Diabète et insuffisance cardiaque : approche épidémiologique par l'analyse croisée de différentes sources de données de santé

Matthieu Wargny

Thèse de doctorat de

NANTES UNIVERSITE

Ecole Doctorale n°605

Biologie Santé

Spécialité : Santé Publique

Par **M. Matthieu WARGNY**

Diabète et insuffisance cardiaque

Approche épidémiologique par l'analyse croisée de différentes sources de données de santé

Thèse présentée et soutenue à Nantes, le 12 décembre 2022

Unité de recherche : UMR_S 1087 / UMR_C 6291 l'Unité de Recherche de l'Institut du Thorax

Rapporteurs avant la soutenance :

Claire BOULETI : Professeure des Universités – Praticien Hospitalier en cardiologie
Karen LEFFONDRE : Professeure des Universités en Mathématiques Appliquées

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Présidente :	Bénédicte GABORIT	Professeure des Universités – Praticien Hospitalier en Endocrinologie
Examinateurs :	Marc CUGGIA	Professeur des Universités – Praticien Hospitalier en Bioinformatique
	Pierre-Antoine GOURRAUD	Professeur des Universités – Praticien Hospitalier en Biologie cellulaire
Dir. de thèse :	Samy HADJADJ	Professeur des Universités – Praticien Hospitalier en Endocrinologie

COMITES DE SUIVI INDIVIDUEL

1ère année

- Madame le Professeure Leïla MORET, PU-PH, Santé Publique et Médecine Sociale
- Monsieur le Professeur Jean-Noël TROCHU, PU-PH, Cardiologue
- Monsieur le Professeur Marc CUGGIA, PU-PH, Biostatistiques, Informatique Médicale et Technologies de Communication

2ème année

- Madame le Professeure Leïla MORET, PU-PH, Santé Publique et Médecine Sociale
- Monsieur le Professeur Jean-Noël TROCHU, PU-PH, Cardiologue
- Monsieur le Professeur Thierry COUFFINHAL, PU-PH, Cardiologue

3ème année

- Madame le Professeure Leïla MORET, PU-PH, Santé Publique et Médecine Sociale
- Monsieur le Professeur Jean-Noël TROCHU, PU-PH, Cardiologue
- Monsieur le Professeur Thierry COUFFINHAL, PU-PH, Cardiologue

REMERCIEMENTS

Aux membres du jury,

A Madame le Professeur Bénédicte Gaborit, pour avoir accepté de présider ce jury de soutenance.

A Mesdames les Professeurs Claire Bouleti et Karen Leffondré, pour avoir accepté d'évaluer ce travail, et en particulier pour leurs retours critiques sur la première version du présent manuscrit.

A Monsieur le Professeur Marc Cuggia, pour avoir accepté de co-encadrer ce travail, et en particulier pour son soutien dans la mise en place du projet GAVROCHE.

A Monsieur le Professeur Pierre-Antoine Gourraud, d'abord en sa qualité de co-encadrant, mais surtout pour avoir su me remettre le pied à l'étrier des données de santé, pour sa confiance et la liberté accordée dans mes orientations techniques et scientifiques.

Et, naturellement, à Monsieur le Professeur Samy Hadjadj, pour avoir accepté de diriger cette thèse jusqu'aux prolongations, et pour avoir su tolérer mon esprit de contradiction peut-être par trop systématique.

Aux membres des comités de suivi individuel, qui ont su apporter un précieux regard critique extérieur à ce travail, et garantir sa bonne conduite.

A toute l'équipe du CIC-EC 1413 « Clinique des Données » du CHU de Nantes, pour m'avoir accompagné et soutenu depuis mon arrivée dans l'équipe en 2018, pour nos échanges scientifiques et la camaraderie. Une pensée particulière à Pauline, pour tant nous faciliter le quotidien.

A l'équipe IV de l'institut du Thorax « Maladies cardiométaboliques », d'abord pour leur bon accueil suite à mon parachutage Toulouso-Poitevin en 2017, ensuite pour leur bienveillance sans faille, particulièrement Cédric, Xavier, Mikaël, Wieneke, Simon, Claire, Sarra, Samuel – j'en oublie beaucoup.

A l'équipe du CIC d'Endocrinologie (etc. !) du CHU de Nantes, avec qui je serais très heureux de faire encore un bon bout de chemin dans l'investigation, si « elles et il » continuent à accepter un médecin de Santé Publique dans leur Clinique.

A l'équipe juridique du CHU de Nantes, et en particulier Maxime Caillier et Philippe Boucher.

A l'équipe du CIC de Poitiers, et en particulier Elise Gand, Stéphanie Ragot et Pierre-Jean Saulnier, pour leur appui dans la compréhension et l'exploitation de la cohorte SURDIAGENE.

Aux membres de la CNAM, du GT REDSIAM « Endocrino » et de Santé publique France, et particulièrement Marjorie Boussac et Philippe Tuppin, Sandrine Fosse-Edorh et Clara Piffaretti, pour leur aide indispensable face au dédale du SNDS dans DMC.

Aux équipes du RiCDC et aux membres des Centres de Données Cliniques du groupe HUGO, pour nos échanges autour du déploiement d'eHOP et des projets inter-régionaux. Je tiens particulièrement à remercier Julien Herbert du CHU de Tours pour son aide dans la réplication des procédures de GAVROCHE.

A mes parents, à ma famille, à mes amis.

A Mathilde.

TABLE DES MATIERES

LISTE DES FIGURES

ABREVIATIONS

AAP	Appel A Projet
ALD	Affection de Longue Durée
AMM	Autorisation de Mise sur le Marché
AOMI	Artériopathie Oblitérante des Membres Inférieurs
ARA-2	Antagonistes des Récepteurs à l'Angiotensine 2
ATC	Classification Anatomique, Thérapeutique et Chimique des médicaments
ATIH	Agence Technique de l'Information sur l'Hospitalisation
AVC	Accident Vasculaire Cérébral
BCMD	Base des Causes Médicales de Décès (gérée par l'INSERM)
BDMA	Bases de Données Médico-Administratives
BEH	Bulletin Epidémiologique Hebdomadaire
CCAM	Classification Commune des Actes Médicaux
CESREES	Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé
CHU	Centre Hospitalier Universitaire
CIF	*Cumulative Incidence Function,* fonction d'incidence cumulée utilisée pour représenter la survie à un événement, en particulier en cas d'événements compétitifs
CIM-10	Classification Internationale des Maladies, 10ème révision
CIP	« Club Inter Pharmaceutique », code d'identifiant de présentation (à 7 ou 13 chiffres) unique pour une forme donnée de délivrance médicamenteuse
CNAM	Caisse Nationale d'Assurance Maladie
CNIL	Commission Nationale de l'Informatique et des Libertés
CONSTANCES	CONSulTANts des Centres d'Examens de Santé (cohorte nationale)
CR	Compte-Rendu
CSE	Comité Scientifique et Ethique
DAG	*Directed Acyclic Graph*, ou graphe dirigé acyclique
DCI	Dénomination Commune Internationale
DCIR	Datamart Consommation Inter-Régimes : regroupe certaines tables SNDS d'intérêt pour l'étude DMC, parmi lesquelles la délivrance de médicaments en pharmacie de ville (codage CIP), des actes médicaux (codage CCAM) et l'information sur les ALD
DDD	*Defined Daily Dose* – dose médicamenteuse standard quotidienne pour une spécialité donnée, tel que proposée par l'OMS [1]
DIM/SIM	Département d'Information Médicale
DMC	*Diabetes Multiple Complications* (Etude – cf. section 4)
DP/DR/DA	Diagnostic Principal/Relié/Associé codé dans le PMSI, associé par exemple à un RUM
DPP4i	Inhibiteurs de la Dipeptidyl Peptidase de type 4
DRI	Direction de la Recherche et de l'Innovation
DT	Diabète (utilisé ici indifféremment pour désigner tout diabète sucré)
eCRF	*electronic Case-Report Form* - formulaire électronique utilisé dans le recueil de données de santé
ECG	Electrocardiogramme
EDS	Entrepôts de données de santé (souvent sous-entendus « hospitaliers », dans le présent manuscrit)
ERG	Evénement Rétinien Grave (traduit par SRE – *Serious Retinal Event*)
FHS	Framingham Heart Study
GAVROCHE	*Glycemia and its Variability in Regards tO Congestive Heart FailurE* (Etude – cf. section 5)
GHM	Groupe Homogène de Malades : le GHM est la catégorie élémentaire de la classification médico-économique propre au PMSI–MCO (médecine, chirurgie, obstétrique et odontologie)[1]

[1] https://www.atih.sante.fr/glossaire

GIRCI-GO	Groupement Inter-régional de Recherche Clinique et d'Innovation du Grand Ouest.
GLP1a	Agoniste du récepteur au *Glucagon-Like Peptide-1*
HDH	Health Data Hub
HFrH	*Heart Failure requiring Hospitalization* – insuffisance cardiaque nécessitant une hospitalisation
HUGO	Hôpitaux Universitaires du Grand Ouest : CHU d'Angers, Brest, Nantes, Rennes et Tours, et Institut de Cancérologie de l'Ouest
IC(A)	Insuffisance Cardiaque (Aiguë)
ICa	Inhibiteurs Calciques (classe médicamenteuses)
IDF	*International Diabetes Federation*
IDM	Infarctus Du Myocarde
IEC	Inhibiteurs de l'Enzyme de Conversion (classe médicamenteuses)
IMC	Indice de Masse Corporelle (= rapport poids (kg) / taille² (m))
IPP	Identifiant Personnel Patient (identifiant unique d'un patient dans le SIH du CHU de Nantes)
IRC(T)	Insuffisance Rénale Chronique (terminale)
LPPR	Liste des Produits et Prestations Remboursables
MCO	Médecine, Chirurgie, Obstétrique et odontologie
MR-004	4ème Méthodologie de Référence de la CNIL, permettant de dispenser d'une autorisation spécifique tout projet respectant la Méthodologie.
NIR	Numéro d'Inscription au Répertoire (encore appelé N° de sécurité sociale)
ODH	*Ouest Data Hub,* plate-forme (« Hub ») inter-régionale regroupant les centres participants du groupe HUGO, et hébergée par le CHU de Nantes
PIA	*Privacy Impact Assessment*, qui pourrait se traduire en français par « Analyse d'impact relative à la vie privée » [2]
PMSI	Programme de Médicalisation des Systèmes d'Information
RAC(U)	Ratio albumine/créatinine urinaire
RD(P)	Rétinopathie Diabétique (Proliférante)
REDSIAM	«Réseau données SNIIRAM » – Réseau national visant à proposer des algorithmes communs afin de mieux utiliser les données du SNDS, en particulier pour identifier certaines pathologies comme le diabète
RIPH/RNIPH	Recherche impliquant/n'impliquant pas la personne humaine
RUM	Résumé d'Unité Médicale : un RUM est produit à la fin de chaque séjour de malade dans une unité médicale assurant des soins de médecine, chirurgie, obstétrique et odontologie, quel que soit le mode de sortie de cette unité
RWE/RWD	*Real-world evidence / Real-world data*, Preuve ou données de « vie réelle »
SI(H)	Système d'Informations (Hospitalier)
SIM	Service d'Information Médicale
SNDS	Système National des Données de Santé, regroupant notamment des données en lien avec le remboursement de soins ambulatoires (DCIR) et hospitaliers (PMSI/ATIH)
SNIIRAM	Système National d'Information Inter-Régimes de l'Assurance Maladie
SQL	*Structured Query Language*, Langage informatique normalisé servant à l'exploitation des bases de données relationnelles
SRE	*Serious Retinal Event* (cf. « ERG »)
SURDIAGENE	SUivi Rénal, DIAbète de type 2 et GENEtique (Etude – cf. section 3)
TALN	Traitement Automatique du Langage Naturel
TMAO	Oxyde de triméthylamine
T2A	Tarification A l'Activité
VG	Variabilité Glycémique. Ce terme ne se rapporte pas à une méthode de calcul univoque de la variabilité glycémique mais correspondra, dans le projet GAVROCHE, au rapport écart-type/moyenne de la glycémie, calculé sur au moins 3 valeurs

GLOSSAIRE

Plusieurs termes utilisés ne répondent pas à une définition consensuelle et peuvent donc être sources de confusion. Je propose ici une définition personnelle de ces termes, issue de mon expérience.

Anonymisation (cf. désidentification)

Plus aboutie que la désidentification, l'anonymisation peut être définie par la suppression à l'échelle d'une variable, d'une base de données ou d'un document, du niveau d'information suffisant pour réidentifier de façon unique un individu. Garantir formellement l'anonymat d'une base de données est cependant difficile, ce qui est une limite fondamentale à la mise à disposition des données de santé à un tiers.

Datamart et étude dans les entrepôts

Dans le cadre de l'utilisation du logiciel eHOP, un *datamart* désigne un sous-ensemble logique ou physique de patients et de données issus d'une étude, qui peut être directement mis à disposition de l'utilisateur sous la forme d'une nouvelle étude ou venir alimenter une étude existante.

Désidentification (cf. anonymisation)

A l'échelle d'une variable, d'une base de données ou d'un document, la désidentification consiste en la suppression des informations permettant l'identification directe d'un unique individu : nom, prénom et date de naissance, mais aussi NIR, numéro de téléphone portable, adresse physique ou électronique. Cela n'empêche cependant pas toujours la réidentification de l'individu dans la base par croisement avec des données externes. Par exemple, un individu désidentifié, mais dont on sait qu'il était âgé de 64 ans lors de son hospitalisation au CHU de Nantes le 11 février 2019 pour une insuffisance cardiaque aiguë, est réidentifiable par croisement avec d'autres sources (SNDS, SIH, entrepôts de données de santé, notamment).

Dénombrement

Comme son nom l'indique, le dénombrement consiste à obtenir le nombre de patients présentant un profil précis. Un exemple est donné **Annexe 2**.

Entrepôts de données de santé (EDS) vs plates-formes ou « hubs »

Un entrepôt de données de santé (EDS) vise à centraliser et structurer les données issues de différentes sources (soin, recherche ou médico-administratives) et à en faciliter l'accès pour des activités liées au soin, à la recherche ou aux tâches administratives hospitalières. Le recueil est pensé de façon systématique, et les données peuvent ou non être qualifiées (l'information de certaines variables ou comptes rendus peut ne pas être caractérisée – *on peut stocker sans savoir ce qui est stocké*). La finalité de l'usage des données n'est pas acquise. Certaines données sont susceptibles de figurer dans l'entrepôt sans jamais être interrogées.

A la différence des EDS, les plates-formes de données de santé ou « *hubs* », tels le *Health Data Hub* (HDH) national français ou le Ouest Data Hub (ODH) inter-régional du Grand Ouest (groupe HUGO), ne permettent pas le stockage des données sans finalité. Ces plates-formes sont destinées à héberger des données en lien avec un projet spécifique (pas nécessairement un projet de recherche) et de façon limitée dans le temps. Cela sous-tend en particulier un principe de parcimonie, inapplicable aux EDS.

Harmoniser des données

Dans une approche multi-source, il s'agit de « modifier ces données pour les rendre identiques quelle qu'en soit la source ».[3] Comme le soulignent Goldberg et Zins, ce choix est dicté par les objectifs scientifiques des analyses envisagées.

Mapping ou « mise en correspondance »

Le *mapping*, dans l'acception informatique du terme, est l'opération de transcodage d'une variable. Dans le cadre de l'exploitation des EDS, c'est une étape indispensable d'abord pour l'exploitation locale des données, ensuite pour la centralisation de données issues de plusieurs EDS, par exemple vers une plate-forme telle que l'ODH. Ce transcodage permet de partir d'une donnée non aisément interprétable pour arriver à une donnée plus qualifiée et directement interrogeable. Prenons le cas des intitulés des documents textes associés aux hospitalisations : les médecins vont leur donner un

nom, parfois prédéfini, parfois rempli manuellement (« CR de consultation lipidologie », « CRC lipido », etc.). Le transcodage va consister à catégoriser ces documents selon une typologie, qui pourra éventuellement avoir plusieurs niveaux de complexité (CRC dans une colonne, lipidologie dans une autre). A l'échelle de l'ODH, le travail de mise en commun de la stratégie de *mapping* va nous permettre d'interroger de façon similaire les différents EDS, et donc d'uniformiser les données de sortie. Au niveau actuel de déploiement, cela est particulièrement utile pour les comptes rendus médicaux et les codes biologiques, volontiers hétérogènes entre les centres en dépit d'un effort de standardisation.

Screening

Le screening est un cas plus évolué de dénombrement. La liste des patients identifiés est mise à disposition d'un investigateur, en vue notamment d'une utilisation à visée de soin ou de recherche. Un exemple est donné **Annexe 3.**

PLAN DU MANUSCRIT

Dans la partie I, je présente la notion de donnée de santé et propose une typologie de trois grandes sources selon leur finalité – recherche, médico-administrative et soin. Les deux dernières correspondent à des données dites de « vie réelle », pour lesquelles je décrirai en particulier les données de l'Assurance Maladie accessibles (SNDS) et celles des entrepôts hospitaliers du Grand Ouest, issues d'EHOP, avant de discuter leur qualité, leurs liens et les modalités d'accès.

Dans la partie II, je présente séparément l'épidémiologie de l'insuffisance cardiaque et de la microangiopathie diabétique, puis leurs liens épidémiologiques et physiopathologiques supposés.

La partie III est une illustration de l'exploitation de données issues de la recherche à partir d'une analyse de la cohorte SURDIAGENE, s'intéressant aux liens entre des biomarqueurs nutritionnels (méthylamines) et l'insuffisance cardiaque. Elle a fait l'objet d'un article accepté dans la revue *Cardiovascular Diabetology* en février 2022.

La partie IV est une illustration de l'exploitation des données issues des bases de données médico-administratives. Il s'agit de l'étude DMC (*Diabetes Multiple Complications*) menée à partir des données de l'ensemble des personnes identifiées comme diabétiques dans les bases du SNDS sur la période 2012-2018. Le travail s'intéresse au risque d'insuffisance cardiaque associé à une atteinte rétinienne attribuable à la microangiopathie diabétique. Les points méthodologiques les plus importants sont détaillés, et le produit final est résumé sous forme d'article scientifique, en soumission au moment de l'envoi du présent manuscrit.

La partie V est une illustration de l'exploitation des données issues du soin à partir des entrepôts de données de santé (EDS) du groupe HUGO. Il s'agit de l'étude GAVROCHE (*Glycemia And its Variability in Regards tO Congestive Heart failurE*) menée à partir des données de l'ensemble des adultes hospitalisés pour insuffisance cardiaque aiguë identifiables dans les EDS des 5 CHU de HUGO, afin d'étudier l'association entre la variabilité glycémique et le risque de décès hospitalier. Ce travail n'est pas achevé au temps de l'envoi du manuscrit, je présente notre démarche en mettant l'accent sur les hypothèses de travail, le circuit des données et l'approche de traitement automatique du langage naturel ou TALN.

La partie VI propose une discussion sur la gestion des données selon leur source, les capacités de croisement, la place accordée au consentement des patients, et conclut sur une proposition de formalisation de bonnes pratiques d'analyses sur données dites « en vie réelle ».

PARTIE I : SOURCES ET FINALITES DES DONNEES DE SANTE

1. Introduction – données de santé et typologie

a. Quelques définitions des données de santé, importance du contexte et de l'utilisation

Une première définition de la donnée de santé peut être lue sur le site Internet de la CNIL[2] : « *Les données à caractère personnel concernant la santé sont les données relatives à la santé physique ou mentale, passée, présente ou future, d'une personne physique (y compris la prestation de services de soins de santé) qui révèlent des informations sur l'état de santé de cette personne.* »

Cette définition est très proche de celle du texte officiel du RGPD (Règlement Général pour la Protection des Données à l'échelle européenne, chapitre 1, article 4) qui définit les données concernant la santé comme *« toute donnée personnelle associée à l'état physique ou mental d'une personne physique, incluant les remboursements de soins, et révélant une information sur l'état de santé de la personne*[3]. » On notera cependant la différence entre le « *relatives à l'état* […] » de la CNIL et le *« associé à l'état* […] » du RGPD, ce dernier ouvrant la voie à l'association « statistique », et donc à toute donnée informant même très indirectement de l'état de santé d'un individu.

Cette définition très large est rendue d'interprétation difficile par les exemples donnés ensuite par la CNIL : la mesure du nombre de pas par jour n'est plus une donnée de santé « si cette donnée n'est pas croisée avec d'autres ». Par ailleurs, un certificat *d'aptitude* à la pratique du sport n'est pas considéré comme une donnée de santé, mais *l'inaptitude* à l'exercice d'une activité en est une. Pour sortir du paradoxe au moins apparent de ces exemples très sommaires, et qui mériteraient d'être débattus, il nous faut acter que la donnée n'est rien sans son contexte. Celui-ci permet de mieux la qualifier : l'information « déplacement de zéro pas par jour » fournie par un téléphone peut être signe d'un individu grabataire comme d'une inactivité de la fonction de géolocalisation. L'information « nationalité française » ne sera pas interprétée de la même manière dans un cabinet de médecine générale à Paris qu'à Pékin, au temps du COVID-19 ou de la variole du singe. Par ailleurs, un corollaire est qu'une donnée individuelle peut re-qualifier en donnée de santé l'information associée à d'autres individus dans une même base. En effet, la variable « aptitude à la pratique du sport » complétée

[2] Consulté en 2020 sur https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante ; les exemples cités étaient encore présents au 20 septembre 2022

[3] Traduction personnelle de l'anglais : « *Data concerning health means personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveals information about his or her health status* »

uniquement positivement dans une population fait suspecter en creux l'information de l'inaptitude pour le reste de la population.

Une première conclusion est qu'il est souvent facile d'établir qu'une donnée est une donnée de santé mais très difficile de prétendre le contraire, puisque cela dépendra du contexte de recueil, des données associées et de la finalité. Dans le cadre de ce manuscrit, nous partirons du postulat prudent que toute donnée individuelle est une donnée de santé avérée ou qui s'ignore.

Ces données de santé peuvent avoir de multiples utilisations (soins courants ou organisation des soins, pilotage, recherche, veille, alerte, administration, économique, etc.), pour de multiples utilisateurs (patients, soignants, décideurs, structures de recherche, assurances - acteurs publics comme privés, etc.), et de multiples modes d'analyse (autocontrôle glycémique chez une personne diabétique, application directe pour le soin sous la forme d'aide à la décision, exploitation épidémiologique et statistique, ou encore des usages plus récents que sont les « -omiques », etc.), comme cela a pu être décrit ailleurs.[4] Nous assumerons ici notre perspective, celle d'un laboratoire de recherche public ayant pour objectif d'exploiter des données à des fins de recherche pour répondre à des questions d'intérêt collectif.

b. Typologie des données de santé

Les données de santé peuvent être classées selon de nombreuses typologies :

- **Selon la méthode de recueil :** données issues d'interrogatoire, d'examen clinique ou d'examens complémentaires (biologiques, radiologiques, etc.), voire moins classiquement données de déplacement issues de la géolocalisation, données d'exposition au reste de la population (comme le célèbre *COVID tracking* mis en place en Corée du Sud [5]) ou encore exposition météorologique (données des stations météorologiques exploitables par exemple dans une étude portant sur la bronchopneumopathie chronique obstructive [6])

- **Selon la nature de leur codage : données structurées,** lorsque l'information est déjà catégorisée et accessible au niveau élémentaire, quel que soit le niveau de complexité associé (poids exprimé en kg, mais aussi codages plus complexes définissant exhaustivement le contenu d'une boîte de médicaments (codage international CIP) ou un dispositif médical comme un stent artériel (codage français LPPR)) ; **et données non structurées,** lorsque la source impose une étape de transformation pour extraire une ou plusieurs informations :

comptes rendus médicaux pouvant être interprétés par un individu ou une machine (traitement automatique du langage naturel ou TALN), mais aussi tracé ECG ou image radiologique, n'apportant aucune information directe mais nécessitant là aussi une interprétation humaine ou machine avant toute exploitation diagnostique, thérapeutique ou statistique

- **Selon le caractère prospectif ou rétrospectif du recueil :** cette définition n'est pas universelle, mais une étude prospective peut être définie par un recueil de l'information avant que l'événement de santé soit survenu chez les sujets.[7] Cette classification est parfois mise en avant pour opposer un recueil de qualité (prospectif) d'une approche potentiellement dégradée (rétrospective) puisque le prospectif permet la standardisation de l'approche, en pensant au plus précis le mode de recueil de la donnée et en y apportant une attention particulière au temps du recueil. Par exemple, l'investigateur mesurera la pression artérielle au repos en respectant certains standards, ou pensera à interroger systématiquement le patient sur son exposition à tel produit toxique. Cela permet *a priori* une meilleure qualité, par comparaison avec une approche rétrospective qui doit se contenter d'une information sans contrôle du mode de recueil, voire sans pouvoir s'assurer de l'existence même du recueil.
  Ainsi de la lecture d'un compte-rendu de consultation : l'affirmation « pas d'hypertension » n'indique pas si la pression a été réellement contrôlée, ni comment. Et si la présence d'une hypertension artérielle n'est pas discutée dans le CR, en déduire son absence revient à concéder un biais de mesure non contrôlé. Mais le caractère prospectif du recueil n'est une condition ni nécessaire ni suffisante à une donnée de qualité. Une étude rétrospective menée dans un service de nutrition appliquant en routine une mesure standardisée du poids pourra prétendre à des données de meilleure qualité qu'une étude prospective menée en service d'urgence, où l'exactitude du poids du malade n'est pas la priorité du personnel soignant. Un exemple est donné par différentes études prospectives, canadiennes et états-uniennes, visant à apprécier la prévalence de l'obésité en population : les auteurs observent un biais de sous-estimation du poids et de surestimation de la taille entre les valeurs déclarées et mesurées, biais d'autant plus difficile à apprécier qu'il était lié au genre et au pays.[8]

Ces trois typologies ont leur intérêt, mais nous prenons le parti d'aborder la question des données sous un 4ème angle, celui de leur finalité, pour en analyser les conséquences sur leur qualité et leur accessibilité. Nous entendons par-là leur finalité au temps de leur « production » - le but poursuivi lors du premier recueil de la donnée, qu'il s'agisse d'un recueil effectué par un être humain (saisie de

température, rédaction de compte-rendu, etc.) ou automatisé (moniteur cardiaque, transmission de délivrances médicamenteuses à l'Assurance Maladie, etc.). Il pourra s'agir de données primitivement recueillies pour la recherche scientifique, ou des données dites de « vie réelle » et secondairement exploitables à des fins de recherche comme les données médico-administratives ou les données issues du soin (sans caractère limitant de cette définition, puisqu'on aurait aussi pu s'intéresser aux finalités commerciales, environnementales, judiciaires, etc.). En effet, le diagnostic de diabète ou d'insuffisance cardiaque ne sera ni recueilli, ni détaillé de la même façon dans un compte-rendu (CR) de consultation, un codage diagnostique issu du PMSI-MCO ou au terme d'un examen clinique standardisé et anticipé dans le cadre d'une étude de cohorte prospective.

Nous proposerons d'abord une description de chacune de ces trois sources, en détaillant le cas du SNDS pour les bases de données médico-administratives (BDMA) et celui de l'entrepôt de données de santé (cas d'usage avec le logiciel eHOP au CHU de Nantes) pour les données issues du soin. La section suivante sera consacrée à la qualification de la donnée, discutée pour chacune des sources, puis l'accès à ces sources et, brièvement, les questions de gouvernance et de réglementation. La question du lien entre ces sources sera abordée dans la dernière section, afin de paver la voie aux trois projets de recherche conduits dans le cadre de ce travail : SURDIAGENE, DMC et GAVROCHE.

2. Données recueillies à des fins de recherche

C'est le cas des essais cliniques interventionnels comme des études de cohorte prospectives, définissant en amont une population d'intérêt, la nature des données recueillies et le mode de recueil (papier, tableur, questionnaire électronique (eCRF)). Habituellement, elles posent au moins un objectif associé à une question scientifique, ce qui permet d'écrire précisément une méthode et en particulier de détailler les données nécessaires, pensées conjointement aux analyses statistiques. Un exemple fondateur est la *Framingham Heart Study (FHS)* lancée en 1948 et encore active 74 ans après, abordant la 3ème génération de participants, et qui a fortement contribué à l'identification des principaux facteurs de risque cardiovasculaire en population.[9]

Plus récemment, la cohorte nationale française CONSTANCES, elle aussi menée en population générale et ayant inclus depuis 2011 plus de 200 000 personnes adultes (18-80 ans) rattachées au Régime général de la Sécurité Sociale, a permis le recueil prospectif de données de santé, comportant interrogatoire médical et paramédical, examen clinique et examens complémentaires standardisés, avec des questionnaires annuels et une visite médicale tous les 5 ans.[10] Elle illustre bien également

la porosité recherche/soin, avec des examens de dépistage ophtalmologiques et ORL, et nous permettra de faire le lien avec les BDMA et la question de l'accès aux données.

Le principal défaut de cette démarche de recherche est son coût : le travail des équipes de recherche, le temps donné par les patients, les ressources techniques et réglementaires nécessaires sont très importants, précisément du fait de cette finalité première de la recherche. Ce coût pourra être amorti en étant associé à des soins courants, comme c'est le cas pour l'étude de cohorte prospective COHPT (NCT05469087, s'intéressant à l'hyperparathyroïdie primaire), la prise en charge « normale » du patient consultant à l'hôpital étant enrichie sur la base du volontariat, voire du bénévolat soignants/patients par un recueil de données pensé pour la recherche (interrogatoire, questionnaire SF-36, biocollection).[11]

Ce cadre sera celui de l'étude SURDIAGENE, dont une analyse est proposée dans la partie III du manuscrit.

Rappelons également que le recueil de données à des fins de recherche crée une situation artificielle, qui, du fait des contraintes éthiques (consentement, acceptabilité de la recherche par les patients et les investigateurs) et du caractère interventionnel a minima (par définition, charge supplémentaire aux soins courants) entraîne des biais de sélection des patients et des biais de mesure de l'information.

Enfin, au même titre que les données de « vie réelle » que sont les données médico-administratives ou du soin, le recueil de données à des fins de recherche n'exclut pas des usages secondaires non anticipés, impliquant par exemple une harmonisation des données (pour l'interopérabilité) ou des dosages imprévus, voire impensables, au temps de la constitution de la biocollection.

3. Données recueillies à des fins médico-administratives

Leur recueil est prospectif car il répond à un standard défini avant la production de la donnée, avec cette spécificité que ce standard est défini par un besoin non médical mais administratif. L'exemple français le plus classique est celui des bases de données de l'Assurance Maladie, regroupées dans le SNIIRAM (Système National d'Information Inter-Régimes de l'Assurance Maladie) et mises en lien avec plusieurs autres BDMA dans le SNDS (Système National des Données de Santé) géré par la CNAM, où l'on trouve (liste non limitative[4]) :

[4] Pour aller plus loin, une description du SNDS par lui-même sur https://www.snds.gouv.fr/SNDS/Qu-est-ce-que-le-SNDS

- le PMSI (Programme de Médicalisation des Systèmes d'Information)
- la BCMD (Base des Causes Médicales de Décès, issue du CépiDc et administrée par l'INSERM)

Le SNDS (ex-SNIIRAM-PMSI) est une base complexe en constante évolution, déjà décrite ailleurs [12,13], qui contient plus de 3000 variables occupant plus de 450 téraoctets. J'en propose un schéma très simplifié **Figure 1**. Le SNDS contient des informations médicales principalement issues du remboursement des soins, pour plus de 99% de la population française, avec un niveau d'exhaustivité en constante amélioration depuis 2005. Brièvement, il permet en particulier d'accéder aux données individuelles désidentifiées (mais non anonymisées)

- de réalisation des consultations médicales et des actes médicaux et paramédicaux
- des diagnostics associés aux affections de longue durée (ALD)
- du détail des délivrances médicamenteuses remboursées par la sécurité sociale (et donc sans les délivrances non remboursées, qu'elles soient prescrites ou en vente libre)
- des diagnostics et des actes médicaux associés aux séjours hospitaliers, publics comme privés, avec une qualité et une précision très différentes pour ses quatre champs que sont le PMSI-MCO (médecine, chirurgie, obstétrique et odontologie), -HAD (hospitalisation à domicile), -SSR (soins de suite et de réadaptation) ou -PSY (psychiatrie).

La finalité administrative, pensée via le prisme du financement (des centres hospitaliers et des soignants) et des remboursements (des bénéficiaires et des prestataires) détermine la présence et la nature de l'information.

Par exemple, un médicament délivré en pharmacie de ville pourra être tracé de façon exhaustive (metformine, censément toujours prescrite et toujours remboursée), partielle (anti-inflammatoires non stéroïdiens type ibuprofène, parfois prescrits et remboursés, parfois sans prescription) ou jamais ou exceptionnellement (cas des inhibiteurs de la phosphodiestérase de type 5). De même, le diagnostic d'insuffisance cardiaque associé à une ALD sera inconstamment déclaré si un patient bénéficie déjà d'une ALD pour coronaropathie. En effet, la déclaration ne sera pas justifiée économiquement pour le patient si les traitements de la première sont déjà pris en charge par la seconde, conduisant à une sous-déclaration par le médecin traitant. Enfin, la grande déception du clinicien accédant au SNDS concerne les examens au sens large : s'il peut savoir que le patient a bénéficié d'une prise de sang, d'une épreuve d'effort ou d'une IRM, le résultat n'est pas accessible puisque le coût de l'examen n'est pas fonction de son caractère normal ou pathologique.

Dans la perspective d'une exploitation à des fins de recherche, ces données déjà recueillies présentent l'immense avantage de leur centralisation (rendue possible ici par l'effort d'harmonisation nationale) et d'un coût en apparence nul pour l'investigateur rattaché à certains organismes publics (équipe INSERM ou CHU), qui peut en demander l'accès sans contrepartie financière. Cependant, ce serait faire l'impasse sur le temps humain nécessaire à l'accès aux données (démarches administratives, plate-forme d'exploitation type portail CNAM voire systèmes-fils) et sur les compétences nécessaires pour qualifier l'usage scientifique de ces bases.

C'est ce travail qui sera illustré par l'étude DMC proposée dans la partie IV de ce manuscrit.

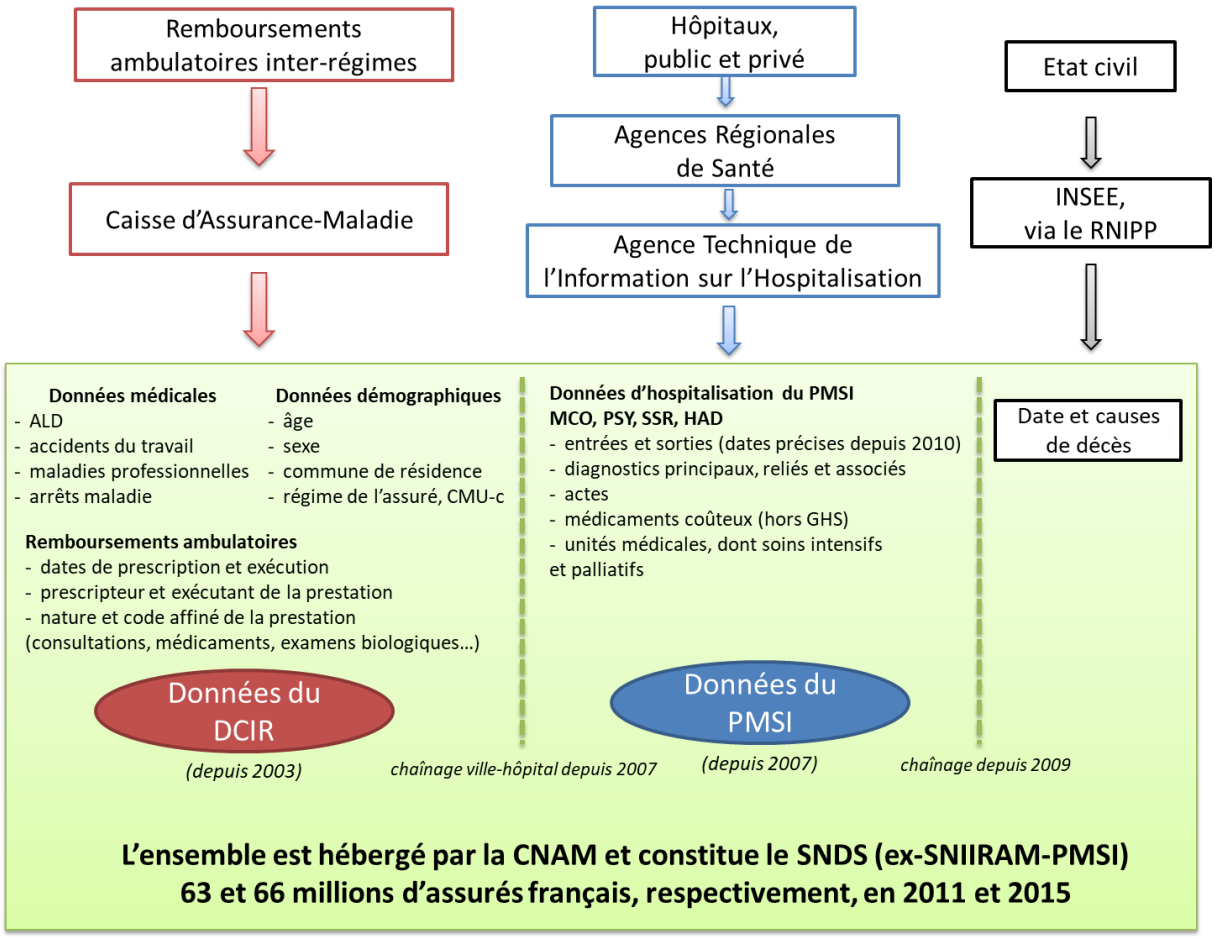


Figure 1. Représentation simplifiée de l'architecture du SNDS (ex-SNIIRAM-PMSI)

ALD : Affection de Longue Durée. CMU-c : Couverture Maladie Universelle complémentaire. CNAMTS : Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés. DCIR : *Datamart* Consommation Inter-Régimes. GHS : Groupe Homogène de Séjours. HAD : Hospitalisation A Domicile. INSEE : Institut National de la Statistique et des Etudes Economiques. MCO : Médecine, Chirurgie, Obstétrique et Odontologie. PMSI : Programme de Médicalisation des Systèmes d'Information. PSY : services de psychiatrie. RNIPP : Registre National d'Information des Personnes Physiques. SNDS : Système National des Données de Santé. SNIIRAM : Système National d'Information Inter-Régimes. SSR : Soins de Suite et de Réadaptation.

Schéma inspiré des articles de Moulis *et al.* [12] et de Tuppin *et al.*[13], et enrichi de l'exposé de Dominique Polton et Philippe Ricordeau[14]

4. Données recueillies pour le soin – cas particulier des EDS et d'eHOP

a. Un accès inégal aux données

Toutes les données produites dans le cadre du soin peuvent *a priori* faire l'objet d'une transformation à des fins de recherche. C'est le cas de la situation déjà discutée de cohortes prospectives où certaines données sont recueillies en marge des soins courants. C'est également le cas de nombreuses études rétrospectives, par exemple lorsqu'une équipe décide de recueillir et analyser les données passées d'une population spécifique qu'elle a prise en charge, comme, pour prendre un exemple parmi d'autres, le dépistage des facteurs de risque cardiovasculaire après une grossesse compliquée de pré-éclampsie, en hospitalisation de jour de médecine interne.[15]

La démarche d'identification de la population, ou *screening*, peut être anticipée par la tenue de listes de patients par pathologie, mais cela entraîne une surcharge de travail de l'équipe soignante, sans garantie que l'information soit un jour exploitée. Il est également possible d'interroger un système d'information local : requête sur codes PMSI via les départements d'information médicale (DIM), ou sur valeurs biologiques dans les bases de biologie. Le *screening* peut être enrichi de l'extraction d'informations structurées en vue de la constitution d'une étude (âge, sexe, valeurs biologiques ou codes issus du PMSI). Dans la perspective d'une recherche scientifique, et en particulier si la recherche n'est pas limitée à un service, ces démarches supposent un cadre réglementaire et sont limitées par la capacité de réponse du service, ainsi que par la qualité des données consultables dans le SIH.

b. La révolution des EDS

Ces limitations sont en passe d'être dépassées par la lente mais prometteuse révolution de l'accès aux SIH que constituent les entrepôts de données de santé, ou EDS. Il s'agit, à l'échelle d'un ou plusieurs centres, de favoriser la création d'une base de données unique concentrant différents flux d'information du SIH (comptes rendus, résultats biologiques, données PMSI…), associée à des moyens dédiés à l'interrogation de la base. Au CHU de Nantes, cela a débuté par une réflexion sur la gouvernance et la conformité légale de l'approche. Depuis 2017, nous avons commencé à déployer la solution eHOP (ex-Roogle) développée par l'équipe DOMASIA (DOnnées MAssives et Systèmes d'Information Apprenants en santé) du Laboratoire du Traitement du Signal et de l'Image de Rennes (LTSI – UMR 1099), équipe dirigée par le Pr Marc Cuggia. Avec l'appui du Réseau Inter-régional des Centres de Données Cliniques (Ri-CDC), eHOP est progressivement déployé au sein du groupe HUGO (Hôpitaux Universitaires du Grand Ouest) qui regroupe les CHU d'Angers, Brest, Nantes, Rennes

et Tours ainsi que l'Institut de Cancérologie de l'Ouest.[16] Il permet ainsi l'intégration de données hétérogènes issues du SIH grâce à des connecteurs spécifiques aux différentes sources.

Ce déploiement à l'échelle inter-régionale permet la mise en commun de l'expérience de l'usage de l'outil, la généralisation simple de procédures de dénombrement ou de *screening* directement partageables sous forme de code entre les centres et, à terme, l'export de données standardisées vers une plate-forme unique. Cette plate-forme ou « *hub* » est appelée ODH pour *Ouest Data Hub*. Elle se distingue en particulier des EDS locaux par le fait qu'elle est destinée à héberger des données spécifiques en lien avec un projet de recherche, et non à stocker systématiquement les données issues du soin et des BDMA. Les éléments techniques du logiciel sont brièvement décrits ici.[17,18] Pour les paraphraser, eHOP respecte différentes normes d'interopérabilité (HL7, CDISC, HPRIM et PN13), ce qui facilite l'intégration de données issues de SIH hétérogènes. Bien qu'utilisant des librairies *OpenSource*, eHOP n'est pas en soi une solution *OpenSource* mais fait l'objet d'un développement industriel en lien avec la société ENOVACOM, filiale d'Orange Business Services[5].

c. Données disponibles dans l'EDS nantais

Une notion centrale dans la compréhension du déploiement d'eHOP est celle de flux. Par un raccourci sémantique, on parlera parfois de flux pour un logiciel, tel que le logiciel Millennium où sont saisis la majorité des CR du CHU de Nantes, ou le logiciel DXLAB pour les données de biologie. En réalité, il s'agit plutôt d'un sous-ensemble de données formatées pour l'un de ces SI, et dont la cohérence *à la fois sémantique et de stockage* donne une forme d'homogénéité au chargement de l'information.

Pour être exploitables, les données saisies sur un logiciel nous intéresseront dans la mesure où elles auront pu être catégorisées. Chaque flux dépend lui-même d'au moins un logiciel du SIH, actif ou non, avec son propre rythme d'alimentation vers eHOP (quotidien, hebdomadaire ou mensuel). A noter qu'il n'y a pas actuellement de véritable « fil de l'eau » qui correspondrait à une intégration des données en temps réel. Une telle alimentation est plus complexe à maintenir et présente un intérêt faible pour les usages actuels d'eHOP. Cette alimentation serait nécessaire dans la perspective d'un appui direct au soin, par exemple sous la forme d'une médecine « personnalisée » dépendante de multiples informations récupérées en temps réel, comme cela a pu être illustré par l'aide à la prescription de drogues vasopressives en réanimation.[19] Mais ceci est hors du champ du présent travail.

[5] Plus d'information sur https://www.enovacom.fr/enjeux/ameliorez-prise-decision-entrepot-donnees

Au 1er septembre 2022, les principaux flux de l'EDS du CHU de Nantes étaient

- **un flux « identité »,** permettant de faire le lien entre les différentes informations associées à un patient
- **des flux issus du PMSI-MCO,** intégrant en particulier les codes de diagnostics (principaux, reliés et associés) CIM-10 associés aux séjours, les actes médicaux (classification commune des actes médicaux ou CCAM) et les codes GHM (Groupe Homogène de Malades), mais pas les codes LPPR (Liste des Produits et Prestations Remboursables) correspondant entre autres aux dispositifs médicaux (ce qui constitue actuellement une limite pour la matériovigilance)
- **un flux des données de biologie,** n'incluant pas toujours les résultats des examens externalisés
- **un flux de comptes rendus** : consultation, hospitalisation, opératoires, etc. issus de différents logiciels métiers
- **un flux « administrations médicamenteuses »,** avec des tables de correspondance (ou *mapping*) permettant leur interrogation par code CIP (Club InterPharmaceutique), DCI (Dénomination Commune Internationale) ou classe ATC (Anatomique, Thérapeutique et Chimique). Ce flux n'intégre cependant pas les chimiothérapies
- **et un flux dit « anthropométrie »,** qui reprend des données cliniques et biologiques élémentaires saisies sous forme structurée (poids, taille, température, etc. dont les précieuses valeurs de glycémie capillaire pour le projet GAVROCHE)

Ces flux sont en constante évolution et nous travaillons actuellement sur l'intégration des données d'anesthésie et de réanimation, d'odontologie et d'ophtalmologie, ce qui demande un effort coordonné de nombreux acteurs : le service concerné par le flux, la Direction des Services Numériques (DSN), l'équipe de la Clinique des Données, et parfois la société distribuant le logiciel métier, le RiCDC ou la société ENOVACOM qui commercialise eHOP.

d. Interroger les données via eHOP

eHOP peut être défini comme une solution logicielle permettant la structuration d'un entrepôt de données biomédicales, autour d'une base construite sur le format Oracle et associée à une interface graphique permettant la consultation directe de ces données. En particulier, cette interface autorise la constitution d'échantillons de données sur des sous-groupes de patients d'intérêt, tout en excluant automatiquement les patients « opposés » et en appliquant des procédures de désidentification des données.

En pratique, il est possible d'interroger ces données via différentes applications :

(i) via un « requêteur » dédié développé par le LTSI, qui offre une interface graphique « clique-boutons » ne nécessitant pas de connaissance en programmation et donc accessible à tous (cf. **Figure 2**),

ou

(ii) via tout logiciel permettant une connexion à la base Oracle, dont nous citerons deux exemples : le logiciel de statistiques *R* [20] et le logiciel *Oracle SQL developer* développé par la société Oracle.[21]

Cette dernière approche offre une plus grande autonomie et une meilleure maîtrise de l'interrogation des données. Elle permet une montée progressive en compétence des utilisateurs, mais elle nécessite des aptitudes en langages de programmation (ici R, langage SQL ou autre). Pour un utilisateur, il apparaît nécessaire d'avoir fait l'expérience de ces différents modes d'interrogation des données d'eHOP afin d'être en mesure de choisir laquelle de ces approches sera la plus sûre et nécessitera le temps de développement le plus court.

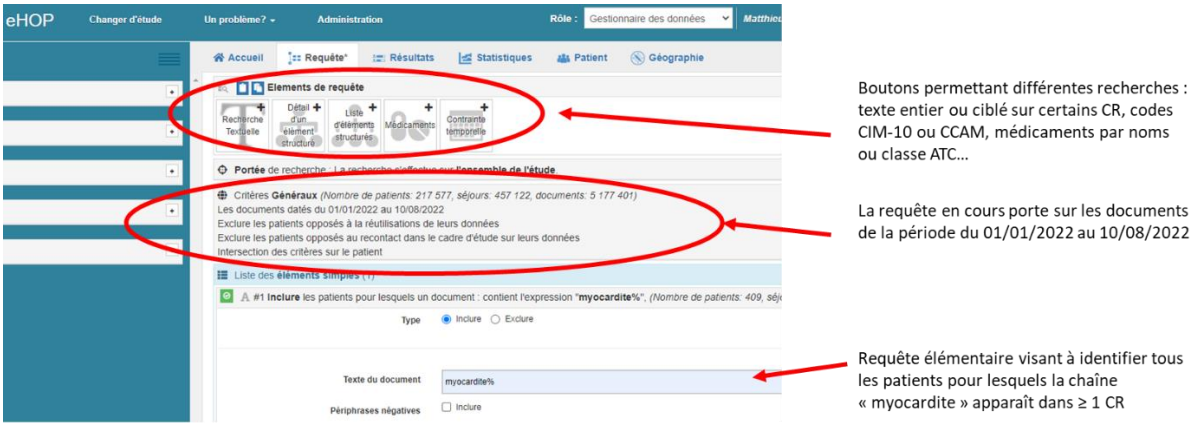


Figure 2. Exemple de requête de l'interface eHOP – capture d'écran sur l'intranet du CHU de Nantes.

Nous illustrerons l'exploitation inter-régionale de ces EDS par le projet GAVROCHE, en partie V de ce manuscrit.

5. Qualifier des données

Par « qualification des données », nous entendons ici notre capacité à connaître la qualité de la donnée, donc à faire correspondre un champ dans une table à une définition précise. Au niveau élémentaire en épidémiologie des populations (se rapportant à des individus humains, mais transposable à toute unité statistique d'intérêt, une souris comme un organe ou une procédure chirurgicale), cette qualification nécessitera quatre informations, avec un niveau de précision très variable :

- l'identification de l'observation (ou unité statistique) : par exemple un code unique désignant un individu
- le temps de mesure : année, jour, voire une précision heure/minute/seconde
- la nature de la donnée : glycémie capillaire à jeun avec la désignation commerciale ou technique de la bandelette et du lecteur glycémique utilisés, et l'unité (mmol/L, g/L ou mg/dL)
- la valeur de la donnée : par exemple la valeur numérique « 5 » pour une glycémie en mmol/L

La qualification constitue un prérequis fondamental dans l'exploitation des bases à des fins de recherche. Sans donnée suffisamment qualifiée, l'analyse statistique de l'information donnera des résultats ininterprétables. Enfin, notre capacité à agir sur cette qualification est intrinsèquement liée à la finalité du recueil.

Les données issues de la recherche permettent d'offrir le cadre idéal de qualification de la donnée : un standard peut être défini à l'avance, par exemple un diagnostic de diabète sucré selon *l'American Diabetes Association.*[22] Lors de l'inclusion des patients dans une étude, les investigateurs se baseront sur cette définition pour qualifier le patient comme étant ou non diabétique. Cela s'applique également à toute mesure, qu'elle s'appuie sur une approche clinique, biologique, radiologique ou autre (électrocardiogramme, électro-encéphalogramme, etc.). Cette standardisation est une condition nécessaire mais non suffisante puisqu'il faudra qu'elle soit précise et compréhensible, et que l'effort de qualification des données soit réellement fourni par les personnes effectuant le recueil, sans erreur de report. Dans l'exemple élémentaire de la mesure du poids d'un patient, le risque est par exemple de se contenter d'un poids déclaratif, de donner ce poids en situation pathologique (tableau œdémateux dans l'insuffisance cardiaque aiguë), ou encore de fonder la mesure sur des balances de mauvaise qualité ou non tarées.

Les données issues des bases médico-administratives peuvent également être recueillies prospectivement et répondre à des standards de qualification. Ainsi, les données issues des PMSI des

centres hospitaliers français, intégrées au niveau national dans l'ATIH et qui se retrouvent pour partie dans le SNDS doivent respecter des standards de qualification permettant la tarification à l'activité (T2A).

L'expérience montre que certains codes se révèlent parfaitement fiables. Par exemple, le résultat d'un *screening* via les codes CCAM réalisé au CHU de Nantes pour les services de chirurgie digestive et endocrinienne et de nutrition (projet NAMICO des Pr Claire Louis-Blanchard et David Jacobi) coïncidait parfaitement aux patients identifiés au fil de l'eau par le service sur l'année 2019 comme ayant bénéficié d'une chirurgie bariatrique de *sleeve-gastrectomie* ou de *bypass* gastrique. Un autre exemple est une étude nationale française portant sur la mortalité et le soin de patients présentant une sclérodermie systémique associée à une pneumopathie interstitielle sur la période 2010-2017.[23] Pour identifier la population, les auteurs s'appuient notamment sur des codes CIM-10, et doivent concéder comme limite leur incapacité à apprécier le sous- et surcodage de ces approches, déterminant leur sensibilité et spécificité. Un screening opéré sur les bases eHOP du CHU de Nantes permet par exemple, en un travail d'une dizaine de minutes et sans accéder aux identifiants directs du patient, de contrôler un échantillon de 20 cas présentant les codes correspondant à la sclérodermie, pour observer que 19 diagnostics sont médicalement confirmés et un cas est fortement suspecté.

Un exemple frappant est celui de la qualité de données *a priori* élémentaires. Comme on le verra dans la partie IV pour l'étude DMC sur les données SNDS, environ 1 individu sur 10,000 présentait des incohérences sur l'âge et le sexe, tandis que pour 1% de ces patients nous ne disposions pas de ces informations. Dans un autre travail en cours portant sur les dispositifs médicaux à élution de paclitaxel, l'étude DETECT (NCT05254106[6]), l'information sur le décès des patients dans le SNDS est récupérée dans au moins quatre sources différentes, ce qui permet de mettre en évidence un recouvrement inégal (données non encore publiées) :

Tableau 1. Population identifiée comme décédée dans l'étude DETECT	
On considère comme *gold standard* l'information du décès issue de la réunion des données DCIR, ATIH et CépiDc entre le 1er janvier 2011 et le 31 décembre 2020, disponibles dans le SNDS en mars 2022	N = 109 598
Décès identifié dans le DCIR (1) : table des bénéficiaires	106 002 (96,7%)
Décès identifié dans le DCIR (2) : table des consommations	62 084 (56,6%)
Décès identifié dans l'ATIH : mode de sortie = décès	58 999 (53,8%)
Décès identifié dans le CépiDc	47 714 (43,5%)

[6] Consultable sur le site *Clinical trials* : https://clinicaltrials.gov/ct2/show/NCT05254106 (article en cours d'écriture)

Il apparaît ainsi impossible d'affirmer la pertinence d'une identification par codage sans une étape de contrôle à la source, qui s'appuiera par exemple sur un échantillonnage. Encore cette approche ne permet-elle que d'affirmer la valeur prédictive positive du codage, tandis que sa sensibilité et sa spécificité ne sont jamais acquises puisqu'elles nécessiteraient un échantillonnage important sur l'ensemble des dossiers des patients, voire des examens complémentaires systématiques en population, afin de pouvoir prétendre à un dépistage uniforme des individus. Les exemples d'échantillonnage donnés ici restent limités car monocentriques et sur une période courte, ne pouvant là encore prétendre à une représentativité géographique ou temporelle, ou à l'absence de biais de sélection vis-à-vis des centres hospitaliers non universitaires, publics ou privés.

Des biais perçus comme faibles peuvent avoir des conséquences majeures sur l'estimation de la prévalence en populations non équilibrées. A titre d'exemple, sur l'hypothèse fictive mais plausible d'un cas de diabète de type 1 (DT1) contre 20 cas de diabète de type 2 (DT2) en France, avec 5% de cas de DT1 classés à tort comme DT2 et 5% de cas de DT2 classés à tort comme DT1, et en faisant l'hypothèse très favorable de l'absence de surdiagnostic chez les personnes non diabétiques, la prévalence du DT2 serait sous-estimée de 4,75% tandis que celle du DT1 serait surestimée de 95% - presque doublée. Cet exemple peut être illustré par sur l'étude CORONADO portant sur les cas de personnes diabétiques hospitalisées pour COVID-19.[24] Bien que la population ait été identifiée par des services de diabétologie, les cas de DT1 avaient d'abord été surestimés d'environ 50%, avant correction suite à une sollicitation ciblée de tous les centres sur les « cas suspects » identifiés par approche algorithmique (travail non publié du Pr Hadjadj et de moi-même).

Quant aux données issues du soin comme celles que l'on peut trouver dans les EDS, leur utilisation brute revient à laisser reposer la qualification sur la seule trace de l'activité de soin, loin des standards de la recherche. Par exemple pour l'EDS du CHU de Nantes, on ne saura généralement pas si un poids relève d'une valeur mesurée ou déclarée ; si une glycémie est mesurée à jeun ou après le repas, voire pour certains champs si elle a été saisie en mmol/L ou en g/L, valeurs pouvant être confondues aux extrêmes (hyperglycémie à 3 g/L ou hypoglycémie à 3 mmol/L). L'EDS pourra s'appuyer sur des données réputées fiables, comme par un exemple un dosage d'hémoglobinémie en laboratoire ou une administration médicamenteuse tracée par un code CIP, ou encore s'appuyer sur des données médico-administratives comme les codes CIM-10 ou CCAM déjà évoqués. Mais, dans certains cas, apprécier une donnée, même bien mesurée, en l'isolant de son contexte clinique, limite son interprétation : ainsi de l'hémoglobine glyquée (HbA_{1c}) si l'on ignore une dialyse, une transfusion sanguine ou une

drépanocytose associée. Naturellement, il existe des démarches d'évaluation de la qualité des données, à l'échelle du SIH ou de l'EDS d'un centre hospitalier.[25] Mais l'effort est considérable et, devant le caractère fluctuant du SIH comme de l'EDS qui lui est associé, l'état des lieux de la qualité risque d'être obsolète avant d'avoir été produit.

