df_icca['person_id'].apply(anonymize)
df_icca['person_source_value'] = df_icca['person_source_value'].apply(anonymize)

df_icca.insert(loc=1, column='gender_concept_id',
               value=df_icca['gender_source_value'].
               map({'Masculin':8507, 'Féminin':8532}))

df_icca.insert(loc=6, column='race_concept_id', value=0)
df_icca.insert(loc=7, column='ethnicity_concept_id', value=0)
df_icca.insert(loc=9, column='provider_id', value=None)
df_icca.insert(loc=10, column='care_site_id', value=0)
df_icca.insert(loc=13, column='gender_source_concept_id', value=None)
df_icca.insert(loc=14, column='race_source_value', value=None)
df_icca.insert(loc=15, column='race_source_concept_id', value=None)
df_icca.insert(loc=16, column='ethnicity_source_value', value=None)
df_icca.insert(loc=17, column='ethnicity_source_concept_id', value=None)
```

```
for row in df_icca.itertuples():
    edsan_dbc.postgresql_execute_query(constants.Database.OMOP,'''
    insert into omop.person(person_id, gender_concept_id,
                            year_of_birth, month_of_birth, day_of_birth, birth_datetime,
                            race_concept_id, ethnicity_concept_id,
                            location_id, provider_id, care_site_id,
                            person_source_value, gender_source_value, gender_source_concept_id,
                            race_source_value, race_source_concept_id,
                            ethnicity_source_value, ethnicity_source_concept_id)
    values (%s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s)''',
    (row.person_id, row.gender_concept_id,
     row.year_of_birth, row.month_of_birth, row.day_of_birth, row.birth_datetime,
     row.race_concept_id, row.ethnicity_concept_id,
     row.location_id, row.provider_id, row.care_site_id,
     row.person_source_value, row.gender_source_value, row.gender_source_concept_id,
     row.race_source_value, row.race_source_concept_id, row.ethnicity_source_value, row.ethnicity_source_concept_id))
```

Bibliographie

1. Lind J. A treatise of the scurvy in three parts. Containing an inquiry into the nature causes and cure of that disease, together with a critical and chronological view of what has been published on the subject. London; 1753.

2. Daniel C. La recherche clinique à partir d'entrepôts de données. L'expérience de l'Assistance Publique – Hôpitaux de Paris (AP–HP) à l'épreuve de la pandémie de Covid-19. La Revue de Médecine Interne. 2020;41:303–7.

3. Sauer CM, Dam TA, Celi LA, Faltys M, de la Hoz MAA, Adhikari L, et al. Systematic Review and Comparison of Publicly Available ICU Data Sets—A Decision Guide for Clinicians and Data Scientists. Critical Care Medicine. 2022;50:e581–8.

4. Qu'est-ce ce qu'une donnée de santé ? | CNIL [Internet]. Available from: https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante

5. Rapport d'information déposé en application de l'article 145 du règlement, par la commission des affaires sociales en conclusion des travaux de la mission d'évaluation et de contrôle des lois de financement de la sécurité sociale. Paris: Assemblé nationale; 2020.

6. Miller CJ, Smith SN, Pugatch M. Experimental and quasi-experimental designs in implementation research. Psychiatry Research. 2020;283:112452.

7. Davidoff F, Haynes B, Sackett D, Smith R. Evidence based medicine. BMJ. 1995;310:1085–6.

8. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" The Lancet. 2005;365:82–93.

9. Jannot A-S, Messiaen C, Khatim A, Pichon T, Sandrin A, the BNDMR infrastructure team. The ongoing French BaMaRa-BNDMR cohort: implementation and deployment of a nationwide information system on rare disease. Journal of the American Medical Informatics Association. 2022;29:553–8.

10. Enrique B, Marta B. Efficacy, Effectiveness and Efficiency in the Health Care: The Need for an Agreement to Clarify its Meaning. Int Arch Public Health Community Med [Internet]. 2020. Available from: https://www.clinmedjournals.org/articles/iaphcm/international-archives-of-public-health-and-community-medicine-iaphcm-4-035.php?jid=iaphcm

11. Bégaud B, Polton D, Franck von L. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé. 2017.

12. Lignot-Leloup M, Merlière Y. Vers un dossier médical partagé pour favoriser la coordination des acteurs. Soins. 2016;61:53.