Parmi ces données issues du soin, nous avons jusqu'ici surtout discuté du cas des données structurées, où, dans un tableau, une case renvoie précisément à une mesure ou à un codage issu d'un travail de standardisation, aussi modeste soit-il. La complexité sera accrue pour les données non structurées, par exemple un compte-rendu (CR) médical. Ces données non structurées sont très importantes pour la recherche puisqu'elles sont parfois le seul moyen de connaître les antécédents du patient ou le résultat d'une échographie transthoracique. Certaines informations peuvent alors être extraites automatiquement par traitement automatique du langage naturel ou TALN. Un cas « simple » est celui de CR répondant à un effort de structuration systématique, comme peuvent l'être les CR opératoires (CRO) : c'est le cas pour le projet inter-régional HACRORTHO, porté par le CHU de Tours (Dr Leslie Grammatico-Guillon), s'intéressant en particulier aux chirurgies de prothèse de hanche. Il apparaît raisonnable de penser que certaines informations d'intérêt seront toujours présentes et exprimées de façon similaire dans les CRO, récupérables par des approches dites par « expression régulière », par exemple la latéralité de la prothèse et la voie d'abord chirurgicale.

Mais dans le cas du projet GAVROCHE qui va s'appuyer sur des CR hétérogènes (urgences, hospitalisation, avis d'équipe mobile), nous n'aurons la garantie ni de la présence des antécédents détaillés du patient, ni de la façon dont ils seront effectivement rapportés dans les CR.

6. Accès, gouvernance et aspects réglementaires

a. Généralités

La transposition dans le droit français des dispositions du RGPD a imposé un nouveau cadre pour le recueil, le stockage et l'usage des données individuelles, en particulier de santé.[26] En pratique, chaque source doit se voir définir son propre mode de recueil et de stockage, sa gouvernance, son cadre réglementaire et ses conditions d'accès, selon qu'il s'agit de données individuelles ou anonymes, issues de recherche observationnelle ou interventionnelle, et sa finalité recherche, médico-administrative ou de soin. Un guide pédagogique simple produit par le HDH et disponible en ligne résume la démarche réglementaire selon que les données intéressent ou non la personne humaine

(RIPH/RNIPH), et que le projet doit ou non faire appel à une autorisation CNIL. Celle-ci n'est pas nécessaire si les données sont anonymes ou si la recherche respecte le cadre donné par une méthodologie de référence (dites MR001, 2, etc.)[7].

Brièvement, les données de recherche et celles issues du soin peuvent être hébergées par l'établissement « producteur » de la donnée, comme un CHU dans le cas d'une étude de cohorte sur les patients consultant au CHU, ou être hébergées par un centre non directement impliqué mais agréé ou certifié HDS (hébergeur de données de santé).[27] Les transferts entre centres doivent être assurés de façon sécurisé, par exemple via la solution Bluefiles (Orange Healthcare®) pour le CHU de Nantes. Dans le soin, la gouvernance est décrite dans un protocole de recherche rédigé par le porteur du projet, ainsi que les conditions d'accès aux données et les règles de publication, éventuellement au sein d'un consortium. L'information et le consentement des patients sont régis par une méthodologie de référence de la CNIL ou selon la catégorie RNIPH (recherche n'impliquant pas la personne humaine) ou RIPH (recherche impliquant la personne humaine) 1 à 3.

b. Le cas particulier d'eHOP : cadre réglementaire, conditions d'accès et gouvernance de l'EDS nantais

Pour le cas particulier des entrepôts de données de santé, il est nécessaire de respecter une méthodologie de référence[8] (adoptée en novembre 2021 mais encore fortement susceptible d'évoluer) ou d'obtenir une autorisation spécifique de la CNIL. Le partage des données est possible via une plateforme de données de santé, l'ODH, avec une procédure particulièrement lourde nécessitant des autorisations successives du CSE de l'ODH, du CESREES puis de la CNIL, et incluant un PIA (analyse d'impact sur la vie privée).

L'EDS du CHU de Nantes a fait l'objet en 2018 d'une autorisation spécifique de la CNIL (cf. **Annexe 1**), qui définit précisément le cadre réglementaire et le périmètre d'exploitation : modalités de l'information apportée aux patients, population concernée, données, finalités autorisées pour l'exploitation de l'entrepôt (qui dépassent largement le cadre de la recherche scientifique), et conditions d'accès selon qu'il s'agit d'un dénombrement, d'un screening ou d'extraction de données de santé, et selon que les données accessibles aux investigateurs sont anonymisées, désidentifiées ou

[7] Rechercher « Guide pédagogique » sur https://www.health-data-hub.fr/documents (consulté le 20/09/2022)

[8] https://www.cnil.fr/sites/default/files/atoms/files/referentiel_entrepot.pdf

directement identifiantes. Le schéma de gouvernance, impliquant la direction de la recherche, la direction des services numériques, et d'autres représentants des instances hospitalières, dont la Commission Médicale d'Etablissement, expose les prochains axes de développement de l'EDS. Ces éléments sont résumés en annexe, avec les détails du réglementaire et de la gouvernance (**Annexe 1**), un bref exemple de dénombrement (**Annexe 1**), de screening (**Annexe 3)** et d'extraction de données pour enrichir une étude préexistante (**Annexe 4**).

c. Un cas d'école en cours : le Health Data Hub

Un espoir fédérateur a été porté par le lancement du chantier de la plateforme nationale de données de santé, le *Health Data Hub* ou HDH, sous la forme d'un groupement d'intérêt public depuis 2019. Cet espoir a rapidement été mêlé à des craintes quant aux conséquences de la centralisation et du mode d'hébergement [3], avec en particulier la remise en question de la solution de « Cloud » de Microsoft®.[28] Le HDH est un excellent exemple pour résumer les problématiques et solutions associées à la centralisation des données de santé en vue de leur exploitation[29] : (i) guichet unique des demandes d'accès, pour une simplification réglementaire et administrative ; (ii) guichet unique également pour les patients, qui pourraient s'opposer globalement à l'exploitation de leur données, ou « par étude » ; (iii) « catalogue » de données consultables par tous (toute donnée hébergée par le HDH serait une donnée partagée potentielle), afin de favoriser techniquement les échanges mais aussi limiter les réflexes d'appropriation individuelle voire d'accès discrétionnaire ; (iv) standardisation des formats (par exemple LOINC pour la biologie), avec une vertu de cette approche dite *top-down* encourageant toutes les échelles de l'interopérabilité ; (v) extension du SNDS actuel à « toute donnée du système de santé », afin de faire partager un socle légal commun à la donnée, quelle que soit son origine ; (vi) favoriser les plateformes locales, parmi lesquelles l'ODH a vocation d'exemple, toujours dans un esprit d'harmonisation *top-down*.

7. Enrichissements inter-sources

Comme on l'a vu, ces trois grandes sources de données sont liées, et chacune est susceptible d'être enrichie par les deux autres. Cet enrichissement peut prendre la forme d'un apport direct d'information ou de précisions sur la qualité de l'information.

Par exemple, des données recueillies dans le cadre d'une recherche peuvent être enrichies de données médico-administratives ou issues du soin. C'est ce que nous avons mis en place pour l'étude MOTHIF-II portant sur la prise en charge de l'hémophilie et portée par les Dr Marc Trossaërt et Valérie Horvais.[30] Sept centres de traitement de l'hémophilie ont recueilli de façon prospective auprès de plus de 2000 patients des caractéristiques de la maladie absentes des bases « vie réelle » actuelles. Ces données ont ensuite été associées aux données nationales du SNDS, par appariement individuel déterministe basé sur le NIR, ce qui a permis d'identifier et quantifier les délivrances médicamenteuses ainsi que les hospitalisations à risque d'accident hémorragique, et ainsi évaluer les conséquences en « vraie vie » d'une modification de prise en charge (traitements à longue durée d'action vs courte durée), tant cliniques (hospitalisations) qu'économiques (coût des médicaments). Pour prendre un autre exemple, la cohorte de patients diabétiques EDIT supervisée par le Pr Hadjadj est d'abord basée sur un phénotypage systématique des patients, avec une visite hospitalière répétée tous les 3 ans, mais sera enrichie des données SNDS pour mieux connaître les délivrances médicamenteuses et les complications (médicaments traceurs, ALD, hospitalisations) au fil du temps.

L'appariement de ces différentes sources de données permet aussi de contrôler la qualité d'algorithmes d'identification de maladies. Nous avons proposé plus haut des exemples simples concernant l'identification des cas de chirurgie bariatrique, ou de sclérodermie systémique : la structuration de l'EDS nantais permettait ainsi une première évaluation rapide d'approches respectivement recherche (contrôle par codages CCAM vs inclusion consécutive par les soignants du service) et médico-administrative (codage CIM-10 de données remontées au SNDS vs diagnostic clinique exprimé localement dans les CR).

Lors de l'épidémie de COVID-19, j'ai été contacté par Sandrine Fosse-Edorh de Santé publique France, qui souhaitait éprouver la qualité du codage CIM-10 utilisé pour l'identification des cas d'hospitalisations pour COVID-19 chez les personnes diabétiques, travail qui a fait l'objet d'une publication accélérée dans le BEH.[31] L'identification de ces patients était déjà pour le CHU de Nantes une priorité, et la qualité du codage a pu être confirmée localement par les services du SIM (Dr Christophe Leux). Evidemment, cette affirmation doit être tempérée : comme déjà souligné, la qualité d'un codage monocentrique sur une période donnée n'autorise pas la généralisation nationale directe sur une autre période. Pour aller plus loin, un réseau de « CH-sentinelles » serait nécessaire, mais avec le risque d'un biais de sélection par les « bons élèves », voire d'une optimisation des seuls indicateurs en vogue.

Un autre exemple de cet enrichissement inter-sources est donné par un travail issu de la plus grande cohorte nationale française en population, CONSTANCES, appariée aux données du SNDS. CONSTANCES permet la validation par un médecin, au temps de la visite, du statut diabétique du patient (oui/non), ce qui a permis de tester différentes approches algorithmiques identifiant les cas de diabète dans le SNDS.[32] Ce contrôle-qualité des algorithmes pour qualifier une information binaire (sensibilité, spécificité, valeurs prédictives positive et négative) est un premier pas pour affirmer le sérieux d'études basées uniquement sur le SNDS, et a été le socle de l'étude DMC décrite partie IV. Cela permet également d'acter la non fiabilité de certaines informations : disposant d'accès à la base CONSTANCES dans le cadre de l'étude HYPOBETA.fr portée par le Pr Bertrand Cariou[9], j'ai pu contrôler que le type de diabète (ici type 1 ou 2) n'était pas caractérisable de façon simple par les données issues du SNDS. En effet, une approche combinant naïvement les diagnostics CIM-10 associés aux ALD et aux hospitalisations se révéla vouée à l'échec, un même patient se voyant volontiers associer les deux diagnostics, presque sans lien statistique avec le *gold standard* de l'information issue de CONSTANCES. Cela ne préjuge bien sûr pas de la performance d'approches plus complexes basées par exemple sur l'âge de début de la maladie et les traitements médicamenteux, mais qui méritent un travail scientifique dédié - par ailleurs en cours[10].

8. Mes trois projets et le croisement des sources

L'objectif général de ce travail de thèse est de mettre à l'épreuve notre capacité à interroger trois différentes sources de données – recherche, médico-administrative et soin – afin de répondre à des questions épidémiologiques liées à une problématique médicale commune.

Cette problématique médicale est le lien entre les spécificités phénotypiques des patients diabétiques et le risque d'insuffisance cardiaque et son pronostic. L'examen de cette question épidémiologique sera doublé d'un retour d'expérience portant sur les difficultés méthodologiques rencontrées, en particulier concernant nos capacités d'accès aux données, la qualification des données, le consentement et l'information des patients.

[9] Cf. https://www.casd.eu/project/epidemiologie-de-lhypobetalipoproteinemie-en-population-generale-hypobeta , consulté le 9 septembre 2022

[10] Cf. https://www.constances.fr/actualites/2020/Intelligence-artificielle-diabete.php, consulté le 21 septembre 2022

La partie II de ce manuscrit propose une description générale du diabète d'une part et de l'insuffisance cardiaque d'autre part, et les arguments épidémiologiques et physiopathologiques qui les réunissent. Ensuite, pour chacune des questions spécifiques traitées dans les trois projets de recherche présentés dans les parties III à V, le rationnel sera exposé en début de partie

Le projet SURDIAGENE (partie III) s'appuie sur des données issues de la recherche mais avec un lien très étroit au soin : il s'agit d'une cohorte prospective de patients diabétiques bénéficiant d'un suivi hospitalier de leur diabète, au CHU de Poitiers, inclus entre 2001 et 2012. Les données de SURDIAGENE ont été enrichies des données médico-administratives locales, par appariement de la cohorte avec le PMSI du CHU, avec une réévaluation clinique systématique des événements par un comité d'adjudication, en particulier les suspicions d'hospitalisation et/ou de décès secondaires à une insuffisance cardiaque. A noter que la cohorte a également été enrichie de données issues des laboratoires de ville (dont le suivi de la créatininémie) mais que ces informations n'ont pas été exploitées pour le besoin du manuscrit.

Le projet DMC (partie IV) présentera exclusivement des données issues des bases de l'Assurance Maladie. Cependant, ce projet illustrera combien notre capacité à exploiter les données est tributaire (i) des données de recherche : appariement CONSTANCES / SNDS validant l'algorithme d'identification, et (ii) des données issues du soin : pertinence de l'algorithme d'identification de l'IC aiguë dans GAVROCHE, échanges avec les experts locaux cliniciens et l'informatique médicale, voire contrôles ponctuels par échantillonnage de la validité des algorithmes des complications, au moins au niveau du centre.

Le projet GAVROCHE (partie V) présentera les conditions d'exploitation des données issues du soin par le prisme du déploiement du logiciel eHOP au CHU de Nantes, dans une situation rendue complexe (i) par la nécessité de mettre en place une approche TALN pour automatiser l'extraction de l'information des CR hospitaliers et (ii) par l'ambition inter-régionale visant à déployer ces méthodes dans les différents centres avant de regrouper les données dans une plate-forme unique. GAVROCHE nous permettra d'illustrer les questions réglementaires et la gouvernance, l'accès aux données et leur circuit, avec un accent particulier sur la question de la qualification des données et le TALN.

PARTIE II : INSUFFISANCE CARDIAQUE ET MICROANGIOPATHIE DIABETIQUE

1. Insuffisance cardiaque : définition et épidémiologie

a. Définition et classification de l'insuffisance cardiaque

Nous retenons ici deux définitions de l'insuffisance cardiaque (IC), l'une européenne et l'autre américaine, traduites des recommandations de prise en charge de l'IC produites par les principales sociétés savantes.[33,34]

En 2021, la société européenne de cardiologie (ESC) propose ainsi : « L'insuffisance cardiaque n'est pas une unique entité pathologique, mais un syndrome clinique qui regroupe des symptômes cardinaux (p. ex. dyspnée, asthénie) pouvant être accompagnés d'autres signes (turgescence jugulaire, crépitants pulmonaires, œdème périphérique). Elle est due à une anomalie structurelle et/ou fonctionnelle du cœur, qui a pour résultat une élévation des pressions intracardiaques et/ou un débit cardiaque inadapté au repos et/ou lors de l'exercice ».

En 2022, la définition proposée par les sociétés savantes américaines est plus concise : « L'insuffisance cardiaque est un syndrome clinique complexe, dont les signes et symptômes résultent d'un affaiblissement de la fonction de remplissage ventriculaire ou d'éjection du sang ».

Outre la clinique, le diagnostic d'IC et sa gravité se fondent sur des critères biologiques (dont l'élévation des peptides natriurétiques) et d'imagerie (dont l'échographie transthoracique), ainsi que sur l'étiologie.

Différentes classifications de l'IC ont été proposées :

- Selon la présentation clinique : aiguë ou chronique
    - **L'IC aiguë** peut être définie par « une augmentation rapide ou graduelle du début des symptômes et/ou des signes d'IC, de sévérité suffisante pour que le patient sollicite une aide médicale en urgence, et conduisant à une admission hospitalière non planifiée » [33]
    - **Le diagnostic d'IC chronique,** parfois difficile dans les formes légères, est basé sur la présence chronique de symptômes et/ou signes de dysfonction cardiaque. La démarche diagnostique et de classification, traduite à partir des recommandations européennes et résumée **Figure 3**, est basée sur l'examen clinique et les examens complémentaires que sont l'électrocardiogramme, la biologie (peptides natriurétiques, numération formule sanguine, ionogramme sanguin et créatininémie, glycémie et hémoglobine glyquée, bilan lipidique, et les fonctions hépatique et

thyroïdienne, notamment pour le diagnostic différentiel), la radiographique thoracique, et l'examen clef que constitue l'échocardiographie transthoracique

- Selon le stade : à risque d'IC (stade A), pré-IC (B), IC symptomatique (C) et IC avancée (D)[34]
- Selon la sévérité de la dyspnée hors de la phase aiguë : la classification NYHA va du stade I à IV, allant d'une dyspnée présente lors des efforts exceptionnels jusqu'à une dyspnée permanente au repos
- Selon la fraction d'éjection ventriculaire gauche (FEVG) : FEVG réduite (≤40%), modérément réduite (41-49%) ou préservée (≥50%).[35] Cette classification revêt une grande importance clinique car elle est associée au pronostic et à l'efficacité de la prise en charge, les essais cliniques ayant montré les bénéfices de traitement chronique d'abord dans la population à FEVG réduite mais aussi plus récemment dans la population avec à FEVG préservée (dite HFpEF) avec la classe des inhibiteurs des SGLT2 [36]

Si la forme à FEVG réduite est classiquement symptomatique, l'IC à FEVG modérément réduite ou préservée peut être asymptomatique et son diagnostic s'appuie également sur des anomalies biologiques (élévation des peptides natriurétiques), et des anomalies cardiaques structurelles (dilatation de l'oreillette gauche, hypertrophie du VG) ou fonctionnelles (dysfonction VG diastolique, augmentation des pressions de remplissage du VG).[35]

Les facteurs de risque d'IC sont nombreux [35], au premier rang desquels se trouvent les maladies cardiovasculaires (coronaropathie, hypertension, trouble du rythme cardiaque, valvulopathie, cardiomyopathies et cardiopathies congénitales) mais également infectieuses (myocardites, SIDA, mais aussi maladie de Chagas en Amérique du Sud), toxiques (alcool, cocaïne, mais aussi médicamenteuses notamment dans les thérapies anti-cancéreuses (anthracyclines) et radio-induites), plus rarement rhumatologiques (lupus érythémateux disséminé, sclérodermie systémique), mais aussi sarcoïdose, amylose, ou encore la cardiomyopathie Takotsubo.

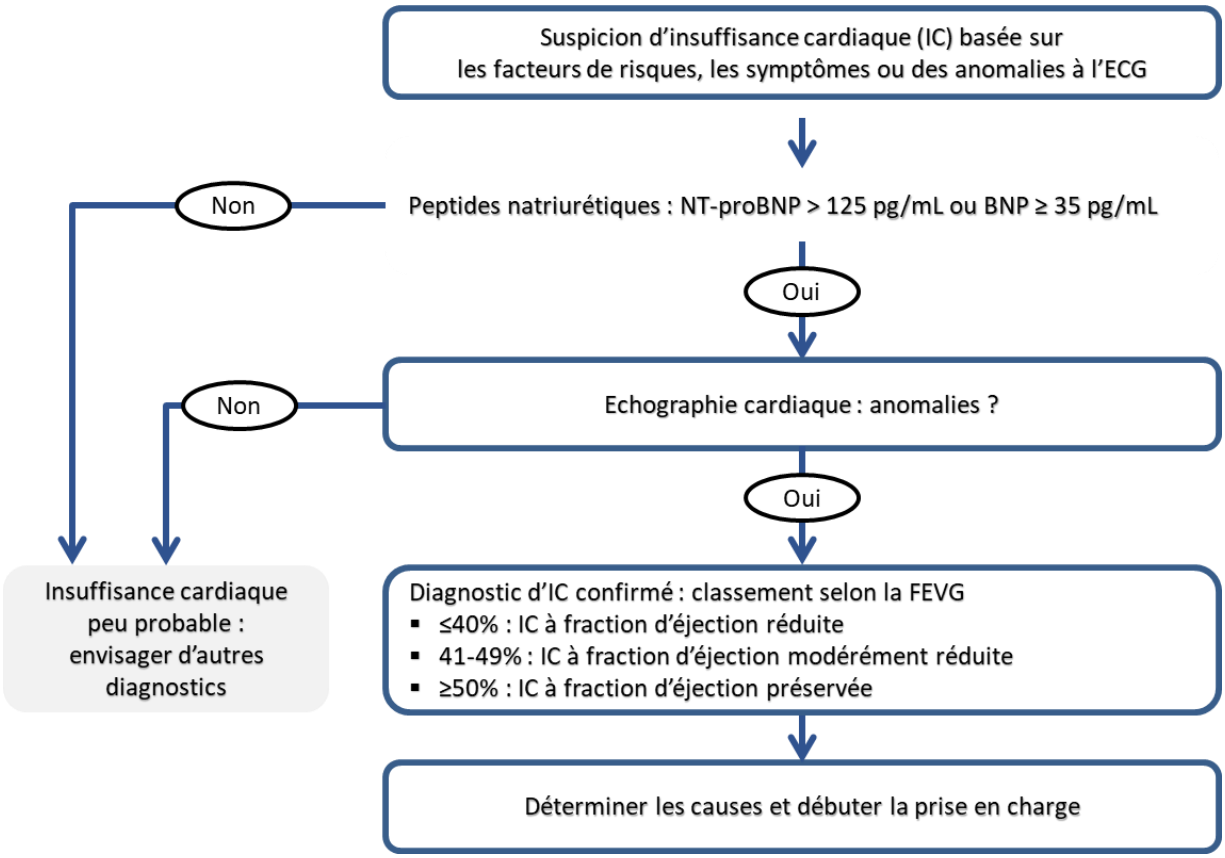


Figure 3. Algorithme diagnostique de l'insuffisance cardiaque chronique et classification selon la fraction d'éjection ventriculaire gauche (FEVG) – adapté des recommandations ESC 2021 [33]

b. **Prévalence, incidence et mortalité associée à l'insuffisance cardiaque**

L'IC présente toutes les caractéristiques d'un problème majeur de santé publique. Elle est fréquente et grave, et en hausse constante dans le monde, en lien notamment avec le vieillissement de la population.

Aux Etats-Unis, la prévalence dans la population âgée de plus de 18 ans était de 2,4% en 2012 avec des projections l'estimant à 3,0% en 2030, passant de 5,8 à 8,5 millions d'individus.[37] L'incidence était évaluée entre 6 et 8/1000 après 45 ans, et 21/1000 après 65 ans.[38] En Asie, les chiffres de prévalence de l'IC varient entre 1,3 et 6,7%.[39] En France, en 2008-2009, sa prévalence est estimée à environ 2,3% et atteint 10 % chez les sujets de plus de 75 ans.[40] En France, en 2008, on comptait 195,800 séjours hospitaliers avec un Diagnostic Principal (DP) d'insuffisance cardiaque.[41] Plus d'un million de Français vivent avec une IC, la mortalité étant évaluée à environ 70 000 morts par an.[42]

Après le diagnostic d'IC, une étude américaine sur les données de Medicare de 2008 retrouvait une mortalité à 1 an de 29,6%.[43] Le pronostic de l'IC aiguë (ICA) est particulièrement sombre. L'étude ARIC montrait une mortalité à 30 jours, 1 an et 5 ans après hospitalisation pour ICA de respectivement 10,4%, 22,0% et 42,3%.[44] La mortalité à 30 jours d'une hospitalisation pour ICA reste proche de 10% dans d'autres études de grande dimension, américaine (9,1%, période 1998-2001) comme internationale (10,0%, années 2000).[45,46] Une revue systématique internationale récente fait état chez l'adulte d'une mortalité moyenne, non standardisée, de 24% à 1 an après un épisode d'ICA, comparable à la mortalité à 1 an d'une fracture de hanche (30,6%).[47,48] En France, en 2009, dans une population de près de 70,000 individus issue du Système National des Données de Santé (SNDS), après hospitalisation pour ICA, la mortalité respectivement hospitalière, à 30 jours, à 1 an et à 2 ans était de 6,4%, 11%, 29% et 40%, similaire aux données de la littérature internationale.[41]

L'impact économique de la prise en charge de l'IC est également considérable, avec un coût total estimé aux Etats-Unis à plus de 30 milliards d'US$ en 2012, et une projection à 70 milliards en 2030.[37]

2. Microangiopathie diabétique : définition et épidémiologie

a. Epidemiologie du diabète

Le diabète sucré, que l'on désignera simplement par diabète dans la suite de ce manuscrit, est une maladie chronique, fréquente et grave.

A l'échelle mondiale, l'*International Diabetes Federation* (IDF) estime que 537 millions de personnes vivaient avec un diabète en 2021, et encore cette estimation est-elle limitée aux personnes âgées de 20 à 79 ans, avec une projection à 643 millions en 2030[49], bien que l'incidence semble se stabiliser dans de nombreux pays à hauts revenus. La prévalence augmente avec l'âge, passant de 2,2% chez les 20-24 ans à 24,0% chez les 75-79 ans, avec une légère prédominance masculine (10,8 vs 10,2% chez les femmes, tous âges confondus) qui s'inverse après 70 ans. Bien qu'il soit difficile d'apprécier dans quelle mesure un décès est attribuable ou non au diabète, l'IDF estimait cette proportion à 12,2% des décès toutes causes chez les 20-79 ans, soit plus de 6,7 millions de morts annuels à l'échelle mondiale.

En France, la prévalence du diabète traité pharmacologiquement - donc tous types confondus - était estimée à 2,8% en 2000[50] pour atteindre 5% en 2015 puis 5,3% en 2020, soit 3,5 millions d'individus.[51] La prévalence du diabète ne semble toujours pas stabilisée chez les plus de 45 ans[52] avec en outre une augmentation importante de l'incidence annuelle des cas de diabète de type 1 chez l'enfant entre 2010 et 2015.[53] A noter que les chiffres français s'appuient presque exclusivement sur les données du SNDS, l'algorithme de détection de la maladie ayant pu être validé en admettant comme *gold standard* les données de la cohorte CONSTANCES.[32] Il n'existe cependant pour l'instant pas d'algorithme validé pour distinguer les diabètes de type 1 et de type 2, ou les autres types de diabète moins fréquents. Par construction, ces chiffres sous-estiment la prévalence et l'incidence réelles puisqu'ils ne tiennent pas compte des diabètes non diagnostiqués ou non pris en charge.

Une analyse de l'étude des cohortes françaises ENTRED (2001 et 2007) a montré une surmortalité persistante par rapport à la population générale, de l'ordre de 20 à 100%, même si la différence de surmortalité globale a légèrement diminué entre 2001-2006 et 2007-2012, en particulier chez les personnes âgées.[54]

b. Les complications du diabète

Le diabète est associé à de multiples complications, parmi lesquelles :

- Des complications aiguës, conséquences directes du déséquilibre glycémique aigu : coma hyperosmolaire ou acido-cétosique

- Des complications au long-terme liées au déséquilibre glycémique chronique
    - microangiopathie ou « maladies des petits vaisseaux » : rétinopathie diabétique (RD), œdème maculaire, néphropathie, neuropathie périphérique et autonome - troubles vésicaux ou érectiles, sudoraux, gastroparésie ou neuropathie autonome cardiaque. Ces complications sont considérées comme fortement associées au niveau glycémique, facteur de risque du retentissement microangiopathique
    - macroangiopathie ou « maladie des gros vaisseaux » : coronaropathie, artériopathie oblitérante des membres inférieurs associée à un important risque d'amputation, accident vasculaire cérébral (AVC) ischémique. Ces complications sont considérées comme peu associées au niveau glycémique

- D'autres complications potentiellement liées au déséquilibre glycémique mais aussi à l'excès pondéral ou aux perturbations métaboliques fréquemment associés, dont le risque global de cancer qui est augmenté tant dans le cas du diabète de type 1 que de type 2 [55], et le risque d'IC qui sera vu plus loin. Il existe aussi une augmentation de la fréquence des infections et un pronostic plus défavorable de celles-ci. L'actualité récente a rappelé qu'il s'agissait d'une population plus vulnérable puisque le diabète a été identifié comme un facteur de risque indépendant de forme de grave d'infection au SARS-CoV-2, qu'il s'agisse d'hospitalisation comme de décès [56,57], et qu'il demeure un sur-risque de mortalité chez les patients hospitalisés par rapport à une population contrôle non diabétique [58]

- Des complications dues aux traitements antidiabétiques, en particulier l'acidose lactique sous metformine ou l'hypoglycémie sous sulfamides ou insuline, pouvant aller jusqu'au coma voire au décès, ainsi qu'être associées à un retentissement chronique (troubles cognitifs en particulier en cas d'épisodes répétés d'hypoglycémie sévère)

Certaines de ces complications sont quantifiées ponctuellement dans le bulletin épidémiologique hebdomadaire (BEH) publié par Santé publique France, notamment à partir de données PMSI ou SNDS. Des indicateurs peuvent être consultés en accès libre sur la plateforme *Géodes.*[59] En 2003, une étude menée à partir des données PMSI rapportait une incidence de l'amputation des membres inférieurs de 3,8/1000 chez les personnes diabétiques, avec un risque standardisé sur l'âge et le sexe douze fois supérieur à celui de la population non diabétique.[60] En 2013, chez les personnes diabétiques françaises traitées pharmacologiquement, l'incidence des hospitalisations pour AOMI était estimée à 2,5/1000, et à 6,7/1000 pour les plaies du pied.[61] La même année, le taux d'incidence (standardisé sur la structure d'âge et de sexe de l'Union Européenne) de la coronaropathie était estimé à 3,7/1000 (2,2 fois plus que chez les non-diabétiques), celle de l'AVC à 4,7/1000 (1,6 fois plus) et celle de l'insuffisance rénale terminale à 0,91/1000 (9 fois plus).[62]

Certaines atteintes microangiopathiques sont cependant difficiles à capturer dans le SNDS : c'est le cas de la maladie rénale chronique non sévère (insuffisance rénale modérée, élévation isolée de l'excrétion urinaire d'albumine) ou de la rétinopathie. L'analyse de la population d'ENTRED 2007 chez des adultes diabétiques français retrouvait une prévalence de maladie rénale chronique (définie à partir du débit de filtration glomérulaire et de l'excrétion urinaire d'albumine) de 29% [63], 19% d'insuffisance rénale modérée à sévère et 0,3% d'insuffisance rénale terminale (estimations sous-estimée du fait de 15% de données manquantes).[64] D'après cette même étude, 16,6% des personnes diabétiques avaient bénéficié d'un traitement ophtalmologique par laser, et 3,9% présentaient une cécité au moins unilatérale. Une analyse SNDS dans la population française traitée pharmacologiquement met en avant une incidence des injections intravitréennes de 1,5% et du traitement par laser de 0,96% en 2017, par ailleurs en hausse sur la période 2014-2017.[65] Après revue de la littérature, les données internationales disponibles ne sont pas récentes : une méta-analyse publiée en 2012 sur les données 1990-2008, réunissant plus de 22,000 individus diabétiques rapporte une prévalence de la rétinopathie diabétique (RD) de 34,6%, ainsi que 6,8% d'œdème maculaire, et estime à 10,2% les atteintes menaçant le pronostic visuel.[66] En Chine, cette prévalence a été estimée à 18,5% sur une méta-analyse d'études publiées entre 1990 et 2017, dont 0,99% pour la RD proliférante.[67] Une étude suédoise en population à l'échelle d'un comté, appariée au registre national suédois et portant sur une population de plus de 12 000 diabétiques retrouvait une prévalence de la RD (toutes formes) de 29,2%, dont 2,0% pour la RD proliférante et 2,5% pour la maculopathie, mais sans possibilité d'identifier les cas de cécité, probablement classés par défaut comme sans RD.[68] A contrario, une étude israélienne en population, qui s'appuie sur les « certificats de cécité » de 2003 (21 685 d'individus aveugles sur 6,8 millions, soit 0,32%), donne le diabète (RD et maculopathie) comme seconde cause de cécité, pour 14,4%, derrière la dégénérescence maculaire liée

à l'âge (28,0%) mais devant le glaucome (11,8%), les maculopathies non diabétiques (7,4%), l'atrophie optique et la cataracte (6,5% pour chacune de ces dernières).[69]

A noter pour les résultats français obtenus à partir du SNDS que les complications du diabète sont définies par des algorithmes reposant sur le croisement de différents codages, essentiellement des actes médicaux CCAM et des diagnostics codés CIM-10 issus des ALD et des hospitalisations PMSI. Cependant, contrairement à ce qui a été vu précédemment pour le diabète, leur qualité repose sur des discussions d'experts mais n'a généralement pas été évaluée. La définition et l'évaluation de ces algorithmes est un objectif du groupe de travail REDSIAM Maladies endocriniennes, nutritionnelles et métaboliques, dirigé par Sandrine Fosse-Edorh.

A noter que l'appariement du SNDS à CONSTANCES trouve ici ses limites : par exemple, pour la rétinopathie diabétique comme pour l'IC, l'absence de recherche ciblée de ces pathologies par l'interrogatoire et les examens complémentaires (rétinographie, fond d'œil, biomarqueurs, échographie transthoracique) la disqualifie pour tenir le rôle de *gold standard*.

c. **Lien entre définition du diabète et microangiopathie diabétique**

Le diabète est défini par une glycémie à jeun supérieure ou égale à 1,26 g/L (7 mmol/L) ou à une fraction A_{1c} de l'hémoglobine glyquée (HbA_{1c}) supérieure à 6,5%, ce qui correspond à une glycémie moyenne à 1,41 g/L.[70] Les valeurs cibles sont fondamentales pour la question microangiopathique puisqu'elles sont fondées non pas sur un sur-risque cardiovasculaire, qui est déjà présent pour des valeurs glycémiques plus faibles, mais sur la microangiopathie [71], et en particulier une complication spécifique du diabète : la rétinopathie diabétique.

3. Lien entre insuffisance cardiaque et diabète

a. Epidémiologie combinée[11]

Une large méta-analyse (47 études de 1966 à 2018 rassemblant plus de 12 millions d'individus) a confirmé le lien statistique entre diabète et IC, pour les deux sexes et les deux types de diabète, avec un sur-risque plus marqué chez les femmes et chez les personnes diabétiques de type 1 par rapport à des populations sans diabète.[73] L'âge et les comorbidités communes aux deux affections ne semblent pas être des facteurs confondants suffisants pour expliquer ce lien, lié au moins partiellement au contrôle glycémique [74] et à l'insulinorésistance.[75] De plus, même si la durée du diabète augmente le risque d'IC, les sujets jeunes sont d'emblée plus à risque qu'une population contrôle non diabétique.[76]

Je propose de considérer particulièrement deux études récentes menées à l'aide du registre national suédois du diabète (SNDR).[77] Ce registre inclut environ 460 000 personnes présentant un diabète de type 2, appariées 1:5 à des sujets contrôles sur l'âge, le sexe et la division administrative (=le comté) et incluses sur la période 1998-2012. Le SNDR permet une caractérisation fine des patients incluant des données de registres similaires à celles trouvées dans le SNDS mais associées à des données d'examen clinique (tabagisme, poids, pression artérielle) et de biologie (HbA_{1c} et créatininémie, mais aussi LDL-c et albuminurie).[78]

Dans l'article de Rawshani *et al.* [79] s'intéressant à différents événements cardiovasculaires (coronaropathie, AVC et IC), avec un suivi médian de 5,7 ans, les auteurs ont montré un important sur-risque d'IC chez les diabétiques par rapport à la population contrôle, présent notamment chez les plus jeunes (<55 ans) sans autre facteur de risque identifié (HbA_{1c} élevée, hypertension artérielle, albuminurie, tabagisme et hypercholestérolémie), avec un rapport de risque (RR) à 2,40 ($IC_{95\%}$, 1,63-3,54) augmentant avec le nombre de facteurs de risque, ce RR dépassant 11 chez ceux cumulant tous les facteurs de risque par rapport à la population contrôle. Dans la figure 3.D du même article, on observe par ailleurs que le risque d'IC semble augmenter linéairement avec le déséquilibre glycémique apprécié par l'HbA_{1c}, tandis que le risque d'infarctus du myocarde ou celui d'AVC semble plafonner.

Dans l'article de Rosengren *et al.*, focalisé sur la question de l'IC, l'incidence annuelle pour l'hospitalisation pour IC était de 11,9/1000 dans la population diabétique contre 6,2/1000 dans la

[11] Les éléments bibliographiques doivent beaucoup à la récente conférence de consensus proposée par l'ADA dans *Diabetes care* [72]

population contrôle.[74] Dans le modèle ajusté, comprenant notamment les principaux facteurs de risque importants d'IC (coronaropathie, fibrillation atriale, dialyse ou greffe rénale), un sur-risque d'IC était observé chez les personnes diabétiques par rapport aux contrôles, avec pour les âges < 55, 55-75 et > 75 ans des HR respectivement à 4,59, 1,74 et 1,11 chez les hommes, et 2,07, 1,44 et 1,11 chez les femmes. Pour les deux premières tranches d'âge ce risque était proportionnel au déséquilibre glycémique apprécié par l'HbA_{1c}, mais cette assertion était moins évidente chez les plus de 75 ans. Le risque était également plus haut en cas d'albuminurie ou de dégradation de la fonction rénale.

Ces analyses, certes menées sur la même population, accréditent fortement l'hypothèse d'une IC liée au déséquilibre diabétique qui n'est que partiellement expliquée par les causes classiques que sont la cardiopathie ischémique ou les troubles du rythme cardiaque. A noter qu'ils n'analysaient pas l'IC selon la fraction d'éjection ventriculaire gauche (conservée ou non), et que la rétinopathie diabétique n'était pas étudiée.

Nous n'avons pas trouvé de données françaises à grande échelle traitant ces questions. Dans l'étude ENTRED 2007, la prévalence de l'insuffisance cardiaque chez les personnes diabétiques était estimée à 6,3%.[64] Dans la population de l'étude CORONADO incluant plus de 5000 adultes hospitalisés pour COVID-19, après appariement 1:1 de populations diabétique et non diabétique (sur âge, sexe, période et centre hospitalier), 11,5% des personnes diabétiques présentaient une insuffisance cardiaque connue et 18.4% recevaient habituellement un traitement par diurétique de l'anse, contre 7,8% et 9,4% des patients non diabétiques, respectivement (OR = 1,67, $IC_{95\%}$ 1,35-2,08 et 2.18, $IC_{95\%}$ 1,82-2,62, respectivement - données non publiées). Sur données médico-administratives, une analyse CépiDc 2008-2010 montrait que les certificats de décès faisant mention d'IC mentionnaient environ deux fois plus la présence d'un diabète que ceux sans mention d'IC.[80]

b. Hypothèses physiopathologiques

Le lien statistique entre diabète et IC peut être expliqué par [72] :

- des facteurs de risque communs à ces deux affections (excès pondéral, HTA non contrôlée, dyslipidémie)
- des facteurs déclenchant l'IC aiguë plus fréquents chez les personnes diabétiques (infections respiratoires ou urinaires, fibrillation auriculaire)
- des facteurs jouant le rôle de médiateurs sur la chaîne causale du diabète vers l'IC chronique (déséquilibre glycémique, coronaropathie, AOMI, microalbuminurie et insuffisance rénale)

Outre ces éléments, une atteinte spécifique au diabète a été décrite, la cardiomyopathie diabétique, pouvant être définie comme une dysfonction ventriculaire en l'absence de coronaropathie et d'hypertension artérielle.[81] Au-delà de la seule perturbation de l'homéostasie glucidique définissant le diabète, cette entité vise à expliquer des modifications fonctionnelles et structurelles du cœur en lien avec d'autres perturbations métaboliques plus fréquentes chez les diabétiques.[72]

Parmi celles-ci, il est décrit une dysfonction du ventricule gauche avec un rôle central de la dysfonction endothéliale et de l'atteinte microvasculaire.[82] Cette dysfonction est multifactorielle, liée à la perturbation du système rénine-angiotensine aldostérone,[81] à la dysfonction mitochondriale et au stress oxydatif,[83] à une altération de l'homéostasie du calcium intracellulaire,[84,85] ainsi qu'à une perte de flexibilité métabolique avec une diminution de la consommation de glucose (du fait de l'insulinorésistance) et une hausse de la consommation d'acides gras libres, également accumulés sous la forme de triglycérides et d'autres métabolites lipidiques dans le myocarde.[86]

Il résulte de ces multiples facteurs un remaniement fibrotique et une dysfonction du VG (aussi attribuable à l'atteinte dysautonomique avec anomalies de la relaxation diastolique)[87] associée à des atteintes microvasculaires coronaires entraînant une perfusion myocardique insuffisante en distalité.[72,88] Cette atteinte est commune aux personnes diabétiques de type 1 et de type 2.[72] Ce profil, particulier à la maladie diabétique, n'empêche naturellement pas les formes plus classiques d'IC.

c. Le lien entre diabète et IC tel qu'étudié dans les 3 projets

L'analyse de la cohorte prospective issue de la recherche SURDIAGENE (partie III) s'intéressera, dans une population diabétique de type 2, à l'association entre certains biomarqueurs nutritionnels et l'insuffisance cardiaque aiguë (ou le décès), au-delà du diabète et de ses complications rénale et cardiaque.

Dans le projet DMC (partie IV) l'analyse des données médico-administratives nous permettra, à partir d'une population identifiée comme diabétique dans les bases de l'Assurance Maladie, d'étudier l'association entre atteinte rétinienne du diabète et insuffisance cardiaque, en contrôlant les principaux facteurs de risque d'IC (dont la coronaropathie, l'HTA, la valvulopathie et les troubles du rythme). L'atteinte rétinienne est prise ici comme témoin de l'atteinte microvasculaire théoriquement commune à la cardiomyopathie diabétique.

Pour le projet GAVROCHE (partie V) construit sur les données issues du soin rendues exploitables par le déploiement des EDS, nous nous intéresserons à toutes les personnes adultes hospitalisées pour

insuffisance cardiaque aiguë, afin d'étudier l'intérêt pronostique de la variabilité glycémique chez ces patients. Le statut diabétique sera considéré comme un facteur d'interaction dans le modèle, afin d'être en mesure d'apprécier le caractère pronostique de la variabilité glycémique également chez les personnes non diabétiques.

PARTIE III : SURDIAGENE - BIOMARQUEURS NUTRITIONNELS ET INSUFFISANCE CARDIAQUE

1. Résumé

L'étude proposée est une analyse ancillaire de la cohorte SURDIAGENE (SUivi Rénal, DIAbète de type 2 et GENEtique), une cohorte prospective et monocentrique de 1468 patients suivis au CHU de Poitiers pour un diabète de type 2 et inclus entre 2001 et 2012. Dans notre typologie, il s'agit donc de données recueillies à des fins de recherche.

SURDIAGENE a déjà fait l'objet de nombreuses publications. Nous nous intéressons ici au lien entre diabète et insuffisance cardiaque dans une perspective nutritionnelle. Cette analyse est exclusivement fondée sur des biomarqueurs puisque l'étude ne comprenait malheureusement pas d'informations exploitables concernant les apports alimentaires. Nous avons étudié des biomarqueurs associés à la consommation de protéines, et en particulier de viande rouge, l'oxyde de triméthylamine (TMAO) et ses précurseurs (bétaine, carnitine, choline), ainsi que la cystéine et l'homocystéine, et des biomarqueurs associés à la consommation de légumes (thio-amino-acides : cystéine, homocystéine et méthionine). Le critère de jugement principal est le risque d'hospitalisation et/ou de décès pour insuffisance cardiaque aiguë.

Concernant les données, nous disposions uniquement d'informations à l'inclusion, sans mesure répétée des marqueurs d'exposition, et d'événements survenus jusque décembre 2015. Nous avons donc utilisé un modèle classique, le modèle de Cox basé sur l'hypothèse des risques proportionnels. Concernant le critère de jugement principal, aucun des biomarqueurs ne lui était significativement associé après ajustement sur les paramètres cardiaques et rénaux. Seul persistait un risque associé à l'homocystéinémie pour la mortalité toutes causes, étudiée ici comme critère de jugement secondaire (rapport de risque à 1,16 pour une déviation standard, $IC_{95\%}$ 1,06-1,27).

Après présentation du déroulement du travail et des contributions, cette partie est essentiellement présentée sous la forme de l'article paru dans *Cardiovascular Diabetology* en juin 2022, avant une conclusion sur l'apport de cette analyse dans le cadre de ma thèse.

2. Déroulement et contributions respectives

Le Professeur Samy Hadjadj, investigateur principal puis depuis 2018 co-investigateur principal, a mis en place la cohorte SURDIAGENE promue par le CHU de Poitiers, cohorte monocentrique dont l'objectif est d'analyser les variables génétiques et non-génétiques associées à l'apparition des complications du diabète de type 2.

Elise Gand a assuré la gestion des données et la création de l'ensemble de la base de l'étude exploitée depuis 2016. Elle m'a transmis le jeu de données nécessaires à notre analyse ancillaire, ainsi qu'un dictionnaire des variables, ce qui signifie que je n'ai eu à effectuer sur les données aucun travail préalable aux analyses statistiques, contrairement aux projets DMC et GAVROCHE.

Les expertises suivantes ont été sollicitées : le Professeur Jean-Noël Trochu pour la lecture cardiologique, les Pr Stéphanie Ragot et Pierre-Jean Saulnier, co-investigateur SURDIAGENE depuis 2018, à la fois pour leur connaissance de la cohorte et leur expertise diabétologique, le Pr Bertrand Cariou pour la diabétologie, le Pr David Jacobi pour la nutrition.

Le Dr Mikaël Croyal a effectué les dosages des biomarqueurs et a apporté son expertise de biologiste, ainsi que les Dr Cédric Le May et Xavier Prieur.

J'ai effectué l'ensemble des analyses statistiques, avec le soutien de Thomas Goronflot pour les « *forest plots* » figurant l'analyse stratifiée (*Supplemental Fig. 3* de l'article) et sa relecture attentive pour le reste des analyses.

Le Pr Hadjadj, le Dr Croyal et moi-même avons rédigé la première version du manuscrit, manuscrit qui a ensuite été revu et critiqué par l'ensemble des co-auteurs.

3. Graphes dirigés acycliques associés (DAG)

Ma connaissance des DAG (pour *Directed Acyclic Graph* ou graphes dirigés acycliques) est essentiellement fondée sur le cours « *Edx* » proposé par le Pr Miguel Hernán, disponible en ligne sous format MOOC,[12] que j'avais mise en application suite à la *review* d'un article de l'étude CORONADO portant sur le lien entre diabète et pronostic chez les patients hospitalisés pour COVID-19.[58]

Les DAG permettent une représentation schématique intuitive des hypothèses de causalité liées à une question épidémiologique. Cet outil propose d'interroger le lien entre deux éléments « A » et « B », et aide en particulier à différencier un lien causal d'une simple association statistique. La représentation graphique résume les hypothèses posées, et la théorie associée aux DAG permet d'en déduire les paramètres à « contrôler » (par des méthodes d'ajustement, d'appariement, de standardisation ou encore de stratification) par le statisticien, et donc de mieux interpréter les résultats des modèles statistiques, par exemple les coefficients d'un modèle de régression logistique ou d'un modèle de Cox.

Pour une étude donnée, plusieurs DAGs peuvent être proposés, correspondant à différentes hypothèses. L'interprétation des modèles a été recontrôlée par le codage direct sur le site DAGitty.[13]

Dans le cadre de l'analyse des données de SURDIAGENE, nous avons distingué cinq groupes de variables mesurées à l'inclusion dans la cohorte :

- L'exposition : un biomarqueur nutritionnel, ici la TMAO
- L'événement : ici l'hospitalisation pour insuffisance cardiaque (HFrH dans l'article), ou le décès
- Les variables candidates comme potentiels facteur de confusion ou médiateurs (la question de la stratification n'est pas abordée ici), à savoir :
    - Age, sexe
    - Les facteurs cardiaques : coronaropathie, peptide natriurétique (NT-proBNP)
    - Les facteurs rénaux : le débit de filtration glomérulaire estimé (DFGe ou eGFR en anglais, considéré comme proxy de la fonction rénale) et l'albuminurie (ratio albumine/créatinine urinaire ou RAC(U), ou UACR en anglais)

[12] https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your

[13] http://www.dagitty.net/development/dags.html

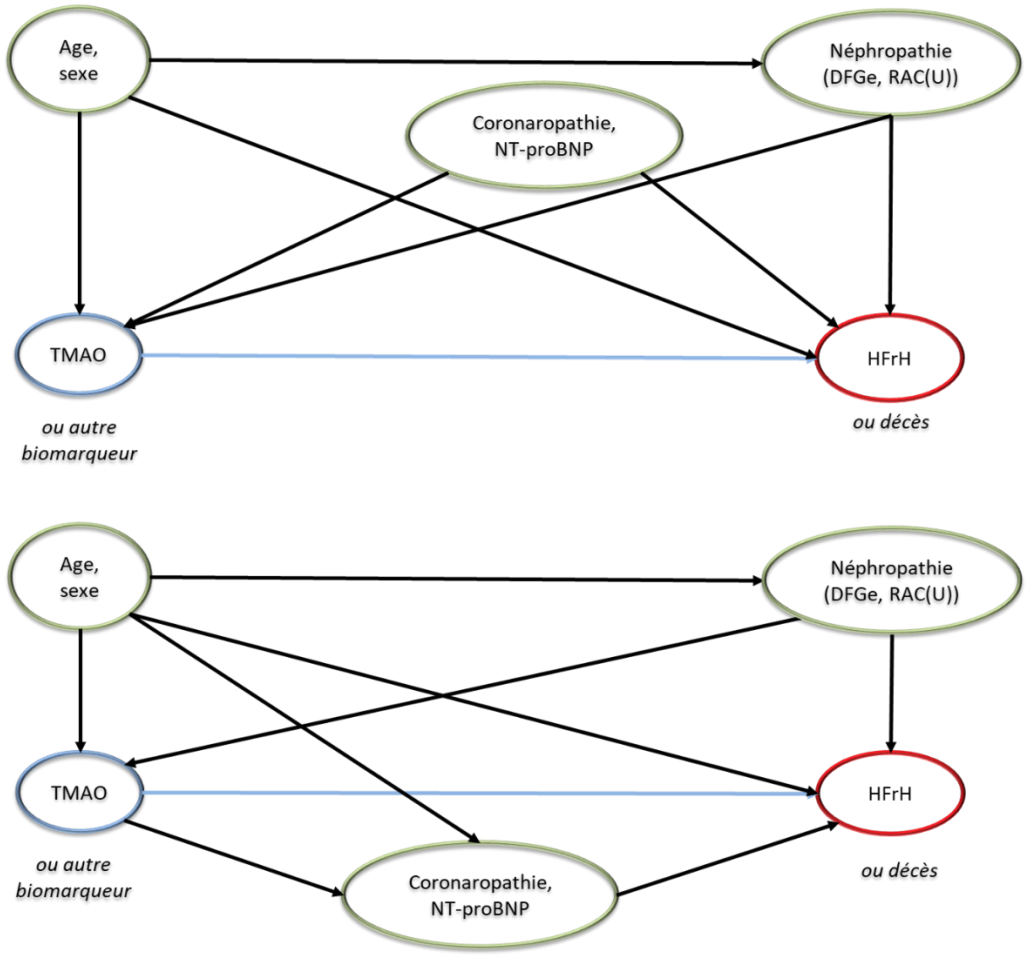


Figure 4. Graphes dirigés acycliques (DAG) résumant des hypothèses de l'étude SURDIAGENE, avec en particulier l'atteinte cardiaque considérée comme un facteur de confusion entre TMAO et HFrH (graphe supérieur), ou comme un médiateur (graphe inférieur)

La **Figure 4** propose deux DAGs résumant certaines hypothèses de causalité dans notre analyse des données de SURDIAGENE. **Dans la première représentation (graphe supérieur),** les trois groupes de variables sont considérés comme des facteurs de confusion dans l'étude du lien entre TMAO et HFrH. Il faudra donc les « contrôler » en ajustant sur ces paramètres dans les modèles de régression, comme cela est proposé dans le modèle complet de l'article. **La seconde représentation (graphe inférieur)** suppose que les facteurs cardiaques (coronaropathie, NT-proBNP) peuvent être une conséquence de la concentration de TMAO, puisqu'on les retrouve sur le chemin causal entre TMAO et HFrH. Le risque total d'HFrH associé à TMAO peut être décomposé en deux sous-risques : un risque direct (flèche bleue) et un risque indirect médié par les facteurs cardiaques. Un modèle ajusté uniquement sur âge, sexe et facteurs rénaux estimera le risque total attribuable à la concentration en TMAO. Le modèle complet, ajusté également sur les facteurs cardiaques, estimera uniquement le risque direct attribuable à la concentration de TMAO, sans la voie du risque médié par les facteurs cardiaques.

4. Article paru dans Cardiovascular Diabetology

Wargny *et al. Cardiovascular Diabetology* (2022) 21:101
https://doi.org/10.1186/s12933-022-01505-9

Cardiovascular Diabetology

RESEARCH Open Access

Nutritional biomarkers and heart failure requiring hospitalization in patients with type 2 diabetes: the SURDIAGENE cohort

Matthieu Wargny[1,2], Mikaël Croyal[1,4,5], Stéphanie Ragot[3], Elise Gand[3], David Jacobi[1,5], Jean-Noël Trochu[1], Xavier Prieur[6], Cédric Le May[6], Thomas Goronflot[2], Bertrand Cariou[1], Pierre-Jean Saulnier[3], Samy Hadjadj[1,4,5*] for the SURDIAGENE study group

Abstract

Background: Heart failure (HF) is a growing complication and one of the leading causes of mortality in people living with type 2 diabetes (T2D). Among the possible causes, the excess of red meat and the insufficiency of vegetables consumption are suspected. Such an alimentation is associated with nutritional biomarkers, including trimethylamine *N*-oxide (TMAO) and its precursors. Here, we aimed to study these biomarkers as potential prognostic factors for HF in patients with T2D.

Methods: We used the SURDIAGENE (SURvival DIAbetes and GENEtics) study, a large, prospective, monocentric cohort study including 1468 patients with T2D between 2001 and 2012. TMAO and its precursors (trimethylamine [TMA], betaine, choline, and carnitine) as well as thio-amino-acids (cysteine, homocysteine and methionine) were measured by liquid chromatography-tandem mass spectrometry. The main outcome was HF requiring Hospitalization (HFrH) defined as the first occurrence of acute HF leading to hospitalization and/or death, established by an adjudication committee, based on hospital records until 31st December 2015. The secondary outcomes were the composite event HFrH and/or cardiovascular death and all-cause death. The association between the biomarkers and the outcomes was studied using cause-specific hazard-models, adjusted for age, sex, history of coronary artery disease, NT-proBNP, CKD-EPI-derived eGFR and the urine albumin/creatinine ratio. Hazard-ratios (HR) are expressed for one standard deviation.

Results: The data of interest were available for 1349/1468 of SURDIAGENE participants (91.9%), including 569 (42.2%) women, with a mean age of 64.3 ± 10.7 years and a median follow-up of 7.3 years [25th–75th percentile, 4.7–10.8]. HFrH was reported in 209 patients (15.5%), HFrH and/or cardiovascular death in 341 (25.3%) and all-cause death in 447 (33.1%). In unadjusted hazard-models, carnitine (HR = 1.20, 95% CI [1.05; 1.37]), betaine (HR = 1.34, [1.20; 1.50]), choline (HR = 1.35, [1.20; 1.52]), TMAO (HR = 1.32, [1.16; 1.50]), cysteine (HR = 1.38, [1.21; 1.58]) and homocysteine (HR = 1.28, [1.17; 1.39]) were associated with HFrH, but not TMA and methionine. In the fully adjusted models, none of these associations was significant, neither for HFrH nor for HFrH and/or CV death, when homocysteine only was positively associated with all-cause death (HR = 1.16, [1.06; 1.27]).

*Correspondence: samy.hadjadj@univ-nantes.fr

[1] Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, 44000 Nantes, France
Full list of author information is available at the end of the article

Conclusions: TMAO and its precursors do not appear to be substantial prognosis factors for HFrH, beyond usual cardiac- and kidney-related risk factors, whereas homocysteine is an independent risk factor for all-cause death in patients with T2D.

Keywords: TMAO, Nutritional biomarkers, Diabetes mellitus, Heart failure, Cohort study, Homocysteine

Introduction

Diabetes is a globally-increasing condition and is expected to affect more than 10% of humans worldwide in 2030 [1]. In addition to microvascular complications, cardiovascular disease is one of the leading causes of morbidity and mortality in patients with diabetes [2, 3]. Subjects living with diabetes present an excess risk of coronary artery disease (CAD) and diabetic cardiomyopathy, which are the two main causes of heart failure (HF). Recently, a large meta-analysis considering over 12 million individuals evidenced a two-fold greater risk of HF in individuals with type 2 diabetes (T2D) compared to those without, both in men and women [4].

Among the multiple factors associated with such complex diseases, nutrition appears to be essential since it is a possible target for preventive and/or therapeutic actions, as a modifiable risk factor for both T2D and HF. In that respect, glycemic index and load were recently shown to be related to cardiovascular disease in a large-scale global epidemiological approach. However, when scrutinizing the effect of different outcomes, it turned out that glycemic index and load were associated with atherosclerosis-related cardiovascular events but not with HF [5].

So far, the impact of nutrition in HF is often under-estimated. In the guidelines of the ESC (European Society of Cardiology), nutrition in HF mainly relates to malnutrition and obesity as contributor of HF. Dietary recommendations are mainly focused on salt intake, healthy eating, maintenance of body weight and refraining from excessive alcohol intake in case of toxic cardiomyopathy [6]. We previously evidenced the association between markers of red meat consumption (trimethylamine *N*-oxide [TMAO] and related metabolites) and the occurrence of major adverse cardiovascular events (MACE), but also mortality in patients with T2D [7]. However, we did not evaluate the impact of nutritional biomarkers on HF, a critical outcome in persons living with diabetes.

A 2019 initiative, the EAT-Lancet commission on healthy diet, recommended to consider 14 key items for universal healthy diet [8]. Being able to identify nutritional biomarkers associated with HF is an important issue especially in T2D. This could in particular enable the early identification of patients susceptible to develop HF and for whom specific management including nutritional counselling could be considered in a preventive manner. Such an approach could pave the way for personalized nutrition of patients with T2D.

Thus, we aimed to assess how baseline nutritional biomarkers related to red meat intake (TMAO and related compounds, i.e. trimethylamine [TMA], betaine, choline and carnitine) and to vegetable intakes (thio-amino-acids: cysteine, homocysteine and methionine) were associated with the incidence of HF requiring Hospitalization (HFrH) in patients with T2D, regardless of their history of HF.

Methods

SURDIAGENE cohort and study population

The design of the SURvival DIAbetes and GENEtics (SURDIAGENE) cohort has already been described elsewhere [9]. Briefly, SURDIAGENE is a large, prospective, monocentric cohort study with the consecutive inclusion of 1468 T2D patients taken care at the Diabetes Department at Poitiers University Hospital, France, between 2001 and 2012. The study was primarily designed to identify the genetic determinants of micro- and macrovascular diabetic complications. At baseline, clinical and biological data were collected and blood/urine samples were drawn. Clinical events corresponding to endpoints of interest were collected during follow-up, based on consultations with general practitioner and hospital records.

Renal function was assessed using estimated glomerular filtration rate (eGFR) calculated with the CKD-EPI 2009-formula [10]. For the present analysis, we excluded patients with: baseline eGFR < 30 mL/min. 1.73 m^2; renal replacement therapy (need for dialysis or history of renal transplant); missing data for ≥ 1 of the following: NT-proBNP, urine albumin/creatinine ratio (uACR), nutritional biomarkers of red meat intake (methylamines: carnitine, betaine, choline, TMAO, TMA) and of vegetable intake (amino-acids: cysteine, homocysteine and methionine).

Definition of clinical history

History of CAD was defined as history of angina pectoris and/or coronary revascularization and/or myocardial infarction. Cerebrovascular disease (CVD) was defined as history of stroke and/or transient ischemic attack. Lower

limb artery disease was defined as lower limb revascularization and/or amputation.

Biology assays

The methylamines (TMAO and its precursors, TMA, betaine, choline and carnitine) as well as the thio-amino-acids (cysteine, homocysteine and methionine) were analysed in baseline fasting plasma samples by liquid chromatography-tandem mass spectrometry as detailed in Additional file 1. Samples were stored at − 80 °C until final use with only 2 freeze/thaw cycle. The intra- and inter-assay imprecisions of the analytical method were assessed throughout experiments and were below 10.2% for all compounds. All compounds were found stable in a set sample of 10 patients with diabetes provided by CHU Nantes ("*maladies métaboliques*" collection) after 3 freeze/thaw cycles with mean recovery ranging from 93.7% to 111.1%. At completion of the study, a representative set of samples (~ 10% of the cohort) was arbitrarily re-analysed 6 months after the initial determination. The new plasma concentrations did not vary by more than 7.2% (from − 6.8% to 7.2%; median: 3.2% [− 4.3%; 5.5%]) in comparison with the first analysis.

NT-proBNP was measured in baseline plasma-EDTA samples by an electrochemiluminescence automated assay (Roche Diagnostics, Mannheim, Germany).

Outcomes

The primary outcome of the study was the first occurrence of HFrH during follow-up. HFrH was defined as the first occurrence of one of the following events, whichever came first: acute HF requiring hospitalization or leading to death, validated by an adjudication committee including both diabetologists and cardiologists, after careful evaluation of hospital and discharge records. We proposed the study of two secondary outcomes: (i) all-cause death, established after linking French national death registry in SURDIAGENE participants; and (ii) HFrH and/or cardiovascular (CV) death, the latter validated by an adjudication committee. Follow-up data were collected until 31st December, 2015.

Statistical analyses

For baseline analysis, patients' characteristics were presented as numbers (%) for categorical parameters, and mean ± SD or median (25th–75th percentile) for quantitative parameters. They were compared according to the final follow-up event for HFrH and all-cause death. Independence between categorical parameters was tested using Fisher's exact test. For testing difference for quantitative parameters between two groups, we proposed Student's t-test or Mann–Whitney U-test according to variable distribution, as deemed appropriate.

For longitudinal analysis, we proposed 5 models for each nutritional biomarker of interest, with different adjustment levels. The nutritional biomarkers were tested one-by-one, separately. Model 1 (M_1): biomarker only; Model 2 (M_2): M_1 adjusted for age and sex; Model 3A (M_{3A}): M_2 adjusted for cardiac covariates (history of CAD and NT-proBNP); Model 3B (M_{3B}): M_2 adjusted for renal covariates (eGFR and uACR); Model 4 (M_4, full model): M_2 adjusted for both cardiac and renal covariates. We considered all-cause death as a competing risk in the analysis of HFrH and followed the recommendation summarized by Austin et al. [11]. So, we calculated HR for (i) cause-specific hazards using Cox regression models, and (ii) relative incidences, using subdistribution hazard models [12]. For quantitative parameters, log-transformation was applied when appropriate and HR were calculated for an increase of 1 SD. Additionally, plots of the cumulative incidence functions (CIF, here with a quartile-based approach) are proposed for HFrH and all-cause death.

A global p-value < 0.05 was considered as statistically significant. Considering 8 parameters studied in 3 main analyses (survival for HFrH, HFrH and/or CV death, and all-cause death), and disregarding the different models used (M1 to M4, cause-specific and subdistribution hazards models) considered as heavily correlated, we proposed a conservative threshold = 0.0021 (≈0.05/24) for individual p-value, following a Bonferroni approach.

We challenged the linearity assumption using fractional polynomials of degree 2 (FP2) [13] to test for other potential shapes of the HR function linking each biomarker and HFrH. However, even with an alpha value for FP2 as high as 0.50, no transformation was proposed, supporting therefore the linearity.

As an exploratory analysis, we also proposed subgroups analyses of the population according to baseline status for CAD, obesity and NT-proBNP level (below or above 125 pg/mL), using cause-specific hazard models adjusted for age and sex for the study of HFrH.

All results are presented using available data, without imputation. All statistical analyses were performed using R version 4.0.0., particularly with "*cpmrsk*" package [14, 15].

Results

The SURDIAGENE study included 1468 patients with T2D, of which 106 patients (7.2%) were secondarily excluded because of eGFR < 30 mL/min. 1.73 m^2 and/or renal replacement therapy at baseline. In order to ensure data consistency and nested multivariable regression models, 13 patients (1.0%) were also removed from the present analysis because of missing data for methylamines and/or NT-proBNP and/or uACR. Finally, 1349

patients (91.9%) were included in the present analysis, with a median follow-up of 7.3 years [25th–75th percentile, 4.7–10.8]. HFrH occured in 209 patients (15.5%), HFrH and/or CV death in 341 patients (25.3%) and all-cause death in 447 patients (33.1%). Flow-chart details can be found in Fig. 1.

A comparison between baseline characteristics according to the occurrence of the outcomes is proposed Table 1. Patients who met the HFrH event were older (70.3 ± 9.4 vs 63.2 ± 10.5 years, $p < 0.0001$) and had a lower body mass index (BMI, 30.4 ± 6.1 vs 31.6 ± 6.3 kg/m^2, $p = 0.012$) than patients without HFrH. They presented more characteristics of microangiopathy, with a greater proportion of macroalbuminuria (32.0 vs 16.5%, $p < 0.001$), a lower eGFR (66.1 ± 20.1 vs 78.7 ± 20.5 mL/min. 1.73 m^2, $p < 0.0001$) and a greater proportion of severe non-proliferative or proliferative retinopathy (21.8 vs. 11.1%, $p < 0.0001$). They also presented a greater proportion of macroangiopathy-related history, such as CAD (50.2 vs 27.2%, $p < 0.0001$), CVD (17.7 vs 11.8%, $p = 0.024$) and lower limb artery disease (17.2 vs 6.7%, $p < 0.0001$).

The relationship between the baseline concentrations of nutritional biomarkers of dietary components is available in Additional file 2: Fig. S1. As expected, betaine and choline concentrations were positively correlated, as well as homocysteine and cysteine concentrations.

Regarding methylamines-related biomarkers, compared to participants who did not develop HFrH, those who developed HFrH had higher concentrations of carnitine (45.8 ± 14.8 µmol/L vs. 43.7 ± 12.0, $p = 0.044$), betaine (36.5 ± 16.1 µmol/L vs. 33.0 ± 13.6, $p = 0.0030$), choline (1.57 ± 0.43 µmol/L vs. 1.46 ± 0.37, $p < 0.0001$) and TMAO (8.8 µmol/L [5.3; 17.0] vs. 6.6 [4.0; 12.3], $p < 0.0001$), but not for TMA (0.75 ± 0.27 vs. 0.77 ± 0.27) (Table 2).

Regarding nutritional biomarkers of vegetable intake, among the 3 considered thio-amino-acids, homocysteine was higher in patients ultimately yielding HFrH (10.9 µmol/L [5.7; 18.0] vs. 8.6 [4.7; 14.1], $p < 0.0001$), but no statistical difference was found for cysteine and methionine concentration.

The results of the cause-specific hazard models for HFrH, HFrH and/or CV death and all-cause death are presented in Table 3, and the subdistribution hazard models are presented in Additional file 2 Table S1. In unadjusted hazard-models, carnitine (HR = 1.20, 95% CI [1.05; 1.37], $p = 0.0065$), betaine (HR = 1.34, [1.20; 1.50], $p < 0.0001$), choline (HR = 1.35, [1.20; 1.52], $p < 0.0001$), TMAO (HR = 1.32, [1.16; 1.50], $p < 0.0001$), cysteine (HR = 1.38, [1.21; 1.58], $p < 0.0001$) and homocysteine (HR = 1.28, [1.17; 1.39], $p < 0.0001$) were associated with HFrH, but not TMA and methionine. These associations remained significant after adjustment for age and sex. After further adjustment for eGFR and uACR, only betaine remained statistically associated with HFrH (HR = 1.33, [1.17; 1.50], $p < 0.0001$). In the fully adjusted models, none of these associations remained significant. The curves for cumulative incidence functions for HFrH are plotted in Fig. 2, and Additional file 2: Fig. S2 for all-cause death. For betaine and homocysteine, the increased risk of HFrH seemed associated with the upper quarter of concentrations.

When studying the risk of all-cause death, betaine (HR = 1.28, [1.18; 1.39], $p < 0.0001$), choline (HR = 1.26, [1.16; 1.38], $p < 0.0001$), TMAO (HR = 1.20, [1.10; 1.31], $p < 0.0001$), cysteine

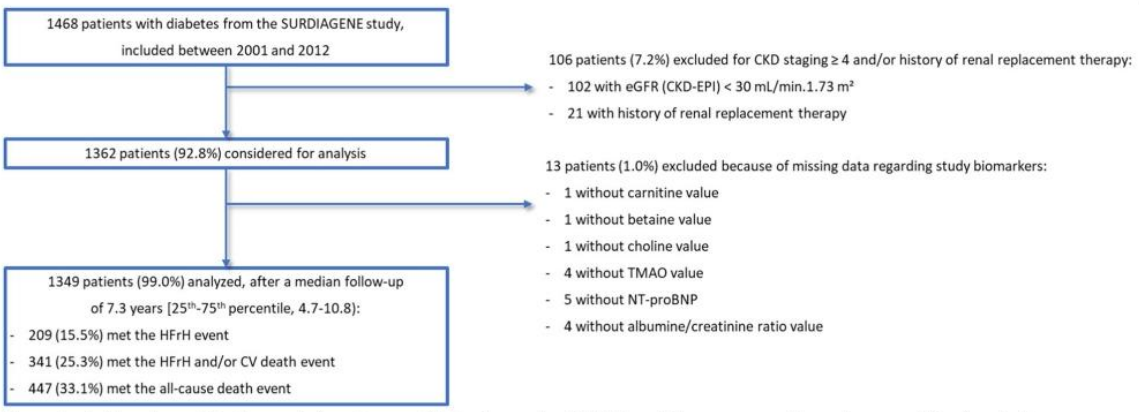


Fig. 1 Study flow-chart. CKD: Chronic Kidney Disease; CV: Cardiovascular; HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; NT-proBNP: N-terminal prohormone of brain natriuretic peptide; TMAO: trimethylamine *N*-oxide

Table 1 Baseline characteristics of the cohort classified by status for HFrH and all-cause death during follow-up

Baseline characteristics	All (n = 1349)	Event: HFrH			Event: All-cause death		
		No (n = 1140)	Yes (n = 209)	P-value	No (n = 902)	Yes (n = 447)	P-value
Sex (female)	569/1349 (42.2%)	480/1140 (42.1%)	89/209 (42.6%)	0.94	414/902 (45.9%)	155/447 (34.7%)	< 0.0001
Age (y)	64.3 ± 10.7	63.2 ± 10.5	70.3 ± 9.4	< 0.0001	61.5 ± 10.4	70 ± 8.8	< 0.0001
Weight (kg)	86.4 ± 18.5	87.2 ± 18.5	81.9 ± 17.8	< 0.0001	87.2 ± 18.5	84.8 ± 18.5	0.025
BMI (kg/m.^{2})	31.4 ± 6.3	31.6 ± 6.3	30.4 ± 6.1	0.012	31.7 ± 6.3	30.8 ± 6.2	0.012
Diabetes duration (y)	12 [6; 20]	11 [5; 19]	18 [12; 27]	< 0.0001	11 [5; 17]	16 [10; 25]	< 0.0001
HbA_{1c} (%)	7.8 ± 1.6	7.8 ± 1.6	7.9 ± 1.4	0.61	7.8 ± 1.6	7.9 ± 1.5	0.15
Smoker				0.0090			0.37
Never	689/1331 (51.8%)	584/1125 (51.9%)	105/206 (51.0%)		471/890 (52.9%)	218/441 (49.4%)	
Former	493/1331 (37.0%)	404/1125 (35.9%)	89/206 (43.2%)		318/890 (35.7%)	175/441 (39.7%)	
Active	149/1331 (11.2%)	137/1125 (12.2%)	12/206 (5.8%)		101/890 (11.3%)	48/441 (10.9%)	
Heart rate (bpm)	70.9 ± 13.6	70.9 ± 13.5	71.0 ± 14.0	0.90	70.5 ± 13.1	71.6 ± 14.4	0.19
Systolic BP (mmHg)	132 ± 17.4	131.4 ± 16.8	135.1 ± 19.8	0.013	130.1 ± 16.5	135.7 ± 18.5	< 0.0001
Diastolic BP (mmHg)	72.5 ± 11.1	72.7 ± 10.9	71.0 ± 12.2	0.059	72.6 ± 11	72.2 ± 11.4	0.47
Albuminuria stage				< 0.0001			< 0.0001
Normal to mildly increased	547/1204 (45.4%)	482/1010 (47.7%)	65/194 (33.5%)		423/797 (53.1%)	124/407 (30.5%)	
Moderately increased	428/1204 (35.5%)	361/1010 (35.7%)	67/194 (34.5%)		273/797 (34.3%)	155/407 (38.1%)	
Severely increased	229/1204 (19.0%)	167/1010 (16.5%)	62/194 (32.0%)		101/797 (12.7%)	128/407 (31.4%)	
uACR (mg/mmol)	3 [1; 10]	2 [1; 9]	7 [2; 31]	< 0.0001	2 [1; 7]	7 [2; 30]	< 0.0001
eGFR (CKD-EPI, mL/min/1.73 m.^{2})	76.7 ± 21.0	78.7 ± 20.5	66.1 ± 20.4	< 0.0001	80.8 ± 19.7	68.5 ± 21.1	< 0.0001
Coronary artery disease	364/1349 (27.0%)	259/1140 (22.7%)	105/209 (50.2%)	< 0.0001	187/902 (20.7%)	177/447 (39.6%)	< 0.0001
Cerebrovascular disease	172/1349 (12.8%)	135/1140 (11.8%)	37/209 (17.7%)	0.024	88/902 (9.8%)	84/447 (18.8%)	< 0.0001
Carotid revascularisation	30/1349 (2.2%)	22/1140 (1.9%)	8/209 (3.8%)	0.12	16/902 (1.8%)	14/447 (3.1%)	0.12
Lower limb artery disease	112/1349 (8.3%)	76/1140 (6.7%)	36/209 (17.2%)	< 0.0001	44/902 (4.9%)	68/447 (15.2%)	< 0.0001
Total cholesterol (mmol/L)	4.78 ± 1.14	4.78 ± 1.15	4.75 ± 1.11	0.70	4.75 ± 1.09	4.83 ± 1.24	0.23
LDL-c (mmol /L)	2.73 ± 0.95	2.75 ± 0.96	2.65 ± 0.88	0.14	2.72 ± 0.93	2.77 ± 0.99	0.39
HDL-c (mmol /L)	1.21 ± 0.41	1.20 ± 0.40	1.26 ± 0.46	0.065	1.21 ± 0.39	1.20 ± 0.46	0.68
Triglycerides (mmol/L)	1.89 ± 1.43	1.88 ± 1.44	1.91 ± 1.37	0.81	1.85 ± 1.22	1.96 ± 1.77	0.28
NT-proBNP (pg/mL)	102 [47; 261]	82 [41; 202]	339 [161; 828]	< 0.0001	70 [36; 165]	234 [97; 578]	< 0.0001

Data are expressed using number (%) for categorical data, and mean ± SD or median [25th–75th percentile] for quantitative data, as appropriate. P-values are calculated using Fisher's exact test for categorical data, and Student T-test or Mann–Whitney U-test for quantitative data

Coronary artery disease was defined as any of the following: angina, coronary revascularization, myocardial infarction. Cerebrovascular disease was defined as any of the following: stroke, transient ischaemic attack. Lower limb artery disease was defined as lower limb revascularization and/or amputation

BMI: Body Mass Index; BP: blood pressure; CKD: Chronic Kidney Disease. HDL-c: high-density-lipoprotein-cholesterol; HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; LDL-c: low-density-lipoprotein-cholesterol; NT-proBNP: N-terminal prohormone of brain natriuretic peptide; uACR: urine albumin/creatinine ratio

(HR = 1.34, [1.22; 1.48], p < 0.0001) and homocysteine (HR = 1.30, [1.23; 1.38], p < 0.0001) were associated with all-cause death in univariate model, but not carnitine, TMA and methionine. After adjustment for age, sex, eGFR and uACR, the association with all-cause death remained significant for betaine (HR = 1.18, [1.08; 1.30], p = 0.0004) and homocysteine (HR = 1.18, [1.09; 1.27], p < 0.0001). In the fully adjusted models, only homocysteine remained significant (HR = 1.16, [1.06; 1.27], p = 0.0011). For HFrH and/or CV death, no association with the biomarkers remained significant in the fully adjusted models.

In an exploratory analysis, we analyzed whether the relationship between nutritional biomarkers and outcomes was modified when stratifying on relevant sub-groups (Additional file 2: Fig. S3A–C). Interestingly, the association between nutritional biomarkers of red-meat and vegetable intakes with clinical outcomes was not strongly influenced by history of CAD, obesity (BMI ≥ 30 kg/m^2) and possible history of HF (NT-propBNP ≥ 125 pg/mL).

Table 2 Baseline values for the nutritional biomarkers of interest, classified by status for follow-up events

Baseline characteristics	All (n = 1349)	Event: HFrH			Event: All-cause death		
		No (n = 1140)	Yes (n = 209)	P-value	No (n = 902)	Yes (n = 447)	P-value
Methylamines							
Carnitine (µmol/L)	44.0 ± 12.4	43.7 ± 12.0	45.8 ± 14.8	0.044	43.9 ± 12.0	44.2 ± 13.3	0.66
Betaine (µmol/L)	33.5 ± 14.1	33.0 ± 13.6	36.5 ± 16.1	0.0030	32.5 ± 13.0	35.5 ± 15.8	0.001
Choline (µmol/L)	1.48 ± 0.38	1.46 ± 0.37	1.57 ± 0.43	< 0.0001	1.44 ± 0.34	1.54 ± 0.44	< 0.0001
TMAO (µmol/L)	6.8 [4.2; 12.8]	6.6 [4.0; 12.3]	8.8 [5.3; 17.0]	< 0.0001	6.5 [4.0; 11.7]	7.9 [4.8; 15.3]	< 0.0001
TMA (µmol/L)	0.76 ± 0.27	0.77 ± 0.27	0.75 ± 0.27	0.53	0.76 ± 0.26	0.76 ± 0.28	0.93
Thio-amino-acids							
Cysteine (µmol/L)	23 [13; 39]	23 [13; 39]	23 [14; 40]	0.38	24 [13; 40]	21 [14; 37]	0.57
Homocysteine (µmol/L)	8.9 [4.7; 14.8]	8.6 [4.7; 14.1]	10.9 [5.7; 18.0]	< 0.0001	8.2 [4.5; 13.5]	10.9 [5.7; 17.5]	< 0.0001
Methionine (µmol/L)	26.1 ± 7.1	26.1 ± 6.8	26.1 ± 8.2	0.93	26.3 ± 6.8	25.7 ± 7.5	0.20

Data are expressed using mean ± SD or median [25th–75th percentile] for quantitative data, as appropriate. P-values are calculated using Fisher's exact test for categorical data, and Student T-test or Mann–Whitney U-test for quantitative data

HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; TMA: trimethylamine; TMAO: trimethylamine *N*-oxide

Discussion

In this monocentric cohort of patients with T2D, we were able to establish that nutritional biomarkers of red meat consumption (methylamines, TMAO and related compounds) and of vegetables (thio-amino-acids, homocysteine and related compounds) were associated with incident acute HFrH, HFrH and/or CV death, and all-cause death, in univariate analysis. However, when performing adjustment for renal parameters, betaine was the only biomarker remaining associated with incident HFrH. When adjusting on cardiac biomarkers (NT-proBNP and history of CAD), we found no remaining effect of those nutritional biomarkers. Likewise, no biomarkers were significantly associated with HFrH and/or CV death in the fully adjusted models. Interestingly, we found that homocysteine concentration was a significant and independent risk factors for all-cause death after multiple adjustments including renal and cardiac biomarkers, and after a conservative correction for multiple testing.

Characteristics of patients with HF—external validation

In this study, we confirmed that established risk factors for HFrH were present such as older age, longer diabetes duration, increased systolic blood pressure (SBP) and NT-proBNP concentrations, and renal parameters (CKD and macroalbuminuria). Of note, CAD was found in approximately half of the participants with incident HFrH, in agreement with the data from large cohorts studying this condition [16]. SURDIAGENE cohort is of peculiar value as it is specific of T2D and its related complications. We namely found that those participants who had HFrH during the follow-up had higher renal and retinal complications compared to those who remained free of HFrH during follow-up. This point was previously suggested in the EMPAREG Outcome study where the greater risk of HF was significantly associated with a greater number of microvascular complications [17]. Particularly, we found a greater risk of HFrH in those participants with diabetic retinopathy, in accordance with previous reports on microvascular disease and HFrH [18, 19].

Dietary approach to prevent HF

The current approach considered nutritional biomarkers potentially indicative of red meat (TMAO, betaine, choline and carnitine) and of vegetable consumption, more specifically folate intake (homocysteine, cysteine and methionine). An intake of red meat (specifically ≤ 28 g/day of pork, lamb and beef) and of vegetables ≥ 200 g/day is part of the EAT-Lancet score. So far, to our knowledge, the EAT-Lancet score was not reported in the field of HF. Red (and processed) meat consumption and cardiovascular disease was reviewed by Ferreira et al. and altogether evidenced a deleterious effect of red meat consumption, while plant protein intakes were seen very positively. [20] However, no specific mention of HF was available to our knowledge.

The DASH (Dietary Approach to Stop Hypertension) emphasizes a diet focusing on decreased intake of red meat and increased intake of vegetables, compared to the usual American diet [21]. Of interest, the DASH diet was largely tested and provided rather positive effects in the context of established HF, with the meta-analysis of two

Table 3 Survival analysis for HFrH, the composite HFrH and/or CV death event and all-cause death

	Unadjusted model		Fully adjusted model	
	HR ($95_{\%CI}$)	*P*-value	HR ($95_{\%CI}$)	*P*-value
Cause-specific HM for HFrH				
Carnitine	1.20 [1.05; 1.37]	0.0065	1.13 [0.99; 1.29]	0.061
Betaine	1.34 [1.20; 1.50]	< 0.0001	1.11 [0.97; 1.27]	0.13
Choline	1.35 [1.20; 1.52]	< 0.0001	0.94 [0.82; 1.08]	0.39
TMAO*	1.32 [1.16; 1.50]	< 0.0001	1.09 [0.94; 1.26]	0.24
TMA	1.01 [0.89; 1.15]	0.86	0.98 [0.86; 1.13]	0.78
Cysteine	1.38 [1.21; 1.58]	< 0.0001	1.10 [0.95; 1.28]	0.20
Homocysteine	1.28 [1.17; 1.39]	< 0.0001	1.05 [0.91; 1.21]	0.49
Methionine	1.02 [0.89; 1.18]	0.73	1.06 [0.93; 1.22]	0.38
Cause-specific HM for HFrH and/or CV death				
Carnitine	1.12 [1.01; 1.25]	0.037	1.06 [0.95; 1.17]	0.32
Betaine	1.27 [1.16; 1.40]	< 0.0001	1.04 [0.93; 1.17]	0.46
Choline	1.28 [1.17; 1.42]	< 0.0001	0.91 [0.82; 1.02]	0.093
TMAO*	1.31 [1.19; 1.45]	< 0.0001	1.10 [0.98; 1.23]	0.11
TMA	1.02 [0.93; 1.13]	0.65	0.99 [0.89; 1.10]	0.81
Cysteine	1.31 [1.17; 1.46]	< 0.0001	1.04 [0.92; 1.18]	0.49
Homocysteine	1.28 [1.20; 1.37]	< 0.0001	1.08 [0.97; 1.21]	0.16
Methionine	0.96 [0.85; 1.08]	0.46	1.00 [0.89; 1.11]	0.93
Cause-specific HM for all-cause death				
Carnitine	1.05 [0.95; 1.15]	0.32	1.01 [0.92; 1.11]	0.84
Betaine	1.28 [1.18; 1.39]	< 0.0001	1.07 [0.97; 1.18]	0.20
Choline	1.26 [1.16; 1.38]	< 0.0001	0.94 [0.85; 1.04]	0.23
TMAO*	1.20 [1.10; 1.31]	< 0.0001	1.03 [0.94; 1.14]	0.52
TMA	1.05 [0.97; 1.14]	0.26	1.01 [0.93; 1.10]	0.83
Cysteine	1.34 [1.22; 1.48]	< 0.0001	1.08 [0.97; 1.20]	0.15
Homocysteine	1.30 [1.23; 1.38]	< 0.0001	1.16 [1.06; 1.27]	0.0011
Methionine	0.96 [0.87; 1.06]	0.37	0.98 [0.89; 1.07]	0.66

* TMAO was natural-log transformed before standardization. Cause-specific hazard models were fitted using two adjustment models, unadjusted and fully adjusted. Covariates added in the fully adjusted models were age, sex, history of CAD, log transformed NT-proBPNP, eGFR and log transformed uACR. The nutritional biomarkers were tested separately from each other in the different adjustment models. All HR are given per 1 SD of the given parameter

Abbreviations: CV: cardiovascular; eGFR: estimated glomerular filtration rate calculated with the CKD-EPI 2009-formula; HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; HM: Hazard model; HR: Hazard-ratio; NT-proBNP: N-terminal prohormone of brain natriuretic peptide; TMA: trimethylamine; TMAO: trimethylamine N-oxide; uACR: urine albumin/creatinine ratio

studies concluding to a HR of HF at 0.71 suggesting a significant benefit of this dietary approach, in an analysis in which subjects with T2D were excluded [22]. However, the DASH diet is also associated with a restriction in salt intake, which limits the conclusion to a specific effect of this diet on HF.

The Mediterranean diet including the standardized PREDIMED diet also insists on the reduction of the red meat intake and the increase in vegetables consumption compared to usual diet [23]. Of interest, the Mediterranean diet was associated with a reduction in HF incidence and mortality in cohort studies. Conversely, using a randomized controlled strategy, the PREDIMED trial found no significant effect of the evaluated nutritional interventions on incident HF. The study population included 7403 participants without established HF, also including 3610

(See figure on next page.)

Fig. 2 Cumulative Incidence Function for HFrH. Quartile values for the different parameters of interest: carnitine (median = 42.6, [25th–75th] percentile = [35.4–50.7]); betaine (31.5, [24.6–38.9]); choline (1.43, [1.22–1.68]); TMAO (6.8, [4.2–12.8]); cysteine (23, [13–39]); homocysteine (8.9, [4.7–14.8]). HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; TMAO: trimethylamine *N*-oxide

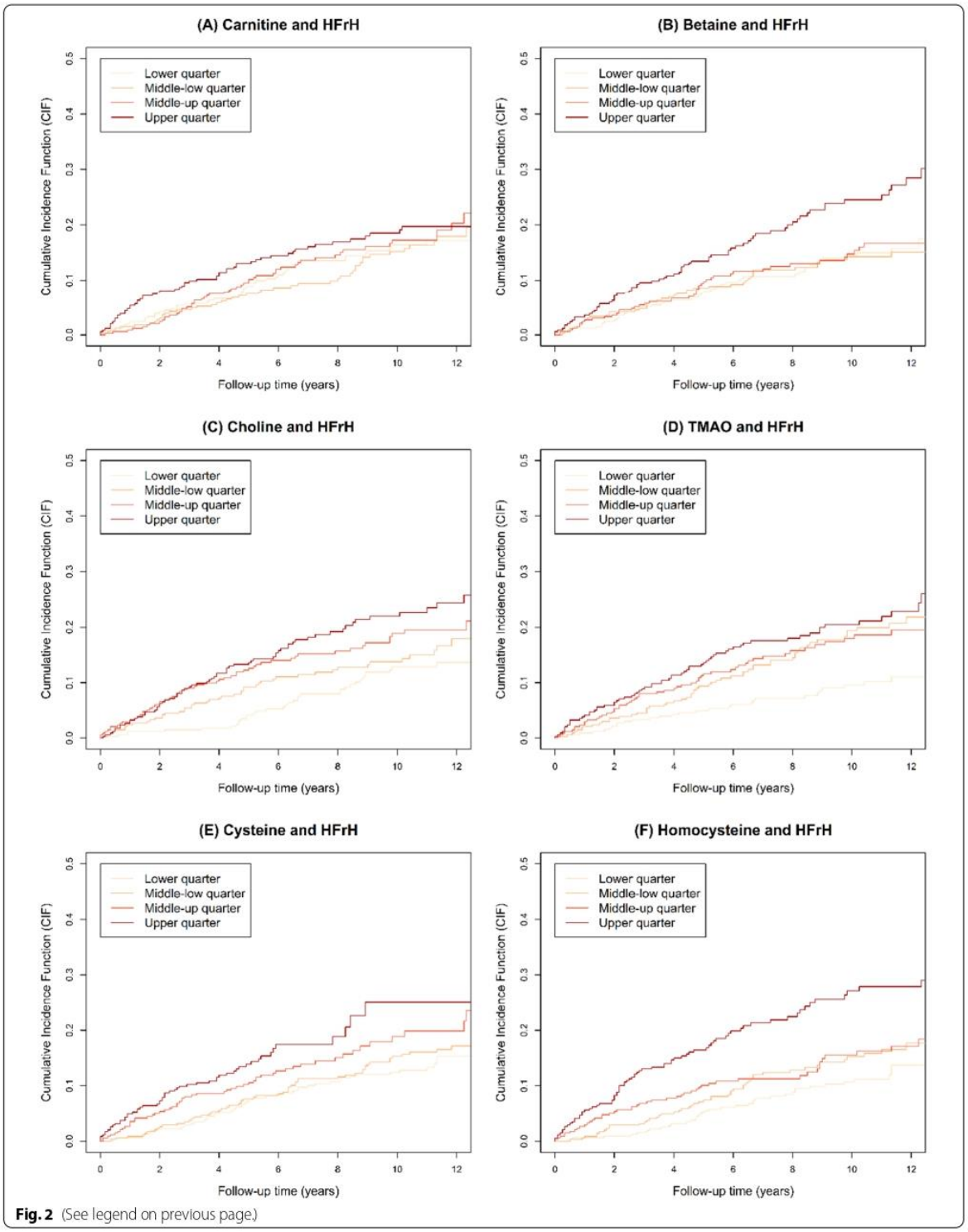


Fig. 2 (See legend on previous page.)

with T2D. However, the incidence of HF was rather low with a total of 94 events during a follow-up of 4.8 years. In the HOPE-2 trial, homocysteine lowering with folate was not associated with differences in hospitalization for HF [24]. In addition, a Cochrane meta-analysis suggested no role of vitamin B-based homocysteine-lowering interventions on all-cause death [25].

Methylamines and health outcomes

To our knowledge, the association between TMAO and related metabolites and incident HFrH was poorly examined in the literature. The majority focused on populations with established HF, which was summarized in a recent meta-analysis, showing that TMAO was a risk factor for MACE and all-cause death [26]. We previously found an association between plasma TMAO concentrations and MACE/death in the SURDIAGENE cohort, which was not restricted to patients with established HF [7]. The results presented here, considering multivariable models, did not reach statistical significance regarding an impact of TMAO on risk of HF in persons living with T2D. Of interest, the plasma concentration of TMAO was reported to be higher in those patients with HF compared to the others [27, 28]. This could partly explain our finding of an association between TMAO and HFrH, which did not persist when adjusting for other cardiac biomarkers, even if plasma concentrations of TMAO and linked metabolites were not correlated to NT-proBNP concentrations. Introducing data on gut microbiota, which is clearly a key explanation when studying this metabolite, is well-above the scope of this paper but should be addressed in future studies [26].

Thio-amino-acids and health outcomes

Our results suggest no obvious and strong relationship between thio-amino-acids and HFrH. Particularly, homocysteine was associated with the HFrH in univariate model but this relationship was not sustained when renal and/or cardiac biomarkers were added to the models. Very similar results were found in patients from the IDNT trial, with T2D and overt nephropathy [29]. However, we found that homocysteine concentration was associated with all-cause death, even when adjusting on age, sex, and, notably, renal and cardiac biomarkers. Of interest, a recent meta-analysis found that homocysteine concentrations were higher in patients with HF compared with those without [30]. Our results suggest against a strong effect of folate consumption regarding HF. However, it can also be argued that vegetables are part of a dietary pattern which proved to be beneficial regarding all-cause death in the PREDIMED trial [31]. This clearly illustrates the complexity of nutritional intervention and the difficulty to isolate the effect of one specific nutrient.

The interpretation of our findings regarding homocysteine association with all-cause death remains an open question. Our data were observational, while interventions altogether tend to be negative. This implies that we do not have enough scientific evidence to suggest for an increase in nutrients able to decrease homocysteine to lower all-cause death. However, our results can also be viewed as an important finding to establish biomarkers associated with key outcomes, in terms of epidemiology. In a computerized era, whether inclusion of homocysteine will lead to a better classification of HFrH-hazard will require continuous efforts, but our results are a strong impetus for such a step on the way to 4P medicine.

Limitations and strengths

The current study has limitations to acknowledge. The primary outcome considered in the current analysis can be questioned. Indeed, in HF trials, a composite endpoint combining CV death and hospitalization for HF is very consistently used. However, as the majority of the SURDIAGENE cohort was not affected by HF at baseline, we focused on HFrH, defined as the first occurrence of acute HF leading to hospitalization and/or death. CV death was not specific enough at variance with HF trials where CV death is mostly secondary to the ominous evolution of the condition. Still, the study of the composite event of HFrH and/or CV death was given along with the main outcome. One obvious question in our present approach was whether the lack of strong effect was related to a weak statistical power. After multiple adjustment, we found that the upper limit of the 95% confidence interval of the HR was 1.27 for betaine and 1.21 for homocysteine (for an increase of 1 SD of the given parameter), which does not support a strong relationship between these proxies of nutritional intakes and incident heart failure. So, even if we must acknowledge for a limited statistical power, considering a larger population and a greater number of events is unlikely to lead to highly clinically relevant association of red meat and folates intakes with HF. Also, the deleterious effect of homocysteine on all-cause death can be challenged. Whether this is due to a spurious result secondary to multiple testing was examined. When applying a very conservative Bonferroni correction (24 tests), homocysteine was still associated with all-cause death, suggesting that it truly represents a relevant risk factor for death in patients with T2D. Moreover, the analyses presented here relied on previously established observations of the link between TMAO and derivates and red meat intake, on one hand, and between homocysteine and folate intake, on the other hand.

Unfortunately, this study could not confront nutritional biomarkers to individual nutritional habits. Also, the exposure (dosing of nutritional biomarkers) was measured only once, at baseline. Also, the exposure (dosing of nutritional biomarkers) was measured only once, at baseline. Therefore, the study results are based on the hypothesis that these determinations were representative of the mean values of the biomarkers. Repeated data would be needed to increase the study accuracy. Thirdly, the observational design of our study leads to low-grade guidelines, even though randomized clinical trials might be challenging requiring some comparisons between animal meat and plant-derived meat, as mentioned by Ferreira et al. [20]. Lastly, a history of HF prior to inclusion in the cohort was not established. No questionnaire is available, to our best knowledge, to establish chronic or previous HF, in a similar fashion as the Rose questionnaire for CAD. However, when we stratified on the recommended threshold of NT-proBNP to indicate potential HF ($\geq$ 125 pg/mL) [32], we did not evidence a strong difference regarding the association of TMAO with HFrH between the two groups. Of note, caution must be taken as our subgroup analysis was not pre-specified.

This study also has some strengths including its long-term follow-up, the adjudication of clinical endpoints by an independent adjudication committee and the use of state of the art methodological determinations using mass spectrometry showing good stability and reproducibility.

Conclusion

To summarize, our study searched for an association between methylamines but also between thio-amino-acid plasma concentrations and severe HF. We did not evidence any major effect of these nutritional biomarkers associated with red meat consumption and folate intakes on incident HF in patients with T2D. The research strategy applied here remains rarely used while it could be considered as an alternative to time- and resource-consuming food questionnaire, to establish the impact of nutritional environment on health outcomes, such as HF. Our results could prove interesting with regard to other specific nutritional biomarkers for the prevention of HF. In addition, multiple approaches adding nutritional biomarkers to other exposome markers are surely a relevant research strategy in complex conditions such as cardiovascular metabolic diseases.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12933-022-01505-9.

Additional file 1. Biological determinations. Nutritional BM and HFrH in T2D. Details of the quantification of methylamines and amino-acids

Additional file 2. Nutritional BM and HFrH in T2D. Supplemental Figures and Tables.

Additional file 3. SURDIAGENE ORGANISATION (Committees and staff).

Acknowledgements
Participants and their general practitioners are warmly thanked. The centres and personnel involved in SURDIAGENE recruitment and adjudications, and the members of the SURDIAGENE study group are shown in the Additional file 3. Mikael Croyal and Samy Hadjadj acknowledge the Centre de Recherche en Nutrition Humaine- CRNH-Ouest for fruitful discussions related to nutritional biomarkers.

Author contributions
SH designed the SURDIAGENE study. MW, MC, SR, EG, JNT, XP, CLM, BC, PJS and SH designed the present analysis. EG performed the data management. MW performed all data analyses. MW, MC and SH wrote the main manuscript text. MW, MC, TG, BC, PJS and SH prepared the tables and figures. All authors significantly contributed to the drafting of the manuscript and reviewed the manuscript. All authors read and approved the final manuscript.

Funding
The SURDIAGENE cohort was supported by grants from the French Ministry of Health (PHRC-Poitiers 2004; PHRC-IR 2008), Association Française des Diabétiques (Research Grant 2003), Groupement pour l'Etude des Maladies Métaboliques et Systémiques (GEMMS Poitiers, France). Biological determinations were supported by research grants from Fondation de France (recipient: MC) and from Société Francophone du Diabète (recipient: PJS).

Availability of data and materials
The French regulatory authorities do not allow sharing of individual health data which can lead to patients' reidentification. Therefore, the entire dataset supporting the conclusions of this article cannot be made publicly available. However, specific subsets of the data may be provided by the corresponding author, on reasonable request.

Declarations

Ethics approval and consent to participate
The SURDIAGENE (SURvie, DIAbète de type 2 et GENEtique) study was approved by the Poitiers University Ethics Committee. All participants gave written informed consent.

Consent for publication
All the authors consent to this publication.

Competing interests
The authors declare no conflict of interest related to this work.

Author details
[1]Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, 44000 Nantes, France. [2]CHU de Nantes, INSERM CIC 1413, Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, Nantes, France. [3]Université de Poitiers, INSERM CHU de Poitiers, Centre d'Investigation Clinique, CIC 1402, Poitiers, France. [4]Université de Nantes, CHU Nantes, Inserm, CNRS, SFR Santé, Inserm UMS 016, CNRS UMS 3556, 44000 Nantes, France. [5]CRNH-Ouest Mass Spectrometry Core Facility, 44000 Nantes, France. [6]Nantes Université, CNRS, INSERM, l'institut du thorax, 44000 Nantes, France.

Received: 19 February 2022 Accepted: 5 April 2022
Published online: 09 June 2022

References

1. IDF Atlas 9th edition and other resources [Internet]. https://diabetesatlas.org/en/resources/. Accessed 30 Apr 2021.
2. Rawshani A, Rawshani A, Franzén S, Eliasson B, Svensson A-M, Miftaraj M, et al. Mortality and cardiovascular disease in type 1 and type 2 diabetes. N Engl J Med. 2017;376:1407–18.
3. Rawshani A, Rawshani A, Franzén S, Sattar N, Eliasson B, Svensson A-M, et al. Risk factors, mortality, and cardiovascular outcomes in patients with type 2 diabetes. N Engl J Med. 2018;379:633–44.
4. Ohkuma T, Komorita Y, Peters SAE, Woodward M. Diabetes as a risk factor for heart failure in women and men: a systematic review and meta-analysis of 47 cohorts including 12 million individuals. Diabetologia. 2019;62:1550–60.
5. Jenkins DJA, Dehghan M, Mente A, Bangdiwala SI, Rangarajan S, Srichaikul K, et al. Glycemic Index, Glycemic Load, and Cardiovascular Disease and Mortality. N Engl J Med. 2021;384:1312–22.
6. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail. 2016;18:891–975.
7. Croyal M, Saulnier P-J, Aguesse A, Gand E, Ragot S, Roussel R, et al. Plasma Trimethylamine N-Oxide and Risk of Cardiovascular Events in Patients With Type 2 Diabetes. J Clin Endocrinol Metab. 2020;105.
8. Willett W, Rockström J, Loken B, Springmann M, Lang T, Vermeulen S, et al. Food in the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food systems. Lancet Lond Engl. 2019;393:447–92.
9. Hadjadj S, Fumeron F, Roussel R, Saulnier P-J, Gallois Y, Ankotche A, et al. Prognostic value of the insertion/deletion polymorphism of the ACE gene in type 2 diabetic subjects: results from the Non-insulin-dependent Diabetes, Hypertension, Microalbuminuria or Proteinuria, Cardiovascular Events, and Ramipril (DIABHYCAR), Diabete de type 2, Nephropathie et Genetique (DIAB2NEPHROGENE), and Survie, Diabete de type 2 et Genetique (SURDIAGENE) studies. Diabetes Care. 2008;31:1847–52.
10. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150:604–12.
11. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation. 2016;133:601–9.
12. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94:496–509.
13. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. Comput Stat Data Anal. 2006;50:3464–85.
14. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.R-project.org/
15. Gray B. cmprsk: Subdistribution Analysis of Competing Risks [Internet]. 2020 [cited 2021 Jan 1].: https://CRAN.R-project.org/package=cmprsk
16. Low Wang CC, Hess CN, Hiatt WR, Goldfine AB. Clinical update: cardiovascular disease in diabetes mellitus: atherosclerotic cardiovascular disease and heart failure in type 2 diabetes mellitus—mechanisms, management, and clinical considerations. Circulation. 2016;133:2459–502.
17. Verma S, Wanner C, Zwiener I, Ofstad AP, George JT, Fitchett D, et al. Influence of microvascular disease on cardiovascular events in type 2 diabetes. J Am Coll Cardiol. 2019;73:2780–2.
18. Tromp J, Lim SL, Tay WT, Teng T-HK, Chandramouli C, Ouwerkerk W, et al. Microvascular disease in patients with diabetes with heart failure and reduced ejection versus preserved ejection fraction. Diabetes Care. 2019;42:1792–9.
19. Sandesara PB, O'Neal WT, Kelli HM, Samman-Tahhan A, Hammadah M, Quyyumi AA, et al. The prognostic significance of diabetes and microvascular complications in patients with heart failure with preserved ejection fraction. Diabetes Care. 2018;41:150–5.
20. Ferreira JP, Sharma A, Zannad F. The future of meat: health impact assessment with randomized evidence. Am J Med. 2021;134:569–75.
21. Sacks FM, Svetkey LP, Vollmer WM, Appel LJ, Bray GA, Harsha D, et al. Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. DASH-Sodium Collaborative Research Group. N Engl J Med. 2001;344:3–10.
22. Salehi-Abargouei A, Maghsoudi Z, Shirani F, Azadbakht L. Effects of Dietary Approaches to Stop Hypertension (DASH)-style diet on fatal or nonfatal cardiovascular diseases–incidence: a systematic review and meta-analysis on observational prospective studies. Nutr Burbank Los Angel Cty Calif. 2013;29:611–8.
23. Estruch R, Ros E, Salas-Salvadó J, Covas M-I, Corella D, Arós F, et al. primary prevention of cardiovascular disease with a Mediterranean diet supplemented with extra-virgin olive oil or nuts. N Engl J Med. 2018;378: e34.
24. Lonn E, Yusuf S, Arnold MJ, Sheridan P, Pogue J, Micks M, et al. Homocysteine lowering with folic acid and B vitamins in vascular disease. N Engl J Med. 2006;354:1567–77.
25. Martí-Carvajal AJ, Solà I, Lathyris D. Homocysteine-lowering interventions for preventing cardiovascular events. Cochrane Database Syst Rev. 2015;1:CD006612.
26. Li W, Huang A, Zhu H, Liu X, Huang X, Huang Y, et al. Gut microbiota-derived trimethylamine N-oxide is associated with poor prognosis in patients with heart failure. Med J Aust. 2020;213:374–9.
27. Trøseid M, Ueland T, Hov JR, Svardal A, Gregersen I, Dahl CP, et al. Microbiota-dependent metabolite trimethylamine-N-oxide is associated with disease severity and survival of patients with chronic heart failure. J Intern Med. 2015;277:717–26.
28. Tang WHW, Wang Z, Fan Y, Levison B, Hazen JE, Donahue LM, et al. Prognostic value of elevated levels of intestinal microbe-generated metabolite trimethylamine-N-oxide in patients with heart failure: refining the gut hypothesis. J Am Coll Cardiol. 2014;64:1908–14.
29. Friedman AN, Hunsicker LG, Selhub J, Bostom AG, Collaborative Study Group. Total plasma homocysteine and arteriosclerotic outcomes in type 2 diabetes with nephropathy. J Am Soc Nephrol JASN. 2005;16:3397–402.
30. Jin N, Huang L, Hong J, Zhao X, Chen Y, Hu J, et al. Elevated homocysteine levels in patients with heart failure: a systematic review and meta-analysis. Medicine. 2021;100: e26875.
31. Martínez-González MA, Sánchez-Tainta A, Corella D, Salas-Salvadó J, Ros E, Arós F, et al. A provegetarian food pattern and reduction in total mortality in the Prevención con Dieta Mediterránea (PREDIMED) study. Am J Clin Nutr. 2014;100(Suppl 1):320S-S328.
32. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2021;42:3599–726.

Publisher's Note

ADDITIONAL FILE 1: BIOLOGICAL DETERMINATION

Nutritional biomarkers and heart failure requiring hospitalization in patients with type 2 diabetes - the SURDIAGENE cohort.

Matthieu Wargny, Mikaël Croyal, Stéphanie Ragot, Elise Gand, David Jacobi, Jean-Noël Trochu, Xavier Prieur, Cédric Le May, Thomas Goronflot, Bertrand Cariou, Pierre-Jean Saulnier, Samy Hadjadj for the SURDIAGENE study group

Quantification of methylamines – Trimethylamine N-oxide (TMAO), trimethylamine (TMA), betaine, choline, and carnitine concentrations were determined by liquid chromatography-tandem mass spectrometry (LC-MS/MS). All solvents used were LC-MS grade and purchased from Biosolve (Valkenswaard, Netherlands). Standard compounds were obtained from Sigma Aldrich (Saint-Quentin Fallavier, France). A pool of reference standard solutions was prepared and serially diluted in acetonitrile to obtain seven standard solutions ranging from 0.05 to 100 µmol/L. Exogenous internal standards (10 µL) diluted at 25 µmol/L in acetonitrile ($^{2}H_9$-choline, $^{2}H_9$-carnitine, $^{13}C_2$-betaine, [$^{13}C_3$,^{15}N]-TMA and $^{2}H_9$-TMAO) were added to 20 µL of standard solutions and plasma samples. All samples were then treated with 75 µL of tert-butyl-bromoacetate (TMA derivatization) diluted at 50 mmol/L in acetonitrile and 10 µL of 70% ammonium hydroxide solution before mixing and incubation in the dark, at room temperature, for 30 min. Then, 50 µL of acetonitrile containing 1% formic acid were added and samples were centrifuged for 10 min at 10,000× g (20°C). Supernatants were then transferred to vials for LC-MS/MS analyses, performed on a Xevo® TQD mass spectrometer with an electrospray interface and an Acquity H-Class® UPLC™ device (Waters Corporation, Milford,

MA, USA). Samples (5 µL) were injected onto an HILIC-BEH column (1.7 µm, 2.1 × 100 mm, Waters Corporation) held at 35 °C. Compounds were separated using a linear gradient of mobile phase B (98% acetonitrile, 0.1% formic acid) in mobile phase A (10 mmol/L ammonium acetate, 0.1% formic acid) at a flow rate of 400 µL/min. Mobile phase A was kept constant for 1 min at 1%, linearly increased from 1% to 45% for 6.5 min, kept constant for 1 min, returned to the initial condition over 1 min, and kept constant for 1.5 min before the next injection. Targeted compounds were then detected by the mass spectrometer with the electrospray interface operating in the positive ion mode (capillary voltage, 1.5 kV; desolvatation gas (N_2) flow and temperature, 650 L/h and 350°C; source temperature, 150 °C). The multiple reaction monitoring mode was applied for MS/MS detection as detailed below. Chromatographic peak area ratios between unlabeled compounds and their respective internal standards constituted the detector responses. Standard solutions were used to plot calibration curves for quantification. The linearity was expressed by the mean R^2 which was greater than 0.998 for all compounds (linear regression, 1/x weighting, origin excluded).

Quantification of amino-acids – Cysteine, homocysteine and methionine plasma concentrations were performed by LC-MS/MS on a Xevo® Triple-Quadrupole mass spectrometer with an electrospray ionization interface equipped with an Acquity H-Class® UPLC™ device (Waters Corporation, Milford, MA, USA). All solvents used were LC-MS grade and purchased from Biosolve. Standard compounds were obtained from Sigma Aldrich. Individual stock solutions (10 mmol/L) of amino-acid and $^2H_3$-cysteine were prepared in 0.1 M HCl. A pool of unlabeled standard solutions was prepared and serially diluted in water to obtain seven standard solutions ranging from 0.1 to 50 µmol/L. A pool solution of labeled $^2H_3$-cysteine and [^{13}C,$^2H_3$]-methionine (50 µmol/L), was prepared in water. The standard solutions and serum samples (20 µL) were then extracted with 100 µL of methanol and 50 µL of the $^2H_3$-cysteine solution. The samples were mixed and centrifuged at 10 000 × *g* and 10 °C for 15 min

to remove the precipitated proteins. The supernatants were collected and dried under a gentle stream of nitrogen (45 °C). The derivatization step was performed by dissolving the dried extract in 100 µL of a freshly prepared butanol solution containing 5% acetyl chloride and kept at 60 °C for 30 min. The solvent was then removed under a gentle stream of nitrogen (60 °C). The dried samples were dissolved in 100 µL of water containing 0.1% formic acid and 50 µmol/L TCEP and injected into the LC-MS/MS system. Samples (10 µL) were injected onto an Acquity BEH-C_{18} column (1.7 µm; 2.1 × 100 mm, Waters Corporation) held at 60 °C, and compounds were separated with a linear gradient of mobile phase B (0.1% formic acid in methanol) in mobile phase A (0.1% formic acid in water) at a flow rate of 400 µL/min. Mobile phase B was kept constant at 1% for 0.5 min, linearly increased from 1% to 95% for 4.5 min, kept constant for 1 min, returned to the initial condition over 0.5 min, and kept constant for 1.5 min before the next injection. Target compounds were then detected by the mass spectrometer with the electrospray interface operating in the positive ion mode (capillary voltage, 3 kV; desolvatation gas (N_2) flow, 650 L/h; desolvatation gas temperature, 350 °C; source temperature, 120 °C). The multiple reaction monitoring mode was applied for MS/MS detection as detailed below. Chromatographic peak area ratios between unlabeled compounds and $^{2}H_3$-cysteine (cysteine and homocysteine) and [^{13}C,$^{2}H_3$]-methionine (methionine) constituted the detector responses. Standard solutions were used to plot the calibration curves for quantification. The assay linearity was expressed by the mean R^2, which was greater than 0.994 for all compounds (linear regression, 1/x weighting, origin excluded).

Multiple reaction monitoring (MRM) transitions used for LC-MS/MS detection.

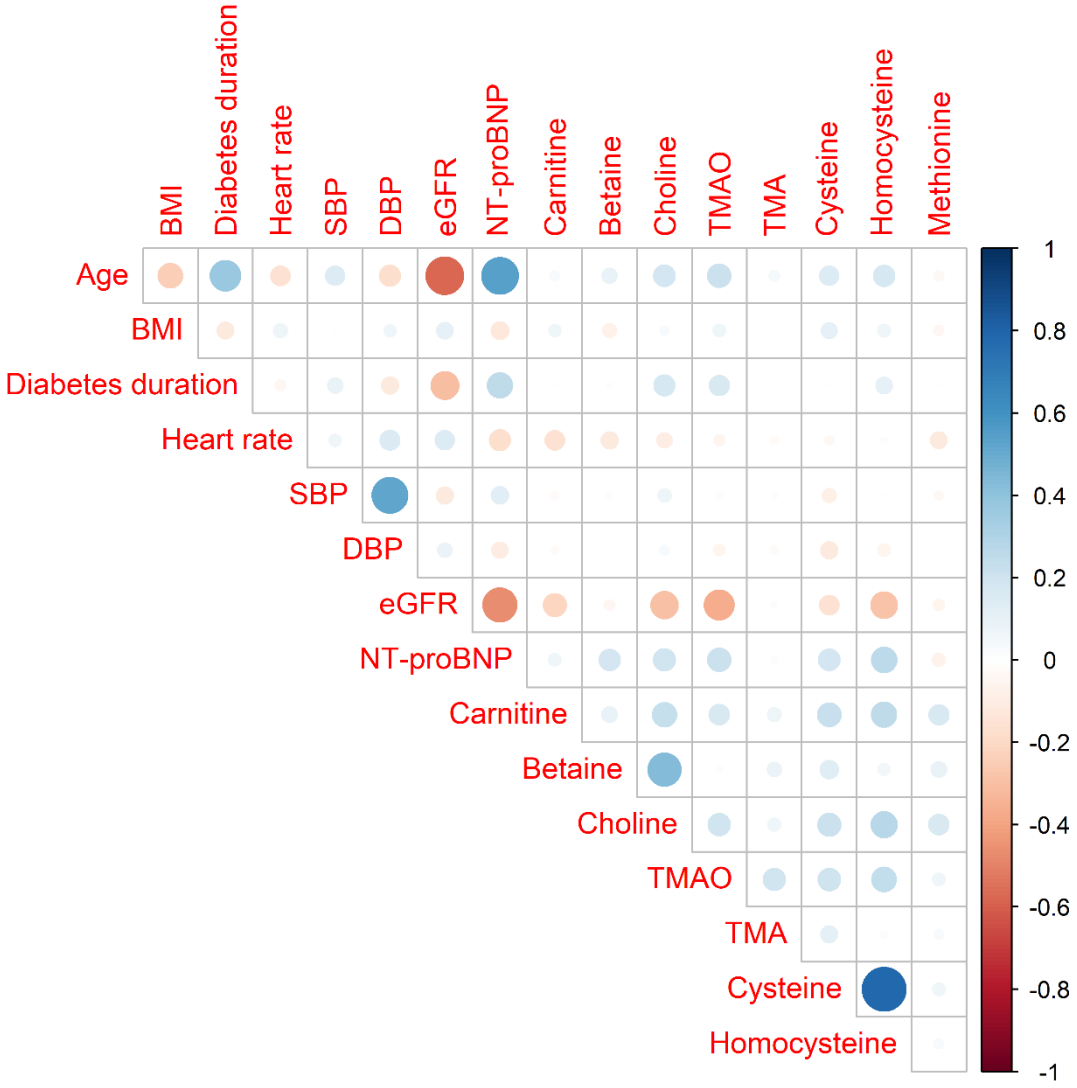
Compound	MRM transition (*m/z*)	Cone/collision (V)
TMAO	75.9 → 58.9	20/11
$^{2}H_{9}$-TMAO	85.0 → 68.0	20/11
TMA	174.1 → 118.0	35/18
[$^{13}C_{3}$, ^{15}N]-TMA	178.1 → 122.0	35/18
Betaine	118.1 → 58.1	40/22
$^{13}C_{2}$-betaine	120.1 → 58.1	40/22
Choline	104.1 → 60.1	40/15
$^{2}H_{9}$-choline	113.2 → 69.1	40/15
Carnitine	162.1 → 103.0	25/14
$^{2}H_{9}$-carnitine	171.1 → 112.0	25/14
Cysteine	178.1 → 75.9	30/15
Homocysteine	192.1 → 89.9	30/15
$^{2}H_{3}$-cysteine	181.1 → 78.9	30/15
Methionine	206.2 → 103.9	30/15
[^{13}C,$^{2}H_{3}$]-methionine	210.2 → 107.9	30/15

ADDITIONAL FILE 2: SUPPLEMENTAL FIGURES AND TABLES

Nutritional biomarkers and heart failure requiring hospitalization in patients with type 2 diabetes - the SURDIAGENE cohort.