13. Hill AB. The Environment and Disease: Association or Causation? Proc R Soc Med. 1965;58:295–300.

14. Irgens LM. The origin of registry-based medical research and care. Acta Neurol Scand. 2012;126:4–6.

15. the Traumabase® Group, Hamada SR, Rosa A, Gauss T, Desclefs J-P, Raux M, et al. Development and validation of a pre-hospital "Red Flag" alert for activation of intra-hospital haemorrhage control response in blunt trauma. Crit Care. 2018;22:113.

16. Scailteux L-M, Droitcourt C, Balusson F, Nowak E, Kerbrat S, Dupuy A, et al. French administrative health care database (SNDS): The value of its enrichment. Therapies. 2019;74:215–23.

17. Documentation technique | SNDS [Internet]. Available from: https://www.snds.gouv.fr/SNDS/Documentation-technique

18. Lamer A, Ficheur G, goire, Rousselet L, van Berleere M, Chazard E, et al. From Data Extraction to Analysis: Proposal of a Methodology to Optimize Hospital Data Reuse Process. Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth [Internet]. IOS Press; 2018 [cited 2023 Jun 30]. p. 41–5. Available from: https://ebooks.iospress.nl/doi/10.3233/978-1-61499-852-5-41

19. Entrepôts de données de santé hospitaliers en France. Haute autorité de santé; 2022.

20. Goldberg M, Zins M. [Health Data Hub: Why and how?]. Med Sci (Paris). 2021;37:271–6.

21. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1.

22. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178.

23. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. Crit Care Med. 2021;49:e563–77.

24. Faltys, Martin, Zimmermann, Marc, Lyu, Xinrui, Hüser, Matthias, Hyland, Stephanie, Rätsch, Gunnar, et al. HiRID, a high time-resolution ICU dataset [Internet]. PhysioNet. Available from: https://physionet.org/content/hirid/1.1.1/

25. Sanchez-Pinto LN, Luo Y, Churpek MM. Big Data and Data Science in Critical Care. Chest. 2018;154:1239–48.

26. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19:64.

27. Chiolero A, Buckeridge D. Glossary for public health surveillance in the age of data science. J Epidemiol Community Health. 2020;74:612–6.

28. Chico V. The impact of the General Data Protection Regulation on health research. British Medical Bulletin. 2018;128:109–18.

29. 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. 2016;3:160018.

30. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.

31. Mark R. The Story of MIMIC. Secondary Analysis of Electronic Health Records [Internet]. Cham: Springer International Publishing; 2016.p. 43–9. Available from: http://link.springer.com/10.1007/978-3-319-43742-2_5

32. Syed M, Syed S, Sexton K, Syeda HB, Garza M, Zozus M, et al. Application of Machine Learning in Intensive Care Unit (ICU) Settings Using MIMIC Dataset: Systematic Review. Informatics. 2021;8:16.

33. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 2005;8:19–32.

34. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–73.

35. Chapter 11: Scoping reviews. JBI Manual for Evidence Synthesis [Internet]. JBI; 2020. Available from: https://jbi-global-wiki.refined.site/space/MANUAL/4687342/Chapter+11%3A+Scoping+reviews

36. Joon Lee, Scott DJ, Villarroel M, Clifford GD, Saeed M, Mark RG. Open-access MIMIC-II database for intensive care research. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. Boston, MA: IEEE; 2011. p. 8315–8. Available from: http://ieeexplore.ieee.org/document/6092050/

37. Rigby AS. Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient. Disability and Rehabilitation. 2000;22:339–44.

38. Martignene N, Amad A, Bellet J, Tabareau J, D'Hondt F, Fovet T, et al. Goupile: A New Paradigm for the Development and Implementation of Clinical Report Forms. In: Séroussi B, Weber P, Dhombres F, Grouin C, Liebe J-D, Pelayo S, et al., editors. Studies in Health Technology and Informatics [Internet]. IOS Press; 2022. Available from: https://ebooks.iospress.nl/doi/10.3233/SHTI220517

39. Van De Sande D, Van Genderen ME, Huiskens J, Gommers D, Van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med. 2021;47:750–60.