Matthieu Wargny, Mikaël Croyal, Stéphanie Ragot, Elise Gand, David Jacobi, Jean-Noël Trochu, Xavier Prieur, Cédric Le May, Thomas Goronflot, Bertrand Cariou, Pierre-Jean Saulnier, Samy Hadjadj for the SURDIAGENE study group

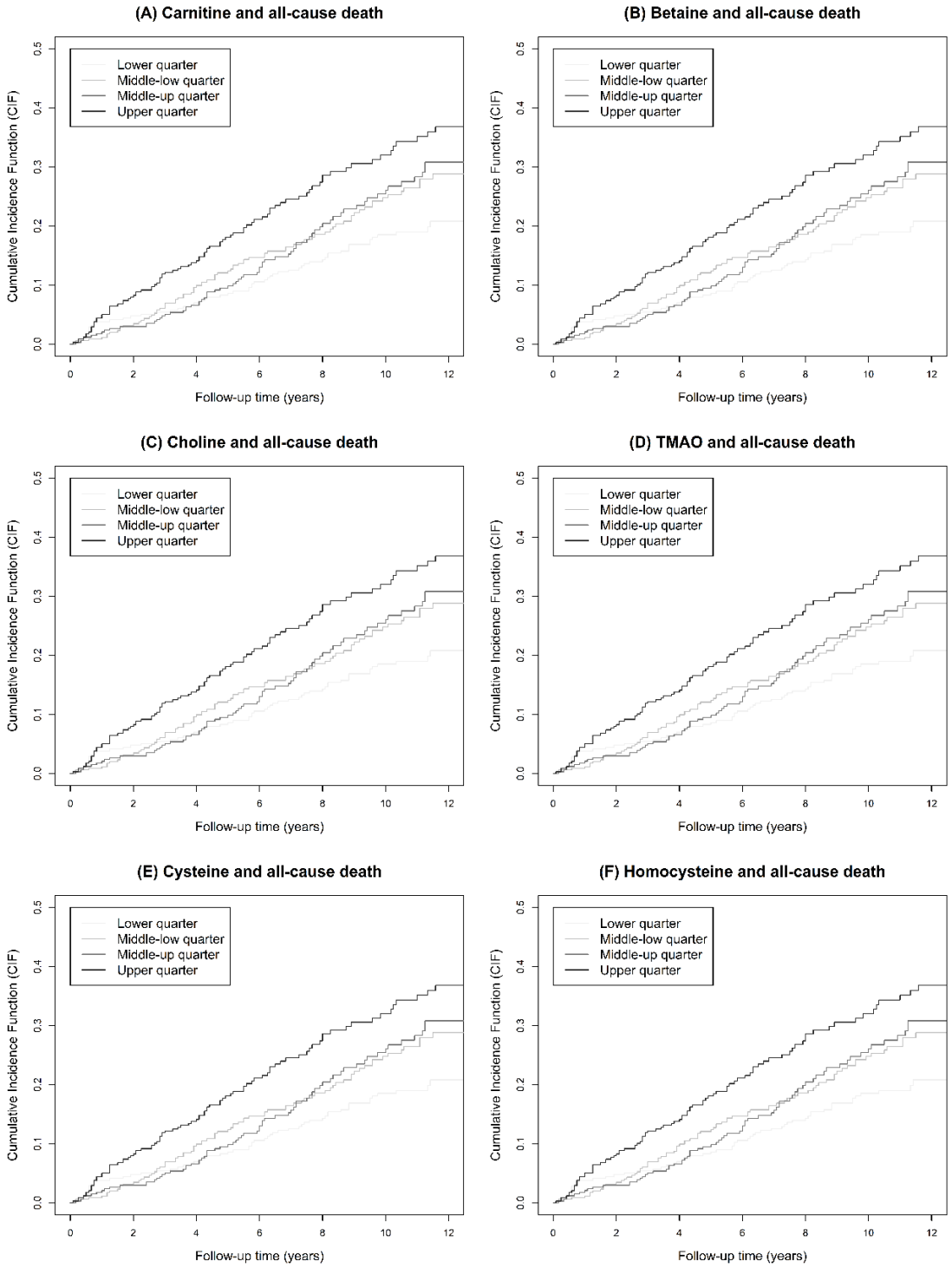
Supplemental Figure 1. Pairwise complete correlations plot for clinical and biological parameters



Spearman's rho correlation coefficients are computed, using all complete pairs of observations for a given pair of variables.

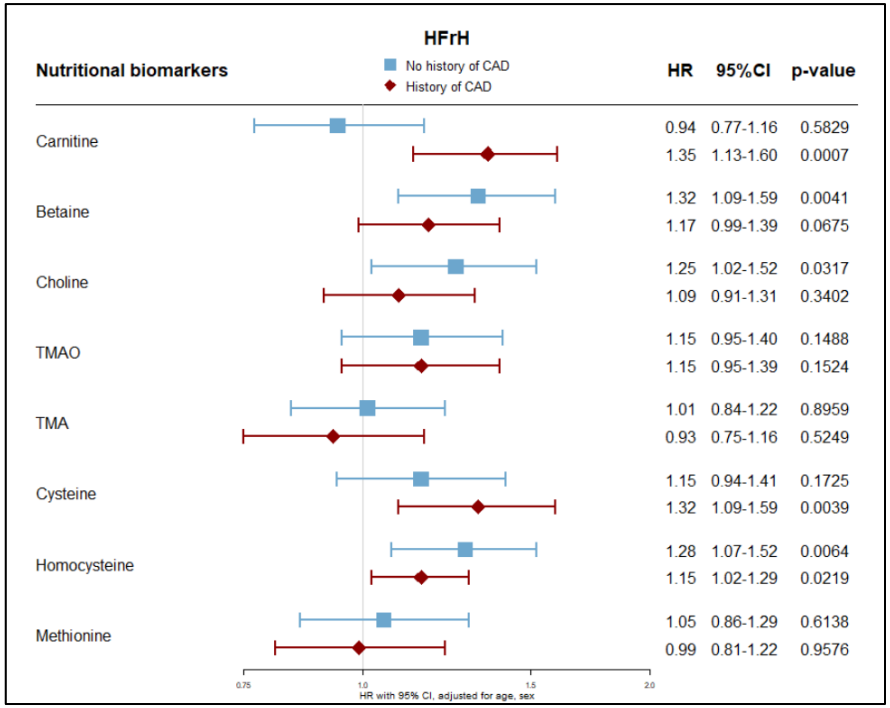
Abbreviations: BMI: body mass index; DBP: diastolic blood pressure; eGFR: estimated glomerular filtration rate calculated with the CKD-EPI 2009-formula; NT-proBNP: N-terminal prohormone of brain natriuretic peptide; SBP: systolic blood pressure; TMA: trimethylamine; TMAO: trimethylamine N-oxide

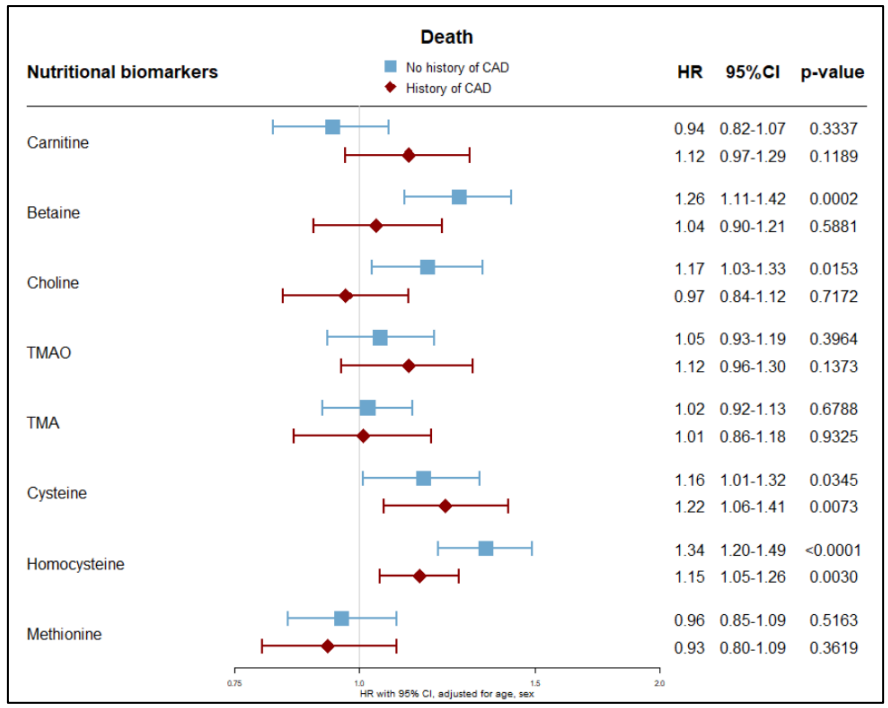
Supplemental Figure 2. Cumulative Incidence Function for all-cause death



Quartile values for the different parameters of interest: carnitine (median = 42.6, [25th-75th] percentile = [35.4-50.7]); betaine (31.5, [24.6-38.9]); choline (1.43, [1.22-1.68]); TMAO (6.8, [4.2-12.8]); cysteine (23, [13-39]); homocysteine (8.9, [4.7-14.8]). Abbreviations: TMAO: trimethylamine N-oxide.

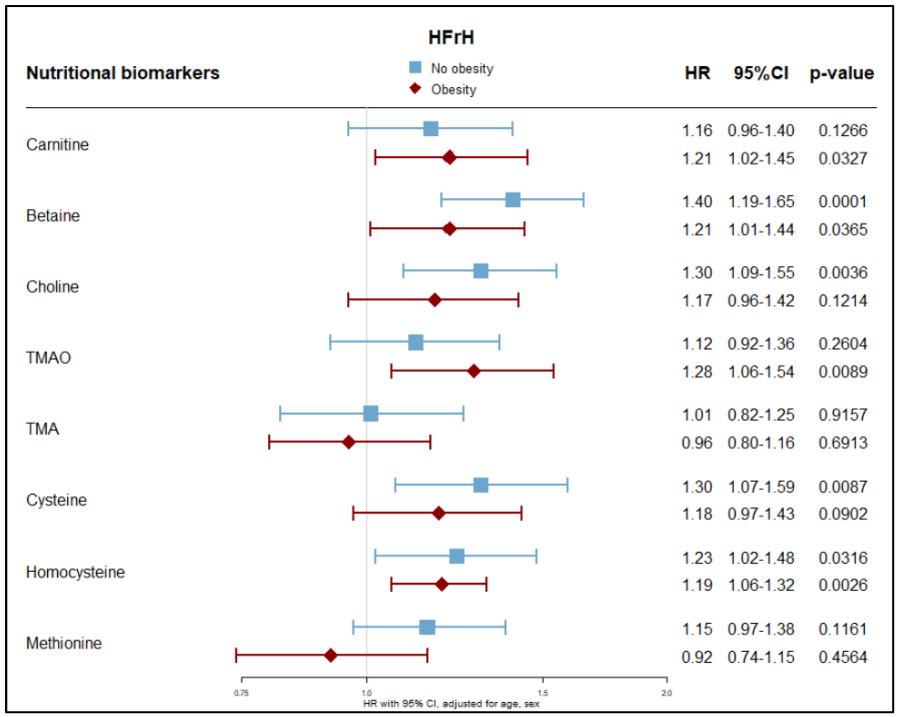
Supplementary Fig 3.A Survival analysis for HFrH and all-cause death stratified by CAD status

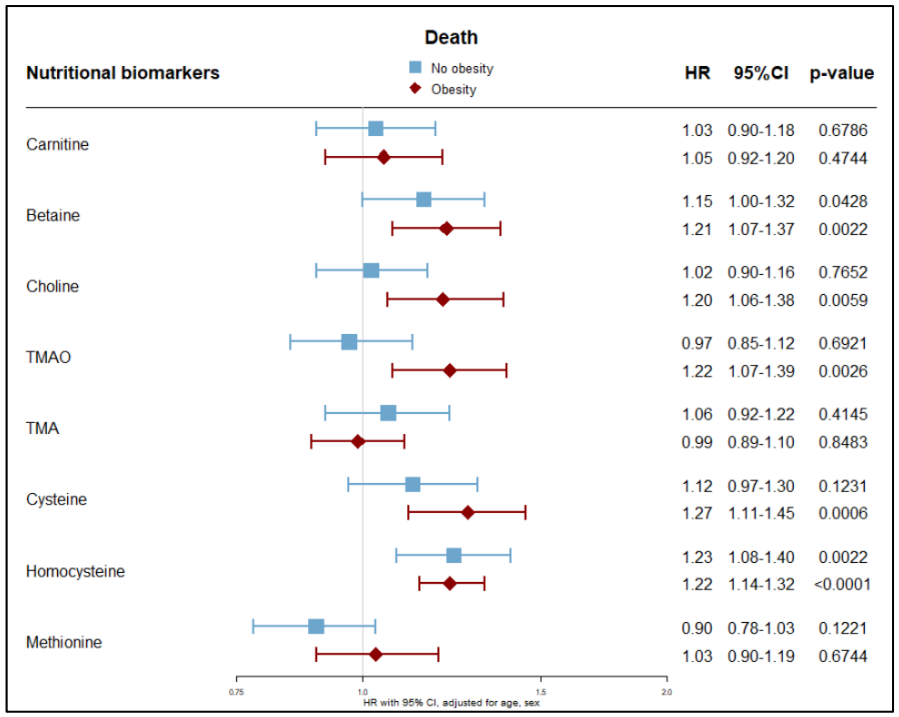




Cause-specific hazard models adjusted for age and sex. All HR are given per 1 SD of the given parameter. TMAO was natural-log transformed before standardization. The nutritional biomarkers were tested separately from each other in the different models. Abbreviations: CAD: coronary artery disease; HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; HR: Hazard-ratio; TMA: trimethylamine; TMAO: trimethylamine N-oxide

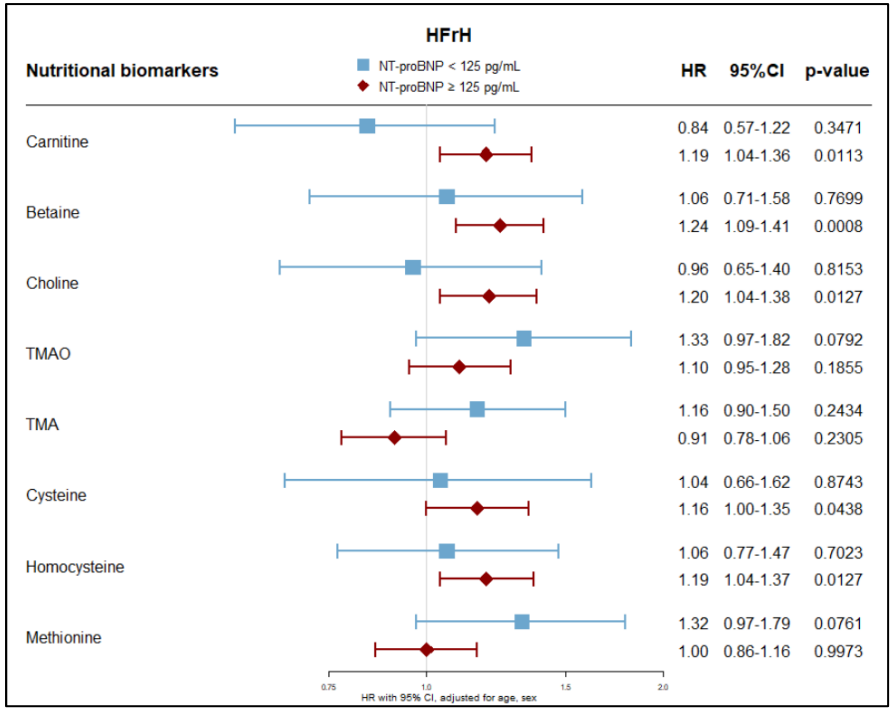
Supplementary Fig 3.B Survival analysis for HFrH and all-cause death stratified by obesity status

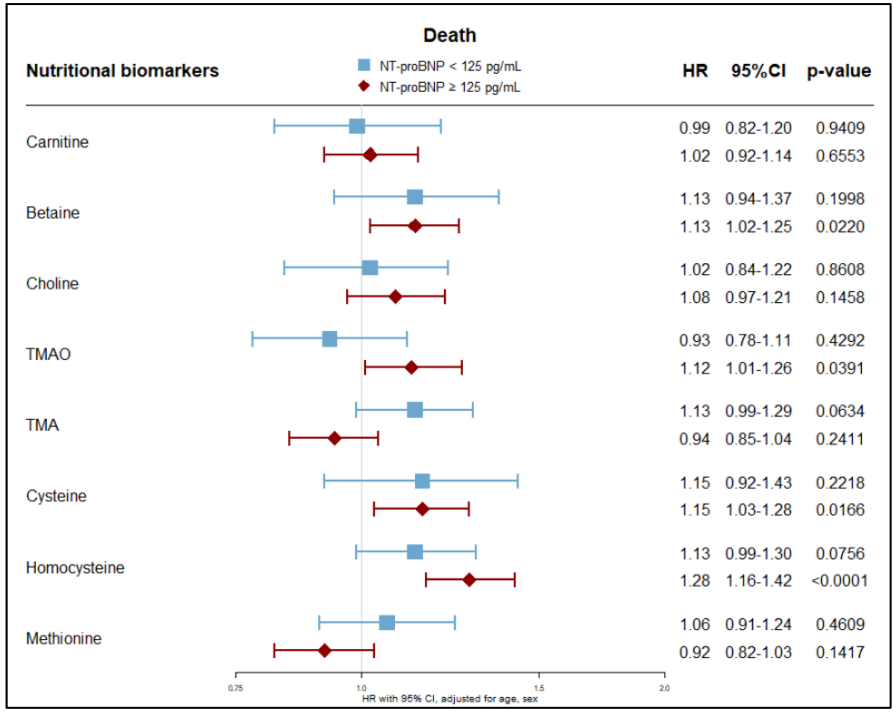




Cause-specific hazard models adjusted for age and sex. All HR are given per 1 SD of the given parameter. TMAO was natural-log transformed before standardization. The nutritional biomarkers were tested separately from each other in the different models. Abbreviations: HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; HR: Hazard-ratio; TMA: trimethylamine; TMAO: trimethylamine N-oxide

Supplementary Fig 3.C Survival analysis for HFrH and all-cause death stratified by NT-proBNP value





Cause-specific hazard models adjusted for age and sex. All HR are given per 1 SD of the given parameter. TMAO was natural-log transformed before standardization. The nutritional biomarkers were tested separately from each other in the different models. Abbreviations: HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; HR: Hazard-ratio; TMA: trimethylamine; TMAO: trimethylamine N-oxide

Supplemental Table 1. Survival analysis for HFrH – relative incidences (subdistribution hazard models)

	M_1		M_2		M_{3A}		M_{3B}		M_4	
	HR (95%CI)	***P*-value**	**HR (95%CI)**	***P*-value**	**HR (95%CI)**	***P*-value**	**HR (95%CI)**	***P*-value**	**HR (95%CI)**	***P*-value**
Methylamines										
Carnitine	1.20 [1.03; 1.40]	0.019	1.19 [1.02; 1.38]	0.025	1.14 [0.99; 1.31]	0.075	1.13 [0.97; 1.32]	0.11	1.14 [0.98; 1.32]	0.085
Betaine	1.25 [1.13; 1.39]	<0.0001	1.22 [1.09; 1.37]	0.0007	1.10 [0.96; 1.25]	0.18	1.24 [1.10; 1.40]	0.0004	1.10 [0.96; 1.27]	0.16
Choline	1.29 [1.15; 1.45]	<0.0001	1.20 [1.06; 1.36]	0.0036	1.00 [0.88; 1.13]	1	1.08 [0.95; 1.23]	0.26	0.97 [0.85; 1.11]	0.70
TMAO*	1.28 [1.13; 1.44]	<0.0001	1.16 [1.01; 1.33]	0.03	1.04 [0.90; 1.19]	0.63	1.06 [0.91; 1.24]	0.44	1.01 [0.86; 1.19]	0.88
TMA	0.99 [0.86; 1.14]	0.91	0.96 [0.83; 1.11]	0.57	0.96 [0.82; 1.13]	0.62	0.95 [0.82; 1.10]	0.52	0.96 [0.82; 1.12]	0.61
Thio-amino-acids										
Cysteine	1.28 [1.12; 1.46]	0.0002	1.15 [1.01; 1.31]	0.036	1.01 [0.87; 1.18]	0.87	1.07 [0.93; 1.23]	0.32	1.00 [0.86; 1.17]	0.96
Homocysteine	1.24 [1.12; 1.37]	<0.0001	1.16 [1.05; 1.28]	0.0025	1.01 [0.88; 1.15]	0.93	1.07 [0.95; 1.20]	0.27	0.98 [0.84; 1.13]	0.77
Methionine	1.01 [0.87; 1.19]	0.86	1.05 [0.90; 1.23]	0.51	1.08 [0.93; 1.25]	0.32	1.08 [0.93; 1.25]	0.31	1.09 [0.93; 1.27]	0.28

*TMAO was natural-log transformed before standardization. All HR are given per 1 SD of the given parameter. M_1: Model 1, univariate; M_2: M_1 covariates + age, sex; M_{3A}: M_2 covariates + history of CAD and log transformed NT-proBPNP; M_{3B}: M_2 covariates + eGFR and log transformed uACR; M_4: M_{3A} covariates + eGFR and log transformed uACR. The nutritional biomarkers were tested separately from each other in the different adjustment models.

Abbreviations: CAD: coronary artery disease; eGFR: estimated glomerular filtration rate calculated with the CKD-EPI 2009-formula; HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; HR: Hazard-ratio; NT-proBNP: N-terminal prohormone of brain natriuretic peptide; TMA: trimethylamine; TMAO: trimethylamine N-oxide; uACR: urine albumin/creatinine ratio

Supplemental Table 2 Survival analysis for HFrH, the composite HFrH and/or CV death event and all-cause death

	M_1		M_2		M_{3A}		M_{3B}	
Cause-specific HM for HFrH	**HR (95%CI)**	***P*-value**	**HR (95%CI)**	***P*-value**	**HR (95%CI)**	***P*-value**	**HR (95%CI)**	***P*-value**
Carnitine	1.20 [1.05; 1.37]	0.0065	1.19 [1.04; 1.35]	0.0090	1.13 [1.00; 1.29]	0.052	1.13 [0.99; 1.29]	0.073
Betaine	1.34 [1.20; 1.50]	<0.0001	1.30 [1.15; 1.47]	<0.0001	1.09 [0.96; 1.25]	0.19	1.33 [1.17; 1.50]	<0.0001
Choline	1.35 [1.20; 1.52]	<0.0001	1.23 [1.08; 1.40]	0.0016	0.98 [0.87; 1.12]	0.79	1.09 [0.95; 1.25]	0.21
TMAO*	1.32 [1.16; 1.50]	<0.0001	1.20 [1.05; 1.37]	0.0073	1.12 [0.97; 1.28]	0.12	1.10 [0.95; 1.27]	0.20
TMA	1.01 [0.89; 1.15]	0.86	0.99 [0.86; 1.13]	0.88	0.97 [0.85; 1.12]	0.73	0.98 [0.86; 1.12]	0.79
Cysteine	1.38 [1.21; 1.58]	<0.0001	1.24 [1.08; 1.42]	0.0022	1.12 [0.97; 1.30]	0.13	1.13 [0.98; 1.31]	0.092
Homocysteine	1.28 [1.17; 1.39]	<0.0001	1.20 [1.09; 1.32]	0.0002	1.09 [0.96; 1.24]	0.17	1.11 [0.98; 1.26]	0.10
Methionine	1.02 [0.89; 1.18]	0.73	1.05 [0.91; 1.20]	0.54	1.04 [0.91; 1.20]	0.53	1.08 [0.94; 1.24]	0.26
Cause-specific HM for HFrH and/or CV death event								
Carnitine	1.12 [1.01; 1.25]	0.037	1.11 [1.00; 1.23]	0.056	1.07 [0.96; 1.18]	0.21	1.05 [0.95; 1.17]	0.35
Betaine	1.27 [1.16; 1.40]	<0.0001	1.21 [1.09; 1.34]	0.00029	1.03 [0.92; 1.15]	0.62	1.23 [1.10; 1.36]	0.00016
Choline	1.28 [1.17; 1.42]	<0.0001	1.16 [1.04; 1.29]	0.0055	0.96 [0.86; 1.07]	0.43	1.03 [0.92; 1.15]	0.64
TMAO*	1.31 [1.19; 1.45]	<0.0001	1.20 [1.08; 1.34]	0.00048	1.13 [1.01; 1.25]	0.026	1.11 [0.99; 1.24]	0.068
TMA	1.02 [0.93; 1.13]	0.65	1.00 [0.90; 1.11]	1	0.98 [0.88; 1.10]	0.78	0.99 [0.90; 1.10]	0.86
Cysteine	1.31 [1.17; 1.46]	<0.0001	1.18 [1.05; 1.32]	0.0043	1.08 [0.96; 1.21]	0.22	1.07 [0.95; 1.21]	0.27
Homocysteine	1.28 [1.20; 1.37]	<0.0001	1.20 [1.12; 1.30]	<0.0001	1.13 [1.03; 1.25]	0.013	1.12 [1.01; 1.23]	0.024
Methionine	0.96 [0.85; 1.08]	0.46	0.97 [0.87; 1.09]	0.64	0.97 [0.87; 1.09]	0.63	1.01 [0.91; 1.14]	0.81
Cause-specific HM for all-cause death								
Carnitine	1.05 [0.95; 1.15]	0.32	1.04 [0.95; 1.15]	0.37	1.01 [0.92; 1.11]	0.80	1.01 [0.92; 1.11]	0.87
Betaine	1.28 [1.18; 1.39]	<0.0001	1.18 [1.08; 1.29]	0.0004	1.05 [0.96; 1.16]	0.28	1.18 [1.08; 1.30]	0.0004
Choline	1.26 [1.16; 1.38]	<0.0001	1.11 [1.01; 1.22]	0.029	0.98 [0.89; 1.08]	0.72	1.01 [0.91; 1.11]	0.91
TMAO*	1.20 [1.10; 1.31]	<0.0001	1.10 [1.00; 1.21]	0.044	1.06 [0.96; 1.17]	0.23	1.03 [0.93; 1.14]	0.54
TMA	1.05 [0.97; 1.14]	0.26	1.02 [0.94; 1.11]	0.69	1.01 [0.93; 1.10]	0.81	1.01 [0.93; 1.10]	0.75
Cysteine	1.34 [1.22; 1.48]	<0.0001	1.20 [1.09; 1.32]	0.0003	1.11 [1.00; 1.23]	0.045	1.11 [1.00; 1.24]	0.041
Homocysteine	1.30 [1.23; 1.38]	<0.0001	1.22 [1.15; 1.30]	<0.0001	1.19 [1.10; 1.28]	<0.0001	1.18 [1.09; 1.27]	<0.0001
Methionine	0.96 [0.87; 1.06]	0.37	0.96 [0.87; 1.05]	0.38	0.95 [0.87; 1.05]	0.32	0.99 [0.90; 1.08]	0.77

*TMAO was natural-log transformed before standardization. Cause-specific hazard models were fitted using different adjustment models. All HR are given per 1 SD of the given parameter. M_1: Model 1, univariate; M_2: M_1 covariates + age, sex; M_{3A}: M_2 covariates + history of CAD and log transformed NT-proBPNP; M_{3B}: M_2 covariates + eGFR and log transformed uACR. The nutritional biomarkers were tested separately from each other in the different models.

Abbreviations: CAD: coronary artery disease; CV: cardiovascular; eGFR: estimated glomerular filtration rate calculated with the CKD-EPI 2009-formula; HFrH: Heart Failure requiring Hospitalization, defined as the first occurrence of acute HF leading to hospitalization and/or death; HR: Hazard-ratio; NT-proBNP: N-terminal prohormone of brain natriuretic peptide; TMA: trimethylamine; TMAO: trimethylamine N-oxide; uACR: urine albumin/creatinine ratio

5. Apport de SURDIAGENE dans le cadre de la thèse

L'analyse proposée de la cohorte SURDIAGENE a permis d'apporter une première réponse à la question de l'intérêt pronostique de certains biomarqueurs nutritionnels (méthylamines, acides aminés) pour le risque d'insuffisance cardiaque chez des personnes présentant un diabète de type 2. Les principales limites sont intrinsèques aux données : coût et faisabilité de l'inclusion, induisant une cohorte de taille intermédiaire (≈1400 individus) et un biais de sélection non évaluable (inclusion hospitalière) ; absence d'enquête alimentaire, empêchant d'établir un lien entre les biomarqueurs et la consommation de viande, œuf, poisson ou légume ; absence de données répétées concernant l'exposition, affaiblissant considérablement l'interprétation du risque associé à distance de la visite d'inclusion ; absence, enfin, de caractérisation répétée de l'insuffisance cardiaque, limitant l'étude des événements majeurs que sont les hospitalisations et le décès.

Cependant, l'étude illustre bien les forces de la cohorte prospective : une biocollection systématique qui, dix-neuf ans après le lancement de l'étude, permet le dosage de biomarqueurs non envisagés lors de la mise en place de la cohorte. Concernant les événements, un enrichissement de ces données de recherche par les données issues du soin, avec la nécessaire tenue d'un comité d'adjudication pour qualifier l'événement cardiaque, ramenant l'investigateur à l'échelle du patient et non à un traitement statistique de données agrégées. Enfin, en creux, une qualité du recueil de données concernant les paramètres essentiels (exposition et événements, facteurs de risque cardiovasculaire et événements associés et néphropathie) qui, une fois actée la question des données manquantes n'est pas remise en question car correspondant à un niveau de qualité équivalent à celui auquel à accès le soignant en soins courants, et permet donc de prétendre à une recherche translationnelle, incompatible avec les possibilités actuelles des données massives en population française.

PARTIE IV : DMC - EVENEMENTS RETINIENS GRAVES ET INSUFFISANCE CARDIAQUE DANS LE SNDS

1. Résumé

Le projet DMC est présenté à la fin de cette partie (section 7) sous la forme d'un article scientifique en anglais, tel qu'il a été soumis au *European Heart Journal* au temps de la transmission du présent manuscrit de thèse.

Quatre éléments sont davantage développés ici : le déroulement général du travail (section 2) ; la population et les données (section 3) ; le travail de gestion des données issues du SNDS (vue générale dans la section 4, cas particulier des médicaments dans la section 5), puisqu'il ne pouvait pas être développé dans l'article et qu'il fait appel à des compétences pluridisciplinaires (épidémiologie, connaissance des données SNDS, data management, informatique SQL/SAS) ; les choix de l'approche statistique (section 6).

2. Déroulement et contributions respectives

Le Professeur Hadjadj et moi-même avons rédigé le protocole de l'étude DMC entre juin 2019 et août 2020, avec le soutien du groupe de travail du REDSIAM dans le champ des maladies endocriniennes, nutritionnelles et métaboliques, et, en particulier, de la relecture attentive de Sandrine Fosse-Edorh et du Dr Clara Piffaretti de Santé publique France.

J'ai constitué le dossier HDH pour soumission au CESREES puis à la CNIL (autorisation n°920426, décision DR-2020-380) sur la plate-forme « démarches simplifiées », soutenu par David Lair de la Direction de la Recherche et de l'Innovation du CHU de Nantes, de septembre à décembre 2020, en faisant valoir l'expérience de l'équipe de la Clinique des Données sur l'analyse de données massives, et en particulier mon expérience personnelle de traitement des données SNDS acquise dans le cadre de ma thèse d'exercice.[89] J'ai ensuite assuré la correspondance avec la CNAM pour formaliser la contractualisation avec le CHU, obtenue en juin 2021. L'accès effectif aux données a été possible sur le portail de la CNAM à partir d'août 2021 pour une période de 18 mois, donc jusqu'à février 2023.

J'ai opéré le *data management* de septembre 2021 à mai 2022 avec l'aide de Sandrine Coudol, épidémiologiste à la Clinique des Données. J'ai réalisé seul les analyses statistiques, sous la supervision du Pr Hadjadj, de janvier à septembre 2022.

Les expertises suivantes ont également été sollicitées, après une réunion d'étape en janvier 2022 présentant les premiers résultats : cardiologique (Pr Jean-Noël Trochu, CHU de Nantes),

diabétologique (Pr Bertrand Cariou, CHU de Nantes), ophtalmologique (Dr Jean-Baptiste Ducloyer, CHU de Nantes), neurovasculaire (Dr Pacôme Constant dit Beaufils, CHU de Nantes), chirurgie vasculaire (Pr Yann Gouëffic, groupe Hospitalier Paris Saint-Joseph), chirurgie bariatrique (Pr Claire Louis-Blanchard, CHU de Nantes), PMSI (Dr Christophe Leux, CHU de Nantes) et SNDS (Dr Philippe Tuppin, CNAM, qui a spécifiquement travaillé sur l'insuffisance cardiaque dans ces bases).

La première version du manuscrit a été rédigée par le Pr Hadjadj et moi-même, puis partagée à l'ensemble du groupe de travail.

La **Figure 5** offre un résumé du calendrier du projet DMC.

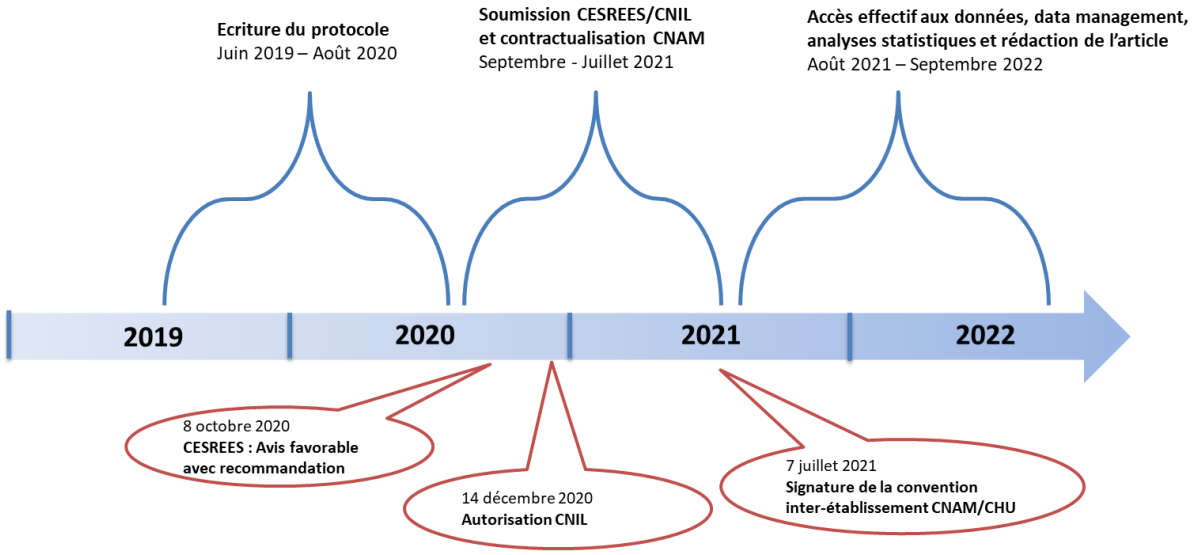


Figure 5. DMC - Calendrier du projet

3. Population et données

a. Identification de la population de DMC

La CNAM a accepté la mise à disposition sur le portail de l'Assurance Maladie des données SNDS de l'ensemble des individus rattachés au régime général et identifiés comme diabétiques sur la période 2012-2018. L'algorithme d'identification se fonde sur un travail préalable publié par Santé publique France, dont la sensibilité et la spécificité ont été estimées par croisement avec les données de la cohorte CONSTANCES.[32,90,91] Brièvement, l'algorithme s'appuie sur trois grandes sources d'information :

- les délivrances de médicaments en lien avec le diabète : tous les médicament de classe ATC débutant par « A10 », à l'exception du benfluorex (MEDIATOR) [1]
- les ALD pour diabète (ALD n°8)
- les hospitalisations rapportant un diagnostic de diabète, qu'il s'agisse de la maladie diabétique *per se* ou de complications jugées spécifiques du diabète

A partir des résultats de cette extraction, nous avons ensuite appliqué les critères d'éligibilité suivants :

- Critères d'inclusion
    - En vie et âge ≥ 25 ans au 31 décembre 2018
    - Absence de discordance sur l'âge et sur le sexe
    - Identifiant individuel unique confirmé par l'INSEE et sans risque d'ambiguïté sur le suivi inter-régimes (en particulier pour les données de l'ATIH)

- Critères d'exclusion pour l'analyse 2019
    - « Perdu de vue » en 2019 : en pratique, absence d'information visible en hospitalisation comme en ambulatoire
    - Critère appliqué pour l'analyse principale : Insuffisance cardiaque identifiée sur 2012-2018, défini par ≥ 1 code CIM-10, GHM ou médicament traceur (éplérénone, sacubitril) spécifique de l'IC

b. Antécédents et Evénements

L'analyse temporelle des comorbidités des patients à partir des informations issues du SNDS est censurée à gauche au 1er janvier 2012. En effet, le remboursement des prestations et les données d'hospitalisation (PMSI-MCO) ne sont disponibles qu'à partir de cette date. On rencontre cependant deux exceptions : (i) Les rares séjours débutés avant le 1er janvier 2012 mais s'étant achevés sur la période 2012-2019 sont disponibles ; (ii) la table des ALD n'est pas annualisée et peut mentionner une ALD débutée avant 2012.

Nous avons proposé ici de définir

- **Comme des antécédents,** sans date de début précise, toutes les pathologies identifiées comme antérieures au 31 décembre 2018
- **Comme des événements,** avec une date de début au jour près, toutes les pathologies

- Non identifiées comme des antécédents au 31 décembre 2018
- Et identifiées comme des événements sur la période du 1er janvier au 31 décembre 2019, à partir du croisement de quatre sources de données d'intérêt : PMSI-MCO (diagnostics CIM-10, GHM, actes CCAM), codes CCAM visibles dans le DCIR, ALD, et plus rarement nouveaux traitements médicamenteux (ex : délivrance d'AVASTIN pour l'injection intravitréenne, ou de sacubitril pour l'insuffisance cardiaque)

c. Codes retenus pour les pathologies d'intérêt

Chacune des pathologies d'intérêt a été définie à partir d'une ou plusieurs sources d'information, en prenant comme point de départ l'approche de la cartographie 2021 de la CNAM [91], cette approche étant ensuite affinée à partir :

- des fiches ALD proposées par la CNAM, pour la coronaropathie [92], l'AVC [93] et l'AOMI [94]

- de publications antérieures sur le SNDS, notamment concernant l'insuffisance cardiaque [40], le pied diabétique [61], la valvulopathie [95], l'insuffisance rénale [40] et les pathologies cardiovasculaires en général [96]

- d'échanges avec des médecins spécialistes : le Dr Jean-Baptiste Ducloyer pour l'ophtalmologie (définition de l'événement rétinien grave), le Pr Jean-Noël Trochu pour la cardiologie, le Dr Pacôme Constant dit Beaufils pour la neurologie (AVC et accident ischémique transitoire), le Pr Yann Gouëffic pour la chirurgie vasculaire (AOMI), le Pr Samy Hadjadj et le Pr Cariou pour la diabétologie, le Dr Christophe Leux pour compléter ces approches par des codes CCAM pertinents (en particulier pour les valvulopathies)

- de notre expérience de GAVROCHE pour l'insuffisance cardiaque

Une description exhaustive des codes utilisés figurait dans le dossier de soumission au CESREES/CNL. Cette description a évolué pour parvenir à la liste disponible dans le *Supplemental File 1* de l'article. Dans le présent manuscrit, nous nous contenterons de donner la définition des deux principales pathologies d'intérêt : l'IC (HF dans l'article, pour *Heart Failure*) et l'événement rétinien grave (SRE dans l'article, pour *Serious Retinal Event)*.

d. Précision sur le lien entre codages et pathologies

Les codages visent à identifier « au plus tôt » une pathologie d'intérêt, à partir des données disponibles dans le SNDS. **Il apparaît essentiel de souligner que nous ne cherchons pas à distinguer un antécédent d'un nouvel événement aigu.** Par exemple, un antécédent de cardiopathie ischémique pourra être codé dans le PMSI-MCO comme diagnostic associé à une hospitalisation pour insuffisance cardiaque aiguë, tandis qu'une hospitalisation pour infarctus du myocarde sera volontiers codée comme diagnostic principal d'une hospitalisation, mais les deux informations seront traitées de la même manière.

C'est sur cette base qu'il ne sera pas fait de distinction entre :

- les codages CIM-10 issus du PMSI-MCO (diagnostic principal, relié ou associé à une hospitalisation, issus des RUM),
- les codes CCAM identifiés comme associés à une hospitalisation (tables MCO_A du PMSI-MCO) ou en ambulatoire (tables CAM du DCIR)
- et les ALD spécifiques et codes CIM-10 associés aux ALD (table IR_IMB_R du DCIR)

e. Définition de la 1ère détection de l'insuffisance cardiaque

L'IC sera définie de façon composite par la 1ère identification d'un code CIM-10 (ALD ou hospitalisation avec diagnostic principal, relié ou associé), d'un code GHM, ou d'un médicament traceur de l'insuffisance cardiaque.

Après validation par l'expert cardiologue, les codes CIM-10 retenus étaient ceux commençant par

- I50 : insuffisance cardiaque
- I11.0 : cardiopathie hypertensive avec IC
- 13.0 : cardionéphropathie hypertensive, avec IC
- 13.2 : cardionéphropathie hypertensive, avec IC et insuffisance rénale
- 13.9 : cardionéphropathie hypertensive, sans précision
- R57.0 : choc cardiogénique

Les codes CIM-10 suivants n'ont pas été retenus : J81 (œdème pulmonaire), I11 (cardiopathie hypertensive) et I13 (cardionéphropathie hypertensive).

Les codes GHM retenus sont ceux commençant par 05M09, « Insuffisances cardiaques et états de choc circulatoires ».

Les médicaments traceurs retenus du fait d'indications spécifiques à l'IC sont l'éplérénone et le sacubitril.

f. Définition de la 1ère détection de l'événement rétinien grave

L'événement rétinien grave (ERG) a été défini par la 1ère détection de l'un des événements suivants :

- ERG 1 : laser rétinien, défini à partir des actes CCAM
  - BGNP001 : séance de photocoagulation choriorétinienne du pôle postérieur, avec laser monochromatique ou laser à colorants
  - BGNP003 : séance de destruction de lésion choriorétinienne par photocoagulation avec laser, à l'aide de verre de contact
  - BGNP004 : séance de destruction de lésion choriorétinienne par photocoagulation transpupillaire avec laser, à l'aide de verre de contact
  - BGNP007 : séance de destruction de lésion choriorétinienne par photocoagulation avec laser, à l'aide d'ophtalmoscope indirect
  - BGNP008 : séance de photocoagulation choriorétinienne du pôle postérieur (pour prise en charge de l'œdème maculaire diabétique), avec laser à argon ou diode

- ERG 2 : décollement rétinien ou hémorragie du corps vitré, défini à partir des codes CIM-10
  - H33.4 : décollement par traction de la rétine
  - H43.1 : hémorragie du corps vitré
  - H45.0 : hémorragie du corps vitré au cours de maladies classées ailleurs

- ERG 3 : injection intravitréenne en lien avec l'atteinte rétinienne (CCAM + médicament)
  - BLGB001 : injection intravitréenne d'agent pharmacologique dans le corps vitré
  - associée à la délivrance d'un médicament d'intérêt (± 60 jours) parmi le dexaméthasone, le ranibizumab, l'aflibercept, la fluocinolone et le bevacizumab

Cette dernière définition (ERG 3) posant le problème d'un recouvrement partiel avec la dégénérescence maculaire liée à l'âge.

4. Gestion des données SNDS : vue d'ensemble

a. Détail des trois grandes sources de données d'intérêt dans le SNDS

Dans le cadre de DMC, les données d'intérêt mises à disposition par la CNAM peuvent être d'abord divisées en 3 grandes sources [97]:

- **les données du SNIIRAM** (Système National d'Information Inter-Régimes), directement issues des bases de l'Assurance Maladie. Elles comprennent en particulier les remboursements de prise en charge ambulatoire, qui nous intéressent ici pour les délivrances médicamenteuses, les consultations médicales (en particulier ophtalmologiques), et certains actes médicaux (comme une chirurgie sur valve cardiaque ou une injection intravitréene)

- **les données de l'ATIH** (Agence Technique de l'Information de l'Hospitalisation), qui contiennent toutes les informations issues des PMSI des centres hospitaliers français, publics comme privés, avec en particulier pour les besoins du projet DMC les codes CIM-10 de diagnostics et les GHM (Groupe Homogène de Malades) des séjours (RUM – Résumé d'unité médicale). Pour DMC, seules les tables MCO (médecine, chirurgie, obstétrique et odontologie) ont été exploitées

- **les données du CépiDc** (Centre d'épidémiologie sur les causes médicales de Décès), transmises par l'INSERM au SNDS, qui informent sur le statut vital des patients, avec le cas échéant le codage des diagnostics des causes de décès telles que rapportées dans le certificat de décès. Cette information n'est cependant disponible que jusqu'à l'année 2017 incluse, et nous nous sommes donc limités à l'information « décès toutes causes »

b. Conséquences de l'approche multi-source sur l'identification des bénéficiaires

L'approche multi-source offerte par le SNDS, avec des tables construites séparément sans identifiant unique commun pour un patient donné, a une conséquence pratique très importante pour l'analyste. Celui-ci doit en effet s'assurer que l'identifiant individuel associé à une donnée renvoie sans ambiguïté à un patient unique, ce qui n'est pas systématique du fait du partage d'identifiant pour différents bénéficiaires couverts par le même ouvrant-droit, et en particulier pour les jumeaux de même sexe. Ce point nous empêche de lier sans ambiguïté les identifiants des tables DCIR et ATIH. Ces notions

sont détaillées en ligne sur l'aide en ligne mise à disposition sur la plateforme du HDH [98] et résumées dans une fiche thématique.[99]

A noter que dans les différentes tables SNDS, une variable « rang bénéficiaire » est proposée pour différencier plusieurs bénéficiaires d'un même ouvrant-droit. Son intérêt est cependant limité en cas d'approche inter-régimes (= si un individu change de régime de sécurité sociale) puisque le rang bénéficiaire n'a pas la même signification selon le régime. De plus, cette variable n'est présente dans les tables ATIH que depuis 2014-2015.

Pour résumer, nous ne disposons pas donc dans toutes les tables SNDS d'un identifiant « vraiment » propre à un individu donné. En effet, nous sommes parfois contraints d'utiliser un pseudo-NIR potentiellement partagé avec d'autres individus. Sa capacité discriminante est donc incertaine, en particulier lors du chainage aux données PMSI. Nous avons jugé qu'une telle ambiguïté était inacceptable pour l'analyse de DMC, puisque le risque est important d'attribuer en défaut ou en excès les informations associées à une hospitalisation. Nous avons donc pris le parti d'exclure secondairement les individus partageant un même identifiant, en suivant une procédure mise à disposition par différents experts de la CNAM et de SPF dans l'indispensable document « SNDS, ce qu'il faut savoir » (cf. [100], se reporter annexe 3 page 49 du document).

Cette approche conservatrice a conduit à l'exclusion de 2,8% des patients de la base (cf. Figure 2 de l'article, diagramme de flux), la perte de puissance et le biais de sélection ayant été jugés acceptables pour notre question d'étude.

c. Tables SNDS exploitées dans le cadre de DMC – vue d'ensemble

Les tables SNDS mises à disposition sur le portail ont été traitées via l'interface *SAS Enterprise Guide*, par une approche combinant le codage SAS et des procédures SQL interprétées par SAS[101], en vue de constituer la base finale contenant les seules données nécessaires à l'analyse statistique. Pour ce travail de *data management*, en plus des sources déjà citées [98,100], nous nous sommes largement appuyés sur les informations en ligne mises à disposition par le HDH via une application Shiny R.[102]

La **Figure 6** propose un schéma simplifié des différentes tables mises à contribution pour la récupération des données de DMC et du schéma de récupération des données.

Ce schéma est centré sur l'individu : la première étape est l'identification des individus d'intérêt, puis vient la création de la table de correspondance avec les autres sources, un même individu pouvant être identifié par différents pseudo-NIR, comme expliqué précédemment. Nous extrayons ensuite les données d'intérêt à partir des différents jeux de tables fournies :

- **Pour les prestations ambulatoires,** les tables sont annualisées, et la table prestation (PRS) permet :
    - Le lien avec les tables de délivrances médicamenteuses ambulatoires (PHA, pour pharmacie) via un thésaurus médicamenteux permettant de lier les codes CIP et les classes ATC et donc de connaître les détails du médicament délivré, ici les médicaments du diabète et les médicaments à visée cardiovasculaire (antihypertenseurs, antiagrégants plaquettaires, anticoagulants, hypolipémiants, traitements spécifiques de l'insuffisance cardiaque comme l'éplérénone et le sacubitril)
    - Le lien avec les actes médicaux codés en ambulatoire (codage CCAM, tables ER_CAM_F_20aa)
    - L'identification directe des consultations ophtalmologiques

- **Pour les hospitalisations PMSI-MCO,** nous avons exploité les données des tables suivantes, également annualisées :
    - T_MCOaaA : description des actes CCAM
    - T_MCOaaB et T_MCOaaC : description du séjour (B) et identification du patient et dates de soin (C)
    - T_MCOaaD : description des diagnostics associés au séjour

- **Pour les affections de longue durée,** l'approche est plus simple puisque la table SNDS (IR_IMB_R) n'est pas annualisée. Elle regroupe l'ensemble des ALD débutées pour un patient donné, une même ALD pouvant figurer sur plusieurs lignes s'il a été nécessaire de la prolonger. Pour une pathologie donnée, on ne récupérait que la 1ère ALD, en prenant soin de respecter la démarche de contrôle qualité bien détaillée sur le site du HDH et rendue nécessaire par des erreurs de codage à la source [103]

- **Pour le statut vital,** nous avons croisé les données à partir de 3 sources, aucune n'étant exhaustive

- SNIIRAM : tables des bénéficiaires IR_BEN_R/IR_BEN_ARC_R, et tables annualisées CT_IND_AAAA_GN de la cartographie des pathologies, qui mentionnent parfois le décès du patient
- PMSI : tables T_MCOaaB et T_MCOaaC, annualisées, dans le cas où le « mode de sortie » d'hospitalisation est le décès
- CépiDc (table KI_CCI_R, non annualisée), table des causes initiales de décès, lorsque l'information a pu être récupérée à partir de la base nationale des décès

- **Autres informations d'intérêt récupérées mais non détaillées**
  - Date de dernière nouvelle définie par la date de décès du patient ou à défaut la dernière date parmi (i) remboursement de prestation ambulatoire et (ii) fin de séjour hospitalier
  - Indice de déprivation (Fdep) associé à la dernière commune de résidence, finalement non exploité dans l'article

A noter que l'approche multi-source induit un risque de redondance. Par exemple, un même acte CCAM peut être codé à la fois dans les tables ambulatoires du SNIIRAM et dans les tables hospitalières PMSI-MCO. Il est toutefois nécessaire d'interroger les deux sources puisqu'aucune n'est exhaustive. Cela n'a pas eu de conséquence pour DMC puisque nous nous intéressions à la première occurrence d'un acte donné : nous n'avons conservé que la 1ère de ces deux observations. Si nous nous étions intéressés au nombre d'occurrences d'un acte sur une période, il eut été nécessaire de ne compter qu'une fois les événements survenus à la même date.

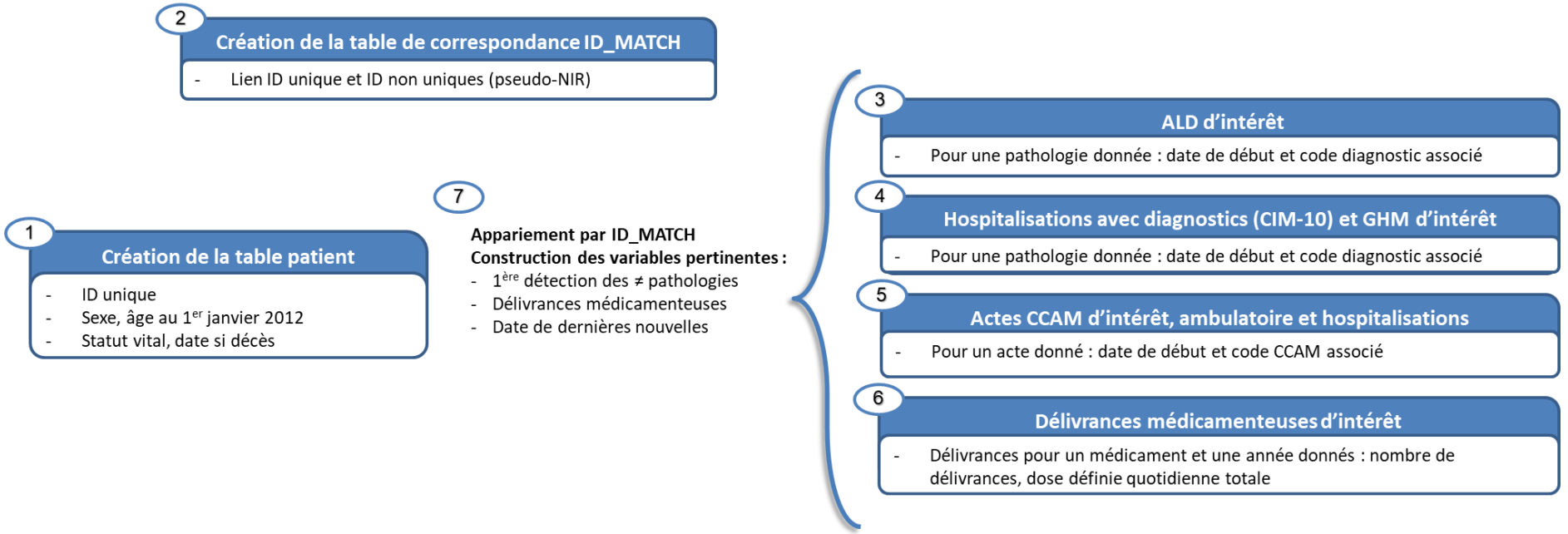


Figure 6. DMC - Schéma récapitulatif pas-à-pas de la construction des données extraites du SNDS pour constituer la base.

Les sept étapes, le contrôle qualité et les tables interrogées sont détaillés page suivante

Description pas-à-pas de la constitution de DMC dans le SNDS

Principales tables impliquées - exemple pour l'année 2018

1 **Constitution de la table patient**
- Sexe, âge à partir des table bénéficiaires
- Statut vital : approche croisant DCIR, ATIH et données CépiDc
- Tests de cohérence associés dont « NIR confirmé »

Identités : DMC_ANO_IDT
Bénéficiaires : IR_BEN_R et IR_BEN_ARC_R
Cartographie des pathologies : CT_IND_2018_GT
Décès CépiDc : KI_CCI_R

2 **Création de la table de correspondance NIR/pseudo-NIR**

Identités : DMC_ANO_IDT

3 **Récupération des ALD d'intérêt**
- Filtre sur les ALD actives sur la période 2012-2019
- Corrections sur les numéros ALD
- Pour une pathologie donnée, on récupère code CIM-10 et date de début

Référentiel médicalisé des bénéficiaires : IR_IMB_R
Thésaurus ALD : IR_CIM_B

4 **Récupération des hospitalisations avec diagnostics d'intérêt via CIM-10 et GHM**
- Extraction de l'ensemble des diagnostics d'intérêt + date associée
- Pour un diagnostic donné, on ne conserve que la première survenue

Séjours MCO : T_MCO_2018_B / T_MCO_2018_C
Diagnostics associés : T_MCO_2018_D

5 **Récupération des actes CCAM d'intérêt**
- Ambulatoire : extraction de l'ensemble des diagnostics d'intérêt + date associée
- Hospitalisation : extraction de l'ensemble des diagnostics d'intérêt + date associée
- Fusion des deux sources -> pour un acte donné, on ne conserve que la première survenue

Prestations : ER_PRS_F_2018
Actes ambulatoires : ER_CAM_F_2018
Codes CCAM associés : T_MCO_2018_A
Tables des séjours MCO : T_MCO_2018_B / T_MCO_2018_C

6 **Récupération des délivrances médicamenteuses d'intérêt (cf. section 5 pour le détail)**
- Extraction des délivrances des médicaments d'intérêt + date associée
- Enrichissement par thésaurus pour lien CIP/ATC + nombre de boîtes, nombre d'unités, grand conditionnement, calcul de la dose définie journalière délivrée
- Plusieurs types d'information sont extraites :
  - 1ère délivrance de certaines classes thérapeutiques (antidiabétique, traitement de l'insuffisance cardiaque)
  - dose définie journalière annuelle pour l'insuline
  - Délivrance définissant le traitement régulier (≥ 3 délivrances à des dates distinctes ou ≥2 avec ≥1 grand conditionnement)
  - Délivrance concomitante à un acte médical (anti-VEGF et injection intravitréenne)

Prestations : ER_PRS_F_2018
Délivrances en pharmacie : ER_PHA_F_2018
Unités commune de dispensation : ER_UCD_F_2018 (pour le sacubitril)

7 **Création des données d'intérêt résumant pathologies, délivrances médicamenteuses et suivi**
- Identification d'une pathologie : 1ère survenue parmi différents codages possibles (diagnostics via CIM-10 et GHM, actes via CCAM, médicaments traceurs via ATC)
- Délivrance médicamenteuse régulière par année
- Date de perte de vue définie par le dernier événement parmi prestations ambulatoires, fin de séjour hospitalier et décès

Eléments de contrôle qualité
1) T_MCO_2018_B : suppression séjours erronés ou fictifs
2) T_MCO_2018_C : suppression des séjours dits « douteux »
3) Délivrances médicamenteuses : somme sur les délivrances annuelles, exclusion si ≤ 0
4) Actes médicaux : somme sur les actes quotidiens, exclusion si ≤ 0

5. Gestion des données SNDS : exemple des délivrances médicamenteuses

a. Informations pharmacologiques d'intérêt

Dans le cadre de DMC, nous nous intéressons aux médicaments délivrés à la population de l'étude, et ce :

- **Pour connaître la date de 1ère délivrance de certains médicaments**, par exemple d'un antidiabétique ou d'un médicament spécifique de l'insuffisance cardiaque. Nous en déduisons que la maladie associée est antérieure à la date de délivrance

- **Pour estimer, pour une année donnée, si un patient était régulièrement traité par ce médicament, ainsi que la dose de totale de médicament délivrée.** Par imitation des pratiques de la CNAM, un traitement régulier a été défini par ≥3 délivrances à des dates distinctes, ou ≥2 délivrances en cas de grand conditionnement.[91] La dose totale délivrée est le produit du nombre de boîtes, par le nombre d'unités par boîte et par la dose unitaire. Elle peut être ramenée en doses définies journalières « DDD » (pour *defined daily doses*) en utilisant les doses définies par l'OMS, par exemple 2g/jour pour la metformine.[1] Pour DMC, nous n'avons exploité que la dose quotidienne moyenne délivrée pour l'insulinothérapie, bien que les DDD aient été calculées pour tous les médicaments

Il nous est donc nécessaire de connaître, pour un individu donné et pour une famille médicamenteuse donnée définie par son code ATC, à quelle date lui a été délivré ce médicament et sous quelle forme : nombre de boîtes, nombre d'unités par boîte et dose par unité. Le cas particulier des médicaments rétrocédés en milieu hospitalier (cas de certaines injections intravitréennes) et des médicaments en Autorisation Temporaire d'Utilisation (ATU - cas du sacubitril pour l'insuffisance cardiaque) a nécessité une approche spécifique, non développée ici.

b. L'information médicamenteuse dans les tables du SNDS : codes CIP et ATC

Dans les tables du SNDS, les médicaments délivrés en ambulatoire, accessibles par les tables PHA annualisées de pharmacie du DCIR, sont renseignés avec 1 ligne = 1 lot de une ou plusieurs boîtes délivrées lors d'un même passage en pharmacie, avec en particulier pour chaque ligne :

- Code CIP (club interpharmaceutique) du conditionnement d'intérêt, à 13 chiffres. Par exemple pour le code 3400938010541 [104] :
    - Désignation
        - Nom commercial, *ex. : GLUCOPHAGE*
        - Dosage, *ex. : 850 mg*
        - Forme pharmaceutique, dont la DCI : *ex. : metformine, COMPRIMES DISPERSIBLES OU ORODISPERSIBLES*
        - Conditionnement : *ex. : 1 boîte de 28 comprimés*
    - Mais aussi nom du laboratoire, éléments de tarification, etc. (non détaillés ici)
- Nombre de boîtes délivrées. A noter que ce nombre peut être négatif, par exemple pour corriger une délivrance codée en excès
- Date de délivrance (ou « exécution de la prestation »)

Notre point de départ dans la base étant la DCI du médicament, celle-ci va renvoyer à un ou plusieurs codes ATC (classification anatomique, thérapeutique et chimique), en particulier pour différencier les monothérapies et les associations médicamenteuses. Par exemple, le code ATC de la metformine en monothérapie est A10BA02, mais il existe 19 autres codes pour les associations, avec différents niveaux de précision : un code unique pour l'association avec les sulfamides hypoglycémiants, mais un code spécifique pour l'association avec chaque gliflozine. L'ensemble des CIP associés peuvent être récupérés à l'aide d'un thésaurus actualisé par la CNAM sur le site du SNDS (la table IR_PHA_R). Dans le cas de la metformine, ce principe actif est retrouvé sous 322 codes CIP différents.

Pour aller plus loin, l'**Annexe 5** liste les pièges rencontrés dans la mise en correspondance des classes ATC et des codes délivrés.

Dans un esprit de parcimonie, la liste exhaustive des classes médicamenteuses d'intérêt était associée à la demande d'autorisation CESREES/CNIL soumise au HDH. Celle-ci a été mise à jour et résumée dans le ***Supplemental File 1*** de l'article soumis.

c. En pratique : extraction des données issues des tables SNDS

Cet exemple illustre également la question de la complexité algorithmique, le terme de « complexité » étant pris ici dans le sens du volume d'opérations informatiques élémentaires nécessaires. Cette extraction met en jeu trois tables du SNDS :

- **la table des prestations (dite PRS),** qui est la table la plus volumineuse de la base DMC avec jusqu'à 1,6 milliard de lignes par année. Elle dépasse largement la question pharmaceutique puisqu'elle porte également sur les prestations de biologie, d'actes médicaux, de transports, de consultations… elle nous est indispensable car elle contient l'identifiant patient et la date d'exécution de la prestation, soit ici la date de délivrance en pharmacie

- **la table PHARMACIE (dite PHA),** qui porte également sur des gros volumes (environ 240 millions de lignes), et où chaque ligne correspond à un code CIP, avec entre autres l'information sur le nombre de boîtes délivrées et le détail du conditionnement (petit ou grand)

- **un thésaurus fourni par la CNAM,** qui détaille toutes les caractéristiques d'un code CIP donné et en particulier la classe ATC

Pour DMC, il va être nécessaire de croiser ces tables via une procédure SQL, sachant que le temps de calcul sera proportionnel au produit des dimensions des tables. Dans un souci d'optimisation, nous allons procéder en deux temps (schématisés **Figure 7**) :

1) **Jointure PHARMACIE et thésaurus :** le thésaurus est filtré sur les seuls codes CIP d'intérêt pour l'étude, permettant de passer de quelques 27300 à 8500 lignes, et la jointure avec la table pharmacie permet d'obtenir les seules 128 millions de lignes associées aux médicaments d'intérêt, enrichies des détails issus du thésaurus (posologie, DDD)

2) **Jointure PRESTATION et PHARMACIE :** la table des prestations est filtrée sur les délivrances médicamenteuses, permettant de passer de 1600 à 370 millions de lignes. Puis une jointure est faite avec la table des prestations, ce qui permet d'obtenir une table de 128 millions de lignes où 1 ligne représente la délivrance d'une forme médicamenteuse (par exemple 3 boîtes d'atorvastatine 40 mg) à une date donnée, avec les informations d'intérêt sur la posologie et le conditionnement

Cela n'est pas détaillé ici mais, en sus d'une sélection sur les lignes (= les observations), les tables ont été filtrées également sur les colonnes d'intérêt (= les variables). En théorie, ce filtre ne doit pas modifier le temps des procédures de jointure SQL des tables puisque la jonction est opérée à partir de la seule clef de jointure sans solliciter les autres variables. Cependant, une fois cette jointure réalisée, un temps non négligeable est nécessaire pour l'écriture (= l'enregistrement sur le serveur) de la table produite, et ce temps sera directement proportionnel au nombre de variables (bien que dépendant de leur type, un booléen étant moins volumineux qu'un nombre, lui-même moins volumineux qu'une longue chaîne de caractères).

Cette démarche permet de ramener le temps de calcul sur le portail à environ 1 heure pour 1 année. Les tables étant annualisées, nous avons pu définir une boucle automatisant la procédure sur 2012-2019. Cette tâche peut ainsi être planifiée sur environ une journée sans solliciter le programmeur. A noter qu'une exécution hors connexion est possible sur le portail CNAM (dite « SAS asynchrone ») mais que je ne suis pas parvenu à l'exploiter pour cette requête, soit par défaut de compréhension de ma part, soit en raison des gros volumes impliqués. En effet, un échec d'exécution sur SAS asynchrone ne générant pas de rapport d'erreur, il m'était plus difficile d'identifier le problème.

Enfin, deux bonnes pratiques d'optimisation ne sont pas détaillées ici mais ont été appliquées pour optimiser le temps d'exécution :

- Les jointures ont été réalisées à partir de tables SAS situées dans la même bibliothèque SAS (« *library* »), et le résultat de ces jointures a été écrit dans cette même bibliothèque

- Dans la mesure du possible, certaines variables d'intérêt directement déduites du code CIP pouvaient être supprimées des tables intermédiaires et ajoutées à la table finale. Le gain d'exécution associé était cependant marginal, tandis que l'écriture du code était un peu plus complexe et donc augmentait le risque d'erreur humaine

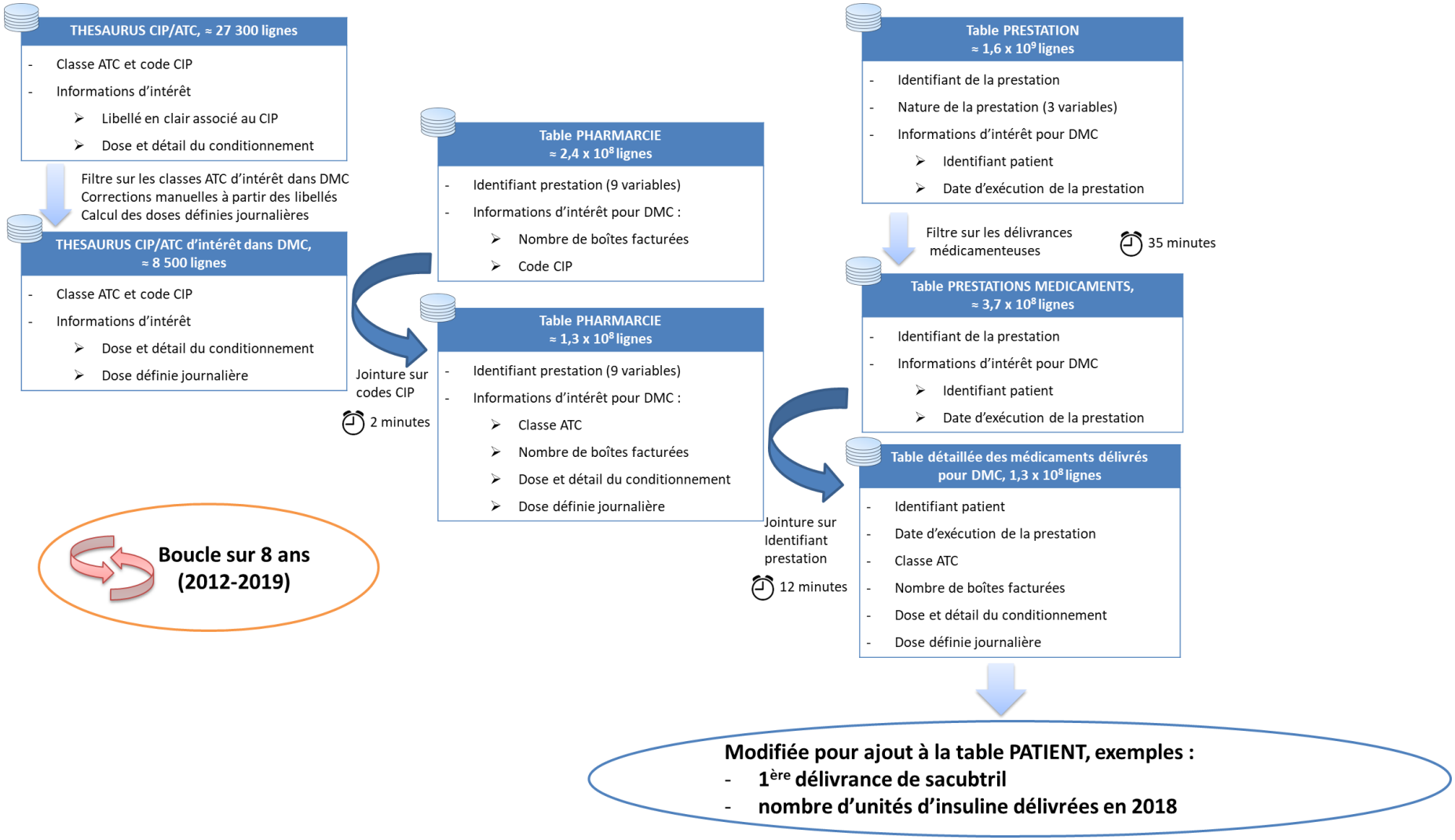


Figure 7. DMC - Schéma d'extraction des données de délivrance médicamenteuse ambulatoire, à partir des tables "prestation" et "pharmacie", et du thésaurus (correspondances entre les codes CIP et les classes ATC), pour une année donnée, répétée sur la période 2012-2019. Temps d'exécution pour 1 année : environ une heure

6. Considérations sur les statistiques

a. Justification du choix des facteurs de confusion

Nous avons retenu comme potentiels facteurs de confusion dans l'association entre l'ERG, d'une part, et l'insuffisance cardiaque, d'autre part, les variables suivantes[14] :

- L'âge et le sexe,
    - Justification : facteurs de risque attendus des deux événements

- Le suivi ophtalmologique récent (≥1 consultation en 2017-2018)
    - Justification : directement lié à la détection de l'ERG

- Les principaux facteurs de risque connus d'insuffisance cardiaque : coronaropathie, trouble du rythme cardiaque, valvulopathie, hypertension artérielle (cette dernière n'étant toutefois appréciée que par les traitements de l'hypertension par limite intrinsèque aux données), AVC et AOMI
    - Justification : facteur de risque d'IC, sans préjuger de leur association éventuelle avec l'ERG
    - AVC et AOMI : comme proxy de l'atteinte ischémique

- Les autres complications microangiopathiques du diabète : neuropathie diabétique, pied diabétique, amputation des membres inférieurs, insuffisance rénale chronique (terminale : dialyse/greffe rénale, ou non),
    - Justification : facteur de risque d'ERG

- Le cancer, comme indicateur de fragilité du patient

- Différents types de traitements :

    - du diabète : metformine, sulfamides hypoglycémiants/répaglinide, inhibiteurs de DPP4, analogues du GLP1, insuline

[14] L'ADA résumait les facteurs de risque communs à l'IC et au diabète les variables suivantes [72] : durée du diabète, équilibre glycémique, contrôle de la pression artérielle, hyperlipidémie, excès pondéral, microalbuminurie, insuffisance rénale, coronaropathie et maladie artérielle périphérique. Nous n'avons pas retenu l'hyperlipidémie et l'excès pondéral.

        - Justification : liés à l'équilibre glycémique et donc potentiellement à des complications du diabète, comme l'ERG et l'IC

    - anti-hypertenseurs : bêtabloquants, inhibiteurs calciques, IEC, ARA-II, diurétiques thiazidiques et diurétiques épargneurs potassiques
        - Justification : facteur de risque d'IC car proxy de la présence d'une HTA, bien que manquant fortement de spécificité

b. Modèle statistique principal

J'ai proposé un modèle de Cox basé sur l'hypothèse des risques proportionnels, certaines variables fixes et d'autre dépendantes du temps [105,106], avec les caractéristiques suivantes :

- Début du suivi au 31 décembre 2018 (« jour 1 » au 1er janvier 2019)
- Fin du suivi lors du 1er événement parmi : insuffisance cardiaque, décès du patient, perte de vue au cours de l'année 2019 ou 31 décembre 2019. En cas de survenue simultanée de l'IC et des autres événements (par exemple décès ou fin de suivi), l'IC était considérée en premier
- Exposition d'intérêt : l'ERG, traitée comme une variable dépendante du temps au cours de la période de suivi de 2019
- Evénement d'intérêt : 1ère détection d'une IC, avec le décès et la perte de vue comme risques compétitifs
- Modèles d'ajustement :
    - M_1 : ajustement sur l'âge, le sexe et le suivi ophtalmologique sur 2017-2018 (binaire)
    - M_2 : M_1 en ajoutant l'ajustement sur les facteurs de risque connus d'insuffisance cardiaque. Du fait de l'importance de ces critères (confirmée par les HR observés dans les résultats), ils étaient encodés comme variables dépendantes du temps. Par exemple, un patient connu pour une coronaropathie survenue au 1er juillet 2019 était associé au groupe indemne de coronaropathie au 1er semestre et comme présentant une coronaropathie au 2nd semestre
    - M_3 : M_2 en ajoutant l'ajustement sur les antécédents de complications du diabète et de cancer

- M_4 : M_3 en ajoutant l'ajustement sur les traitements antidiabétiques et antihypertenseurs

Toutes les analyses statistiques ont été réalisées à partir du logiciel statistique R version 4.0.3 [20], via l'interface RStudio. Il n'y a pas eu d'imputation sur les données manquantes. Une p-value < 0,05 a été considérée comme statistiquement significative. Il n'a pas été proposé de correction de l'inflation du risque alpha due à la multiplicité des tests.

Pour note, l'utilisation de variables dépendant du temps a nécessité la modification de la base de données, en transformant la table initiale (1 ligne = 1 patient) en une table où chaque ligne correspondait à une période de suivi d'un patient sans modification des variables dépendant du temps. La fonction de passage d'une table à l'autre (dite fonction de *stacking* pour « empilement ») a été écrite sous R, sans recourir à une fonction pré-existante.

7. Publication – en soumission au *European Heart Journal* à la remise du présent manuscrit

Au 24 octobre 2022, article proposé sous la forme PDF telle que soumise au *European Heart Journal*, incluse dans le présent manuscrit à partir de la page suivante.

Au 13 novembre 2022, l'article a été refusé par le journal. Les commentaires de l'éditeur et des *reviewers* sont partagés aux membres du jury de thèse.

Suite à ce rejet, la forme de l'article et le journal cible sont encore en discussion, et l'article revu n'était pas encore soumis à la date de la soutenance, le 12 décembre 2022.

European Heart Journal

Association of serious retinal events with incident heart failure in people living with diabetes: a nationwide population study

--Manuscript Draft--

Manuscript Number:	
Full Title:	Association of serious retinal events with incident heart failure in people living with diabetes: a nationwide population study
Article Type:	Clinical Research
Keywords:	French National Healthcare System; Diabetes mellitus; heart failure; Microvascular complications; Multistate model
Corresponding Author:	Matthieu Wargny, M.D. Nantes University Hospital Nantes, Pays de la Loire FRANCE
Order of Authors (with Contributor Roles):	Matthieu Wargny
	Jean-Baptiste Ducloyer
	Pacôme Constant dit Beaufils
	Christophe Leux
	Thomas Goronflot
	Philippe Tuppin
	Sandrine Fosse-Edorh
	Clara Piffaretti
	Jean-Noël Trochu
	Pierre-Antoine Gourraud
	Bertrand Cariou
	Samy Hadjadj
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Nantes University Hospital
Corresponding Author's Secondary Institution:	
First Author:	Matthieu Wargny
Order of Authors Secondary Information:	
Abstract:	Abstract Background and Aims In people with diabetes mellitus, increasing evidence highlights heart failure (HF) as an under-considered complication potentially explained by the microangiopathic burden. Our study aimed to assess the relationship between serious retinal events (SRE) and the incidence of HF in people living with diabetes, with and without established cause of HF. Methods We performed a retrospective cohort study based on 2012-2019 data from the French National Healthcare Data System. A validated algorithm identified 3 027 413 participants [mean age (SD) 67.1 years (13.1), 47.2% women] with diabetes and without identified HF on December 31, 2018. Outcomes and confounding factors were based on data from hospital stays (diagnoses, medical procedures), long-term chronic diseases (eligible for full reimbursement) and outpatient drug deliveries. SRE was based on laser and surgical procedures, and intravitreal drug-related delivery. HF was based on hospital stays, long-term chronic diseases and specific drug delivery.

	Results In 2019, 55 052 new HF cases were observed which corresponded to an incidence rate of 19.3/1000 person-years (95% confidence interval, 19.1-19.4). In a time-to-event analysis adjusted for age, sex, established cardiac diseases (coronary artery disease, heart valve disease, cardiac arrhythmia), stroke/transient ischaemic attack, peripheral artery disease, neuropathy, diabetic foot, lower limb amputation, chronic or end-stage kidney disease, cancer, and treatment (antidiabetic, antihypertensive), participants with SRE had increased risk of HF (hazard-ratio=1.11, 95%CI 1.07-1.14). This risk was higher in those without compared with those with established cardiac diseases (HR=1.42, 1.34-1.51 and 1.04, 1.01-1.08, respectively). Conclusions People with diabetes and SRE should be considered for in-depth HF screening.
Suggested Reviewers:	Kamlesh Khunti, MD, PhD Professor, University of Leicester College of Life Sciences kk22@leicester.ac.uk Expert in the field of diabetes and related complications
	Jasper Tromp, MD, PhD Assistant Professor, NUS SPH: National University Singapore Saw Swee Hock School of Public Health jasper_tromp@nus.edu.sg Expert in the field, author of ref. 20 "Age dependent associations of risk factors with heart failure: pooled population based cohort study. BMJ 2021;372:n461.", PMID 33758001
Opposed Reviewers:	
Additional Information:	
Question	Response
Total Word Count:	3491
Word Count Manuscript-only (excluding references):	4263
As Corresponding Author, I take full responsibility for all information declared in this notification.	Yes
As Corresponding Author, I agree to be the principal correspondent with the Editorial Office, review the edited manuscript and proof, and make decisions about releasing manuscript information to the media, federal agencies, etc.	Yes
All persons who have made substantial contributions to the manuscript (e.g. data acquisition, analysis, or writing / editing assistance), but who do not fulfill authorship criteria, are named with their specific contributions in the Acknowledgements Section of the manuscript.	Yes
All persons named in the Acknowledgements Section have provided the Corresponding Author with written permission to be named in the	Yes

manuscript.	
If an Acknowledgements Section is not included in the paper then no other persons have made substantial contributions to this manuscript.	Yes
Please enter the names of the authors who did anything else on the manuscript other than what we have listed:	N/A
This manuscript represents valid and substantiated work.	Yes
If asked, I will provide or fully cooperate in obtaining and providing the original data on which the manuscript is based so the editors or their designates can examine it.	Yes
All named authors are aware of this submission and fulfil the four criteria for authorship as outlined in the ICMJE criteria	Yes
TWITTER Please provide the Twitter handle of one or more authors (or their institutions) involved in this contribution. The social media team will use these when disseminating the work (if accepted).	institut_thorax
First Author Secondary Information:	

l'institut du thorax

Unité de recherche
Inserm UMR1087/CNRS UMR 6291
IRS – UN
8 quai Moncousu
BP 70721
44007 NANTES Cedex 1
FRANCE

T. +33 (0)2 28 08 01 10
F. +33 (0)2 28 08 01 30
U1087@univ-nantes.fr
www.umr1087.univ-nantes.fr

Pôle médical // Recherche clinique
Hôpital Nord-Laënnec
CHU de Nantes
44093 NANTES Cedex 1

T. +33 (0)2 40 16 57 24
F. +33 (0)2 40 16 57 25
www.chu-nantes.fr

DIRECTEUR
Pr Bertrand Cariou

DIRECTOIRE

Groupe Soins
Dr Delphine Drui
Pr Jean-Noël Trochu

Groupe Recherche Translationnelle
Pr Thierry Le Tourneau
Pr Vincent Probst

Groupe Recherche Fondamentale
Dr Richard Redon
Dr Vincent Sauzeau

Groupe Innovations Technologiques
Pr Yann Gouëffic
Pr Patrice Guérin

Groupe Enseignement
Pr François-Xavier Blanc
Pr Chantal Gauthier

Pr Samy Hadjadj, MD
samy.hadjadj@univ-nantes.fr

Nantes, France

October 24, 2022

To Professor Filippo Crea
Editor-in-Chief
European Heart Journal

Dear Editor,

Please find attached our manuscript entitled "**Association of serious retinal events with incident heart failure in people living with diabetes: a nationwide population study**" for possible publication as an original article in *European Heart Journal*.

Our results support the hypothesis that severe retinopathy is associated with heart failure (HF). Moreover, we found the greatest hazard ratio (HR) in the relationship between severe retinopathy and HF in those with no established cardiac disease, supporting a relationship between diabetic retinopathy and diabetic cardiomyopathy (HR=1.42, 95%CI 1.34 to 1.51).

In this report considering a population of ca. 3 million people living with diabetes in France, we used public health administrative data not only to examine an epidemiological question with evidence on the association between exposure and outcome, but also to address a pathophysiological hypothesis regarding the relationship between severe retinal events and HF. This approach is original and of peculiar relevance at the time when HF is considered as an important and under-considered complication associated with ageing in people with diabetes and when new results are being reported about the life-saving efficiency of SGLT2-inhibitors, particularly in HF with preserved ejection fraction.

The current manuscript was accepted under its present form by all of the authors. The results have not been published or communicated previously and the manuscript is not submitted for

publication elsewhere. We think that these results are highly original and will be of great interest for your readership of clinicians, researchers and patients.

Of course, we remain at your disposal for any changes your reviewers may deem necessary.

We look forward to hearing from you.

Sincerely yours,

Pr Samy Hadjadj

Abstract

Background and Aims

In people with diabetes mellitus, increasing evidence highlights heart failure (HF) as an under-considered complication potentially explained by the microangiopathic burden. Our study aimed to assess the relationship between serious retinal events (SRE) and the incidence of HF in people living with diabetes, with and without established cause of HF.

Methods

We performed a retrospective cohort study based on 2012-2019 data from the French National Healthcare Data System. A validated algorithm identified 3 027 413 participants [mean age (SD) 67.1 years (13.1), 47.2% women] with diabetes and without identified HF on December 31, 2018. Outcomes and confounding factors were based on data from hospital stays (diagnoses, medical procedures), long-term chronic diseases (eligible for full reimbursement) and outpatient drug deliveries. SRE was based on laser and surgical procedures, and intravitreal drug-related delivery. HF was based on hospital stays, long-term chronic diseases and specific drug delivery.

Results

In 2019, 55 052 new HF cases were observed which corresponded to an incidence rate of 19.3/1000 person-years (95% confidence interval, 19.1-19.4). In a time-to-event analysis adjusted for age, sex, established cardiac diseases (coronary artery disease, heart valve disease, cardiac arrhythmia), stroke/transient ischaemic attack, peripheral artery disease, neuropathy, diabetic foot, lower limb amputation, chronic or end-stage kidney disease, cancer, and treatment (antidiabetic, antihypertensive), participants with SRE had increased risk of HF (hazard-ratio=1.11, 95%CI 1.07-1.14). This risk was higher in those without compared with those with established cardiac diseases (HR=1.42, 1.34-1.51 and 1.04, 1.01-1.08, respectively).

Conclusions

People with diabetes and SRE should be considered for in-depth HF screening.

Structured graphical abstract

Key Question

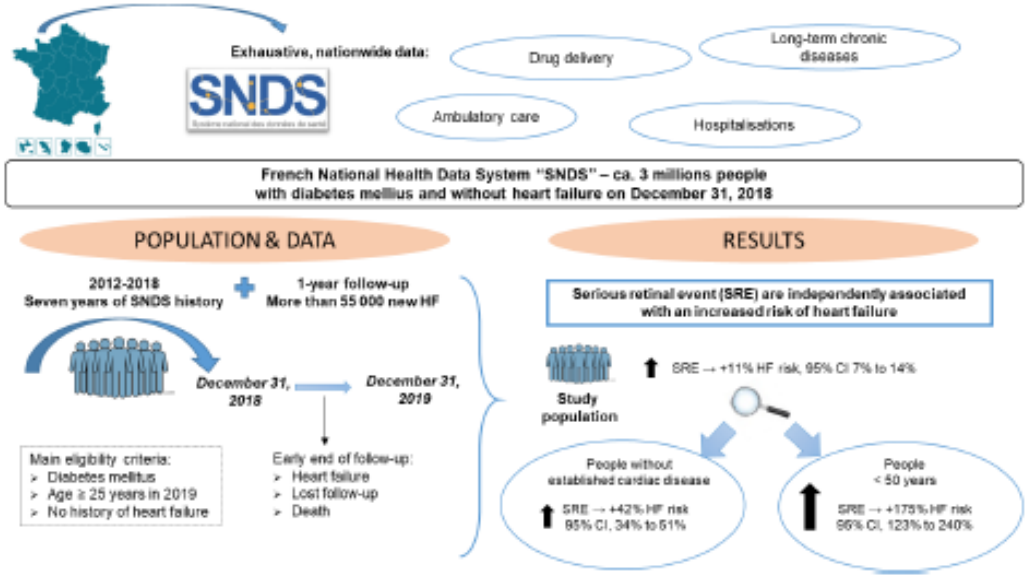
Is retinopathy associated with an increased risk of heart failure in patients living with diabetes, and what is the specific risk in the different strata of the population?

Key Finding

Serious retinal events are associated with an increased risk of heart failure (+11%), especially in people without establish cardiac disease (+42%) or in younger people (+175% if age < 50 years).

Take Home Message

People with diabetes and serious retinal event should be considered for in-depth heart failure screening.



Association of serious retinal events with incident heart failure in people living with diabetes: a nationwide population study

Matthieu Wargny,[1,2] (ORCID n°0000-0001-6027-9486) Jean-Baptiste Ducloyer,[3] (ORCID n°0000-0002-1306-1908) Pacôme Constant dit Beaufils,[1,4] (ORCID n°0000-0002-0466-2816) Christophe Leux,[5] Thomas Goronflot,[2] (ORCID n°0000-0002-1019-9821) Philippe Tuppin,[6] (ORCID n°0000-0001-5698-9215) Sandrine Fosse-Edorh,[7] (ORCID n°0000-0002-0105-5551) Clara Piffaretti,[7] (ORCID n°0000-0003-4999-0028) Jean-Noël Trochu,[1] (ORCID n°0000-0003-4742-281X) Pierre-Antoine Gourraud,[2] (ORCID n°0000-0003-1131-9554) Bertrand Cariou,[1] (ORCID n°0000-0002-1580-8040), Samy Hadjadj,[1] (ORCID n°0000-0001-7110-6994)

Institutions

1. Nantes Université, CHU Nantes, CNRS, Inserm, l'institut du thorax, F-44000 Nantes, France

2. Nantes Université, CHU Nantes, Pôle Hospitalo-Universitaire 11 : Santé Publique, Clinique des données, INSERM, CIC 1413, F-44000 Nantes, France

3. Nantes Université, CHU Nantes, service d'ophtalmologie, F-44000 Nantes, France

4. Nantes Université, CHU Nantes, service de neuroradiologie diagnostique et interventionnelle, F-44000 Nantes, France

5. Nantes Université, CHU Nantes, service d'information médicale, F-44000 Nantes, France

6. Department of Pathologies and Patients, Caisse Nationale d'Assurance Maladie, Paris, France

7. Santé publique France, The French National Public Health Agency, Saint Maurice, France

Corresponding author: Matthieu Wargny, matthieu.wargny@chu-nantes.fr, Nantes Université, CHU Nantes, Pôle Hospitalo-Universitaire 11 : Santé Publique, Clinique des données, INSERM, CIC 1413, 1, Place Alexis Ricordeau, F-44000 Nantes, France

Key words: French National Healthcare System – Diabetes mellitus – Heart Failure – Microvascular complications – Multistate model

Abbreviations: CAD: coronary artery disease; CCAM: French common classification for medical acts; CI: confidence interval; CKD: chronic kidney disease; CNAM: French National Health Insurance Fund; DMC study: Diabetes Multiple Complications study; ESKD: End-stage kidney disease; HF: heart failure; HFpEF/HFrEF: heart failure with preserved/reduced ejection fraction; HR: hazard ratio; ICD-10: International Classification of Diseases, 10th edition; IDR: Incidence density ratio; LTD: long-term chronic diseases; PAD: peripheral artery disease; PY: person-year; SNDS: French National Healthcare Data System; SRE: serious retinal event; T2D: type 2 diabetes

Introduction

Diabetes mellitus is a frequent condition expected to affect more than 10% of humans worldwide in 2030.(1) The operational definition of diabetes is based on the appearance of retinal microaneurysms and diabetes is therefore basically characterized by its risk of long-term microangiopathic complications.(2) In addition to microangiopathic complications, heart failure (HF) has recently been highlighted as an under-considered complication in diabetes.(3) A large meta-analysis of over 12 million people reported a two-fold greater risk of HF in individuals with type 2 diabetes (T2D) in both men and women compared with those without.(4)

The specific relationship between microvascular complications and HF has been previously suggested but needs to be firmly established.(5) From an epidemiological perspective, since microvascular diseases are often associated with macrovascular complications in people with diabetes, such an association will result in a non-causal relationship between microvascular burden and HF secondary to coronary artery disease (CAD). From a pathophysiological perspective, this could typically lead to HF with reduced ejection fraction (HFrEF) associated with systolic dysfunction. Conversely, diabetes microvascular disease could affect microvasculature of the heart with no obvious involvement of epicardial coronary arteries, leading to a stiffer, less compliant heart, which could translate clinically as HF with preserved ejection fraction (HFpEF), associated with diastolic dysfunction.(6)

We therefore hypothesised that the relationship between retinal microangiopathy and HF could be of value to establish HF risk in people with diabetes. To test this hypothesis, we examined the association between serious retinal events (SRE), indicative of specific

microvascular damage, and the first identification of HF in the population of people living with diabetes from the French National Health Data System (SNDS).

Methods

Study design and data access

The DMC (*Diabetes Multiple Complications*) study is a retrospective, medico-administrative, nationwide, near-exhaustive cohort based on individual data from the SNDS (99% of the French population with universal coverage).(7) Briefly, the prominent clinical data are (i) the delivery of all reimbursed drugs with ATC (Anatomical Therapeutic Chemical) classification; (ii) all hospital stays, public or private, with diagnosis codes (essentially ICD-10, International Classification of Diseases, 10th revision); (iii) medical and surgical procedures, both for hospital in- and outpatients, coded using the French CCAM (common classification for medical acts); and (iv) "long-term chronic diseases" (LTDs, eligible for full reimbursement) on a defined list by decree after expertise from the *Haute Autorité de Santé* (French National Authority for Health); (v) causes of deaths by the Epidemiological Centre for medical causes of death.

We submitted a request for access to these data to the National Health Insurance Fund (CNAM). The request was approved by the French Ethics & Scientific Committee for studies and evaluations in the field of health and authorized by the National Commission for Informatics and Freedom (Authorisation number 920426/decision DR-2020-380). The study was conducted in accordance with the principles of the Declaration of Helsinki. An overview of the DMC study design is proposed in *Figure 1*.

Population and follow-up

The request applied to all individuals regardless of age identified with diabetes in the general healthcare insurance scheme (≈86% of the French population) between January 1, 2012 and December 31, 2018. The definition of diabetes follows a pre-existing algorithm based on three sources of data: specific diabetes treatments, LTDs and diagnoses linked to hospital stays.(8) The algorithm was cross-validated with an external cohort.(9)

The following non-inclusion criteria were applied to the screened population: unavailable sex or date of birth; inconsistencies regarding sex, date of birth or date of death; age <25 years in 2019; ambiguous individual identifiers (unique identifier shared by several individuals, lack of information to discriminate between twins of the same sex) – strictly following the CNAM and *Santé publique France* (France Public Health agency) guidelines.[10–12] In order to protect privacy, age was only noted within one year of the exact age: for example, all women born in 1980 were considered to be exactly 32 on 1st January 2012.

The data were made available for full years from 2012 to 2019. When it was not known if participants died or not, they were considered as lost to follow-up after their last visible reimbursement or last end of hospital stay, whichever came last. The years 2012-2018 were used to characterise participant history and 2019 for a full year of follow-up.

Data collection and outcomes

Diabetes duration was not directly available and therefore was defined as the time spent since diabetes identification on January 1, 2019 and encoded with three modalities (≤2 years, 3 to 5 years, >5 years). Participant outcomes and history or events were defined by the first occurrence of a code of interest related to the pathology: HF, established cardiac diseases (CAD, heart valve disease and cardiac arrhythmia), stroke/TIA (transient ischaemic attack), peripheral artery disease (PAD), diabetes-related complications including SRE (laser procedure (SRE-1), retinal detachment/vitreal haemorrhage (SRE-2), and/or intravitreal injection associated with specific drug delivery (corticosteroids, anti-VEGF) (SRE-3)). The choice of codes was established with a systematic approach crossing, when available, recent CNAM guidelines,[8] former articles on the related topic,(12–16) and critical feedback of the algorithm by both a senior clinician in the field (JBD, PCDB, JNT, BC and SH) and a MD specialist in hospital coding (CL). All codes used (CCAM, ICD-10, ATC) are detailed in supplementary file 1. In particular, HF was defined as the first event (hospital stay or LTD) with HF diagnosis codes, or the delivery of specific treatment, eplerenone and/or the association of sacubitril and valsartan.

Statistical analyses

Participant history was extracted from data gathered during the 2012-2018 period and described by SRE status in participants still followed up on December 31, 2018 with a focus on the full year 2018 for treatment description. Categorical variables were described as number (%). Quantitative variables were described as mean (SD) or median (25^{th}-75^{th} percentile) according to the distribution valued by drawing the histogram. The incidence densities for 2019 events were calculated for participants without 2012-2018 history for the event of interest and expressed using person-years (PY) with the mean age of event and the

women/men incidence density ratio (IDR) with 95% confidence interval (95% CI) calculated using the semi-exact method.[17]

For the time-to-event analysis, a multi-state model representation of the chronology of the 2019 events was first proposed in order to evaluate the transition probabilities (notably between SRE only, HF only, mixed cases and death) before the exclusion of participants with known history of HF.[18] A Cox model based on proportional hazards with fixed and time-varying covariates was used with SRE as the main exposure of interest (time-varying), HF as the event, right-censoring in case of death or loss of follow-up, and the following adjustment models based on background knowledge: (M_1) age and sex; (M_2) M_1 with adjustment for time-varying cardiovascular diseases (CAD, heart valve disease, cardiac arrhythmia, CVD, stroke/TIA); (M_3) M_2 with adjustment for diabetes-related complications (neuropathy, foot ulcer, chronic kidney/end stage kidney disease (CKD/ESKD), lower limb amputation) and cancer; (M_4) M_3 with adjustment for 2018 treatment (antidiabetic, antihypertensive).

Sub-analyses were performed using only the model M_4 and stratified according to status for age, sex, CAD, heart valve disease, cardiac arrhythmia, and the following binary combinations: CAD and/or heart valve disease, and CAD and/or heart valve disease and/or cardiac arrhythmia.

We also performed three sensitivity analyses: (i) in order to challenge the SRE definition, the time-to-event analysis was performed separately on the three different components of SRE (laser procedure, hospital ICD-10 diagnoses, intravitreal injection); (ii) stratified analysis according to loop diuretics treatment in 2018, a non-specific approach to consider pre-existing HF not identifiable by other means; (iii) stratified analysis according to ESKD status in order to limit the lack of specificity of participants hospitalised with lower-limb oedema and potentially classified as HF during hospital stays.

All statistical analyses were performed using statistical software SAS Enterprise Guide version 7.15 or R version 4.1.1.[19]

Results

Study flowchart

The CNAM provided 4 184 482 potentially unique identifiers of participants screened as living with diabetes in 2012-2018 (*Figure 2*), of which 4 017 815 (96%) presented with available and consistent data for age, sex and date of death with a unique identifier allowing non-ambiguous association of data between the different information systems. The 40 338 participants (1%) of age <25 years in 2019 were excluded. Finally, 3 977 477 individuals were identified for the 2012-2018 period, with 3 295 758 individuals (83%) still followed up on December 31, 2018 (population analysed in *Table 1* and *Figure S2*). The main analyses were conducted on the 3 027 413 (91.8%) participants without known history of HF (population analysed in *Tables 2 & 3*, *Figure 2* and *Table S1*).

Population characteristics on December 31, 2018

The characteristics of the 3 295 758 participants alive and still followed up on January 1, 2019 are presented in *Table 1*. Briefly, 46.9% were women and the mean age was 67.9 years (13.2) and most were known to have diabetes for more than 5 years (73.5%). History of HF was observed in 268 345 people (8.2%, including 95.9% identified with diagnosis codes) and was associated with CAD (53.6%), heart valve disease (23.4%) and cardiac arrhythmia (52.0%), but 20.4% of participants with HF had none of these three conditions. Participants with HF had a history of SRE for 9.3% versus 4.9% in participants without HF. The population without known history of HF according to SRE status on December 31, 2018 is presented in *Table S1*.

As a control for consistency, we observed that in the population with history of atherosclerosis-related cardiovascular disease (CAD, stroke/TIA and/or PAD), 73.7% of the participants were treated with statin therapy in 2018 and 72.2% with antiplatelet agents, vs. 23.9% and 41% in the population without history of cardiovascular disease. In the population with cardiac arrhythmia, 60.3% were treated with anticoagulant therapy in 2018, vs. 4.1% in the population without.

The course of events during 2019 for SRE and HF is presented in *Figure S2*. The unadjusted yearly transition probability for HF for participants without SRE was 1.8% vs. 3.4% in participants with SRE. The yearly transition probability for death rose from 2.1% for people with neither SRE nor HF to 3.4% for people with SRE, 17.1% in people with HF, and 17.2% in people with both SRE and HF.

Clinical outcomes during 2019

The incidence of the events during the year 2019 for the participants without identified HF on December 31, 2018 is described in *Table 2*. HF developed in 55 052 people (1.8%). The incidence density was 19.3/1000 PY (95% CI, 19.1 to 19.4). During the same period, the incidence densities were 16.1, 6.1 and 16.7/1 000 PY for CAD, heart valve disease and cardiac arrhythmia, respectively, in the study population without history for these events. Death occurred in 72 879 participants (2.4%), with an incidence density of 25.3/1 000 PY.

In the time-to-event analyses presented in *Table 3*, SRE was significantly associated with HF with HR=1.32 (95% CI, 1.28 to 1.36) in a model adjusted for age, sex and ophthalmologist acts, decreased to 1.16 (1.12 to 1.20) after adjustment for cardiovascular diseases (CAD, cardiac arrhythmia, heart valve disease, stroke/TIA and PAD), 1.12 (1.09 to 1.16) after adjustment for other diabetes complications and cancer, and 1.11 (1.07 to 1.14)

after adjustment for treatment (antidiabetic and anti-hypertensive). In the fully adjusted model, the main risk factors for HF were the established cardiac diseases: CAD (HR=2.61, 2.56 to 2.66), heart valve disease (2.63, 2.57 to 2.70) and cardiac arrhythmia (3.87, 3.80 to 3.95).

In the stratified analyses considering the SRE risk for HF in different subgroups presented in *Figure 2*, all of the HRs for SRE were significant except for participants aged ≥70 years, with coronary heart disease, ESKD, or already treated with loop diuretics. In participants without vs. with established cardiac disease, the HR for SRE was 1.42 (1.34 to 1.51) vs. 1.04 (1.01 to 1.08). When the different age categories were considered, the HR associated with SRE increased with decreasing age from 1.02 (0.99 to 1.06) to 2.75 (2.23 to 3.40) in participants aged ≥70 and <50 years, respectively, whereas they were similar in both sexes (1.10 vs. 1.12 for men and women, respectively).

Discussion

Principal findings

Using a nationwide approach on a population of approximately 3 300 000 people living with diabetes, we showed that SRE was associated with a 11% (95% CI, 7% to 14%) increased risk of HF after multiple adjustment considering age, sex, comorbidities, and medications. The deleterious effect of SRE was more obvious in those without established cardiac diseases. This was observed in the younger participants (<50 years) where SRE was associated with a nearly 3-fold higher risk, a differential nature of age-related risk factors already pointed out.[20] This was also observed in those participants of all ages and without established cardiac diseases (possibly corresponding to HFpEF) with an increased risk as high as 42% (34% to 51%). It should be noted that the risk associated with SRE was similar for

both sexes. Given these results, this study underscores the deleterious effect of SRE for incident HF and encourages clinicians to adapt HF screening in this population.

Comparison with other studies

Our approach was based on a medico-administrative database but there were already some relevant data linking microvascular burden to the risk of HF in people with T2D such as in the EMPAREG Outcome trial.[21] Indeed, microvascular complications were associated with an increased risk of hospitalisation for HF by 63%. The question whether the risk of hospitalisation for HF is secondary to its well-established association with kidney disease could not be solved in this study since data from the subgroup of participants with retinopathy specifically (22% of the study population) were not examined. Our results are also consistent with the TOPCAT trial by assessing the role of microvascular burden among diabetes participants where the 739 without microvascular burden were hospitalised less for HF than the 352 with microvascular burden.[22] Moreover, the EMPAREG study also reported that the higher the number of microvascular complications, the higher the risk, suggesting that the increased risk associated with microvascular burden was not all about CKD. In the present study, we observed only a slight drop (HR=1.16 to 1.12) in the risk of HF associated with SRE after adjustment for CKD/ESKD, but the association was no longer significant when it was limited to the population with ESKD. This suggests that renal involvement is part of the relationship with HF but not limited to it. Furthermore, and more closely related to our specific question, recent studies have provided different insights. A recent analysis of approximately 30 000 people with diabetes included in the UK Biobank study showed that retinopathy was associated with an increased risk of HF, both in type 1 diabetes (HR=2.69, 1.75 to 4.14) and T2D (HR=1.24, CI 1.13 to 1.36),[5] whereas in a meta-analysis including

several models performed in nearly 1 million people living with diabetes, diabetic retinopathy was not retained as a predictor for HF.[23] Indeed, the relationship between severe diabetic retinopathy and HF could be obscured by other microvascular complications such as kidney disease, even in large-scale studies with thousands of participants. However, the larger cohort considered here with over 50 000 incident HF enabled us to discount the statistical power issue.

Possible pathophysiological interpretations

Our primary research goal was not only to study the relationship between retinal complications and HF but also to bring statistical support to a pathophysiological hypothesis, notably that SRE would be particularly relevant in the context of HFpEF. Unfortunately, specific information about the form of HF was not available in the database and HFrEF and HFpEF could not be disentangled. We therefore considered the situation with HF occurring without established cardiac disease (CAD, heart valve disease and/or cardiac arrhythmia) as a proxy of HFpEF. SRE was consistently associated with HF regardless of the subgroup, but this risk was much higher in people without known substantial risk of HF. For example, the risk of HF associated with SRE was much greater in those without compared with those with established cardiac disease, the HR dropping from 1.42 (1.34 to 1.51) to 1.04 (1.01 to 1.08). This is an indirect but consistent argument to point out that microvascular disease in the retina could be associated with microvascular disease in the myocardium leading to stiffened heart structure and paving the way to HFpEF. This observation is also in accordance with data from the ASIAN-HF registry where diabetic retinopathy was more often associated with HFpEF than HFrEF.[24]

The mechanistic factors linking severe diabetic retinopathy and HF with a speculated peculiar effect on HFpEF can be questioned. Recently, some studies have reported that coronary microvascular dysfunction was strongly associated with HFpEF, putting light on endothelial dysfunction.[6,25] In their review, Sinha et al. emphasized the role of long diabetes exposure (also a well-established risk factor for diabetic retinopathy[26]) for decreased coronary flow reserve and microvascular dysfunction. Since endothelial dysfunction is an established risk factor for diabetic retinopathy,[27] data in the literature support that endothelial dysfunction could be the missing link between severe diabetic retinopathy and HFpEF, although this remains to be firmly established by dedicated studies.

Strengths and weaknesses of the study

Our work has some limitations, mainly stemming from the pure medico-administrative structure of our data. Firstly, no clinical or biological determinations were available, limiting the ability to adjust to basic clinical covariates such as BMI or blood pressure control, or to check on comorbidities with relevant biomarkers. This limitation has already been mentioned when considering renal complications in the relationship between SRE and HF. For example, only diagnoses of CKD and/or ESKD could be included in multivariable models but not the estimated glomerular filtration rate or albuminuria. Thirdly, the precision of ICD-10 codes used to define outcomes must also be questioned. However, since these codes are used to define reimbursements allocated to hospitals, they are regularly monitored and provide good overall quality to the current data. Regarding the exposure of interest, it must be acknowledged that we could only rely on SRE, while less severe retinopathy leading to mild or moderate diabetic retinopathy were not duly captured, confining the study to more severe phenotypes. Moreover, the specificity of the approach is not warranted since some procedures such as intravitreal injections could also reflect age-related macular

degeneration and not diabetic macular oedema. However, we think that the consistent results found when looking at the three different components separately (laser procedure, retinal detachment/vitreal haemorrhage, intravitreal injections) are a reassuring sign concerning this potential bias. Finally, regarding the event of interest, HFpEF was supported only by a proxy (comparison between HF risk in persons with or without established cardiac disease) but was not directly examined considering imaging techniques and clinician expertise in a dedicated cohort. Ultimately, we were not able to distinguish type 1, type 2, and other types of diabetes. However, regarding the SRE/HF association, it can be argued that the risk is expected to be relatively equivalent according to the type of diabetes provided risk factors are taken into account, such as metabolic control.

However, the ability to examine a large nationwide population was seen as a great asset which made it possible to draw firm conclusions with narrow confidence intervals and no issues on generalisability, at least for western populations. Regarding SRE and HF identifications, the fact that the related medical acts are costly and are not paid if not relayed to the SNDS, gives us reasons to have confidence in the exhaustiveness of their reporting. Moreover, the homogenous system of free healthcare makes us believe that even if no socio-economic proxies were used in statistical models, the gratuitous nature of care minimizes the impact of related selection and confounding bias in the association between SRE and HF.

Conclusion and implications

Our results have important consequences in terms of possible clustering of HF in diabetes and for screening. Such positive findings support the search for myocardial microvascular disease in people with severe diabetic retinopathy, notably in the population under 50 years of age. Regarding their risk of HF, this could lead to a more systematic

examination (natriuretic peptides, transthoracic echography) in order to identify those who could benefit the most from interventions such as the use of SGLT2-inhibitors, shown to decrease HF incidence, including HFpEF.[28,29]

Acknowledgments

We wish to thank the rest of the 2019-2022 REDSIAM team for their time and the discussion which helped us plan and impliment the DMC study, especially Laurence Mandereau-Bruno (Santé publique France) and Audrey Cougnard-Grégoire (Bordeaux population health research centre). We also want to express our gratitude to the CNAM team and especially Marjorie Boussac who patiently accompanied us in the data access process. Additionally, we thank Pr Yann Gouëffic (Groupe Hospitalier Paris Saint-Joseph) for his clinical expertise regarding the review and the consistency of the codes in the field of vascular surgery; Sandrine Coudol (Nantes University Hospital) for her help regarding SNDS data management; and Pr Marc Cuggia (Rennes University Hospital) for his careful monitoring and feedback on the project. Finally, we thank Peter Tucker for his review of the English language of the manuscript.

Funding

This study did not receive specific funding

Conflict of interest

No potential conflicts of interest relevant to this article were reported. BC reports grants and personal fees from Amgen, personal fees from AstraZeneca, personal fees from Akcea, personal fees from Genfit, personal fees from Gilead, personal fees from Eli Lilly,

personal fees from Novo Nordisk, personal fees from Merck (MSD), grants and personal fees from Sanofi, grants and personal fees from Regeneron. SH reports personal fees and non-financial support from AstraZeneca, grants and personal fees from Bayer, personal fees from Boehringer Ingelheim, grants from Dinno Santé, personal fees from Eli Lilly, non-financial support from LVL, personal fees and non-financial support from MSD, personal fees from Novartis, grants from Pierre Fabre Santé, personal fees and non-financial support from Sanofi, personal fees and non-financial support from Servier, personal fees from Valbiotis. The other authors have no conflicts of interest to declare.

Data availability statement

Because of a substantial risk of patient identification, no sharing of individual SNDS data is allowed by French regulatory authorities

References

1. IDF Atlas 9th edition and other resources [Internet]. [cited 2021 Apr 30]. Available from: https://diabetesatlas.org/en/resources/

2. Rawshani A, Rawshani A, Franzén S, Eliasson B, Svensson AM, Miftaraj M, et al. Mortality and Cardiovascular Disease in Type 1 and Type 2 Diabetes. *N Engl J Med* 2017;**376**:1407–18.

3. Pop-Busui R, Januzzi JL, Bruemmer D, Butalia S, Green JB, Horton WB, et al. Heart Failure: An Underappreciated Complication of Diabetes. A Consensus Report of the American Diabetes Association. *Diabetes Care* 2022;**45**:1670–90.

4. Ohkuma T, Komorita Y, Peters SAE, Woodward M. Diabetes as a risk factor for heart failure in women and men: a systematic review and meta-analysis of 47 cohorts including 12 million individuals. *Diabetologia* 2019;**62**:1550–60.

5. Li FR, Hukportie DN, Yang J, Yang HH, Chen GC, Wu XB. Microvascular Burden and Incident Heart Failure Among Middle-Aged and Older Adults With Type 1 or Type 2 Diabetes. *Diabetes Care* 2022;dc220177.

6. Shah SJ, Lam CSP, Svedlund S, Saraste A, Hage C, Tan RS, et al. Prevalence and correlates of coronary microvascular dysfunction in heart failure with preserved ejection fraction: PROMIS-HFpEF. *Eur Heart J*. 2018;**39**:3439–50.

7. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. *Rev Epidemiol Santé Publiqu*e 2017;**65 Suppl 4**:S149–67.

8. Caisse Nationale de l'Assurance Maladie (CNAM). Méthodologie médicale de la cartographie des pathologies et des dépenses, version G8 (années 2015 à 2019, Tous Régimes) [Internet]. 2021 [cited 2022 Jan 20]. Available from: https://assurance-maladie.ameli.fr/sites/default/files/2021_methode-reperage-pathologies_cartographie_1.pdf

9. Fuentes S, Cosson E, Mandereau-Bruno L, Fagot-Campagna A, Bernillon P, Goldberg M, et al. Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort. *Int J Public Health* 2019;**64**:441–50.

10. Documents Sante publique France | Documentation du SNDS [Internet]. [cited 2022 Aug 9]. Available from: https://documentation-snds.health-data-hub.fr/formation_snds/sante_publique_france.html

11. Caisse Nationale de l'Assurance Maladie (CNAM). SNDS Fiche pratique : Thème Bénéficiaires / Notions de bénéficiaires [Internet]. 2020 [cited 2022 Aug 9]. Available from: https://documentation-snds.health-data-hub.fr/files/Cnam/2019-06_CNAM-INDS_SNDS_Fiches_Thematiques_BENEF_MAJ-2020-09_MPL-2.0.pdf

12. Equipe SNDS de la Direction Appui, Traitements et Analyses des données. SNDS - Ce qu'il faut savoir [Internet]. 2021 [cited 2022 Aug 9]. Available from: https://documentation-snds.health-data-hub.fr/files/Sante_publique_France/2021-10-SpF-SNDS-ce-quil-faut-savoir-v3-MPL-2.0.pdf

13. Fosse Edorh S, Mandereau Bruno L, Hartemann Heurtier A. Les hospitalisations pour complications podologiques chez les personnes diabétiques traitées pharmacologiquement en France en 2013. *Bull Epidémiol Hebd* 2015;**34–35**:638–44.

14. Giral P, Neumann A, Weill A, Coste J. Cardiovascular effect of discontinuing statins for primary prevention at the age of 75 years: a nationwide population-based cohort study in France. *Eur Heart J* 2019;**40**:3516–25.

15. Amadou C, Denis P, Cosker K, Fagot-Campagna A. Less amputations for diabetic foot ulcer from 2008 to 2014, hospital management improved but substantial progress is still possible: A French nationwide study. Santanelli F, editor. *PLoS ONE* 2020;**15**:e0242524.

16. Grave C, Tribouilloy C, Juillière Y, Tuppin P, Weill A, Gabet A, et al. Hospitalisations for valvular heart disease in France: patients characteristics and trends 2006-2016. *Bull Epidémiol Hebd* 2020;**4**:70–9.

17. Rothman K, Greenland S. Modern epidemiology. 3rd ed. Philadelphia. Lippincott Williams & Wilkins; 2008.

18. Carstensen B, Plummer M. Using Lexis Objects for Multi-State Models in R. *Journal of Statistical Software* 2011;**38**(1):1–18.

19. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/

20. Tromp J, Paniagua SMA, Lau ES, Allen NB, Blaha MJ, Gansevoort RT, et al. Age dependent associations of risk factors with heart failure: pooled population based cohort study. *BMJ* 2021;**372**:n461.