40. Popoff B, Occhiali É, Grangé S, Bergis A, Carpentier D, Tamion F, et al. Trends in major intensive care medicine journals: A machine learning approach. Journal of Critical Care. 2022;72:154163.

41. Gutierrez G. Artificial Intelligence in the Intensive Care Unit. Crit Care. 2020;24:101.

42. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care. 2019;23:284.

43. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–15.

44. Popoff B. Contribution of open access databases to intensive care medicine research: a scoping review. 2022 [cited 2023 Jul 24]; Available from: https://osf.io/kugaz/

45. JCR Clarivate [Internet]. Available from: https://jcr.clarivate.com

46. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning [Internet]. New York, NY: Springer New York; 2009. Available from: http://link.springer.com/10.1007/978-0-387-84858-7

47. Kaelbling LP, Littman ML, Moore AW. Reinforcement Learning: A Survey. 1996. Available from: https://arxiv.org/abs/cs/9605103

48. Cosgriff CV, Celi LA, Stone DJ. Critical Care, Critical Data. Biomed Eng Comput Biol. 2019;10:117959721985656.

49. Prgomet M, Li L, Niazkhani Z, Georgiou A, Westbrook JI. Impact of commercial computerized provider order entry (CPOE) and clinical decision support systems (CDSSs) on medication errors, length of stay, and mortality in intensive care units: a systematic review and meta-analysis. Journal of the American Medical Informatics Association. 2017;24:413–22.

50. Philips - IntelliSpace Critical Care and Anesthesia [Internet]. Available from: https://www.usa.philips.com/healthcare/product/HCNOCTN332/intellispace-critical-care-and-anesthesia

51. George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, Johan van der Lei, Nicole Pratt, G. Niklas Norén, Yu-Chuan Li, Paul E. Stang, David Madigan, Patrick B. Ryan. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Studies in Health Technology and Informatics. 2015;Volume 216: MEDINFO 2015: eHealth-enabled Health:574–8.

52. Lamer A, Abou-Arab O, Bourgeois A, Parrot A, Popoff B, Beuscart J-B, et al. Transforming Anesthesia Data Into the Observational Medical Outcomes Partnership Common Data Model: Development and Usability Study. J Med Internet Res. 2021;23:e29259.

53. The Book of OHDSI [Internet]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/

54. Accueil - data.gouv.fr [Internet]. Available from: https://www.data.gouv.fr/fr/

55. Paris N, Lamer A, Parrot A. Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study. JMIR Med Inform. 2021;9:e30970.

56. Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre M-M, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. Appl Clin Inform. 2020;11:13–22.

57. Martin-Sanchez F, Verspoor K. Big Data in Medicine Is Driving Big Changes. Yearb Med Inform. 2014;23:14–20.

58. Keloth VK, Banda JM, Gurley M, Heider PM, Kennedy G, Liu H, et al. Representing and utilizing clinical textual data for real world studies: An OHDSI approach. Journal of Biomedical Informatics. 2023;142:104343.

59. Crowson MG, Moukheiber D, Arévalo AR, Lam BD, Mantena S, Rana A, et al. A systematic review of federated learning applications for biomedical data. Mordaunt DA, editor. PLOS Digit Health. 2022;1:e0000033.

SERMENT D'HIPPOCRATE

Au moment d'être admis(e) à exercer la médecine, je promets et je jure d'être fidèle aux lois de l'honneur et de la probité.
Mon premier souci sera de rétablir, de préserver ou de promouvoir la santé dans tous ses éléments, physiques et mentaux, individuels et sociaux. Je respecterai toutes les personnes, leur autonomie et leur volonté, sans aucune discrimination selon leur état ou leurs convictions. J'interviendrai pour les protéger si elles sont affaiblies, vulnérables ou menacées dans leur intégrité ou leur dignité. Même sous la contrainte, je ne ferai pas usage de mes connaissances contre les lois de l'humanité. J'informerai les patients des décisions envisagées, de leurs raisons et de leurs conséquences. Je ne tromperai jamais leur confiance et n'exploiterai pas le pouvoir hérité des circonstances pour forcer les consciences. Je donnerai mes soins à l'indigent et à quiconque me les demandera. Je ne me laisserai pas influencer par la soif du gain ou la recherche de la gloire.

Admis(e) dans l'intimité des personnes, je tairai les secrets qui me seront confiés. Reçu(e) à l'intérieur des maisons, je respecterai les secrets des foyers et ma conduite ne servira pas à corrompre les mœurs. Je ferai tout pour soulager les souffrances. Je ne prolongerai pas abusivement les agonies. Je ne provoquerai jamais la mort délibérément.

Je préserverai l'indépendance nécessaire à l'accomplissement de ma mission. Je n'entreprendrai rien qui dépasse mes compétences. Je les entretiendrai et les perfectionnerai pour assurer au mieux les services qui me seront demandés.

J'apporterai mon aide à mes confrères ainsi qu'à leurs familles dans l'adversité. Que les hommes et mes confrères m'accordent leur estime si je suis fidèle à mes promesses ; que je sois déshonoré(e) et méprisé(e) si j'y manque.

Titre : Réutilisation des données de réanimation : état de lieux des bases existantes et mise en place d'un entrepôt de données de réanimation au CHU de Rouen.

Mots clés : Réanimation, Données massives, Base de données, Entrepôt de données, Accès ouvert, AmsterdamUMC, eICU-CRD, HiRID, MIMIC.

Contexte : Les services des réanimations prennent en charge les patients les plus graves avec un haut risque de mortalité. En raison de l'état critique de ces patients, une surveillance étroite est nécessaire, conduisant à la collecte d'un volume important de données. Des collaborations ont permis l'émergence de grandes bases de données en accès libre à l'origine de nombreuses publications dans le domaine.

Objectif : L'objectif de cette revue de la littérature est d'identifier les caractéristiques des études utilisant des bases de données ouvertes de soins intensifs et de décrire la contribution de ces études à la recherche en soins intensifs.

Méthodes : La recherche a été effectuée à partir de 3 bases de données (PubMed - Medline, Embase, Web of Science) de la création de la base de données jusqu'au 1er août 2022. Ont été inclus les articles originaux basés sur 4 bases de données ouvertes concernant des patients adultes admis en unités de réanimation (Amsterdam University Medical Centers Database (AmsterdamUMC), Collaborative Research Database (eICU-CRD), High time resolution ICU dataset (HiRID), Medical Information Mart for Intensive Care (MIMIC)). Les caractéristiques liées à la description des publications, à la conception des études et aux analyses statistiques ont été extraites et analysées.

Résultats : Nous avons observé une augmentation constante du nombre de publications provenant de ces bases de données depuis 2016. Les bases de données MIMIC ont été les plus fréquemment utilisées, tandis que les pays contribuant le plus étaient la Chine et les États-Unis avec 683 (52.8%) et 367 (28.4%) publications respectivement. Le facteur d'impact médian des publications depuis la création des bases de données ouvertes en soins intensifs est de 3,8 [2,8 - 5,8]. Les sujets cardiovasculaires et infectieux étaient les plus représentés avec 333 (25.7%) et 319 (24.7%) articles respectivement. En ce qui concerne les méthodes statistiques, la régression logistique était le modèle le plus couramment utilisé pour les questions d'inférence et de prédiction avec 383 (55.5%) et 276 (47.3%) études respectivement. La majorité des études d'inférence ont présenté des résultats statistiquement significatifs (84.1%). Dans les études de prédiction, la mesure de performance la plus récurrente était l'AUC, avec une valeur médiane de 0.840 [0.780 – 0.890].

Conclusions : L'abondance des résultats scientifiques issus de ces bases de données et la diversité des sujets abordés mettent en évidence l'importance de ces bases de données en tant que ressources précieuses pour la recherche clinique et suggèrent leur impact potentiel sur la pratique clinique en soins intensifs. Cependant, la qualité des études et leur pertinence clinique restent très hétérogènes, la majorité des articles étant publiés dans des revues à faible impact. Ainsi, cette étude souligne l'importance de la nécessité de mettre à disposition une base de données de réanimation à l'échelle locale et nationale.