21. Verma S, Wanner C, Zwiener I, Ofstad AP, George JT, Fitchett D, et al. Influence of Microvascular Disease on Cardiovascular Events in Type 2 Diabetes. *J Am Coll Cardiol* 2019;**73**:2780–2.

22. Sandesara PB, O'Neal WT, Kelli HM, Samman-Tahhan A, Hammadah M, Quyyumi AA, et al. The Prognostic Significance of Diabetes and Microvascular Complications in Patients With Heart Failure With Preserved Ejection Fraction. *Diabetes Care* 2018;**41**:150–5.

23. Razaghizad A, Oulousian E, Randhawa VK, Ferreira JP, Brophy JM, Greene SJ, et al. Clinical Prediction Models for Heart Failure Hospitalization in Type 2 Diabetes: A Systematic Review and Meta-Analysis. *J Am Heart Assoc* 2022;**11**:e024833.

24. Tromp J, Lim SL, Tay WT, Teng THK, Chandramouli C, Ouwerkerk W, et al. Microvascular Disease in Patients With Diabetes With Heart Failure and Reduced Ejection Versus Preserved Ejection Fraction. *Diabetes Care* 2019 Sep;42(9):1792–9.

25. Sinha A, Rahman H, Webb A, Shah AM, Perera D. Untangling the pathophysiologic link between coronary microvascular dysfunction and heart failure with preserved ejection fraction. *Eur Heart J*. 2021 Nov 14;42(43):4431–41.

26. Wong TY, Cheung CMG, Larsen M, Sharma S, Simó R. Diabetic retinopathy. *Nat Rev Dis Primers* 2016 Mar 17;2:16012.

27. Klein BEK, Knudtson MD, Tsai MY, Klein R. The relation of markers of inflammation and endothelial dysfunction to the prevalence and progression of diabetic retinopathy: Wisconsin epidemiologic study of diabetic retinopathy. *Arch Ophthalmol* 2009 Sep;127(9):1175–82.

28. Li S, Vandvik PO, Lytvyn L, Guyatt GH, Palmer SC, Rodriguez-Gutierrez R, et al. SGLT-2 inhibitors or GLP-1 receptor agonists for adults with type 2 diabetes: a clinical practice guideline. *BMJ* 2021 May 11;373:n1091.

29. Vaduganathan M, Docherty KF, Claggett BL, Jhund PS, de Boer RA, Hernandez AF, et al. SGLT-2 inhibitors in patients with heart failure: a comprehensive meta-analysis of five randomised controlled trials. *Lancet* 2022 Sep 3;400(10354):757–67.

Figure legends

Figure 1. Study design

ATC: Anatomic, therapeutic and chemical classification of drugs; CCAM: French classification for medical and surgical procedures; DMC: Diabetes Multiple Complications; ICD-10: International Classification of Diseases, 10th edition; IS: Information System; SNDS: French National Healthcare Data System.

Figure 2. Association of the different SRE subtypes (SRE-1/-2/-3) with HF, and association of SRE with HF in the different subsets defined using age, sex, established HF risk factors, end stage kidney disease and history of loop diuretics (*N* = 3 027 413)

Cox model based on proportional hazards hypothesis, for the analysis of HF during 2019 in patients still alive and followed up on December 31, 2018 without known HF. All of the models proposed include the full model (M_4) with the following covariates: age, sex, SRE status (serious retinal event, time-dependent), cardiovascular time-dependent covariates (CAD, cardiac arrhythmia and heart valve disease, stroke/TIA, PAD), history of neuropathy, diabetic foot, lower limb amputation, nephropathy (none/chronic/end-stage kidney disease) and cancer, and known treatments in 2018 (metformin, sulfonylurea and/or repaglinide, GLP1-receptor agonists, insulin, beta blockers, calcium-channel blockers, angiotensin-converting enzyme inhibitors, angiotensin-renin blockers, thiazide and potassium-sparing diuretics).

CAD: coronary artery disease; HF: heart failure; PAD: peripheral artery disease; PY: person-years; SRE: Serious retinal event; SRE: serious retinal event SRE-1: retinal laser; SRE-2: retinal detachment or vitreal haemorrhage; SRE-3: intravitreal injection with specific treatment; TIA: Transient Ischaemic attack.

Appendices

Figure S1. Study flow-chart

DMC: Diabetes Multiple Complications; ID: identifier; IS: information system.

Figure S2. Multi-state model for Serious Retinal Event & Heart Failure in 2019 (*N* = 3 295 758)

"Healthy" corresponds to people with diabetes free from HF and SRE. Mixed is defined as history of both SRE and HF, independently of whichever came first.

HF: heart failure; PY: person-year; SRE: serious retinal event

Table S1. Clinical characteristics associated with the DMC population without HF history and followed-up in 2019 (*N* = 3 027 413), according to SRE status on December 31, 2018

Supplementary File 1. List of codes used and their explanation

	All (*N* = 3 295 758)	No HF history (*N* = 3 027 413)	HF history (*N* = 268 345)
Women	1 544 175 (46.9%)	1 429 015 (47.2%)	115 160 (42.9%)
Age on January 1, 2019 (years)	67.9 (13.2)	67.1 (13.1)	76.2 (11.4)
Age categories			
<50 years	298 494 (9.1%)	293 767 (9.7%)	4 727 (1.8%)
50-59 years	504 600 (15.3%)	487 412 (16.1%)	17 188 (6.4%)
60-69 years	932 712 (28.3%)	881 847 (29.1%)	50 865 (19.0%)
≥ 70 years	1 559 952 (47.3%)	1 364 387 (45.1%)	195 565 (72.9%)
Diabetes duration			
≤2 years	325 956 (9.9%)	311 345 (10.3%)	14 611 (5.4%)
3 to 5 years	546 580 (16.6%)	516 790 (17.1%)	29 790 (11.1%)
> 5 years	2 423 222 (73.5%)	2 199 278 (72.6%)	223 944 (83.5%)
Cardiovascular history			
CAD	568 696 (17.3%)	424 788 (14.0%)	143 908 (53.6%)
Heart valve disease	134 119 (4.1%)	71 391 (2.4%)	62 728 (23.4%)
Cardiac arrhythmia	374 818 (11.4%)	235 355 (7.8%)	139 463 (52.0%)
≥ one of the previous 3	827 977 (25.1%)	614 294 (20.3%)	213 683 (79.6%)
Stroke/TIA	193 874 (5.89%)	155 730 (5.1%)	38 144 (14.2%)
PAD	275 618 (8.4%)	214 193 (7.1%)	61 425 (22.9%)
Other diabetes complications			
Neuropathy	166 875 (5.1%)	132 184 (4.4%)	34 691 (12.9%)
Foot ulcer	64 215 (2.0%)	42 063 (1.4%)	22 152 (8.3%)
Lower limb amputation	26 115 (0.8%)	17 408 (0.6%)	8 707 (3.2%)
Chronic kidney disease	308 016 (9.3%)	212 982 (7.0%)	95 034 (35.4%)
End-stage kidney disease	32 658 (1.0%)	19 358 (0.6%)	13 300 (5.0%)
Cancer	517 207 (15.7%)	450 267 (14.9%)	66 940 (24.9%)
Eye-related outcomes*			
SRE-1	114 006 (3.5%)	98 282 (3.3%)	15 724 (5.9%)
SRE-2	10 841 (0.3%)	8 756 (0.3%)	2 085 (0.8%)
SRE-3	86 737 (2.6%)	72 895 (2.4%)	13 842 (5.2%)
Any SRE	172 179 (5.2%)	147 271 (4.9%)	24 908 (9.3%)
≥1 ophthalmologic act in 2017-2018	2 245 398 (68.1%)	2 078 784 (68.7%)	166 614 (62.1%)
Diabetes treatment in 2018			
Metformin	2 099 186 (63.7%)	1 977 885 (65.3%)	121 301 (45.2%)
Sulfonylurea and/or repaglinide	1 056 339 (32.1%)	977 980 (32.3%)	78 359 (29.2%)
DPP4-inhibitors	852 212 (25.9%)	793 624 (26.2%)	58 588 (21.8%)
GLP-1 receptor agonist	238 359 (7.2%)	220 844 (7.3%)	17 515 (6.5%)
Insulin therapy	697 720 (21.2%)	601 550 (19.9%)	96 170 (35.8%)
If insulin therapy: number of IU/day	23.6 (24.1)	23.8 (24.1)	22.4 (23.7)
Diet only	502 430 (15.2%)	454 487 (15.0%)	47 943 (17.9%)
Anti-hypertensive treatment in 2018			
Any anti-hypertensive treatment	2 403 206 (72.9%)	2 148 645 (71.0%)	254 561 (94.9%)
Beta blockers	1 066 202 (32.4%)	888 876 (29.4%)	177 326 (66.1%)
Calcium-channel blockers	977 512 (29.7%)	878 460 (29.0%)	99 052 (36.9%)
ACE-inhibitors	897 133 (27.2%)	787 366 (26.0%)	109 767 (40.9%)
ARBs	1 032 276 (31.3%)	949 181 (31.4%)	83 095 (31.0%)
Loop diuretics	366 932 (11.1%)	208 165 (6.9%)	158 767 (59.2%)
Thiazide diuretics	671 067 (20.4%)	635 412 (21.0%)	35 655 (13.3%)
Potassium-sparing diuretic	176 956 (5.4%)	127 629 (4.2%)	49 327 (18.4%)
Other treatments of interest in 2018			
Antiplatelet agents	1 201 104 (36.4%)	1 061 797 (35.1%)	139 307 (51.9%)
Anticoagulants	345 496 (10.5%)	232 950 (7.7%)	112 546 (41.9%)
Statins	1 611 879 (48.9%)	1 452 572 (48.0%)	159 307 (59.4%)
Ezetimibe	191 347 (5.8%)	172 789 (5.7%)	18 558 (6.9%)

Table 1. Clinical characteristics associated with the DMC population followed-up in 2019 (*N* = 3 295 758), according to HF and SRE status on December 31, 2018

Categorical variables are described using n (%). Quantitative variables are described using mean (standard deviation). HF: heart failure. CAD: coronary artery disease; PAD: peripheral artery disease; SRE: serious retinal event; TIA: transient ischaemic attack.

*SRE-1: retinal laser photocoagulation; SRE-2: retinal detachment or vitreal haemorrhage; SRE-3: intravitreal injection of anti VEGF or corticosteroids.

	All (*N* = 3 027 413)			No SRE (*N* = 2 880 125)		SRE (*N* = 147 288)		
2019 outcomes	**2019 Events/total PY**	**2019 Events /1 000 PY**	**Mean age if event**	**2019 Events/1 000 PY**	**Mean age if event**	**2019 Events /1 000 PY**	**Mean age if event**	**IDR***
Heart Failure								
Any	55 052/2 847 371 PY	19.3/1 000 PY	77.4 y.	18.6/1 000 PY	77.4 y.	34.7/1 000 PY	77.5 y.	1.87 (1.84-1.90)
HF with known CAD	26 216/2 859 322 PY	9.2/1 000 PY	76.5 y.	8.7/1 000 PY	76.5 y.	17.9/1 000 PY	76.6 y.	2.06 (2.01-2.10)
HF with known heart valve disease	10 341/2 866 098 PY	3.6/1 000 PY	79.3 y.	3.5/1 000 PY	79.3 y.	6.6/1 000 PY	79.5 y.	1.91 (1.85-1.98)
HF with known cardiac arrhythmia	25 615/2 859 767 PY	9.0/1 000 PY	79.6 y.	8.7/1 000 PY	79.5 y.	14.8/1 000 PY	80.6 y.	1.71 (1.67-1.75)
HF with none of the previous 3	13 470/2 864 926 PY	4.7/1 000 PY	76.1 y.	4.5/1 000 PY	76.2 y.	8.1/1 000 PY	75.6 y.	1.78 (1.73-1.84)
Cardiovascular history								
CAD	45 983/2 849 189 PY	16.1/1 000 PY	71.4 y.	15.7/1 000 PY	71.2 y.	24.2/1 000 PY	73.0 y.	1.54 (1.51-1.57)
Heart valve disease	17 557/2 862 571 PY	6.1/1 000 PY	77.0 y.	5.9/1 000 PY	76.9 y.	9.9/1 000 PY	77.8 y.	1.66 (1.62-1.71)
Cardiac arrhythmia	47 507/2 849 591 PY	16.7/1 000 PY	76.3 y.	16.3/1 000 PY	76.2 y.	23.3/1 000 PY	78.2 y.	1.43 (1.40-1.45)
Stroke/TIA	24 128/2 860 678 PY	8.43/1 000 PY	75.4 y.	8.16/1 000 PY	75.3 y.	13.84/1 000 PY	76.1 y.	1.70 (1.66-1.74)
PAD	25 272/2 858 718 PY	8.84/1 000 PY	72.3 y.	8.41/1 000 PY	72.4 y.	17.27/1 000 PY	71.7 y.	2.05 (2.01-2.10)
Other diabetes complications								
Neuropathy	20 253/2 860 776 PY	7.08/1 000 PY	68.8 y.	6.61/1 000 PY	68.8 y.	16.34/1 000 PY	68.6 y.	2.47 (2.42-2.53)
Diabetic foot	11 620/2 865 368 PY	4.06/1 000 PY	73.9 y.	3.74/1 000 PY	74.3 y.	10.28/1 000 PY	71.6 y.	2.75 (2.68-2.83)
Chronic kidney disease	4 240/2 868 564 PY	1.48/1 000 PY	71.4 y.	1.28/1 000 PY	72.0 y.	5.37/1 000 PY	68.7 y.	4.20 (4.04-4.38)
End stage kidney disease	42 685/2 851 005 PY	15.0/1 000 PY	74.5 y.	14.3/1 000 PY	74.6 y.	28.6/1 000 PY	73.7 y.	2.01 (1.97-2.04)
Lower limb amputation	3 920/2 868 712 PY	1.37/1 000 PY	70.2 y.	1.17/1 000 PY	71.1 y.	5.12/1 000 PY	66.0 y.	4.36 (4.18-4.54)
Cancer	55 209/2 845 911 PY	19.4/1 000 PY	72.8 y.	19.3/1 000 PY	72.6 y.	21.9/1 000 PY	75.8 y.	1.14 (1.11-1.16)
Death	72 879/2 880 554 PY	25.3/1 000 PY	79.2 y.	24.8/1 000 PY	79.1 y.	35.8/1 000 PY	80.5 y.	1.45 (1.43-1.47)

Table 2. Description of the DMC population without HF history and follow-up in 2019: incidence of the events of interest during 2019 (*N* = 3 027 413)

All patients with a history related to an item are excluded from the item's incidence analyses. CAD: coronary artery disease; HF: heart failure; PAD: peripheral artery disease; PY: person-years; SRE: Serious Retinal Event; TIA: Transient Ischaemic Attack.

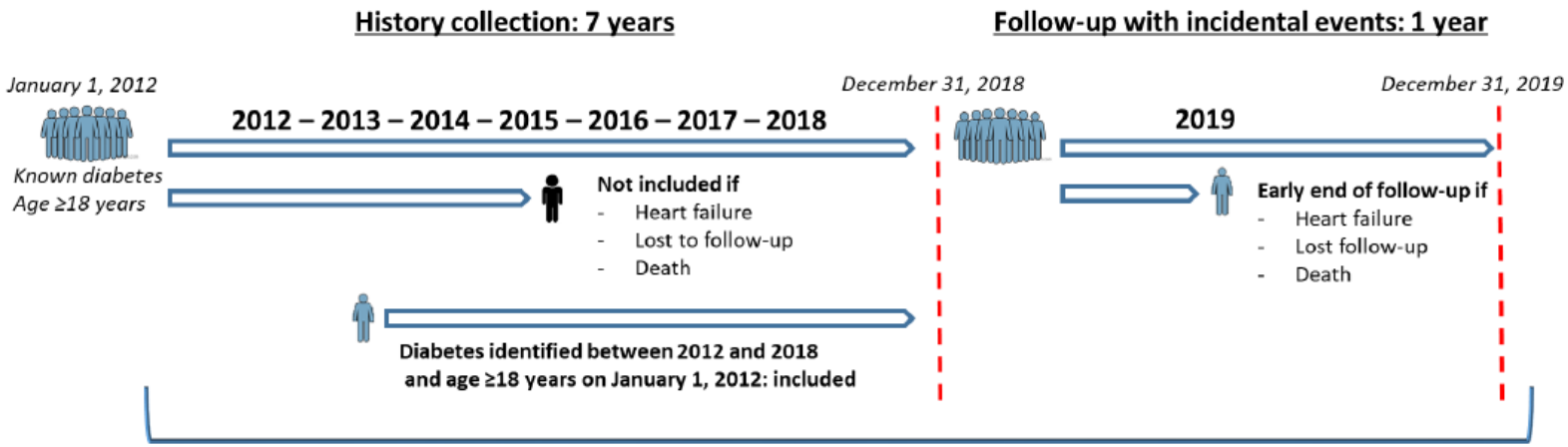
*(SRE/No SRE) incidence density ratio

	M_1		M_2		M_3		M_4	
55 052 events/3 027 413 individuals (1.82%)	**HR (95% CI)**	**P-value**	**HR (95% CI)**	**P-value**	**HR (95% CI)**	**P-value**	**HR (95% CI)**	**P-value**
Age (+1 year)	1.06 [1.06; 1.06]	<0.0001	1.04 [1.04; 1.04]	<0.0001	1.04 [1.03; 1.04]	<0.0001	1.03 [1.03; 1.04]	<0.0001
Women/men	0.66 [0.65; 0.67]	<0.0001	0.95 [0.93; 0.97]	<0.0001	0.96 [0.94; 0.97]	<0.0001	0.94 [0.93; 0.96]	<0.0001
Serious retinal event (any)	1.32 [1.28; 1.36]	<0.0001	1.16 [1.12; 1.20]	<0.0001	1.12 [1.09; 1.16]	<0.0001	1.11 [1.07; 1.14]	<0.0001
Cardiovascular (time-dependent covariates)								
CAD			2.63 [2.58; 2.68]	<0.0001	2.60 [2.55; 2.65]	<0.0001	2.61 [2.56; 2.66]	<0.0001
Heart valve disease			2.66 [2.59; 2.72]	<0.0001	2.63 [2.57; 2.69]	<0.0001	2.63 [2.57; 2.70]	<0.0001
Cardiac arrhythmia			3.93 [3.85; 4.01]	<0.0001	3.86 [3.79; 3.94]	<0.0001	3.87 [3.80; 3.95]	<0.0001
Stroke/TIA			1.14 [1.11; 1.17]	<0.0001	1.13 [1.10; 1.16]	<0.0001	1.13 [1.10; 1.16]	<0.0001
PAD			1.49 [1.45; 1.52]	<0.0001	1.40 [1.37; 1.43]	<0.0001	1.38 [1.35; 1.42]	<0.0001
Other diabetes complications								
Neuropathy					1.00 [0.97; 1.04]	0.97	0.99 [0.95; 1.02]	0.52
Diabetic foot					1.21 [1.15; 1.28]	<0.0001	1.22 [1.16; 1.28]	<0.0001
Lower limb amputation					1.02 [0.95; 1.10]	0.56	1.02 [0.95; 1.10]	0.51
Kidney disease (reference = None)								
Chronic kidney disease					1.25 [1.22; 1.28]	<0.0001	1.22 [1.18; 1.25]	<0.0001
End stage kidney disease					1.26 [1.17; 1.34]	<0.0001	1.24 [1.16; 1.33]	<0.0001
Cancer					1.07 [1.05; 1.09]	<0.0001	1.07 [1.05; 1.09]	<0.0001
Diabetes treatment in 2018								
Metformin							0.95 [0.93; 0.96]	<0.0001
Sulfonylurea and/or repaglinide							1.04 [1.02; 1.06]	<0.0001
DPP4-inhibitors							1.01 [0.99; 1.03]	0.36
GLP-1 receptor agonist							0.92 [0.88; 0.95]	<0.0001
Insulin							1.06 [1.03; 1.08]	<0.0001
Anti-hypertensive treatment in 2018								
Beta blockers							0.94 [0.92; 0.96]	<0.0001
Calcium-channel blockers							1.12 [1.10; 1.14]	<0.0001
Angiotensin-converting enzyme inhibitors							1.06 [1.03; 1.08]	<0.0001
Angiotensin-renin blockers							1.08 [1.05; 1.11]	<0.0001
Thiazide diuretics							1.03 [1.01; 1.06]	0.011
Potassium-sparing diuretics							1.36 [1.32; 1.41]	<0.0001

Table 3. Risk of developing heart failure in 2019 in the DMC population without HF history on December 31, 2018 (*N* = 3 027 413), multivariable Cox model using both fixed and time-dependent covariates

CAD: coronary artery disease; HF: heart failure; PAD: peripheral artery disease; SRE: Serious retinal event; TIA: transient ischaemic attack.

Cox model based on proportional hazards hypothesis, for the analysis of HF during 2019 in patients still alive and followed up on December 31, 2018, without known HF. The models proposed include the following covariates: (M_1): age, sex, SRE status (time-dependent); (M_2): M_1 with cardiovascular covariates, used as time-dependent (CAD, heart valve disease, cardiac arrhythmia, stroke/TIA and PAD); (M_3): M_2 with patient history on December 31, 2018 of neuropathy, diabetic foot, lower limb amputation, kidney disease (none/chronic/end-stage) and cancer; (M_4): M_3 with known treatments during 2018 (metformin, sulfonylurea and/or repaglinide, GLP1- receptor agonists, insulin, beta blockers, calcium-channel blockers, angiotensin-converting enzyme inhibitors, angiotensin-renin blockers, thiazide and potassium-sparing diuretics. The proportional hazards condition was controlled graphically for the SRE.



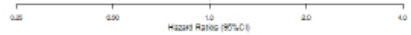
Patient eligibility criteria:

- Age ≥18 years on January 1, 2012
- Diabetes identified between 2012 and 2018
- Consistent data regarding age, sex and death
- No ambiguity on patient identifier for inter-IS link
- For the main analysis: patients neither dead nor lost to follow-up nor history of heart failure on December 31, 2018

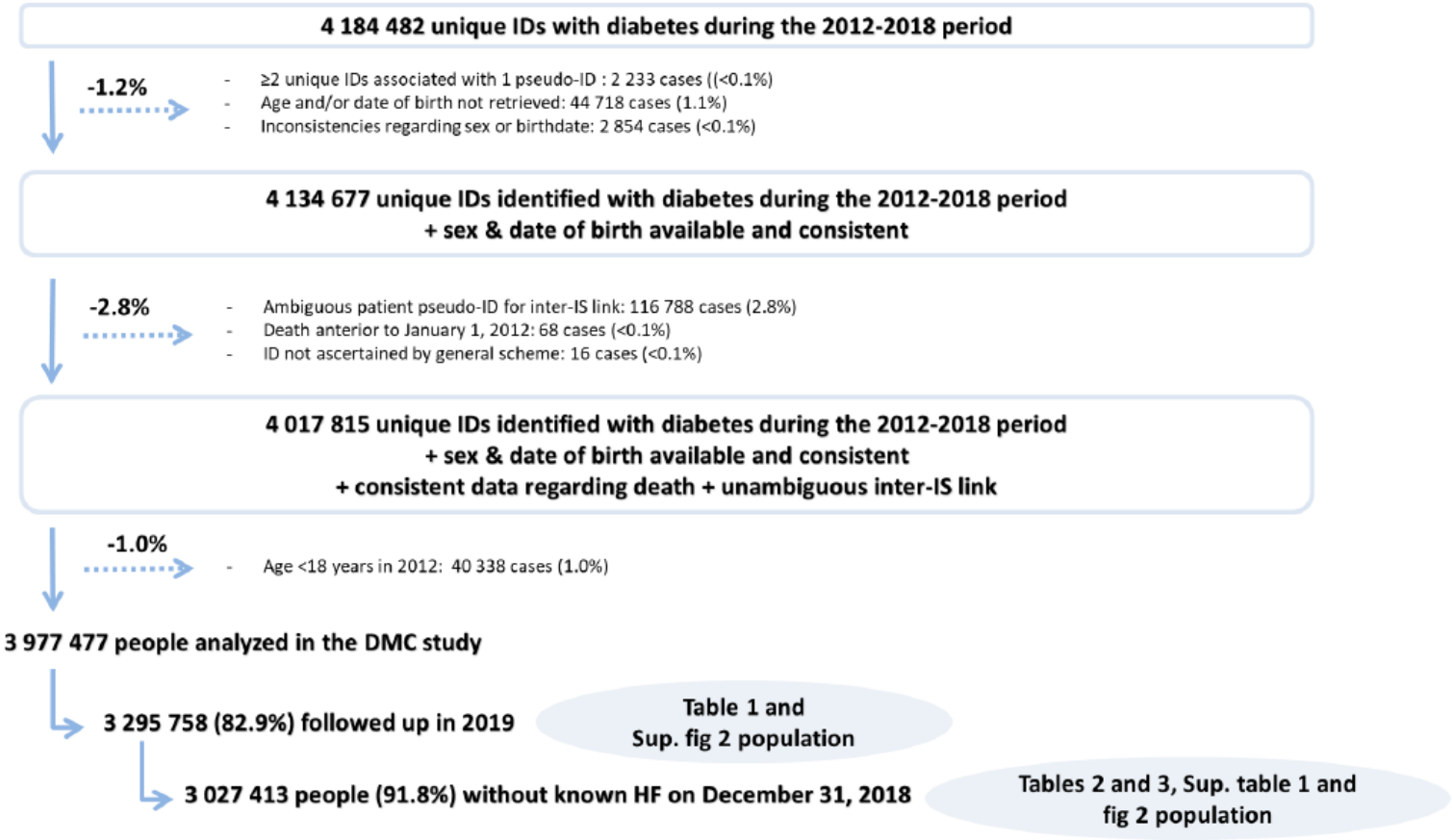
SNDS
Système national des données de santé

Fig 1. Study design

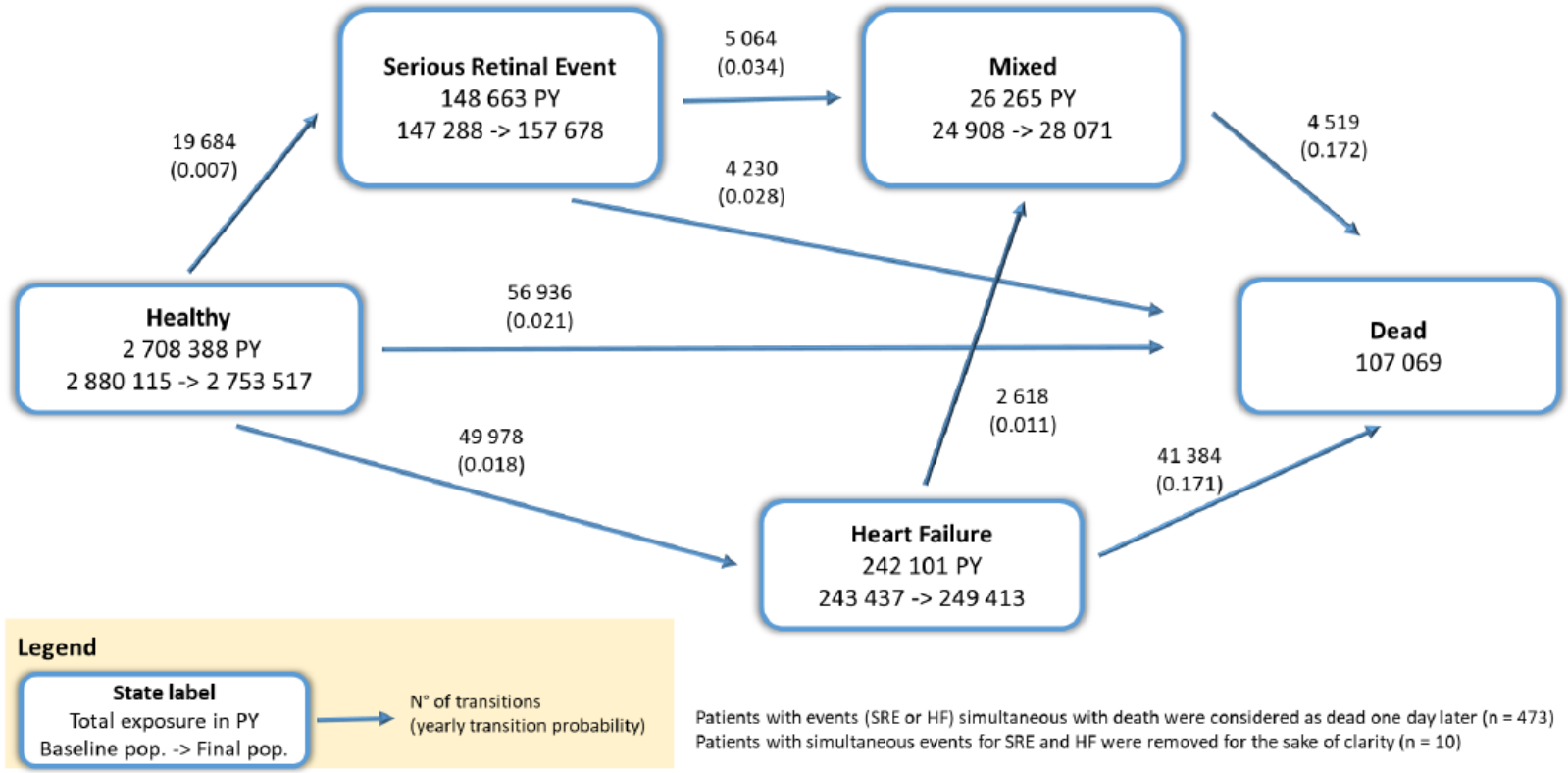
	Groups	HR	95%CI	p-value	Events / PY	Events / 1000 PY
All patients	Any SRE	1.11	1.07-1.14	<0.0001	55 052 / 2 859 013 PY	19.3 / 1000 PY
	SRE-1*	1.20	1.15-1.25	<0.0001		
	SRE-2*	1.18	1.05-1.33	0.0043		
	SRE-3*	1.07	1.02-1.11	0.0020		
Age class	<50 years	2.75	2.23-3.40	<0.0001	817 / 276 290 PY	3.0 / 1000 PY
	50-59 years	1.95	1.72-2.22	<0.0001	3 045 / 465 791 PY	6.5 / 1000 PY
	60-69 years	1.45	1.35-1.55	<0.0001	9 637 / 845 297 PY	11.4 / 1000 PY
	≥ 70 years	1.02	0.99-1.06	0.21	41 553 / 1 295 149 PY	32.1 / 1000 PY
Sex	Men	1.10	1.05-1.14	<0.0001	31 269 / 1 518 915 PY	20.6 / 1000 PY
	Women	1.12	1.07-1.17	<0.0001	23 783 / 1 363 613 PY	17.4 / 1000 PY
CAD	No	1.21	1.16-1.26	<0.0001	28 836 / 2 455 441 PY	11.7 / 1000 PY
	Yes	1.03	0.99-1.08	0.15	26 216 / 427 087 PY	61.4 / 1000 PY
Heart valve disease	No	1.17	1.14-1.21	<0.0001	44 711 / 2 806 963 PY	15.9 / 1000 PY
	Yes	0.91	0.85-0.97	0.0063	10 341 / 75 564 PY	136.9 / 1000 PY
Cardiac arrhythmia	No	1.32	1.27-1.37	<0.0001	29 437 / 2 638 938 PY	11.2 / 1000 PY
	Yes	0.92	0.88-0.96	0.00042	25 615 / 243 589 PY	105.2 / 1000 PY
CAD and/or heart valve disease	No	1.25	1.19-1.31	<0.0001	24 255 / 2 413 815 PY	10.0 / 1000 PY
	Yes	1.04	1.00-1.08	0.045	30 797 / 468 712 PY	65.7 / 1000 PY
CAD, heart valve disease and/or cardiac arrythmia	No	1.42	1.34-1.51	<0.0001	13 470 / 2 264 006 PY	5.9 / 1000 PY
	Yes	1.04	1.01-1.08	0.021	41 582 / 618 522 PY	67.2 / 1000 PY
End stage kidney disease	No	1.12	1.09-1.16	<0.0001	53 927 / 2 864 105 PY	18.8 / 1000 PY
	Yes	0.90	0.76-1.06	0.20	1 125 / 18 423 PY	61.1 / 1000 PY
History of loop diuretics	No	1.18	1.14-1.23	<0.0001	38 191 / 2 685 543 PY	14.2 / 1000 PY
	Yes	0.96	0.91-1.01	0.15	16 861 / 196 984 PY	85.6 / 1000 PY



2. Association of the different SRE subtypes (SRE-1/-2/-3) with HF, and association of SRE with HF in the different subsets defined using age, sex, established HF risk tors, end stage kidney disease and history of loop diuretics (*N* = 3 027 413)



Supplemental fig 1. Study flowchart



Supplemental fig 2. Multi-state model for Serious Retinal Event & Heart Failure, year 2019 (*N* = 3 295 758)

	No SRE (*N* = 2 880 125)	SRE (*N* = 147 288)
Women	1 357 990 (47.2%)	71 025 (47.1%)
Age on January 1, 2019 (years)	66.9 (13.1)	71.1 (12.6)
Age		
<50 years	285 459 (9.9%)	8 308 (5.6%)
50-59 years	471 834 (16.4%)	15 578 (10.6%)
60-69 years	843 813 (29.3%)	38 034 (25.8%)
≥ 70 years	1 279 019 (44.4%)	85 368 (58.0%)
Diabetes duration		
≤2 years	305 379 (10.6%)	5 966 (4.1%)
3 to 5 years	503 973 (17.5%)	12 817 (8.7%)
> 5 years	2 070 773 (71.9%)	128 505 (87.2%)
Cardiovascular history		
CAD	395 476 (13.7%)	29 312 (19.9%)
Heart valve disease	66 368 (2.3%)	5 023 (3.4%)
Cardiac arrhythmia	220 426 (7.7%)	14 929 (10.1%)
Stroke/TIA	144 568 (5.0%)	11 162 (7.6%)
PAD	194 055 (6.7%)	20 138 (13.7%)
Other diabetes complications		
Neuropathy	111 114 (3.9%)	21 070 (14.3%)
Foot ulcer	34 861 (1.2%)	7 202 (4.9%)
Lower limb amputation	13 480 (0.5%)	3 928 (2.7%)
Chronic kidney disease	186 312 (6.5%)	26 670 (18.1%)
End-stage kidney disease	16 420 (0.6%)	2 938 (2.0%)
Cancer	425 868 (14.8%)	24 399 (16.6%)
Eye-related outcomes*		
SRE-1	No history	98 282 (66.7%)
SRE-2	No history	8 756 (5.9%)
SRE-3	No history	72 895 (49.5%)
≥1 ophthalmologic act in 2017-2018	1 944 789 (67.5%)	133 995 (91.0%)
Diabetes treatment in 2018		
Metformin	1 888 864 (65.6%)	89 021 (60.4%)
Sulfonylurea and/or repaglinide	926 341 (32.2%)	51 639 (35.1%)
DPP4-inhibitors	754 842 (26.2%)	38 782 (26.3%)
GLP-1 receptor agonist	205 630 (7.1%)	15 214 (10.3%)
Insulin therapy	535 937 (18.6%)	65 613 (44.5%)
If insulin therapy: number of IU/day	23.3 (23.8)	28.3 (26.4)
Diet only	441 746 (15.3%)	12 741 (8.7%)
Anti-hypertensive treatment in 2018		
Any anti-hypertensive treatment	2 029 344 (70.5%)	119 301 (81.0%)
Beta blockers	8 389 41 (29.1%)	49 935 (33.9%)
Calcium-channel blockers	821 858 (28.5%)	56 602 (38.4%)
ACE-inhibitors	740 931 (25.7%)	46 435 (31.5%)
ARBs	895 033 (31.1%)	54 148 (36.8%)
Loop diuretics	190 202 (6.6%)	17 963 (12.2%)
Thiazide diuretics	600 340 (20.8%)	35 072 (23.8%)
Potassium-sparing diuretic	121 378 (4.2%)	6 251 (4.2%)
Other treatments of interest in 2018		
Antiplatelet agents	991 440 (34.4%)	70 357 (47.8%)
Anticoagulants	218 992 (7.6%)	13 958 (9.5%)
Statins	1 370 379 (47.6%)	82 193 (55.8%)
Ezetimibe	163 276 (5.7%)	9 513 (6.5%)

Supplemental table 1. Clinical characteristics associated with the DMC population without HF history and followed-up in 2019 (*N* = 3 027 413), according to SRE status on December 31, 2018

Categorical variables are described using n (%). Quantitative variables are described using mean (standard deviation). HF: heart failure. CAD: coronary artery disease; PAD: peripheral artery disease; SRE: serious retinal event; TIA: transient ischaemic attack.

*SRE-1: retinal laser photocoagulation; SRE-2: retinal detachment or vitreal haemorrhage; SRE-3: intravitreal injection of anti VEGF or corticosteroids.

Association of serious retinal events with incident heart failure in people living with diabetes: a nationwide population study

Matthieu Wargny, Jean-Baptiste Ducloyer, Pacôme Constant dit Beaufils, Christophe Leux, Thomas Goronflot, Philippe Tuppin, Sandrine Fosse-Edorh, Clara Piffaretti, Jean-Noël Trochu, Pierre-Antoine Gourraud, Bertrand Cariou, Samy Hadjadj

-

Supplementary file 1

List of codes used and their explanation

Glossary

ATC: International drug classification, Anatomic, Therapeutic and Chemical.

CCAM: « *Classification Commune des Actes Médicaux* » French classification of medical procedures, including but no limited to surgical procedures.

GHM: « *Groupe Homogène de Malades*» French classification used to summarise diagnoses related groups for hospital stays.

ICD-10: International Classification of Diseases, 10th revision.

LTD: corresponding to « Long-Term chronic Diseases » (« *Affection de Longue Durée* », eligible for 100% reimbursement) belonging to a defined list by decree after expertise from the HAS (« *Haute Autorité de Santé* », French National Authority for Health).

When a code is expressed partially, with a troncature, this implies the inclusion of all codes starting by this segment.

For example:

- « I50 » for heart failure includes all I500 (Congestive heart failure), I501 (Left ventricular insufficiency), etc.
- « A10A » for insulins and analogues includes A10AB (fast-acting insulins), A10AC (intermediate acting), etc.

PART I. Codes used to identify the different diseases of interest

1. Coronary artery disease

Codes	English	French
LTD n°13	Chronic coronary syndrome	Syndrome coronarien chronique
ICD-10		
I20	Angina pectoris	Angine de poitrine
I21	Acute myocardial infarction	Infarctus du myocarde
I22	Subsequent Myocardial Infarction	Infarctus du myocarde à répétition
I23	Certain current complications following acute myocardial infarction	Complication récente d'un infarctus aigu du myocarde
I24	Other acute ischaemic heart diseases	Autres cardiopathies ischémiques aiguës
I25	Chronic ischaemic heart disease	Cardiopathie ischémique chronique
CCAM		
DDAA002	Patch angioplasty of the common trunk of the left coronary artery, by thoracotomy with extracorporeal circulation	Angioplastie d'élargissement du tronc commun de l'artère coronaire gauche, par thoracotomie avec CEC
DDAF001	Transcutaneous arterial intraluminal coronary vessel dilation without stenting	Dilatation intraluminale d'un vaisseau coronaire sans pose d'endoprothèse, par voie artérielle transcutanée
DDAF003	Intraluminal dilatation of 3 or more coronary vessels with transcutaneous arterial stenting	Dilatation intraluminale de 3 vaisseaux coronaires ou plus avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF004	Intraluminal dilatation of 2 coronary vessels with transcutaneous arterial stenting	Dilatation intraluminale de 2 vaisseaux coronaires avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF006	Intraluminal coronary vessel dilatation with transcutaneous arterial stenting	Dilatation intraluminale d'un vaisseau coronaire avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF007	Intraluminal dilatation of 2 coronary vessels with coronary arteriography, with stenting, by transcutaneous arterial approach	Dilatation intraluminale de 2 vaisseaux coronaires avec artériographie coronaire, avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF008	Intraluminal coronary vessel dilatation with transcutaneous arterial stenting	Dilatation intraluminale d'un vaisseau coronaire avec artériographie coronaire, avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF009	Intraluminal dilation of 3 or more coronary vessels with coronary arteriography, with stenting, by transcutaneous arterial approach	Dilatation intraluminale de 3 vaisseaux coronaires ou plus avec artériographie coronaire, avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF010	Intraluminal coronary vessel dilatation with coronary arteriography, without stenting, by transcutaneous arterial approach	Dilatation intraluminale d'un vaisseau coronaire avec artériographie coronaire, sans pose d'endoprothèse, par voie artérielle transcutanée
DDFF001	Intraluminal coronary artery atherectomy by rotational method using a transcutaneous arterial approach	Athérectomie intraluminale d'artère coronaire par méthode rotatoire [rotationnelle], par voie artérielle transcutanée

Codes	English	French
CCAM		
DDFF002	Intraluminal coronary artery atherectomy by transcutaneous arterial approach	Athérectomie intraluminale d'artère coronaire, par voie artérielle transcutanée
DDMA003	Coronary revascularisation using 3 arterial grafts with 3 distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par 3 greffons artériels avec 3 anastomoses distales, par thoracotomie avec CEC
DDMA004	Coronary vein graft revascularisation with 4 or more distal anastomoses by thoracotomy with extracorporeal circulation	Revascularisation coronaire par greffon veineux avec 4 anastomoses distales ou plus, par thoracotomie avec CEC
DDMA005	Coronary revascularisation with 2 arterial and 3 venous grafts with 3 distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels et par greffon veineux avec 3 anastomoses distales, par thoracotomie avec CEC
DDMA006	Coronary revascularisation using 2 arterial grafts with 3 distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels avec 3 anastomoses distales, par thoracotomie avec CEC
DDMA007	Coronary vein graft revascularisation with distal anastomosis by thoracotomy with extracorporeal circulation	Revascularisation coronaire par greffon veineux avec une anastomose distale, par thoracotomie avec CEC
DDMA008	Coronary revascularisation using 2 arterial grafts with 4 or more distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels avec 4 anastomoses distales ou plus, par thoracotomie avec CEC
DDMA009	Coronary revascularisation with 2 arterial and 4 or more distal anastomoses by thoracotomy with extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels et par greffon veineux avec 4 anastomoses distales ou plus, par thoracotomie avec CEC
DDMA011	Coronary revascularisation by arterial and venous graft with 2 distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par un greffon artériel et par greffon veineux avec 2 anastomoses distales, par thoracotomie avec CEC
DDMA012	Coronary revascularization with 3 arterial and 4 or more distal venous grafts by thoracotomy with extracorporeal circulation	Revascularisation coronaire par 3 greffons artériels et par greffon veineux avec 4 anastomoses distales ou plus, par thoracotomie avec CEC
DDMA013	Coronary revascularisation using 3 arterial grafts with 4 or more distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par 3 greffons artériels avec 4 anastomoses distales ou plus, par thoracotomie avec CEC
DDMA015	Arterial graft coronary revascularisation with distal anastomosis by thoracotomy with extracorporeal circulation	Revascularisation coronaire par un greffon artériel avec une anastomose distale, par thoracotomie avec CEC
DDMA016	Coronary vein graft revascularisation with 3 distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par greffon veineux avec 3 anastomoses distales, par thoracotomie avec CEC
DDMA017	Coronary revascularisation using an arterial graft with 2 distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par un greffon artériel avec 2 anastomoses distales, par thoracotomie avec CEC
DDMA018	Coronary revascularisation by arterial and venous graft with 3 distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par un greffon artériel et par greffon veineux avec 3 anastomoses distales, par thoracotomie avec CEC
DDMA019	Coronary vein graft revascularisation with 2 distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par greffon veineux avec 2 anastomoses distales, par thoracotomie avec CEC
DDMA020	Coronary revascularisation by 2 arterial grafts with 2 distal anastomoses, by thoracotomy with extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels avec 2 anastomoses distales, par thoracotomie avec CEC
DDMA021	Arterial and vein graft coronary revascularisation with 4 or more distal anastomoses by thoracotomy with extracorporeal circulation	Revascularisation coronaire par un greffon artériel et par greffon veineux avec 4 anastomoses distales ou plus, par thoracotomie avec CEC

Codes	English	French
CCAM		
DDMA022	Coronary revascularisation with 2 arterial and 3 venous grafts with 3 distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels et par greffon veineux avec 3 anastomoses distales, par thoracotomie sans CEC
DDMA023	Coronary revascularisation using an arterial graft with 2 distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par un greffon artériel avec 2 anastomoses distales, par thoracotomie sans CEC
DDMA024	Coronary revascularisation using a venous graft with 2 distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par greffon veineux avec 2 anastomoses distales, par thoracotomie sans CEC
DDMA025	Arterial graft coronary revascularisation with distal anastomosis by thoracotomy without extracorporeal circulation	Revascularisation coronaire par un greffon artériel avec une anastomose distale, par thoracotomie sans CEC
DDMA026	oronary revascularisation by 2 arterial grafts with 2 distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels avec 2 anastomoses distales, par thoracotomie sans CEC
DDMA027	Coronary vein graft revascularisation with 3 distal anastomoses by thoracotomy without extracorporeal circulation	Revascularisation coronaire par greffon veineux avec 3 anastomoses distales, par thoracotomie sans CEC
DDMA028	Coronary vein graft revascularisation with distal anastomosis by thoracotomy without extracorporeal circulation	Revascularisation coronaire par greffon veineux avec une anastomose distale, par thoracotomie sans CEC
DDMA029	Coronary revascularisation by arterial and venous graft with 3 distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par un greffon artériel et par greffon veineux avec 3 anastomoses distales, par thoracotomie sans CEC
DDMA030	Coronary revascularisation using 3 arterial grafts with 3 distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par 3 greffons artériels avec 3 anastomoses distales, par thoracotomie sans CEC
DDMA031	Coronary revascularisation using 2 arterial grafts with 3 distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels avec 3 anastomoses distales, par thoracotomie sans CEC
DDMA032	Coronary revascularisation by arterial and venous graft with 2 distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par un greffon artériel et par greffon veineux avec 2 anastomoses distales, par thoracotomie sans CEC
DDMA033	Coronary revascularisation using 2 arterial grafts with 4 or more distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels avec 4 anastomoses distales ou plus, par thoracotomie sans CEC
DDMA034	Coronary revascularisation with 2 arterial and 4 or more distal anastomoses by thoracotomy without extracorporeal circulation	Revascularisation coronaire par 2 greffons artériels et par greffon veineux avec 4 anastomoses distales ou plus, par thoracotomie sans CEC
DDMA035	Coronary revascularisation using 3 arterial grafts with 4 or more distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par 3 greffons artériels avec 4 anastomoses distales ou plus, par thoracotomie sans CEC
DDMA036	Coronary revascularisation with 3 arterial and 4 or more venous grafts with distal anastomoses by thoracotomy without extracorporeal circulation	Revascularisation coronaire par 3 greffons artériels et par greffon veineux avec 4 anastomoses distales ou plus, par thoracotomie sans CEC
DDMA037	Coronary vein graft revascularisation with 4 or more distal anastomoses by thoracotomy without extracorporeal circulation	Revascularisation coronaire par greffon veineux avec 4 anastomoses distales ou plus, par thoracotomie sans CEC
DDMA038	Coronary revascularisation by arterial and venous graft with 4 or more distal anastomoses, by thoracotomy without extracorporeal circulation	Revascularisation coronaire par un greffon artériel et par greffon veineux avec 4 anastomoses distales ou plus, par thoracotomie sans CEC

Codes	English	French
CCAM		
DDQH006	Transcutaneous arterial coronary bypass angiography	Angiographie de pontage coronaire, par voie artérielle transcutanée
DDQH011	Coronary arteriography with coronary bypass angiography and left ventriculography by transcutaneous arterial approach	Artériographie coronaire avec angiographie d'un pontage coronaire et ventriculographie gauche, par voie artérielle transcutanée
DDQH013	Coronary arteriography with angiography of several coronary bypasses without left ventriculography, by transcutaneous arterial approach	Artériographie coronaire avec angiographie de plusieurs pontages coronaires sans ventriculographie gauche, par voie artérielle transcutanée
DDQH014	Coronary arteriography with angiography of a coronary bypass without left ventriculography by transcutaneous arterial approach	Artériographie coronaire avec angiographie d'un pontage coronaire sans ventriculographie gauche, par voie artérielle transcutanée
DDQH015	Coronary arteriography with angiography of several coronary bypasses and left ventriculography, by transcutaneous arterial approach	Artériographie coronaire avec angiographie de plusieurs pontages coronaires et ventriculographie gauche, par voie artérielle transcutanée
DDPF002	Transcutaneous arterial coronary artery recanalisation with stenting	Recanalisation d'artère coronaire avec pose d'endoprothèse, par voie artérielle transcutanée

2. Heart valve disease

Codes	English	French
ICD-10		
I05	Rheumatic mitral valve diseases	Maladies rhumatismales de la valvule mitrale
I06	Rheumatic aortic valve diseases	Maladies rhumatismales de la valvule aortique
I07	Rheumatic tricuspid valve diseases	Maladies rhumatismales de la valvule tricuspide
I08	Multiple valves diseases	Maladies rhumatismales de plusieurs valvules
I34	Nonrheumatic mitral valve disorders	Atteintes non rhumatismales de la valvule mitrale
I35	Nonrheumatic aortic valve disorders	Atteintes non rhumatismales de la valvule aortique
I36	Nonrheumatic tricuspide valve disorders	Atteintes non rhumatismales de la valvule tricuspide
I37	Nonrheumatic pulmonary valve disorders	Atteintes de la valvule pulmonaire : non rhumatismale
I39	Endocarditis and heart valve disorders in diseases classified elsewhere	Endocardite et atteintes valvulaires cardiaques au cours de maladies classées ailleurs
Q220	Pulmonary valve atresia	Atrésie de la valve pulmonaire
Q221	Congenital pulmonary valve stenosis	Sténose congénitale de la valve pulmonaire
Q222	Congenital pulmonary valve insufficiency	Insuffisance congénitale de la valve pulmonaire
Q223	Other congenital pulmonary valve defects	Autres malformations congénitales de la valve pulmonaire
Q224	Congenital tricuspid valve stenosis	Sténose congénitale de la valvule tricuspide
Q225	Ebstein anomaly	Maladie d'Ebstein
Q228	Other congenital tricuspid valve defects	Autres malformations congénitales de la valvule tricuspide
Q229	Congenital malformation of the tricuspid valve, unspecified	Malformation congénitale de la valvule tricuspide, sans précision
Q230	Congenital aortic valve stenosis	Sténose congénitale de la valvule aortique
Q231	Congenital aortic valve insufficiency	Insuffisance congénitale de la valvule aortique
Q232	Congenital mitral stenosis	Sténose mitrale congénitale
Q233	Congenital mitral insufficiency	Insuffisance mitrale congénitale
Q238	Other congenital malformations of the aortic and mitral valves	Autres malformations congénitales des valvules aortique et mitrale
Q239	Congenital malformation of the aortic and mitral valves, unspecified	Malformation congénitale des valvules aortique et mitrale, sans précision
T820	Mechanical complication of a heart valve prosthesis	Complication mécanique d'une prothèse valvulaire cardiaque
T822	Mechanical complications of coronary artery bypass surgery and heart valve transplantation	Complication mécanique d'un pontage coronarien et d'une greffe valvulaire cardiaque
T826	Infection and inflammatory reaction due to a heart valve prosthesis	Infection et réaction inflammatoire dues à une prothèse valvulaire cardiaque
Z952	Presence of a prosthetic heart valve	Présence de prothèse d'une valvule cardiaque
Z953	Presence of a xenogenic heart valve	Présence d'une valvule cardiaque xénogénique
Z954	Presence of another heart valve replacement	Présence d'une autre valvule cardiaque de remplacement

CCAM	Cardiac valvulotomies or valvectomies	Valvulotomies ou valvectomies cardiaques
DBPA002	Right atrioventricular commissurotomy or valvectomy, by thoracotomy with extracorporeal circulation	Commissurotomie ou valvectomie atrioventriculaire droite, par thoracotomie avec CEC
DBPA004	Commissurotomy or pulmonary valvectomy, by thoracotomy with with extracorporeal circulation	Commissurotomie ou valvectomie pulmonaire, par thoracotomie avec CEC
DBPA005	Commissurotomy of the left atrioventricular valve by thoracotomy without extracorporeal circulation	Commissurotomie de la valve atrioventriculaire gauche, par thoracotomie sans CEC
DBPA006	Commissurotomy of the left atrioventricular valve by thoracotomy with extracorporeal circulation	Commissurotomie de la valve atrioventriculaire gauche, par thoracotomie avec CEC
DBPA007	Commissurotomy of the aortic valve, by thoracotomy with extracorporeal circulation	Commissurotomie de la valve aortique, par thoracotomie avec CEC
CCAM	**Cardiac annuloplasties and valvuloplasties**	**Annuloplasties et valvuloplasties cardiaques**
DBMA002	Left atrioventricular valvoplasty, by thoracotomy with extracorporeal circulation	Valvoplastie atrioventriculaire gauche, par thoracotomie avec CEC
DBMA003	Left atrioventricular annuloplasty, by thoracotomy with extracorporeal circulation	Annuloplastie atrioventriculaire gauche, par thoracotomie avec CEC
DBMA008	Right atrioventricular annuloplasty, by thoracotomy with extracorporeal circulation	Annuloplastie atrioventriculaire droite, par thoracotomie avec CEC
DBMA012	Right atrioventricular valvoplasty, by thoracotomy with extracorporeal circulation	Valvoplastie atrioventriculaire droite, par thoracotomie avec CEC
CCAM	**Valve replacements**	**Remplacements valvulaires**
DBKA001	Homograft aortic valve replacement via thoracotomy with extracorporeal circulation	Remplacement de la valve aortique par homogreffe, par thoracotomie avec CEC
DBKA002	Left atrioventricular valve replacement by prosthesis in non-anatomical position, by thoracotomy with extracorporeal circulation	Remplacement de la valve atrioventriculaire gauche par prothèse en position non anatomique, par thoracotomie avec CEC
DBKA003	Aortic valve replacement by bioprosthesis without frame, by thoracotomy with with extracorporeal circulation	Remplacement de la valve aortique par bioprothèse sans armature, par thoracotomie avec CEC
DBKA004	Right atrioventricular valve replacement by mechanical prosthesis or bioprosthesis with armature, by thoracotomy with extracorporeal circulation	Remplacement de la valve atrioventriculaire droite par prothèse mécanique ou bioprothèse avec armature, par thoracotomie avec CEC
DBKA005	Homograft left atrioventricular valve replacement via thoracotomy with with extracorporeal circulation	Remplacement de la valve atrioventriculaire gauche par homogreffe, par thoracotomie avec CEC
DBKA006	Replacement of the aortic valve by mechanical prosthesis or bioprosthesis with armature, by thoracotomy with extracorporeal circulation	Remplacement de la valve aortique par prothèse mécanique ou bioprothèse avec armature, par thoracotomie avec CEC
DBKA007	Pulmonary valve replacement by mechanical prosthesis or bioprosthesis with frame, by thoracotomy with extracorporeal circulation	Remplacement de la valve pulmonaire par prothèse mécanique ou bioprothèse avec armature, par thoracotomie avec CEC
DBKA008	Right atrioventricular valve replacement by homograft, via thoracotomy with with extracorporeal circulation	Remplacement de la valve atrioventriculaire droite par homogreffe, par thoracotomie avec CEC
DBKA010	Replacement of the left atrioventricular valve by mechanical prosthesis or bioprosthesis with armature, by thoracotomy with extracorporeal circulation	Remplacement de la valve atrioventriculaire gauche par prothèse mécanique ou bioprothèse avec armature, par thoracotomie avec CEC
DBKA011	Aortic valve replacement by prosthesis in non-anatomical position, by thoracotomy with extracorporeal circulation	Remplacement de la valve aortique par prothèse en position non anatomique, par thoracotomie avec CEC
DBKA012	Pulmonary valve replacement by homograft or bioprosthesis without armature, by thoracotomy with extracorporeal circulation	Remplacement de la valve pulmonaire par homogreffe ou bioprothèse sans armature, par thoracotomie avec CEC

CCAM	Other procedures on the cardiac orifices	Autres actes sur les orifices du coeur
DBBF198	Transcutaneous venous and transseptal device-based left atrioventricular orifice narrowing with transesophageal ultrasound-doppler guidance	Rétrécissement de l'orifice atrioventriculaire gauche par dispositif par voie veineuse transcutanée et voie transseptale avec guidage par échographie-doppler par voie transœsophagienne
DBEA001	Reinsertion of a cardiac orifice prosthesis, by thoracotomy with extracorporeal circulation	Réinsertion d'une prothèse orificielle cardiaque, par thoracotomie avec CEC
DBLA004	Pose d'une bioprothèse de la valve aortique, par abord de l'apex du cœur par thoracotomie sans extracorporeal circulation	Pose d'une bioprothèse de la valve aortique, par abord de l'apex du cœur par thoracotomie sans CEC
DBLF001	Transcutaneous arterial placement of an aortic valve bioprosthesis	Pose d'une bioprothèse de la valve aortique, par voie artérielle transcutanée
DBLF009	Transcutaneous venous placement of a pulmonary valve bioprosthesis in a prosthetic conduit	Pose d'une bioprothèse de la valve pulmonaire dans un conduit prothétique, par voie veineuse transcutanée
DBSF001	Closure of a dehiscence by transcutaneous vascular insertion of a cardiac orifice prosthesis	Fermeture d'une déhiscence par désinsertion de prothèse orificielle cardiaque, par voie vasculaire transcutanée

3. Cardiac arrhythmia

Codes	English	French
ICD-10		
I47	Paroxysmal tachycardia	Tachycardie paroxystique
I48	Atrial fibrillation and flutter	Fibrillation et flutters auriculaires
I49	Other cardiac arrhythmias	Autres arythmies cardiaques

4. Stroke/TIA

Codes	English	French
LTD n°1	Disabling stroke	Accident vasculaire cérébral invalidant
ICD-10		
I60	Subarachnoid haemorrhage	Hémorragie sous-arachnoïdienne
I61	Intracerebral haemorrhage	Hémorragie intracérébrale
I62	Other nontraumatic intracranial haemorrhage	Autres hémorragies intracrâniennes non traumatiques
I63	Cerebral infarction	Infarctus cérébral
I64	Stroke, not specified as haemorrhage or infarction	AVC, non précisé comme étant hémorragique ou par infarctus
I69	Sequelae of cerebrovascular disease	Séquelles de maladies cérébrovasculaires
CCAM	Cerebrovascular disease (1/2)	Pathologie cérébrovasculaire (1/2)
EAFA001	Thrombectomy or endarterectomy of intracranial vessel, by craniotomy	Embolectomie ou thromboendartériectomie de vaisseau intracrânien, par craniotomie
EANF002	Superselective in situ thrombolysis of intracranial artery by transcutaneous arterial approach	Fibrinolyse in situ suprasélective d'artère intracrânienne, par voie artérielle transcutanée
EAAF002	Intraluminal dilatation of the intracranial internal carotid artery with stenting, by transcutaneous arterial approach	Dilatation intraluminale du tronc de l'artère carotide interne intracrânienne avec pose d'endoprothèse, par voie artérielle transcutanée
EAAF004	Intraluminal dilatation of the trunk of the intracranial internal carotid artery without stenting, by transcutaneous arterial approach	Dilatation intraluminale du tronc de l'artère carotide interne intracrânienne sans pose d'endoprothèse, par voie artérielle transcutanée
EAAF902	Intraluminal dilatation of the intracranial vertebral artery or basilar trunk without stenting, by transcutaneous arterial approach	Dilatation intraluminale de l'artère vertébrale intracrânienne ou de l'artère basilaire sans pose d'endoprothèse, par voie artérielle transcutanée
EAAF903	Intraluminal dilatation of the intracranial vertebral artery or basilar trunk with stenting, by transcutaneous arterial approach	Dilatation intraluminale de l'artère vertébrale intracrânienne ou de l'artère basilaire avec pose d'endoprothèse, par voie artérielle transcutanée
EBAF001	Intraluminal dilatation of the extracranial internal carotid artery with transcutaneous arterial stenting	Dilatation intraluminale de l'artère carotide interne extracrânienne avec pose d'endoprothèse, par voie artérielle transcutanée
EBAF003	Intraluminal dilatation of the extracranial internal carotid artery without transcutaneous arterial stenting	Dilatation intraluminale de l'artère carotide interne extracrânienne sans pose d'endoprothèse, par voie artérielle transcutanée
EBAF004	Intraluminal dilatation of the cervical common carotid artery without stenting by transcutaneous arterial approach	Dilatation intraluminale de l'artère carotide commune cervicale sans pose d'endoprothèse, par voie artérielle transcutanée
EBAF009	Intraluminal dilatation of the carotid bifurcation without stenting by transcutaneous arterial approach	Dilatation intraluminale de la bifurcation carotidienne sans pose d'endoprothèse, par voie artérielle transcutanée
EBAF010	Intraluminal dilatation of the cervical common carotid artery with transcutaneous arterial stenting	Dilatation intraluminale de l'artère carotide commune cervicale avec pose d'endoprothèse, par voie artérielle transcutanée
EBAF011	Intraluminal carotid bifurcation dilatation with transcutaneous arterial stenting	Dilatation intraluminale de la bifurcation carotidienne avec pose d'endoprothèse, par voie artérielle transcutanée
EBAF013	Intraluminal dilatation of the extracranial vertebral artery without stenting by transcutaneous arterial approach	Dilatation intraluminale de l'artère vertébrale extracrânienne sans pose d'endoprothèse, par voie artérielle transcutanée

Codes	English	French (2/2)
CCAM codes	Cerebrovascular disease (2/2)	Pathologie cérébrovasculaire (2/2)
EBFA002	Endarterectomy of the carotid bifurcation with patch angioplasty, by cervicotomy with vascular bypass	Thromboendartériectomie de la bifurcation carotidienne avec angioplastie d'élargissement, par cervicotomie avec dérivation vasculaire
EBFA003	Endarterectomy of the common carotid artery, by cervicotomy	Thromboendartériectomie de l'artère carotide commune, par cervicotomie
EBFA005	Thrombectomy of the common carotid artery, by cervicotomy	Thrombectomie de l'artère carotide commune, par cervicotomie
EBFA006	Endarterectomy of the carotid bifurcation without patch angioplasty, by cervicotomy with vascular bypass	Thromboendartériectomie de la bifurcation carotidienne sans angioplastie d'élargissement, par cervicotomie avec dérivation vasculaire
EBFA008	Eversion endarteriectomy of the carotid bifurcation, by cervicotomy without vascular bypass	Thromboendartériectomie de la bifurcation carotidienne par retournement, par cervicotomie sans dérivation vasculaire
EBFA010	Endarterectomy of the common carotid artery, by cervicotomy and thoracotomy	Thromboendartériectomie de l'artère carotide commune, par cervicotomie et par thoracotomie
EBFA012	Endarterectomy of the carotid bifurcation without patch angioplasty, by cervicotomy without vascular bypass	Thromboendartériectomie de la bifurcation carotidienne sans angioplastie d'élargissement, par cervicotomie sans dérivation vasculaire
EBFA015	Eversion endarterectomy of the carotid bifurcation, by cervicotomy with vascular bypass	Thromboendartériectomie de la bifurcation carotidienne par retournement, par cervicotomie avec dérivation vasculaire
EBFA016	Endarterectomy of the carotid bifurcation with patch angioplasty, by cervicotomy without vascular bypass	Thromboendartériectomie de la bifurcation carotidienne avec angioplastie d'élargissement, par cervicotomie sans dérivation vasculaire
EBFA017	Endarterectomy of the proximal vertebral artery, by cervicotomy	Thromboendartériectomie de l'artère vertébrale proximale, par cervicotomie
EBNF001	Selective or hyperselective in situ thrombolysis of an extracranial artery reaching the cervicocerebral region by transcutaneous arterial approach	Fibrinolyse in situ sélective ou hypersélective d'une artère extracrânienne à destination cervicocéphalique, par voie artérielle transcutanée
EBNF002	Superselective in situ thrombolysis of an extracranial artery reaching the cervicocerebral region by transcutaneous arterial approach fibrinolysis of an extracranial artery to the cervicocerebral region by transcutaneous arterial approach	Fibrinolyse in situ suprasélective d'une artère extracrânienne à destination cervicocéphalique, par voie artérielle transcutanée
EAJF341	Transcutaneous arterial mechanical thrombectomy evacuation of intracranial artery thrombus	Évacuation de thrombus d'artère intracrânienne par voie artérielle transcutanée

Codes	English	French
CCAM	Surgeries to restore continuity in cases of vascular malformation or dissection	Chirurgies de remise en continuité lors de malformation vasculaire ou lors de dissection
DGCA032	Bypass surgery between the aorta and the brachiocephalic trunk, by thoracotomy	Pontage entre l'aorte et le tronc artériel brachiocéphalique, par thoracotomie
ECKA001	Replacement of the brachiocephalic trunk, by thoracotomy	Remplacement du tronc artériel brachiocéphalique, par thoracotomie
EBAA002	Carotid bifurcation patch angioplasty without endarterectomy, by cervicotomy	Angioplastie d'élargissement de la bifurcation carotidienne sans thromboendartériectomie, par cervicotomie
EBCA004	Intercarotid cross-bypass, by cervicotomy	Pontage croisé intercarotidien, par cervicotomie
EBCA005	Distal carotidovertebral or distal subclaviovertebral bypass, by cervicotomy	Pontage carotidovertébral distal ou subclaviovertébral distal, par cervicotomie
EBCA008	Ipsilateral carotidosubclavian or carotidoaxillary bypass, by cervicotomy	Pontage homolatéral carotidosubclavier ou carotidoaxillaire, par cervicotomie
EBCA010	Extra-intracranial arterial bypass without autograft, by craniotomy and cervicotomy	Pontage artériel extra-intracrânien sans autogreffe, par craniotomie et par cervicotomie
EBCA011	Extra-intracranial arterial bypass with autograft, by craniotomy and cervicotomy	Pontage artériel extra-intracrânien avec autogreffe, par craniotomie et par cervicotomie
EBCA013	Carotidohumeral or subclaviohumeral bypass, direct approach	Pontage carotidohuméral ou subclaviohuméral, par abord direct
EBCA014	Proximal carotidovertebral bypass or proximal subclaviovertebral bypass, by cervicotomy	Pontage carotidovertébral proximal ou subclaviovertébral proximal, par cervicotomie
EBCA015	Aortocarotid bypass, cervicotomy and thoracotomy	Pontage aortocarotidien, par cervicotomie et par thoracotomie
EBCA017	Bypass surgery between the common carotid artery and the ipsilateral internal carotid artery, by cervicotomy	Pontage entre l'artère carotide commune et l'artère carotide interne homolatérale, par cervicotomie
EBEA002	Reimplantation of the proximal vertebral artery into the subclavian artery or the common carotid artery by cervicotomy	Réimplantation de l'artère vertébrale proximale dans l'artère subclavière ou dans l'artère carotide commune, par cervicotomie
EBEA003	Reimplantation of the subclavian artery into the common carotid artery by cervicotomy	Réimplantation de l'artère subclavière dans l'artère carotide commune, par cervicotomie
EBEA004	Reimplantation of the distal vertebral artery into the internal or external carotid artery by cervicotomy	Réimplantation de l'artère vertébrale distale dans l'artère carotide interne ou dans l'artère carotide externe, par cervicotomie
EBEA005	Reimplantation of the common carotid artery into the subclavian artery by cervicotomy	Réimplantation de l'artère carotide commune dans l'artère subclavière, par cervicotomie
EBFA014	Resection of the internal carotid artery with re-implantation into the common carotid artery, by cervicotomy	Résection de l'artère carotide interne avec réimplantation dans l'artère carotide commune, par cervicotomie
EBFA018	Resection-anastomosis or replacement of the proximal vertebral artery, by cervicotomy	Résection-anastomose ou remplacement de l'artère vertébrale proximale, par cervicotomie
EBFA019	Resection-anastomosis of the internal carotid artery, by cervicotomy	Résection-anastomose de l'artère carotide interne, par cervicotomie
EBFA020	Resection-anastomosis of the common carotid artery, by cervicotomy and thoracotomy	Résection-anastomose de l'artère carotide commune, par cervicotomie et par thoracotomie
EBFA021	Resection-anastomosis of the common carotid artery, by cervicotomy	Résection-anastomose de l'artère carotide commune, par cervicotomie
EBKA001	Replacement of the common carotid artery, by cervicotomy	Remplacement de l'artère carotide commune, par cervicotomie
EBKA002	Replacement of the internal carotid artery, by cervicotomy and/or thoractomy	Remplacement de l'artère carotide interne, par cervicotomie et/ou par thoractomie
EBKA003	Common carotid artery replacement, by cervicotomy and thoractomy	Remplacement de l'artère carotide commune, par cervicotomie et par thoractomie
EBKA004	Replacement of the carotid bifurcation or the extracranial internal carotid artery, by cervicotomy	Remplacement de la bifurcation carotidienne ou de l'artère carotid interne extracrânienne, par cervicotomie

Codes	English	French
CCAM	Others/indistinct codes	Autres codes/Codages aspécifiques
ECAF003	Intraluminal dilatation of the brachiocephalic trunk or the intrathoracic common carotid artery without stenting, by transcutaneous arterial approach	Dilatation intraluminale du tronc artériel brachiocéphalique ou de l'artère carotide commune intrathoracique sans pose d'endoprothèse, par voie artérielle transcutanée
ECAF004	Intraluminal dilatation of the brachiocephalic trunk or the intrathoracic common carotid artery with stenting, by transcutaneous arterial approach	Dilatation intraluminale du tronc artériel brachiocéphalique ou de l'artère carotide commune intrathoracique avec pose d'endoprothèse, par voie artérielle transcutanée
ECLF004	Covered stenting of the brachiocephalic trunk or the intrathoracic common carotid artery by transcutaneous arterial access	Pose d'endoprothèse couverte dans le tronc artériel brachiocéphalique ou l'artère carotide commune intrathoracique, par voie artérielle transcutanée
ECPF004	Recanalization of the subclavian artery upstream of the vertebral artery ostium without stenting, by transcutaneous arterial approach	Recanalisation de l'artère subclavière en amont de l'ostium de l'artère vertébrale sans pose d'endoprothèse, par voie artérielle transcutanée
ECPF005	Recanalization of the subclavian artery upstream of the vertebral artery ostium with stenting, by transcutaneous arterial approach	Recanalisation de l'artère subclavière en amont de l'ostium de l'artère vertébrale avec pose d'endoprothèse, par voie artérielle transcutanée
ECFA001	Endarterectomy of the brachiocephalic trunk, by thoracotomy	Thromboendartériectomie du tronc artériel brachiocéphalique, par thoracotomie
EAAF900	Intraluminal branch dilatation of internal carotid artery branch with stenting by transcutaneous arterial approach	Dilatation intraluminale de branche de l'artère carotide interne avec pose d'endoprothèse, par voie artérielle transcutanée
EAAF901	Intraluminal branch dilatation of the internal carotid artery without stenting by transcutaneous arterial approach	Dilatation intraluminale de branche de l'artère carotide interne sans pose d'endoprothèse, par voie artérielle transcutanée
EBCA001	Carotidosubclavian or carotidoaxillary cross-bypass, by cervicotomy	Pontage croisé carotidosubclavier ou carotidoaxillaire, par cervicotomie
EASF010	Occlusion of an intracranial arterial saccular aneurysm during an acute haemorrhage, by transcutaneous arterial approach	Oblitération d'un anévrisme sacculaire artériel intracrânien en période aigüe hémorragique, par voie artérielle transcutanée
AAJA004	Evacuation of non-traumatic intracerebral haematoma by craniotomy	Évacuation d'hématome intracérébral non traumatique, par craniotomie
AAJH004	Transcranial evacuation of non-traumatic intracerebral haematoma with radiological guidance	Évacuation d'hématome intracérébral non traumatique, par voie transcrânienne avec guidage radiologique
AAJH002	Transcranial evacuation of non-traumatic intracerebral haematoma with CT guidance	Évacuation d'hématome intracérébral non traumatique, par voie transcrânienne avec guidage scanographique
EASF007	Intraluminal occlusion of an intracranial artery harboring an aneurysm during acute haemorrhage by transcutaneous arterial approach	Oblitération intraluminale d'une artère intracrânienne porteuse d'un anévrisme en période aigüe hémorragique, par voie artérielle transcutanée
EASF013	Occlusion of several intracranial arterial saccular aneurysms in acute haemorrhage, by transcutaneous arterial approach	Oblitération de plusieurs anévrismes sacculaires artériels intracrâniens en période aigüe hémorragique, par voie artérielle transcutanée
ABJA001	Evacuation of cerebral intraventricular haemorrhage by craniotomy	Évacuation d'une hémorragie intraventriculaire cérébrale, par craniotomie
ABJC900	Evacuation of non-traumatic cerebral intraventricular haemorrhage by video surgery	Évacuation d'une hémorragie intraventriculaire cérébrale non traumatique, par vidéochirurgie

5. Heart failure

Codes	English	French
ICD-10		
I50	Heart failure	Insuffisance cardiaque
I11.0	Hypertensive heart disease with (congestive) heart failure	Cardiopathie hypertensive avec IC
I13.0	Hypertensive heart and renal disease with (congestive) heart failure	Cardionéphropathie hypertensive, avec IC
I13.2	Hypertensive heart and renal disease with both (congestive) heart failure and renal failure	Cardionéphropathie hypertensive, avec IC et insuffisance rénale
I13.9	Hypertensive heart and renal disease, unspecified	Cardionéphropathie hypertensive, sans précision
R57.0	Cardiogenic shock	Choc cardiogénique
GHM		
05M09	Heart failure and circulatory shock	Insuffisance cardiaque et états de choc circulatoire
ATC		
C03DA04	Eplerenone	Eplérénone
C09DX04	Sacubitril (associated with valsartan)	Sacubitril (associated with valsartan)

6. Peripheral artery disease

Codes	English	French
LTD n°3	Lower limb arteriopathy	Artériopathie chronique avec manifestations ischémiques
ICD-10		
I70.2	Atherosclerosis of arteries of extremities	Athérosclérose des artères distales
I73.9	Peripheral vascular disease, unspecified	Maladie vasculaire périphérique, sans précision
I74.0	Embolism and thrombosis of abdominal aorta	Embolie et thrombose de l'aorte abdominale
I74.3	Embolism and thrombosis of arteries of lower extremities	Embolie et thrombose des artères distales, sans précision
I74.4	Embolism and thrombosis of arteries of extremities, unspecified	Embolie et thrombose des artères distales, sans précision
I74.5	Embolism and thrombosis of iliac artery	Embolie et thrombose de l'artère iliaque
I79.2	Peripheral angiopathy in diseases classified elsewhere	Angiopathie périphérique au cours de maladie classée ailleurs : angiopathie périphérique diabétique
CCAM		
EEAF001	Intraluminal dilation of several lower limb artery without stent, using percutaneous transluminal angioplasty	Dilatation intraluminale de plusieurs artères du membre inférieur sans pose d'endoprothèse, par voie artérielle transcutanée
EEAF002	Intraluminal dilation of a lower limb artery with dilation of the common iliac artery and/or of the ipsilateral external iliac artery, with stent, using percutaneous transluminal angioplasty	Dilatation intraluminale d'une artère du membre inférieur avec dilatation intraluminale de l'artère iliaque commune et/ou de l'artère iliaque externe homolatérale avec pose d'endoprothèse, par voie artérielle transcutanée
EEAF003	Intraluminal dilation of a lower limb artery without stent, using percutaneous transluminal angioplasty	Dilatation intraluminale d'une artère du membre inférieur sans pose d'endoprothèse, par voie artérielle transcutanée
EEAF004	Intraluminal dilation of a lower limb artery with stent, using percutaneous transluminal angioplasty	Dilatation intraluminale d'une artère du membre inférieur avec pose d'endoprothèse, par voie artérielle transcutanée
EEAF005	Intraluminal dilation of a lower limb artery with dilation of the common iliac artery and/or of the ipsilateral external iliac artery, without stent, using percutaneous transluminal angioplasty	Dilatation intraluminale d'une artère du membre inférieur avec dilatation intraluminale de l'artère iliaque commune et/ou de l'artère iliaque externe homolatérale sans pose d'endoprothèse, par voie artérielle transcutanée
EEAF006	Intraluminal dilation of several lower limb artery with stent, using percutaneous transluminal angioplasty	Dilatation intraluminale de plusieurs artères du membre inférieur avec pose d'endoprothèse, par voie artérielle transcutanée
EELF002	Covered stent placement in a lower limb artery, using percutaneous transluminal angioplasty	Pose d'endoprothèse couverte dans une artère du membre inférieur, par voie artérielle transcutanée
EEPF001	Recanalization of one lower limb artery with stent, percutaneous transluminal angioplasty	Recanalisation d'une artère du membre inférieur avec pose d'endoprothèse, par voie artérielle transcutanée
EEPF002	Recanalization of one lower limb artery without stent, percutaneous transluminal angioplasty	Recanalisation d'une artère du membre inférieur sans pose d'endoprothèse, par voie artérielle transcutanée

Codes	English	French
CCAM	**Iliac artery procedure**	**Chirurgie de l'artère iliaque**
DGPF002	Recanalisation of the aortic bifurcation, with stent, using percutaneous transluminal angioplasty	Recanalisation de la bifurcation aortique avec pose d'endoprothèse, par voie artérielle transcutanée bilatérale
EDAF002	Intraluminal dilation of the common iliac artery and/or of the external iliac artery, without stent, using percutaneous transluminal angioplasty	Dilatation intraluminale de l'artère iliaque commune et/ou de l'artère iliaque externe sans pose d'endoprothèse, par voie artérielle transcutanée
EDAF003	Intraluminal dilation of the common iliac artery and/or of the external iliac artery, with stent, using percutaneous transluminal angioplasty	Dilatation intraluminale de l'artère iliaque commune et/ou de l'artère iliaque externe avec pose d'endoprothèse, par voie artérielle transcutanée
EDAF004	Intraluminal dilation of the internal iliac artery, without stent	Dilatation intraluminale de l'artère iliaque interne sans pose d'endoprothèse, par voie artérielle transcutanée
EDAF006	Intraluminal dilation of the internal iliac artery, with stent	Dilatation intraluminale de l'artère iliaque interne avec pose d'endoprothèse, par voie artérielle transcutanée
EDLF004	Stenting (drug-eluted stent) of the internal iliac artery and/or of the external iliac artery, with embolization of the internal iliac artery	Pose d'endoprothèse couverte dans l'artère iliaque commune et/ou l'artère iliaque externe avec embolisation de l'artère iliaque interne, par voie artérielle transcutanée
EDLF005	Stenting (drug-eluted stent) of the internal iliac artery and/or of the external iliac artery	Pose d'endoprothèse couverte iliaque par voie artérielle transcutanée
EDPF001	Recanalisation of the internal iliac artery, with stent	Recanalisation de l'artère iliaque interne avec pose d'endoprothèse, par voie artérielle transcutanée
EDPF006	Recanalisation of the internal iliac artery or of the external iliac artery, with drug-eluting stent	Recanalisation de l'artère iliaque commune et/ou de l'artère iliaque externe avec pose d'endoprothèse couverte, par voie artérielle transcutanée
EDPF007	Recanalisation of the internal iliac artery, without stent	Recanalisation de l'artère iliaque interne sans pose d'endoprothèse, par voie artérielle transcutanée
EDPF008	Recanalisation of the internal iliac artery or of the external iliac artery, without stent	Recanalisation de l'artère iliaque commune et/ou de l'artère iliaque externe sans pose d'endoprothèse, par voie artérielle transcutanée
EDPF009	Recanalisation of the internal iliac artery or of the external iliac artery, with stent	Recanalisation de l'artère iliaque commune et/ou de l'artère iliaque externe avec pose d'endoprothèse, par voie artérielle transcutanée
CCAM	**Others codes associated with PAD**	Autres codes liés à l'artériopathie oblitérantes des membres inférieurs
DGFA001	Endarterectomy of the trunk of the abdominal aorta, by laparotomy	Thromboendartériectomie du tronc de l'aorte abdominale, par laparotomie
DGFA003	Aortobisilic endarterectomy, by laparotomy	Thromboendartériectomie aortobisiliaque, par laparotomie
DGFA004	Thrombectomy of the abdominal aorta, common iliac artery and/or external iliac artery through bilateral inguinofemoral approach	Thrombectomie de l'aorte abdominale, de l'artère iliaque commune et/ou de l'artère iliaque externe, par abord inguinofémoral bilatéral

7. Lower limb amputation

Codes	English	French
CCAM		
NZFA002	Transtibial amputation	Amputation transtibiale
NZFA004	Amputation or disarticulation of several toes	Amputation ou désarticulation de plusieurs orteils
NZFA005	Amputation or disarticulation of the midfoot or of the forefoot, without stabilisation of the hindfoot	Amputation ou désarticulation au médiopied ou à l'avant-pied, sans stabilisation de l'arrière-pied
NZFA006	Amputation or disarticulation of the lower limb through thigh bone, sacroiliac joint or sacrum	Désarticulation ou amputation du membre inférieur à travers l'os coxal, l'articulation sacro-iliaque ou le sacrum
NZFA007	Transfemoral amputation	Amputation transfémorale
NZFA009	Amputation or disarticulation of the ankle or of the hindfoot	Amputation ou désarticulation à la cheville ou à l'arrière-pied
NZFA010	Amputation or disarticulation of one toe	Amputation ou désarticulation d'un orteil
NZFA013	Amputation or disarticulation of the midfoot or of the forefoot, with stabilisation of the hindfoot	Amputation ou désarticulation du médiopied ou de l'avant-pied, avec stabilisation de l'arrière-pied

8. Cancer – Malignant tumours

Codes	English	French
LTD n°30	Cancer	Cancer
ICD-10 codes		
C	Malignant neoplasms	Néoplasie maligne
D37 to D48	Neoplasms of uncertain and unknown behaviour	Néoplasie d'évolution incertaine ou inconnue

9. Chronic kidney disease

Codes	English	French
LTD n°19	Chronic kidney disease	Maladie rénale chronique
ICD-10 codes		
E10.2, E11.2, E12.2, E13.2, E14.2	Diabetes with renal complication	Diabète sucré avec complication rénale
I12	Hypertensive nephropathy	Néphropathie hypertensive
I13	Hypertensive cardio nephropathy	Cardionéphropathie hypertensive
N08.3	Diabetic nephropathy	Néphropathie diabétique
N18	Chronic kidney disease	Maladie rénale chronique

10. End stage kidney disease (ESKD)

Codes	English	French
ICD-10 codes		
Z94.0	Kidney transplant status	Transplanté rénal
T86.1	Failure and rejection of transplanted kidney	Echec/rejet de transplant rénal
CCAM codes		
JAEA003	Kidney graft	Greffe rénale
HNEA003	Double Kidney-pancreas graft	Double greffe rein-pancréas
GHM codes		
27C06	Kidney graft	Greffe rénale
11M17	Monitoring of kidney graft	Suivi de greffe rénale
11K02, 28Z01Z to 28Z06Z	Training for peritoneal dialysis or haemodialysis or haemodialysis session	Session d'entraînement à la dialyse péritonéale ou à l'hémodialyse, ou session de dialyse

11. Bariatric surgery

Codes	English	French
CCAM	**Gastric bypass**	**Court-circuit gastrique**
HFCA001	Gastric bypass with roux-en-y for morbid obesity, by laparotomy	Court-circuit gastrique avec anse montée en Y [Bypass gastrique en Y] pour obésité morbide, par laparotomie
HFCC003	Gastric bypass with roux-en-y for morbid obesity, by laparoscopy	Court-circuit gastrique avec anse montée en Y [Bypass gastrique en Y] pour obésité morbide, par coelioscopie
CCAM	**Sleeve-gastrectomy**	**Sleeve-gastrectomie**
HFFA001	Gastrectomy with biliopancreatic or intestinal bypass for morbid obesity, by laparotomy	Gastrectomie avec court-circuit biliopancréatique ou intestinal pour obésité morbide, par laparotomie
HFFA011	Longitudinal gastrectomy [Sleeve gastrectomy] for morbid obesity, by laparotomy	Gastrectomie longitudinale [Sleeve gastrectomy] pour obésité morbide, par laparotomie
HFFC004	Laparoscopic gastrectomy with biliopancreatic or intestinal bypass for morbid obesity, by laparoscopy	Gastrectomie avec court-circuit biliopancréatique ou intestinal pour obésité morbide, par coelioscopie
HFFC018	Longitudinal gastrectomy [Sleeve gastrectomy] for morbid obesity, by laparoscopy	Gastrectomie longitudinale [Sleeve gastrectomy] pour obésité morbide, par coelioscopie

12. Foot ulcer

Codes	English	French
ICD-10 codes		
M86.07	Acute haematogenous osteomyelitis, ankle and foot	Ostéomyélite aiguë hématogène, Cheville et pied
M86.17	Other acute osteomyelitis, ankle and foot	Autres ostéomyélites aiguës, Cheville et pied
M86.27	Subacute osteomyelitis, ankle and foot	Ostéomyélite subaiguë, Cheville et pied
M86.37	Chronic multifocal osteomyelitis, ankle and foot	Ostéomyélite multifocale chronique, Cheville et pied
M86.47	Chronic osteomyelitis with draining sinus, ankle and foot	Ostéomyélite chronique avec fistule de drainage, Cheville et pied
M86.57	Other chronic haematogenous osteomyelitis, ankle and foot	Autres ostéomyélites chroniques hématogènes, Cheville et pied
M86.67	Other chronic osteomyelitis, ankle and foot	Autres ostéomyélites chroniques, Cheville et pied
M86.87	Other osteomyelitis, unspecified, ankle and foot	Autres ostéomyélites, Cheville et pied
M86.97	Osteomyelitis, unspecified, ankle and foot	Ostéomyélites, non spécifiées, Cheville et pied
S90	Superficial injury of ankle, foot and toes	Lésion traumatique superficielle, cheville et pied
S91	Open wound of ankle, foot and toes	Plaie ouverte, cheville et pied
L97	Non-pressure chronic ulcer of lower limb, not elsewhere classified	Ulcère du membre inférieur, non classé ailleurs

13. Diabetes neuropathy

Codes	English	French
ICD-10 codes		
E10.4, E11.4, E12.4, E13.4, E14.4	Diabetic neuropathy	Diabète avec neuropathie
G59.0	Diabetic mononeuropathy	Mononévrite diabétique
G63.2	Diabetic polyneuropathy	Polynévrite diabétique
G99.0	Autonomic neuropathy in diseases classified elsewhere	Neuropathie autonome dans les maladies endocrines et métaboliques (dont neuropathie amyloïde et diabétique)

14. Serious retinal events (SRE)

Codes	English	French
ICD-10		
H33.4	Tractional retinal detachment	Décollement par traction de la rétine
H43.1	Vitreal haemorrhage	Hémorragie du corps vitré
H45.0	Vitreal haemorrhage in diseases classified elsewhere	Hémorragie du corps vitré au cours de maladies classées ailleurs
CCAM	**Laser procedures**	**Procédures lasers**
BGNP001	Chorioretinal photocoagulation of the posterior pole, with monochromatic laser or dye laser	Séance de photocoagulation choriorétinienne du pôle postérieur, avec laser monochromatique ou laser à colorants
BGNP003	Chorioretinal lesion destruction session by laser photocoagulation, using a contact lens	Séance de destruction de lésion choriorétinienne par photocoagulation avec laser, à l'aide de verre de contact
BGNP004	Chorioretinal lesion destruction session by transpupillary photocoagulation with laser, using a contact lens	Séance de destruction de lésion choriorétinienne par photocoagulation transpupillaire avec laser, à l'aide de verre de contact
BGNP007	Chorioretinal lesion destruction session by laser photocoagulation, using indirect ophthalmoscope	Séance de destruction de lésion choriorétinienne par photocoagulation avec laser, à l'aide d'ophtalmoscope indirect
BGNP008	Chorioretinal photocoagulation of the posterior pole (for the management of diabetic macular edema), with argon or diode laser	Séance de photocoagulation choriorétinienne du pôle postérieur (pour prise en charge de l'œdème maculaire diabétique), avec laser à argon ou diode
CCAM	**Intravitreal injection**	**Injection intravitréenne**
BGLB001	njection of pharmacological agent into the vitreal	Injection d'agent pharmacologique dans le corps vitré
ATC	**Drugs associated with intravitreal injection**	Médicaments associés à l'injection intravitréenne
S01BA01	Dexamethasone implant	Implant de dexaméthasone
S01LA04	Ranibizumab	Ranibizumab
S01LA05	Aflibercept	Aflibercept
S01BA15	Fluocinolone acetonid implant	Implant d'acétate de fluocinolone
L01CX07	Bevacizumab	Bevacizumab

PART II. ATC codes used for drug delivery identifications

Class	Corresponding ATC codes
Diabetes treatment	A10, except for A10BX06 (benfluorex)
Metformin	- Monotherapy: A10BA02 (monotherapy) - Combinations: A10BD02, A10BD07, A10BD08, A10BD10
Sulfonylurea and/or repaglinide	- Monotherapy: A10BB01, A10BB03, A10BB04, A10BB06, A10BB07, A10BB09, A10BB12, A10BX02 - Combinations: A10BD02
DPP4-inhibitors	- Monotherapy: A10BH01, A10BH02, A10BH03 - Combinations: A10BD07, A10BD08, A10BD10
GLP-1 receptor agonist	- Monotherapy: A10BJ - Combinations: A10AE56
Insulin	A10A
Anti-hypertensive treatment	C02, C03, C07, C08, C09
Beta blockers	- Monotherapy: C07 - Combinations: C09BX02, C09BX04, C09BX05, C09DX05
Calcium-channel blockers	- Monotherapy: C08 - Combinations: C07FB, C09BB, C09DB, C09BX, and C10BX03
ACE-inhibitors	- Monotherapy: C09AA - Combinations: C09BA, C009BB, C09BX
ARBs	- Monotherapy: C09CA - Combinations: C07BA, C09DA, C09DB, C09DX04
Loop diuretics	- Monotherapy: C03CA - Combinations: C03EB
Thiazide diuretics	- Monotherapy: C03A (monotherapy) - Combinations: C03AB, C03AH, C03AX, C02LA01, C03EA01, C03EA, C07BB07, C07BB12, C07DA06, C09BA, C09DA, C09XA52
Potassium-sparing diuretic	- Monotherapy: C03D - Combinations: C03E
Others cardiovascular treatments	
Antiplatelet agents	- Monotherapy: B01AC04, B01AC06, B01AC07, B01AC22, B01AC24 - Combinations: B01AC30
Anticoagulants	- Heparin: B01AB - Vitamin K antagonists: B01AA - Direct thrombin inhibitors B01AE05, B01AE07, B01AF01, B01AF02
Statins	- Monotherapy: C10AA - Combinations: C10BA (except for C10BA10), C10BX
Ezetimibe	- Monotherapy: C10AX09 - Combinations: C10BA02, C10BA05, C10BA06

The following drugs were not examined here:

- Glitazones (thiazolidinediones), which were removed from French market in 2011
- Gliptin (SGLT2-inhibitors), reimbursed in France only since 2020

The international ATC classification can easily be queried online on https://www.whocc.no/atc_ddd_index (last accessed 30rd September, 2022)

STROBE Statement—Checklist of items that should be included in reports of *cohort studies*

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	p. 1-2
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	p. 3-4
Objectives	3	State specific objectives, including any prespecified hypotheses	p. 3-4
Methods			
Study design	4	Present key elements of study design early in the paper	p. 3-4 and fig 1
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	p. 4-6 and fig 1
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	p. 5-6 and fig 1
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	p. 6-7 and S. file 1
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	p. 4, 6 and Sup. file 1
Bias	9	Describe any efforts to address potential sources of bias	p. 7
Study size	10	Explain how the study size was arrived at	p. 8 and Sup. fig 1 (flowchart)
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	p. 6
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	p. 6-8
		(*b*) Describe any methods used to examine subgroups and interactions	
		(*c*) Explain how missing data were addressed	
		(*d*) If applicable, explain how loss to follow-up was addressed	
		(*e*) Describe any sensitivity analyses	

Results			Page No
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	p. 8 and Sup. fig 1 (flowchart)
		(b) Give reasons for non-participation at each stage	
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	p. 8-9, Tables 1-2 and Sup. table 1
		(b) Indicate number of participants with missing data for each variable of interest	N/A
		(c) Summarise follow-up time (eg, average and total amount)	Table 2, fig 2 and Sup. fig 2
Outcome data	15*	Report numbers of outcome events or summary measures over time	Table 2, fig 2 and Sup. fig 2
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	Tables 2-3
		(*b*) Report category boundaries when continuous variables were categorized	Table 1, Sup. table 1
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	Fig 2
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Fig 2, Sup. fig 2
Discussion			
Key results	18	Summarise key results with reference to study objectives	p. 10
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	p. 13-14
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	p. 11-13
Generalisability	21	Discuss the generalisability (external validity) of the study results	p. 11 and 13
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	N/A

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

PARTIE V : GAVROCHE - VARIABILITE GLYCEMIQUE ET MORTALITE LORS D'UNE HOSPITALISATION POUR INSUFFISANCE CARDIAQUE AIGUE

1. Résumé

Le projet GAVROCHE s'intéresse à l'association entre la variabilité glycémique des premiers jours suivant l'admission et le décès chez des patients hospitalisés pour insuffisance cardiaque aiguë.

Des trois pans de ce travail de thèse, c'est le seul à ne pas être proposé sous forme d'article. En effet, ce travail n'était pas assez abouti pour justifier cette forme. Plus qu'une tentative de réponse à une question scientifique formelle, GAVROCHE est proposé ici comme une illustration de la conduite d'un projet multicentrique inter-régional à partir des entrepôts de données de santé (EDS), avec un accent particulier sur les flux de données, la validation des cas et la mise en place du TALN.

Les données ciblées sont issues des EDS du Grand Ouest, soit cinq CHU du groupe HUGO - Angers, Brest, Nantes, Rennes et Tours. Un enjeu majeur dans la structuration de ces EDS est d'être en mesure de transformer et rendre accessibles des données issues du soin afin qu'elles soient exploitables pour la recherche.

Pour certaines données, le principal besoin réside dans leur accessibilité, car le niveau de qualification nécessaire au soin ou aux procédures administratives les rend *a priori* directement exploitables à des fins de recherche. C'est le cas notamment des données biologiques standardisées par le travail des laboratoires d'analyse et structurées pour être interprétables directement par le clinicien. C'est aussi le cas de certaines données cliniques pour lesquelles un champ spécifique est dévolu dans le SIH (âge, sexe, poids, taille, température corporelle, etc.) ou encore de certaines données issues du PMSI, comme le code CCAM nous donnant l'information qu'un patient a bénéficié de tel examen ou de tel geste chirurgical.

Mais pour d'autres données, comme celles issues du texte des comptes rendus hospitaliers (CRH), l'accessibilité est une condition nécessaire mais non suffisante pour qualifier la donnée à grande échelle. Pour un projet s'intéressant à une maladie rare sur quelques dizaines de patients, le travail humain de recueil, qui va consister à lire chaque CR pour en extraire certaines informations, est suffisant, et volontiers plus efficace que le traitement automatisé. Un être humain exécutera un travail non seulement de meilleure qualité, mais plus rapidement. Mais pour un projet à plus grande échelle comme GAVROCHE, qui porte sur 5 centres comptant chacun plusieurs milliers de patients, nous allons devoir recourir à une méthode automatisée d'extraction d'information à partir des CRH : le traitement automatique du langage naturel, ou TALN.

Dans le cas de GAVROCHE, la grande majorité des données brutes est bien accessible dans le SIH, et les relations qui les unissent sont explicitées par un schéma relationnel immédiatement interprétable et presque identique entre chaque centre grâce à un déploiement de concert autour d'un logiciel commun, « eHOP ». Les principales contraintes sont donc ailleurs :

- **Pour les données structurées (PMSI, biologie) :** la qualification à finalité recherche est incomplète et va surtout nécessiter un travail de contrôle (pour les codes CIM-10 et GHM) et d'uniformisation (dosages en laboratoire)
- **Données non structurées :** l'information est présente dans le texte des CRH mais inaccessible à grande échelle sans mettre en œuvre le TALN, et l'expertise d'un ingénieur s'avère indispensable
- **Aspect réglementaire et accès :** les droits et la capacité à accéder et à traiter ces données « sauvages » issues du soin requièrent des démarches réglementaires importantes (CSE inter-régional, CESREES et CNIL au niveau national, dont la rédaction d'un PIA), dans le contexte d'un réseau RiCDC (Réseau inter-régional des centres de données cliniques) n'ayant pas encore d'expérience commune de la recherche, et la nécessité d'exporter les données des 5 centres vers une plateforme unique, l'ODH pour *Ouest Data Hub*

Dans cette partie, je décrirai d'abord les grandes lignes du déroulement du travail et les contributions respectives, avant d'exposer brièvement le rationnel scientifique du projet. Puis, afin de permettre une première vue d'ensemble, je donnerai les dates clefs, ce qui a été réalisé et doit encore l'être, avant de présenter les démarches scientifiques, éthiques et réglementaires. Je reviendrai ensuite au projet GAVROCHE lui-même, la population éligible et les données nécessaires, en distinguant données structurées et non structurées, le circuit des données local et vers l'ODH, puis je décrirai la mise en place du TALN et en particulier la phase d'annotation. Enfin, je discuterai la qualité des données, les analyses statistiques prévues, et conclurai sur un état des lieux de GAVROCHE.

2. Déroulement et contributions respectives

a. Financement et autorisations réglementaires

Du fait du caractère pionnier du projet GAVROCHE et bien que s'agissant d'une étude rétrospective sur données, le travail réglementaire a été considérable. Le lancement du projet a été assuré par

l'obtention de l'AAP-GIRCI Grand Ouest en septembre 2019[15], pour un budget de 131 000€. Ce montant a permis de financer les ressources humaines nécessaires au projet (CDC pour 66% et TALN pour 22% du budget), ainsi que le travail juridique et les frais annexes (12%). La compétence TALN, incarnée par Adrien Bazoge, étudiant ingénieur en alternance professionnelle puis doctorant au LS2N, a pu être pérennisée par l'obtention d'un co-financement 50/50 par une bourse du projet ANR AIBy4 (ANR-20-THIA-0011) et par le CHU de Nantes via le budget de la Clinique des Données, en particulier grâce à un financement supplémentaire de 50 000€ obtenu auprès du laboratoire AstraZeneca.

Le projet a été présenté successivement à l'échelle inter-régionale au CSE (Conseil Scientifique et Ethique du groupe HUGO, cf. **Annexe 6** pour l'avis rendu par le CSE), et à l'échelle nationale au CESREES (cf. **Annexe 7**) puis à la CNIL (cf. **Annexe 8**), cette dernière étape ayant requis sept mois d'instruction. Finalement, la convention inter-établissement et son avenant ont été signés au 5 mai 2022. GAVROCHE est co-porté par le Pr Hadjadj et moi-même, et nous avons rédigé ensemble les principaux éléments des dossiers de soumission, avec l'aide notable :

- **Pour le budget de l'AAP GIRCI-GO 2019** : de Zeineb Lamoureux et David Lair (chefs de projet à la DRI du CHU de Nantes)

- **Pour le PIA demandé par la CNIL :** ce document est particulièrement complexe puisqu'il implique des garanties de protection de la vie privée tenant compte des différents acteurs, à savoir le CH porteur (CHU de Nantes), les CH participants (Angers, Brest, Rennes et Tours), les concepteurs de l'outil eHOP (personnel du LTSI/RiCDC), l'hébergeur de l'ODH (DSN du CHU de Nantes) ainsi que les ingénieurs en charge de l'ODH (personnel du LTSI/RiCDC). Brièvement, un premier modèle de PIA a été proposé par le RiCDC puis complété et actualisé suite aux échanges avec Marie Lebigre et Cédric Cartau (membre DRI et DPO recherche du CHU de Nantes), le Dr Christine Riou (DPO de l'ODH) et Pascal Van Hille (chef de projet LTSI - ODH), sous ma coordination.

- **Pour la convention inter-établissements,** entièrement rédigée par la cellule juridique du CHU de Nantes : Maxime Caillier et Philippe Boucher, sous la direction de Benoît Labarthe, et Emilie Varey pour le lien avec la chefferie de projet

[15] Cf. https://www.chu-hugo.fr/accueil/projets/Ouest-DataHub, dernière consultation le 11 septembre 2022

Des quatre projets retenus à l'AAP-GIRCI Grand Ouest 2019, le projet GAVROCHE a été pionnier pour chacune de ces étapes. Un calendrier plus détaillé est proposé **Figure 8**.

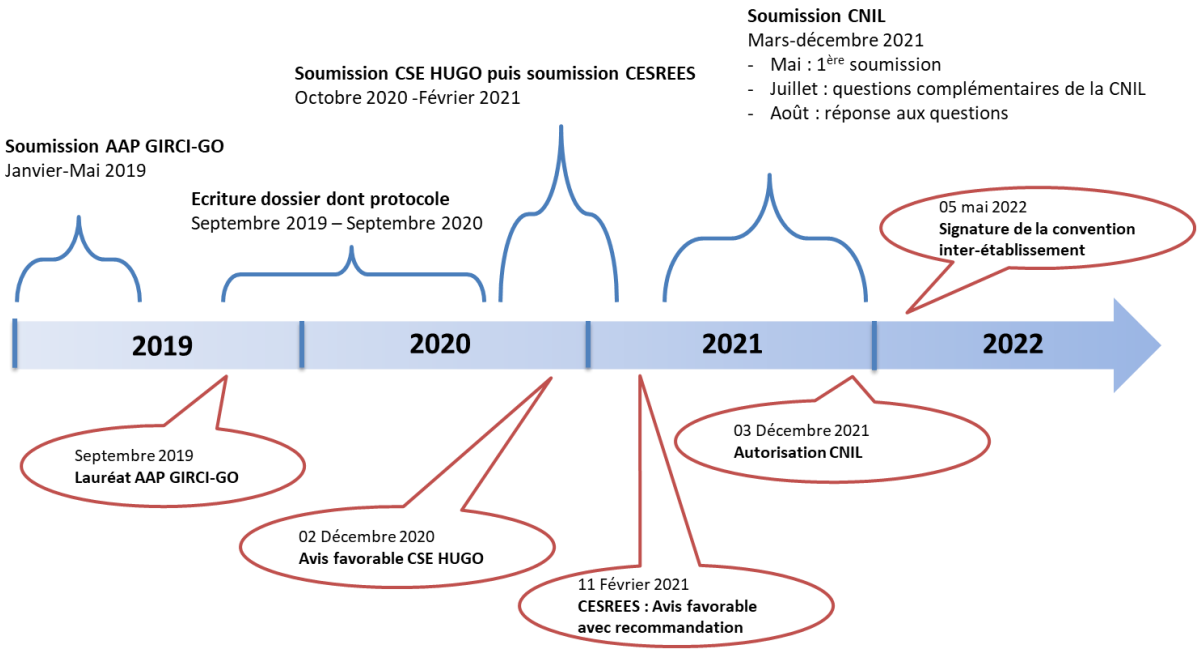


Figure 8. GAVROCHE : calendrier de la soumission à l'appel à projet et réglementaire.

AAP GIRCI-GO : Appel A Projet du Groupement Inter-régional de Recherche Clinique et d'Innovation du Grand Ouest. CNIL : Commission Nationale de l'Informatique et des Libertés. CSE : comité scientifique et éthique. HUGO : Hôpitaux Universitaires du Grand Ouest.

b. Etapes et acteurs du projet

Le protocole initial a évolué avec les soumissions successives CSE/CESREES/CNIL. L'essentiel du travail consiste à présent en l'identification de la population et des données, la validation du circuit des données et la mise en œuvre du TALN, qui s'appuient sur les compétences de Delphine Toublant (DSN, cheffe de projet entrepôt du CHU de Nantes), Adrien Bazoge (ingénieur TALN), des cliniciens associés à l'annotation, des autres CDC participants (en particulier Julien Herbert, sollicité pour déployer les outils d'annotation et répliquer les résultats au CHU de Tours) et de la coordination RiCDC par Pascal Van Hille.

3. Rationnel scientifique

a. Rationnel clinique et épidémiologique

L'épidémiologie de l'insuffisance cardiaque et de sa forme aiguë, en particulier nécessitant une hospitalisation, a été exposée plus tôt dans ce manuscrit (**section II.1.c**). L'ICA est une situation fréquente et grave. Optimiser la prise en charge des patients hospitalisés relève d'un enjeu majeur de santé publique. L'amélioration du pronostic s'appuie en particulier sur l'identification précoce des sujets les plus à risque. L'hyperglycémie à l'admission peut être un facteur de mauvais pronostic dans des situations cardio-vasculaires aiguës telles que le syndrome coronarien aigu [46] et l'accident vasculaire cérébral.[107] Mais la correction de l'hyperglycémie par insulinothérapie n'a pas permis d'améliorer le pronostic dans les essais randomisés [108] et le rapport bénéfice/risque a été jugé défavorable en raison des hypoglycémies induites. Des alternatives thérapeutiques existent pour obtenir, sans hypoglycémie, une amélioration non seulement de la glycémie mais aussi de la variabilité glycémique (VG), mais cette question thérapeutique est hors du champ d'investigation immédiat de ce projet.

Dans le cas de l'ICA, la relation entre la glycémie à l'admission et le pronostic n'est pas aussi bien établie. Ainsi, une large étude en population réelle aux USA n'a pas montré de lien entre hyperglycémie et pronostic, ni à 30 jours ni à 1 an.[45] A contrario, une étude de cohorte multicentrique internationale a retrouvé un sur-risque d'événement cardiovasculaire majeur positivement associé à l'hyperglycémie à l'admission.[46] Dans le contexte des patients de soins intensifs, il a été montré qu'une VG élevée était fortement et positivement associée à la mortalité.[109] De nouvelles données complémentaires récentes montrent que la VG est un facteur de mauvais pronostic dans le syndrome coronarien aigu chez le patient diabétique.[110] D'un point de vue physiopathologique, une faible VG pourrait correspondre à une situation favorable telle qu'une bonne flexibilité métabolique ou une activation adéquate du système nerveux autonome. L'intérêt pronostique de cette variabilité n'a cependant pas encore été étudié dans une large cohorte de patients présentant une ICA, ou hors contexte diabétique. A notre connaissance, seule une étude monocentrique a étudié le lien entre VG et pronostic de l'ICA, ne retrouvant pas d'intérêt pronostique évident pour la glycémie moyenne mais une augmentation de la mortalité en cas de forte VG.[111] La portée de cette étude est cependant limitée par le fait que la très grande majorité des patients étudiés présentaient un syndrome coronarien aigu, ce qui ne représente pas la réalité de l'ICA en France, comme le suggèrent par ailleurs les résultats de DMC (Tableau 1 de l'article proposé partie IV : 58,1% des patients connus pour une insuffisance cardiaque n'ont pas de coronaropathie associée).

De plus, si l'intérêt de la VG comme facteur pronostique est confirmé à partir de nos données, cela pourra justifier des études prospectives plus fines à partir des capteurs de mesure en continu du glucose (de type *flash monitoring*, capteurs FreeStyle Libre ou Dexcom G5).

b. L'enjeu des données

La thématique de l'ICA a déjà été abordée en France par les données massives grâce au SNDS, permettant une excellente représentativité nationale.[41] Mais cette approche reste actuellement limitée par le manque de données cliniques et la non-disponibilité des divers résultats d'examens, qu'ils soient biologiques, d'imagerie ou encore d'explorations fonctionnelles.

A contrario, le projet GAVROCHE se fonde sur l'accès à des données d'une granularité clinique fine, rendu possible au sein du RiCDC par l'outil eHOP. Nous aurons d'abord besoin d'identifier les patients éligibles, puis de recueillir des informations cliniques et biologiques les concernant pour l'analyse statistique qui s'appuiera sur certains facteurs de confusion ou d'interaction, notamment la présence du diabète et son équilibre évalué par l'HbA_{1c}, ou encore la fraction d'éjection ventriculaire gauche (FEVG).

L'identification de la population cible sera possible grâce à l'intégration des données locales issues du PMSI. Les informations biologiques ont été collectées sous forme structurée dans les entrepôts hospitaliers, grâce aux flux des logiciels de soin propres à chaque centre. La standardisation des différents laboratoires (accréditation COFRAC ISO 15189) permettra la transférabilité des résultats d'un centre à l'autre. Des informations cliniques seront récupérables par le PMSI (p. ex. : diagnostic associé de maladie coronaire ou de diabète), mais cette information est par nature parcellaire et insuffisante pour résumer les comorbidités du patient. L'une des difficultés de GAVROCHE réside donc dans l'extraction d'informations précises et de qualité homogène à partir des données non structurées proposées dans les documents liés au soin. Le codage systématique de ces informations par les soignants n'est pas envisageable à l'échelle de la cohorte, puisque cela correspondrait à plus de 20 000 dossiers. Atteindre notre objectif passe donc par l'automatisation de l'extraction des données par TALN.

Ce projet multicentrique permet ainsi :

(i) De par sa dimension, d'atteindre une taille de population permettant une grande précision des indicateurs y compris si le sur-risque associé est faible (« significativité statistique » pour des OR à 1,10 pour une augmentation d'un écart-type du paramètre évalué)

(ii) De par son caractère multicentrique, de tester la robustesse des observations : sous-analyses par centre, puis approche *leave-one-out* (itération des analyses en excluant chaque centre, un à un), ne visant pas à la significativité statistique mais au contrôle de la cohérence des résultats

(iii) De par le caractère transdisciplinaire de la question posée, de réunir des experts cliniciens (cardiologues, diabétologues), biologistes, DIM et méthodologistes, dont l'expertise et la perspective critique permettront d'améliorer l'usage des EDS, tout en les acculturant aux EDS pour des usages futurs

(iv) De par la diversité des informations rendues disponibles par eHOP et l'approche de TALN sur les comptes rendus, de prendre en compte les facteurs de confusion ou d'interaction jugés pertinents par les cliniciens : antécédent de diabète ou de coronaropathie, CRP, HbA_{1c}, peptides natriurétiques, etc.

4. Hypothèses et objectifs de l'étude

a. **Hypothèses et objectifs de recherche en épidémiologie clinique**

Nous faisons l'hypothèse que la VG a un rôle pronostique dans l'ICA.

Notre objectif principal est l'analyse du rôle pronostique de la VG sur la mortalité associée à l'ICA chez les sujets présentant ou non un diabète, VG définie ici par le coefficient de variation de la glycémie des 72 premières heures soit le rapport écart-type/moyenne, qui fait l'objet d'un consensus international publié par l'*American Diabetes Association* en 2017.[112]

Un objectif secondaire de l'étude est l'analyse du rôle pronostique de la 1ère glycémie à l'admission sur la mortalité. Les scores pronostiques dans l'ICA ne sont pas largement diffusés, à l'inverse de ce

qu'on retrouve dans la cardiopathie ischémique.[113,114] Dans un second temps, nous pourrons envisager le développement d'un score pronostique simple à partir des données de GAVROCHE.

Le critère de jugement principal associé à cet objectif correspond au résultat de la régression logistique multivariable détaillée dans la partie 10 – Plan d'analyses statistiques.

b. **Hypothèses de recherche sur données massives en santé**

Notre approche des données massives se fonde sur deux grandes hypothèses.

La première porte sur l'accès aux données et la gouvernance au sein du RiCDC, soit la possibilité technique pour chaque centre de constituer sa base de données propre à l'étude et de la partager au sein d'un consortium, en respectant l'éthique et la réglementation.

La seconde porte sur la possibilité d'extraction d'informations médicales structurées (ici binaires, telles qu'un antécédent de coronaropathie oui/non, ou quantitatives, telles que la fraction d'éjection ventriculaire gauche) à partir de CRH, par un algorithme de TALN construit de façon supervisée.

5. Périmètre (1) : Population

La population d'étude correspond à une population hospitalière de sujets présentant une ICA ayant nécessité une hospitalisation sur la période 2011-2019. Soulignons que l'unité statistique principale sera le séjour pour ICA, et non le patient, qui peut être associé à plusieurs séjours d'intérêt.

a. **Critères d'éligibilité des séjours d'intérêt**

Critères d'inclusion

- Femme ou homme d'âge ≥ 18 ans

- Hospitalisé(e) dans l'un des centres hospitaliers du groupe HUGO (Angers, Brest, Nantes, Rennes et Tours[16]) entre le 1er janvier 2011 et le 31 décembre 2019, dans la limite des données disponibles dans les EDS de chaque centre
- Dont le séjour hospitalier est associé dans le PMSI à un code GHM commençant par « 05M09 » (insuffisance cardiaque - état de choc circulatoire) et/ou à un code CIM-10 d'insuffisance cardiaque (I50, I11.0, I13.0 I13.2, I13.9 et R570) en diagnostic principal ou relié
- Ayant au moins un CRH, une lettre de sortie ou une lettre de synthèse associée au séjour dans l'EDS (ou tout document jugé équivalent par le personnel du CDC local)
- Ayant au moins une valeur de glycémie disponible dans les 24h suivant l'admission. Ce critère est nécessaire pour l'objectif secondaire, même si 3 glycémies sont requises pour le calcul de la variabilité glycémique et donc pour l'objectif principal
- Personnes dont le traitement des données à des fins de recherche dans l'EDS local a été autorisé, dans les limites approuvées par la CNIL

Critères de non-inclusion

- Séjour sans nuitée et non soldé par le décès du patient (remarque : cet ajout est postérieur à la soumission CNIL, afin d'exclure les hospitalisations de jour, non identifiables aisément dans le SI)
- Personne ayant manifesté sa volonté de ne pas voir ses données utilisées à des fins de recherche par l'établissement, ou pour le projet GAVROCHE spécifiquement

b. Justification de l'identification des patients par le critère combiné GHM et/ou CIM-10

Nous avons choisi de réaliser un screening combinant une approche GHM et une approche CIM-10, bien que les experts DIM/PMSI aient souligné le caractère plus spécifique du GHM sur le CIM-10 pour identifier les patients d'intérêt, c'est-à-dire hospitalisés pour ICA. En appui de cet avis d'expert, l'expérience de l'annotation a rapidement confirmé que les hospitalisations identifiées uniquement

[16] L'Institut de Cancérologie de l'Ouest où eHOP est également déployé n'a pas été associé au projet, considérant que les hospitalisations aiguës pour insuffisance cardiaque étaient plus marginales dans ce centre

par le code CIM-10, sans code GHM d'insuffisance cardiaque, étaient moins spécifiques de l'ICA que celles identifiées au moins par GHM.

Cependant, (i) l'expérience a montré que le GHM est loin d'être spécifique, (ii) les patients inclus en excès sont supposés être exclus ensuite par l'approche TALN, (iii) si l'efficacité de l'approche TALN est jugée insuffisante, nous pourrons faire le choix d'exclure dans un second temps les patients identifiés uniquement sur code CIM-10, et (iv) il est probable que les pratiques de codage ne soient pas tout à fait homogènes d'un centre à l'autre. En particulier, pour certains centres comme Nantes, sur la période d'intérêt, ce sont les cliniciens qui codaient les diagnostics pour le PMSI, tandis que d'autres centres comme Brest pratiquaient déjà la « professionnalisation du codage » donc assuré par des experts du DIM, médecins ou techniciens. Il conviendra donc de contrôler par TALN les CRH afin de confirmer la spécificité du code quant au diagnostic d'ICA à l'admission, et éventuellement de le remettre en question selon les résultats observés et l'efficacité du TALN.

Nous ne pourrons pas prétendre à une parfaite sensibilité. Cependant, le volume attendu de patients est conséquent (> 20 000 à l'échelle d'HUGO sur la période concernée) et la spécificité a été privilégiée, en gardant à l'esprit que notre approche génère nécessairement un biais de sélection, en particulier pour les patients admis pour ICA mais pour lesquels d'autres affections ont pesé davantage dans le codage du DP/DR et du GHM.

c. Identification des doublons entre les centres

Un patients admis plusieurs fois dans un même centre pour un séjour répondant aux critères d'éligibilité de l'étude sera facilement identifié à l'échelle du centre, où ses données sont désidentifiées mais pas encore anonymisées avant le transfert vers l'ODH.

Par contre, une fois les données regroupées dans l'ODH, la perte du lien avec les bases initiales et la moindre granularité (intentionnelle) des données ne permettent pas la réidentification déterministe d'un patient qui aurait été admis dans différents établissements.

Cet appariement sans réidentification sera donc permis par une procédure proposée par le RiCDC, basée sur les variables suivantes et explicitée dans la demande d'autorisation soumise auprès de la CNIL : NIR de l'assuré, sexe, date de naissance et rang de naissance. Le NIR lui-même ne sera pas exporté dans l'ODH mais permettra la création d'un identifiant unique par patient, reproductible entre différents centres. L'utilisation locale du NIR est justifiée car il permet un chaînage fiable et simple des

données de ces patients, les autres méthodes d'appariement ne permettant pas un chaînage exhaustif et pouvant conduire à des erreurs d'identification.

d. Flow-chart établi sur l'EDS porteur

La récupération des comptes rendus nécessaires au TALN requiert un traitement « externe » à eHOP. Cette approche suppose l'identification des patients d'intérêt, très simple par l'écriture d'une procédure SQL. Il est tout aussi aisé de la compléter afin qu'elle permette dans le même temps la récupération des seules données d'intérêt, assurant la parcimonie du projet.

Le code SQL produit, disponible **Annexe 10** (population) et **Annexe 11** (données) est directement partageable avec les autres centres. Dans sa version du 11 août 2022, il ne comprend pas encore l'exclusion des oppositions. Pour en faciliter l'appropriation, chaque étape est commentée directement dans le code et un temps d'exécution est fourni à titre indicatif.

Le *mapping* des données n'étant pas encore parfaitement déployé dans l'inter-région, ce code est appelé à être modifié de façon marginale par les autres centres, en particulier si les CRH ou les données biologiques ne répondent pas aux mêmes codes LOINC que pour Nantes.

Les étapes du code SQL sont calquées sur les critères d'inclusion et résumées pas-à-pas **Annexe 12.** Le flow-chart d'identification de la population d'intérêt est décrit **Figure 9**. Au 12 octobre 2022, l'exécution de ces deux procédures SQL durait un peu moins de 300 secondes et permettait l'identification de 7252 séjours et l'extraction de l'ensemble des données structurées nécessaire à GAVROCHE et du texte des CR pour la période 2011-2019, à partir de l'EDS du CHU de Nantes.

Avec l'aimable participation de M. Julien Herbert du CDC de Tours, ce code, après adaptation des codes LOINC de biologie et d'identification des comptes rendus, permet, sur la même période d'intérêt, l'identification de 8456 séjours, à partir de l'EDS du CHU de Tours.

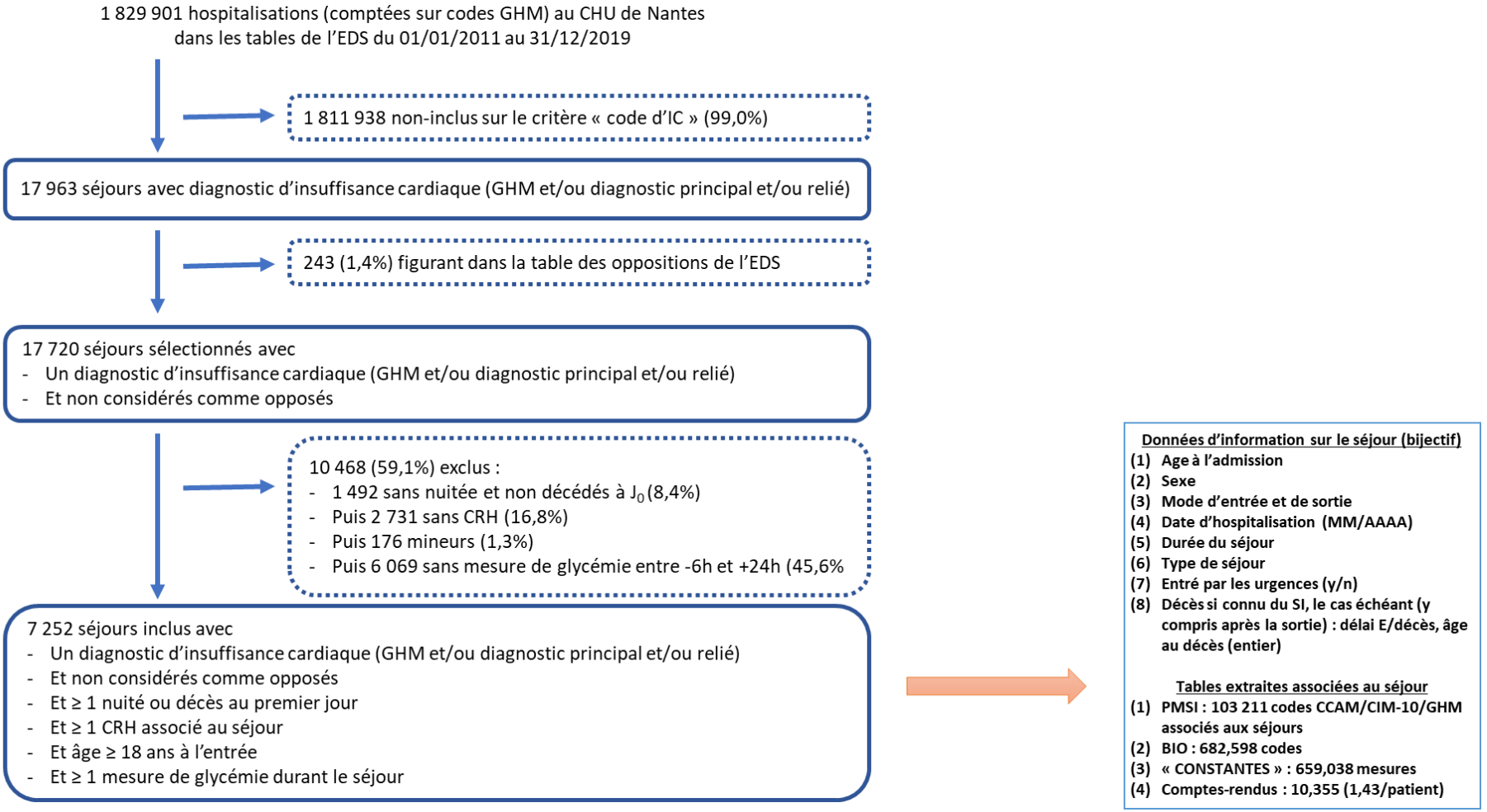


Figure 9. Flow-chart de l'identification des séjours d'intérêt pour GAVROCHE au 11 août 2022, obtenu par procédure SQL (SQL dev) et quantification des données associées

CCAM : Classification Commune des Actes Médicaux. CIM-10 : Classification Internationale des Maladies, 10ème révision. CRH : comptes rendus hospitaliers. EDS : entrepôt de données de santé. GHM : Groupe Homogène de Malades. IC : insuffisance cardiaque. PMSI : Programme de Médicalisation des SI. SI : système d'information.

6. Périmètre (2) : Données et parcimonie

Les données nécessaires au projet GAVROCHE ont été définies en prenant pour point de départ le modèle statistique final. Dès la soumission du projet à l'AAP 2019 notre objectif a été la parcimonie, c'est-à-dire la définition à visée exhaustive des données nécessaires à l'étude, variable par variable. La liste des données jugées nécessaires a été un peu modifiée en particulier suite au travail d'annotation des CR. Le détail de ces données est proposé en **Annexe 13**. Elles peuvent être regroupées ainsi :

- Caractéristiques sociodémographiques : âge, sexe, code INSEE de la commune de résidence en vue du croisement avec des indicateurs socio-économiques
- Hospitalisation : mois et année, durée du séjour
- Causes de l'IC chronique : coronaropathie, trouble du rythme, etc.
- Facteur déclenchant de l'ICA : syndrome coronarien aigu, infection, etc.
- Comorbidités cardiovasculaires et autres comorbidités associées au pronostic (potentiels facteurs de confusion) : accident vasculaire cérébral, cancer, etc.
- Statut vital à différents temps : à l'issue du séjour, à 30 jours et à 1 an

Ces données pourront être structurées ou non, voire « mixtes », c'est-à-dire récupérées sous les deux formats. C'est le cas par exemple le cas de la fréquence cardiaque, qui est partiellement enregistrée sur le flux « anthropométrie » du CHU de Nantes et va être également récupérée par TALN.

Un schéma relationnel faisant le lien entre ces données à l'échelle d'un séjour est proposé **Figure 10.**

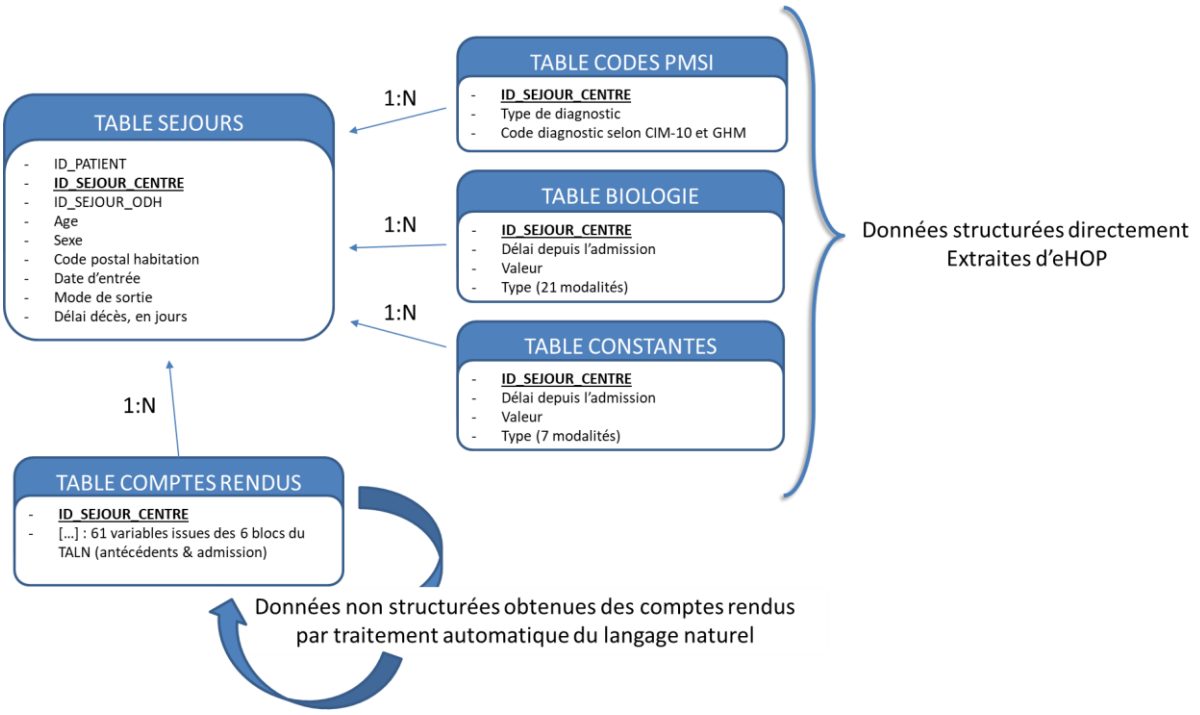


Figure 10. GAVROCHE - Schéma relationnel simplifié de l'ensemble des données du projet.

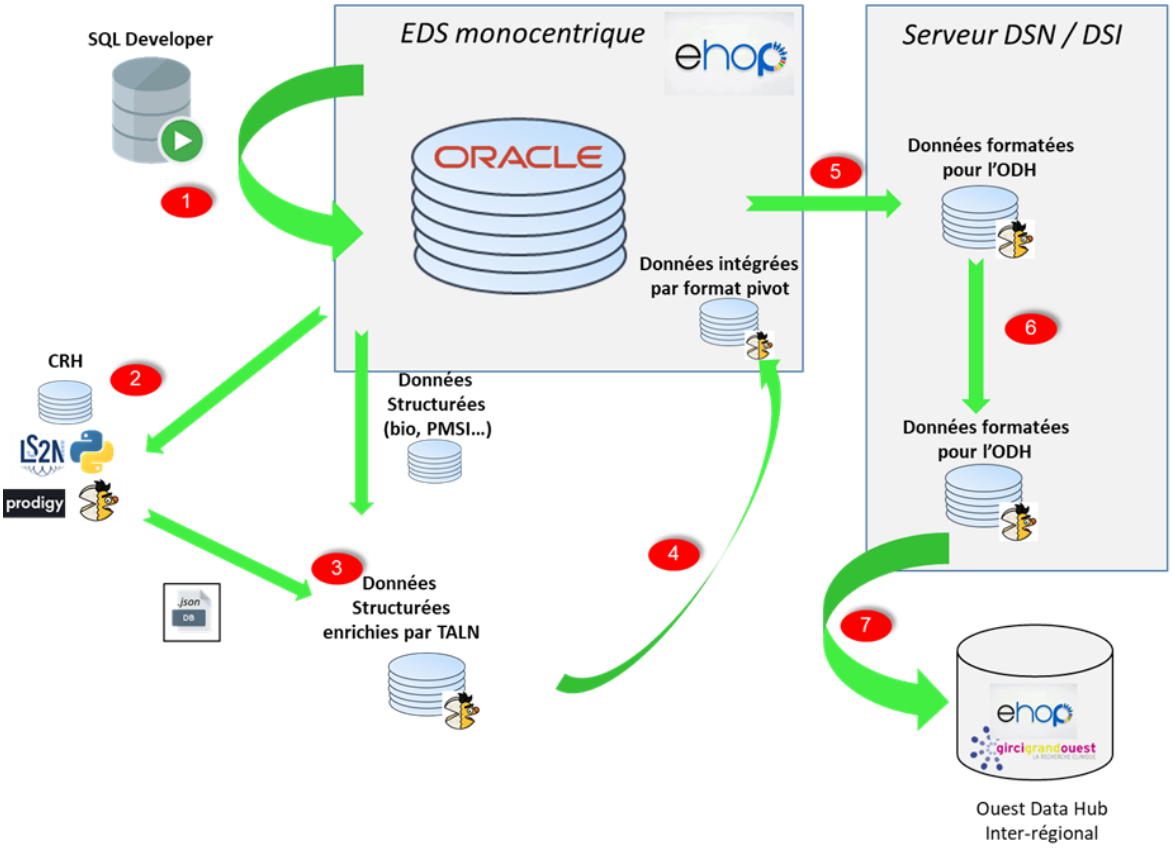
L'écriture « 1:N » signifie qu'à chaque séjour peut être associé un nombre entier d'éléments des autres tables (par exemple 5 codes diagnostics CIM-10, 20 dosages biologiques, 10 constante, 2 CR hospitaliers et les données issues du TALN associées…)

7. Circuit général des données

En raison de la nécessité d'extraire des données localement par TALN, et donc pas seulement directement à partir des bases structurées d'eHOP, nous avons choisi une approche SQL unique pour identifier la population, extraire les données structurées et extraire les comptes rendus nécessaires au TALN. La procédure TALN appliquée aux CR permettra ensuite l'extraction des données supplémentaires à l'étude. Notre centre n'a pas la maîtrise du système d'export des données depuis un centre vers l'ODH puisque celui-ci n'avait pas, au 1er octobre 2022, encore été expérimenté en situation réelle.

Nous sommes donc tenus de respecter la contrainte suivante : les données ne peuvent être exportées directement hors d'eHOP vers l'ODH, mais doivent être intégrées au logiciel avant d'être exportées localement, puis au personnel local de la DSN (qui joue le rôle de tiers de confiance) pour création d'un identifiant unique, avant d'être transformées et exportées vers l'ODH.

Ce schéma global complexe de circulation des données est exposé à l'échelle d'un centre dans la **Figure 11** page suivante.



	Circuit général des données de GAVROCHE vers l'ODH	Acteurs (cas de Nantes)
1	Interrogation de la base Oracle via *SQL Developer* : ▪ Identification de la population ▪ Extraction conjointe de l'ensemble des données ➢ Structurées : PMSI, biologie, constantes… ➢ Non structurées : texte des CRH pour le TALN	CDC (M. Wargny)
2	Etape de TALN (détaillée dans la section suivante) ▪ Chargement des CRH sur serveur *Python* dédié avec accès restreint ▪ Annotations avec le logiciel *Prodigy* ▪ Développement et validation du TALN sur échantillon par apprentissage profond, modèle de langue contextuel CamemBERT ▪ Application du TALN à l'ensemble des CRH ▪ En sortie : données structurées au format *.json*	CDC / DSN (A. Bazoge / D. Toublant)
3	Fusion des données structurées, obtenues ou non par TALN	CDC / DSN (A. Bazoge / D. Toublant)
4	Formatage des données au format pivot pour réintégration dans eHOP	DSN / CDD (D. Toublant / M. Wargny)
5	Formatage des données mises à disposition de la DSN avant export ODH	DSN / CDD (D. Toublant / M. Wargny)
6	Identifiant unique par méthode de hachage : ▪ Lien séjour-IPP-NIR et création d'un identifiant unique ▪ Suppression des séjours-IPP-NIR	DSN (D. Toublant)
7	Export vers l'ODH, où sera appliqué un nouveau chiffrement des identifiants avant la mise à disposition des données à l'équipe du CDC	DSN / ODH (D. Toublant / Personnel ODH)

Figure 11. GAVROCHE - Schéma général du circuit des données

L'étape de TALN est détaillée dans la section suivante. CDC : Centre de Données Cliniques. DSN/DSI : Direction des Services numériques/informatiques. EDS : Entrepôt de Données de Santé. ODH : Ouest Data Hub. PMSI : programme médicalisé des systèmes d'information. TALN : traitement automatique du langage naturel.

8. Mise en place du TALN

Le projet GAVROCHE nécessite donc d'appliquer un TALN afin d'extraire certaines informations d'intérêt des CRH. Il est rendu possible par le travail commun avec M. Adrien Bazoge, ingénieur-doctorant en TALN associé à notre équipe depuis novembre 2020, sous la direction du Pr Emmanuel Morin du LS2N, et co-encadré par les Prs Béatrice Daille et Pierre-Antoine Gourraud.

L'objet ici n'est pas de la détailler de façon complète l'approche TALN. Nous reprendrons donc (a) les grandes phases du TALN et (b) les éléments nécessaires à une bonne annotation.

a. Les trois grandes phases du TALN

Le travail peut être décomposé en trois phases successives. Il débute après extraction des CRH liés aux séjours satisfaisant aux critères d'inclusion de l'étude (cf. étape 2 de la **Figure 11**).

On distinguera la phase d'apprentissage du TALN (monocentrique à Nantes), la phase de validation sur un échantillon de CRH annotés par chaque centre, et enfin la phase de déploiement visant à qualifier à grande échelle les données extraites de tous les CRH par TALN. Ce circuit est décrit ci-dessous et également schématisé dans la **Figure 12**.

(1) Une 1ère phase d'apprentissage, monocentrique, basée sur les annotations des seuls CRH du CHU de Nantes

- Annotation nantaise d'un échantillon de CRH pour ICA
- Apprentissage du TALN à partir de l'échantillon : qualification des items (cause de décompensation de l'ICA, AVC oui/non…) en particulier les tournures de phrase négatives et autres éléments pouvant engendrer des erreurs (antécédents familiaux)
- Une partie de l'échantillon sert de base de test pour éprouver la robustesse du TALN afin de limiter le sur-ajustement aux données

(2) Une 2nde phase de validation, qui sera conduite indépendamment dans chaque centre

- Comme pour la 1ère phase, les experts locaux annotent un échantillon de CRH pour la même série de variables prédéfinies
- Sans apprentissage préalable, la méthode de TALN est appliquée à l'échantillon
- Confrontation des résultats obtenus par les experts et par le TALN : quantification des résultats discordants, adjudication des résultats discordants

- A l'issue de cette phase, les performances du TALN auront pu être quantifiées dans les différents centres. Nous nous attendons à de fortes variations de la qualité entre ces items (il est a priori plus facile d'identifier un antécédent de diabète ou de tabagisme que de déterminer le facteur causal d'une insuffisance cardiaque chronique). Selon les performances nous pourrons décider pour chaque variable (i) d'abandonner l'espoir de la qualifier par TALN, (ii) de revenir à la phase 1 pour annoter davantage de CRH ou (iii) en cas de bonnes performances, de considérer le TALN comme valide pour la variable en question

(3) Une 3ème étape de qualification automatique à grande échelle

- C'est l'objectif final du TALN dans le projet GAVROCHE : une fois la procédure validée, par variable et par centre, elle peut être appliquée à l'ensemble des CRH et produire les données nécessaires (cf. étape 3 de **la Figure 11**).

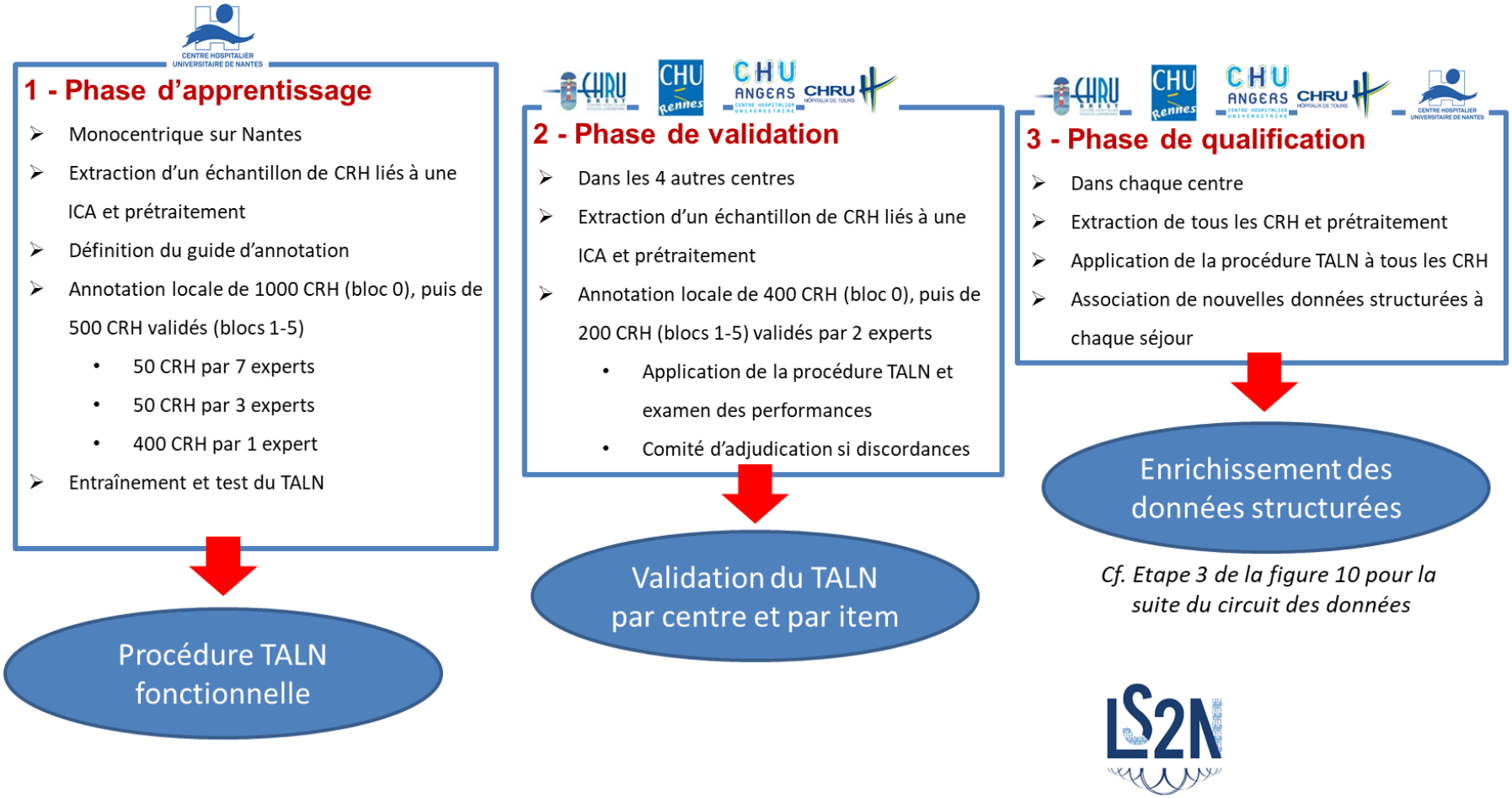


Figure 12. GAVROCHE - Les différentes phases de la mise en place du TALN

CRH : comptes rendus hospitaliers. ICA : Insuffisance cardiaque aiguë. TALN : traitement automatique du langage naturel.

b. Éléments nécessaires à l'annotation

L'annotation peut se définir comme « *le fait d'associer une information à une séquence textuelle, ou plus simplement à un segment de phrase* ». Une annotation est définie par une catégorie (ou un « type ») qui énumère l'ensemble des séquences textuelles possibles pour cette catégorie.

Par exemple, la catégorie « diabète de type 2 » pourra être définie par les séquences textuelles « DT2 », « diabète de type 2 », voire « diabète non insulino-dépendant ».

L'objectif de l'annotation est dans un premier temps de constituer des ressources permettant d'apprendre à une machine à lire les CR pour en extraire automatiquement les informations nécessaires. Par exemple, répondre à la question « est-ce que le patient a un antécédent d'infarctus du myocarde ? » à la lecture d'un CR.

Dans un second temps, l'annotation sert à évaluer la pertinence des informations extraites par la machine. Par exemple, pour 100 CR, on identifiera la proportion pour lesquels la machine aurait répondu correctement à cette question.

L'annotation a nécessité :

(i) **De choisir un logiciel d'annotation.** Plusieurs logiciels ont été testés par l'équipe. Après six mois de tests, nous avons retenu comme solution le logiciel Prodigy [115] (installation possible sur serveur CHU, ergonomie de l'interface, coût raisonnable de 2500€ pour les licences des cinq CHU)

(ii) **De mettre à disposition un serveur (virtuel) interne dédié à l'annotation :** lien avec les CRH désidentifiés, installation de Python nécessaire à l'interprétation du produit de l'annotation

(iii) **De définir précisément l'ensemble des catégories à annoter, formalisées par un guide d'annotation.** Nous en avons retenu 48, réparties en 6 blocs d'annotations. Un bloc d'annotation correspond à un ensemble comprenant un maximum de 10 catégories à annoter à la lecture d'un CRH donné, limite impérative pour des raisons d'ergonomie. Six blocs ont été définis :

- BLOC 0 : validation de l'ICA à l'admission (oui/non/incertain), contrôle de la date d'admission

- BLOC 1 : facteur déclenchant de l'ICA (6 modalités), 1er épisode d'ICA (2 modalités), traitement habituel (annotation globale, sans individualiser les médicaments)
- BLOC 2 : facteur étiologique de la cardiopathie causale (6 modalités), type d'ICA (4 modalités)
- BLOC 3 : antécédent d'insuffisance respiratoire chronique, de BPCO, de syndrome d'apnée obstructive du sommeil, d'AVC, d'AC/FA, de diabète (4 modalités)
- BLOC 4 : caractéristiques cliniques à l'admission : fréquence cardiaque, pression artérielle (systolique et diastolique), poids, taille, IMC, tabagisme
- BLOC 5 : autres antécédents : troubles du rythme, dépression, troubles cognitifs, cancer ; caractéristiques cliniques à l'admission : AC/FA, FEVG (3 modalités)

(iv) **De constituer un groupe d'annotateurs et un circuit d'annotation.** Comme il nous était très difficile de constituer clairement ce pool sans certitude initiale sur le logiciel d'annotation, le mode de déploiement sur serveur et les blocs d'annotation, et que ce travail nécessite de solliciter fortement des cliniciens par ailleurs peu disponibles, ce groupe a beaucoup évolué au cours de mon travail de thèse. Le bloc 0 a fait l'objet d'un travail séparé, avec un total d'environ 1000 CR annotés par au moins 1 annotateur. En juillet 2022, nous avons décidé de la configuration suivante pour les blocs 1 à 5 à partir des CRH validés par annotation du bloc 0 :

- Annotation commune de 50 CRH pour tous les annotateurs (sept en tout : trois endocrinologues (S. Smati, P. Morcel, E. Scharbarg), un médecin neurovasculaire (P. Constant dit Beaufils), un cardiologue (D. Stévant), le Pr Hadjadj et moi-même), afin de contrôler la qualité de leur annotation en les confrontant aux autres experts
- Annotation commune de 50 CRH pour PCDB, SH et MW, afin d'améliorer le guide d'annotation
- Puis annotation de 400 CRH par un seul annotateur, selon les disponibilités de chacun.

(v) **Le tout coordonné et accompagné par M. Bazoge, Mme Toublant, le Pr Hadjadj et moi-même**

9. Qualité des données

Nous discuterons ici de façon brève mais systématique de la qualité attendue des données issues du PMSI-MCO et des informations extraites des CR par TALN, et de la biologie.

a. Données issues du PMSI-MCO et des CR

Les données exploitées correspondent aux codages (CIM-10, GHM et CCAM) et aux CR associés aux hospitalisations d'intérêt. Nous avons fait le choix de ne pas récupérer de façon systématique les données des hospitalisations antérieures, craignant d'induire un biais de mesure entre les populations, les données n'étant accessibles qu'au niveau des CHU de HUGO, sans les données des CH locaux ou des CHU hors de HUGO. Notre caractérisation se limite donc à l'information jugée pertinente par les soignants (visible dans les CR) et nécessaire à la T2A (PMSI-MCO). Contrairement à l'approche « recherche » applicable dans SURDIAGENE, on ne peut prétendre à une caractérisation systématique du patient à son admission.

En particulier, les données indispensables à la compréhension de l'épisode sont réputées visibles dans les CR associés à l'hospitalisation, le codage CIM-10 permettant un « rattrapage » partiel de cette information. Les résultats obtenus par les deux approches seront comparés secondairement à l'échelle de HUGO. Notre démarche en deux temps, (i) confirmation de l'éligibilité (« vrai » épisode d'ICA) puis (ii) caractérisation du cas, avec un contrôle par échantillonnage aléatoire de ces informations, nous permettra d'apprécier la viabilité de la procédure.

Les lacunes mises en évidence dans notre démarche sont les suivantes. **Premièrement,** les cas cliniques indéterminés ne pouvant être définitivement adjudiqués par les cliniciens experts, par exemple les cas d'insuffisance rénale chronique avec tableau œdémateux et dyspnée, sans autre signe de défaillance cardiaque ni d'affirmation de l'IC dans le CR. **Deuxièmement,** l'information disponible dans les CR est très variable, en particulier pour les patients âgés décédés très rapidement après l'admission, pour lesquels les CR sont volontiers concis. De plus, les antécédents de patients bien connus des services sont parfois décrits très sommairement, du fait de CR liés à des consultations ou hospitalisations antérieurs, non pris en compte dans notre démarche. Aussi, certaines données plus « secondaires », telles que les facteurs de risque cardiovasculaire, figurent de façon très variable, y compris pour des patients bien décrits. **Troisièmement,** la question de la temporalité de l'information est primordiale (p. ex. ICA dès l'admission ou secondairement au cours du séjour), et nous n'avons encore aucune

garantie de l'efficacité de l'approche TALN pour extrapoler les annotations à d'autres situations ambiguës. **Quatrièmement,** le nombre de CR à annoter est un enjeu crucial dans la faisabilité du projet, et nos experts TALN associés ne pouvaient prédire précisément ce nombre, qui dépend à la fois de la fréquence d'apparition de l'information et de la variabilité de son expression dans les CR.

La qualité inégale de l'identification de l'ICA par les codes CIM-10 et/ou GHM, mise en évidence par l'échantillonnage, nous rappelle qu'un tel travail est nécessaire pour toutes les pathologies déduites de ces codages, et donc de la coronaropathie, des troubles du rythme, *etc*.

Enfin, à ces difficultés d'extraction des CR s'ajoute au moins pour le site nantais le problème de CR associés par erreur au séjour du fait des limites du SIH (défaut déjà rapporté dans le soin mais non réglé par l'éditeur).

b. Données issues de la biologie

Les données de biologie requises pour le projet sont élémentaires, et les standards des laboratoires sollicités dans les CHU (accréditation COFRAC ISO 15189) sont garants de leur qualité. Nous serons bien sûr dépendants des données manquantes éventuelles. Concernant l'exposition principale d'intérêt, la variabilité glycémique (VG), sa mesure sera dépendante de la fréquence des valeurs. Pour éviter un « biais d'immortalité » sur la période d'hospitalisation elle-même, l'analyse principale devra être bornée aux patients ayant survécu un temps minimal (48h ou 72h, selon les données effectivement disponibles) tout en permettant le calcul de cette VG. Nous ne disposerons pas de certaines données importantes modifiant la glycémie (prise alimentaire et médicamenteuse, syndrome infectieux, modalités de mesure capillaire, veineuse ou interstitielle) mais nous avons accepté ce biais de mesure, l'objectif étant d'étudier l'intérêt prédictif de la VG en « situation réelle ».

La principale difficulté associée aux données de biologie est le « *mapping* » sur un standard international « LOINC » permettant leur intégration uniforme au niveau de l'ODH, les SIH étant différents d'un CHU à un autre. Ce travail a profité sur la période 2020-2022 de celui d'un projet plus important, HUGOSHARE, portant sur plusieurs centaines de variables biologiques, avec l'appui d'un biologiste indépendant, Philippe Chatron de la société C2BIO. Cet effort de standardisation participe d'un cercle vertueux entre la recherche et le soin, puisque le codage LOINC des données de biologie

est aussi en passe de devenir une obligation légale, portée par l'Agence du Numérique en Santé[17], afin d'améliorer l'interopérabilité des SI traitant des données de biologie.

Enfin, s'agissant de données des EDS encore jamais analysées pour la plupart des centres de HUGO, nous proposerons une mesure de contrôle qualité supplémentaire, à appliquer avant ou après le chargement des données dans l'ODH selon la disponibilité des équipes des CDC. Il s'agira d'une analyse quantitative temporelle systématique des données intégrées, avec pour objectif (i) d'identifier des valeurs aberrantes et (ii) de contrôler leur distribution ainsi que l'absence de point rupture importante (« *breakpoint* ») au cours de la période d'étude (2011-2019), pouvant par exemple être attribuable à une modification de la méthode de mesure.

En pratique, il s'agira d'indicateurs de position et de dispersion présentés par année, et de représentation graphique de l'ensemble des mesures (représentation temporelle en nombre de jours depuis l'origine, modulo une fonction de transformation tel ici le logarithme népérien pour les NT-proBNP). Le tout est applicable sous R à partir de deux variables, le temps (en jours depuis le 1er janvier 2011) et la valeur de la mesure (quantité).

A l'échelle de Nantes, cela donne les résultats ci-après pour trois variables biologiques (glycémie, HbA_{1c} et NT-proBNP) et trois variables cliniques (IMC, taille et température corporelle), sur les périodes 2012-2019 (7364 hospitalisations) et 2015-2019 (3699 hospitalisations), respectivement. Ce premier aperçu appelle les commentaires suivants :

- Sur la population : on observe une diminution rapide du nombre de séjours entre 2015 et 2018 (1208 à 473), a priori non attribuable à un changement d'activité, et qui pourra correspondre à une modification des habitudes de codage ou à un défaut d'intégration des données (cf. **Tableau 3**)
- Pour toutes les variables, aucune mesure n'est observée sur une période de quelques dizaines de jours correspondant approximativement à août 2022 (cf. **Figure 13 et Figure 14**). Après retour aux données, on observe qu'aucune hospitalisation n'a été retenue entre le 17 juillet 2018 et le 15 août 2018, ce qui fait craindre un défaut de remontée d'informations sur cette période
- Quelques valeurs de glycémie sont aberrantes (> 10 g/L) et justifieront d'un retour à la donnée de soin afin d'identifier si le problème vient de la saisie de la donnée, de sa transformation par

[17] Consultable sur https://esante.gouv.fr/jeu-de-valeurs-loinc

le système, ou d'une erreur d'unité (10 g/L pouvant en réalité correspondra à 10 mmol/L, par exemple). Ce commentaire est vrai également pour l'IMC (quelques valeurs à zéro)

- De façon plus positive, les données observées pour l'HbA_{1c} ou les NT-proBNP ne sont pas aberrantes, en qualité comme en quantité

Dans notre démarche d'exploitation des entrepôts de données de santé, cela illustre l'intérêt des projets à grande échelle pour améliorer la qualité de l'entrepôt, mais aussi et surtout le caractère essentiel de la formalisation du contrôle de chaque information.

Pour le passage à l'échelle inter-régionale, les codes SQL et R écrits pour le contrôle de ces données sont pensés de façon à être facilement réplicables à l'échelle de chaque centre, dans un temps raisonnable (évalué approximativement à 1 journée, édition du rapport final comprise), mais ils supposent ensuite une démarche active pour identifier les valeurs aberrantes et la stratégie de correction, qui sera, elle, fonction du mode de déploiement de l'entrepôt et de l'origine des flux de données, et donc propre à chaque centre. Comme nous le voyons pour Nantes – et ce sera différent, mais pas nécessairement plus simple, pour les quatre autres CHU - le travail est rendu plus complexe par la bascule progressive d'un SI à un autre en 2014-2015 (logiciel Clinicom vers Millennium).

La fréquence des erreurs retrouvées localement et la complexité du circuit des données pour permettre leur centralisation à l'ODH plaide pour consolider localement la qualité des données avant leur centralisation.

Tableau 3. Description de six paramètres biologiques et cliniques d'intérêt extraits dans le cadre du projet GAVROCHE, sur la période 2012-2019								
Année	**Nb de mesures/Nb de séjours**	**Nb moyen de mesures/séjour**	**Moyenne (DS)**	**Médiane (25ème-75ème percentile)**	**Min/Max si dans les seuils**	**Nb zéros**	**Nb ≤ seuil**	**Nb ≥ seuil**
Glycémie (g/L)							**0.10**	**10**
2012	8967/1094	8.2	1.31 (0.49)	1.21 (0.99-1.49)	0.20/9.41	0	0	0
2013	9100/1258	7.2	1.30 (0.51)	1.19 (0.97-1.48)	0.18/6.50	0	0	1
2014	12140/1248	9.7	1.38 (0.58)	1.24 (1.01-1.58)	0.13/9.38	0	0	1
2015	7822/1194	6.6	1.36 (0.54)	1.24 (1.03-1.57)	0.18/6.19	0	1	0
2016	4215/917	4.6	1.40 (0.57)	1.28 (1.04-1.58)	0.22/7.88	0	0	0
2017	3907/593	6.6	1.45 (0.57)	1.33 (1.10-1.64)	0.31/8.12	0	0	0
2018	2668/466	5.7	1.44 (0.58)	1.31 (1.08-1.64)	0.20/8.10	0	0	0
2019	1380/482	2.9	1.39 (0.51)	1.28 (1.06-1.58)	0.18/5.62	0	0	0
HbA_{1c} (%)							**3**	**25**
2012	194/1094	0.2	7.0 (1.5)	6.6 (6.0- 7.5)	4.5/13.5	0	0	0
2013	196/1258	0.2	7.1 (1.4)	6.8 (6.2- 7.7)	4.3/15.7	0	0	0
2014	213/1248	0.2	7.0 (1.4)	6.7 (6.0- 7.7)	4.6/12.8	0	0	0
2015	178/1194	0.1	7.2 (1.4)	7.0 (6.1- 7.9)	5.0/12.7	0	0	0
2016	194/917	0.2	7.0 (1.4)	6.8 (5.9- 7.7)	4.5/12.3	0	0	0
2017	111/593	0.2	6.7 (1.2)	6.3 (5.8- 7.3)	5.1/10.8	0	0	0
2018	78/466	0.2	7.0 (1.3)	6.5 (6.1- 7.8)	4.6/10.5	0	0	0
2019	94/482	0.2	6.5 (1.3)	6.3 (5.7- 7.1)	4.1/12.2	0	0	0
NT-proBNP (ng/L)							**<0**	**10^5**
2012	1281/1094	1.2	6762 (8592)	3790 (1707- 8207)	33/68175	0	0	1
2013	1481/1258	1.2	6805 (6877)	4374 (2045- 8908)	11/34762	0	0	0
2014	1446/1248	1.2	6689 (7016)	4179 (1948- 8719)	27/43070	0	0	0
2015	1442/1194	1.2	6382 (6795)	3774 (1824- 8752)	19/34888	0	0	0
2016	1120/917	1.2	6526 (6871)	3972 (1846- 8514)	28/34981	0	0	0
2017	658/593	1.1	6265 (6762)	3859 (1932- 8122)	27/34664	0	0	0
2018	628/466	1.3	6754 (7155)	4376 (1922- 8837)	93/55472	0	0	0
2019	714/482	1.5	6674 (6990)	3942 (1772- 9430)	34/34951	0	0	0
IMC (kg/m^2)							10	100
2015	1641/1194	1.4	27.3 (6.8)	26.0 (22.9-30.5)	10.6/54.5	4	6	12
2016	1993/917	2.2	27.2 (6.3)	25.9 (22.8-30.7)	12.8/62.1	0	0	63
2017	1159/593	2.0	26.6 (7.2)	25.5 (21.9-30.0)	12.6/94.5	0	0	13
2018	874/466	1.9	26.5 (6.8)	25.2 (21.3-30.0)	14.8/57.8	0	0	5
2019	724/482	1.5	27.4 (6.3)	26.8 (23.7-30.3)	15.3/50.6	0	0	0
Taille (cm)							50	220
2015	853/1194	0.7	164 (11)	164 (158-170)	57/192	0	3	0
2016	1171/917	1.3	164 (11)	165 (160-170)	60/190	0	37	0
2017	760/593	1.3	166 (13)	167 (160-174)	53/198	0	3	0
2018	540/466	1.2	166 (13)	167 (160-172)	56/194	0	4	0
2019	486/482	1.0	167 (11)	169 (160-173)	68/196	0	0	0
Température (°C)							30	45
2015	5266/1194	4.4	36.7 (0.63)	36.6 (36.3-37.1)	32.2/40.8	0	3	1
2016	18933/917	20.6	36.7 (0.68)	36.6 (36.2-37.0)	31.0/41.3	0	17	0
2017	15829/593	26.7	36.8 (0.66)	36.7 (36.3-37.2)	32.0/41.6	0	5	0
2018	14942/466	32.1	36.8 (0.7)	36.8 (36.4-37.2)	30.6/40.8	0	3	1
2019	20338/482	42.2	37.1 (0.73)	37.0 (36.6-37.5)	31.5/41.0	0	4	1

Figure 13. GAVROCHE - Représentation chronologique de l'ensemble des mesures des données biologiques de glycémie, HbA_{1c} et NT-proBNP du projet, au CHU de Nantes sur la période 2012-2019. Le NT-proBNP a été transformé (logarithme népérien) du fait de son augmentation d'allure exponentielle en cas d'insuffisance cardiaque aiguë

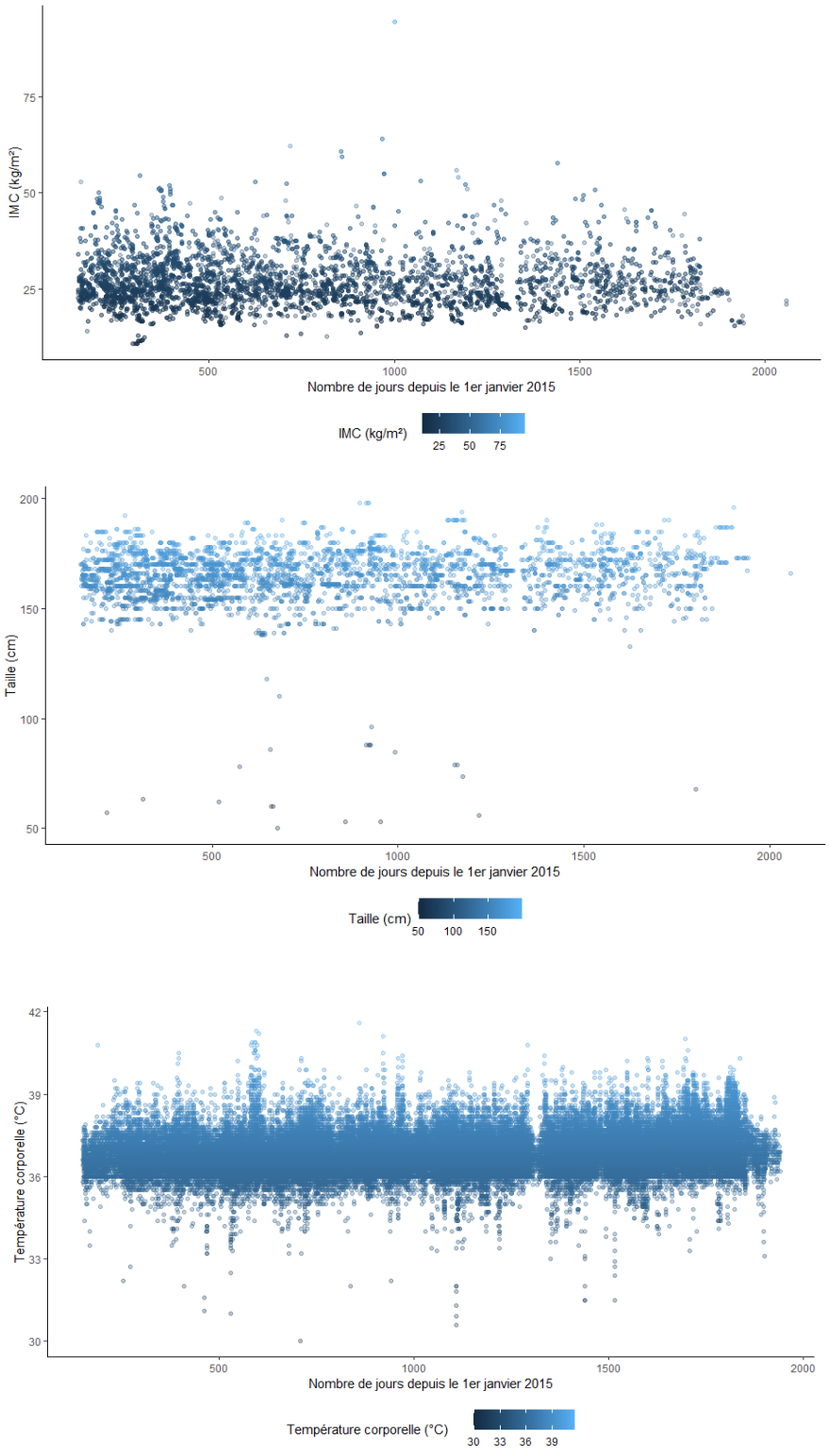


Figure 14. GAVROCHE - Représentation chronologique de l'ensemble des mesures des données cliniques d'indice de masse corporelle (IMC), taille et température corporelle, au CHU de Nantes sur la période 2015-2019 (pas de données structurées avant 2015)

10. Plan d'analyses statistiques

Toutes les variables seront décrites pour l'ensemble des centres puis pour chacun d'entre eux. Les variables catégorielles seront décrites par l'effectif et le pourcentage de chaque modalité. Les variables quantitatives seront décrites par des paramètres de position (moyenne, médiane) et de dispersion (écart-type, $25^{ème}$ – $75^{ème}$ percentile), selon la distribution appréciée graphiquement sur histogramme.

L'analyse principale sera proposée sous la forme d'un modèle de régression logistique (non conditionnelle) multivariable, avec les caractéristiques suivantes :

- **Population :** tous les patients adultes inclus dans l'étude GAVROCHE, avec un diagnostic d'ICA à l'admission confirmée par le TALN, et sans données manquantes parmi les variables nécessaires au modèle d'ajustement le plus complet (cf. ci-dessous modèle M_4)
- **Variable à expliquer :** décès au cours de l'hospitalisation pour ICA
- **Variable explicative d'intérêt principal :** la variabilité glycémique (VG) des 72 premières heures, définie par le coefficient de variabilité de la glycémie (écart-type/moyenne, donc nécessitant au moins 3 mesures de la glycémie)
- **Facteurs confondants :** quatre modèles d'ajustement, emboîtés, seront proposés. Le choix des variables est basé sur une connaissance au préalable des facteurs confondants potentiels (« *background knowledge* » [116]), c'est-à-dire les facteurs cliniques et biologiques connus comme associés à la mortalité pour ICA, et potentiellement associés à la VG [113,117] :
    - M_0 : VG seule
    - M_1 : M_0 avec âge et sexe
    - M_2 : M_1 avec les antécédents connus :
        - notion d'épisode d'ICA avant l'admission,
        - cause possible d'IC : trouble du rythme, valvulopathie, coronaropathie
        - autres antécédents cardiovasculaires : HTA, diabète (non/type 1/type 2/autre ou indéterminé), AVC et/ou AIT
        - autres antécédents : insuffisance respiratoire chronique, BPCO, cancer, troubles cognitifs, dépression
        - traitements médicamenteux (cf. **Annexe 13** listant les variables d'intérêt pour GAVROCHE, table 5.C)
    - M_3 : M_2 avec les caractéristiques à l'admission (≤ J1) :
        - Admission par les urgences

        - Facteur déclenchant de l'ICA (ischémique, trouble du rythme/ACFA à l'admission, poussée hypertensive, infection, modification de traitement ou de régime)
        - Type d'ICA parmi 4 modalités : droite isolée, ICG décompensée, OAP, choc cardiogénique
        - Constantes cliniques : fréquence cardiaque maximale, pressions artérielle systolique et diastolique maximale, poids ou IMC)
        - Biologie des 24 premières heures : CRP, BNP ou NT-proBNP, hémoglobinémie, natrémie, kaliémie, estimation du débit de filtration glomérulaire (CKD-EPI si disponible), hémoglobinémie, plaquettes, leucocytose
        - FEVG au cours du séjour : réduite (<41%) / modérément réduite (41-49%) / préservée (≥50%)
    - M_4 : M_3 avec recherche d'interaction entre l'antécédent de diabète et la VG

Avant d'être utilisées dans le modèle, les variables quantitatives pourront être standardisées (« centrées-réduites », soit soustraction de la moyenne puis division par l'écart-type) afin de permettre la comparaison directe des coefficients du modèle.

Les analyses de sensibilité suivantes seront réalisées :

- Analyse par centre (représentée par *Forest plot*)
- Approche *leave-one-out*, avec la même analyse excluant systématiquement chacun des centres en laissant les autres dans le modèle
- VG définie par l'écart-type de la glycémie, en ajoutant alors l'ajustement sur la première glycémie à l'admission
- Décès considéré à 28 jours et à 1 an, selon la disponibilité des données, en envisageant l'utilisation d'un modèle temps-événement type modèle de Cox basé sur l'hypothèse des risques proportionnels.

Toutes les analyses seront proposées sur données disponibles, sans imputer les valeurs manquantes. La population non analysée du fait de données manquantes sera décrite en regard de la population analysée. Une p-value < 0,05 sera considérée comme statistiquement significative. Les analyses seront réalisées sur la plate-forme de l'ODH, à partir de la dernière version du logiciel libre de statistiques R (interface RStudio).

11. Etat des lieux de GAVROCHE au 1er octobre 2022

Au temps de restitution du présent manuscrit, les étapes suivantes avaient été réalisées pour le projet GAVROCHE :

- Obtention de l'autorisation CNIL et contractualisation inter-CHU signée par les différents acteurs
- Ecriture du protocole, avec une définition précise de la population d'étude, des données recueillies (SQL, TALN) et du circuit des données
- En particulier pour la population et l'extraction des données (codes reproductibles sous forme SQL et R, SQL ayant été testé à Tours) : identification de la population et des tables de données d'intérêt
- En particulier pour les données quantitatives : automatisation d'un contrôle quantitatif et visuel des données, définition des valeurs aberrantes
- En particulier pour le TALN : choix des logiciels (annotation sous Prodigy, TALN sous Python), formalisation de la procédure d'annotation avec guide d'annotation, automatisation du « reporting » sur la qualité de l'annotation (travail d'Adrien Bazoge, non présenté ici : confrontation inter-annotateurs, humain/humain et humain/machine), procédure d'adjudication des cas discordants non encore éprouvée

Fondamentalement, trois éléments majeurs ont été sous-estimés lors de mise en place de ce travail : la qualification locale des données, le TALN, et le circuit des données.

En particulier, l'effort consenti à l'échelle du CHU porteur par des équipes directement liées au projet (Clinique des Données et DSN), et avec des enjeux universitaires individuels (la thèse d'Adrien Bazoge et le présent travail) ne pourra être exigé auprès des autres centres participants, essentiellement pour des questions des ressources humaines. La démarche de qualification de la population et des données quantitatives, de par son caractère « vertueux » pour la qualité locale des données, peut être acceptable avec l'accompagnement nantais. Mais il n'en sera pas de même pour le TALN, qui fait appel à une ressource rare, l'expertise clinique, et avec une moindre garantie de retombée locale pour les CHU hors Nantes. L'enjeu pour l'année 2023 sera donc la formalisation d'une procédure « dégradée » du TALN, autant que possible automatisée, au moins pour la phase d'adjudication et le contrôle qualité, afin d'extraire des données de qualité à partir des CR, données qui pourront ensuite être remontées à l'ODH.

PARTIE VI : DISCUSSION GENERALE

1. Synthèse

Dans cette discussion, j'aborderai d'abord la question de la qualité des données, en lien avec la finalité de leurs sources et les conditions du gel de base (ou de consolidation des données), et avec les enjeux de l'interopérabilité. Je proposerai ensuite une réflexion sur le consentement des patients, le mode de recueil de ce consentement et ses implications pour les patients et la recherche, avant de présenter brièvement ma position de chercheur dans un CHU français et ce qu'elle m'a permis, en terme d'accès aux données. Enfin, je donnerai ma vision d'une démarche idéale de conduite de projet de recherche sur les données en vie réelle, avant de conclure sur l'ensemble de ce travail.

2. Gestion des données selon leurs sources

Nous en avons déjà tracé les grandes lignes : ce sont toutes les dimensions de la gestion des données qu'il faut reconsidérer selon leur finalité, afin de mieux comprendre les forces et les limites de leur exploitation pour la recherche. Pour le chercheur, c'est un enjeu majeur afin d'anticiper la faisabilité technique (la simple possibilité de constitution de la base) mais aussi la faisabilité humaine (moyens humains nécessaires pour la réalisation de l'étude) - ce point me semblant très largement sous-estimé.

a. Qualité des données et gel de base

Avant l'analyse des données d'une étude, une étape majeure classique est le « gel de base » : la population d'analyse est figée, et toutes les données associées ont été revues et corrigées. Les données disponibles sont soit connues avec des valeurs cohérentes ou actées comme aberrantes, soit inconnues et actées comme non récupérables. Ce moment est daté, sans possibilité de retour en arrière.

Nous avons déjà abordé cette question pour les données issues de la recherche, avec l'exemple idéal des essais cliniques bien conduits, dont l'intégrité est susceptible d'être contrôlée par les agences. Toutes les données et leurs modalités de recueil peuvent être prédéfinies et les examens (cliniques, biologiques ou autres) standardisés. Après le dernier examen du dernier patient, un temps est consacré au gel de base. Les données manquantes ou aberrantes entraînent des requêtes, résolues par des aller-retours entre l'équipe ayant accès aux données sources et le *data manager*, par l'intermédiaire d'un attaché de recherche clinique dédié au *monitoring*. L'effort de recueil peut être

proportionné à l'importance de la donnée pour l'étude. Si un événement indésirable grave nécessite une attention particulière, l'expertise du médecin investigateur responsable est sollicitée et lui-même pourra si nécessaire solliciter d'autres expertises, revoir le patient ou prescrire de nouveaux examens. La qualité de l'information recueillie atteindra voire dépassera celle ordinairement disponible dans le soin courant.

Pourquoi insister autant sur cette démarche, naturelle à tous les pratiquants de la recherche clinique, en particulier interventionnelle ? Pour prendre la mesure de la différence d'avec les données de vie réelle à grande échelle, que ce soit par le biais d'une source médico-administrative comme le SNDS, ou par celui d'une source issue du soin comme les EDS.

Dans le cas des données médico-administratives, le gel de base dépend du temps de remontée de l'information dans le système. Pour le SNDS, cette durée est différente selon qu'il s'agit d'informations du DCIR (ambulatoire) ou du PMSI/ATIH (hospitalisations). Ainsi, une feuille de soins étant valable deux ans et son traitement pouvant prendre 3 mois, l'Assurance Maladie propose de considérer un délai de 27 mois pour considérer qu'une table mensuelle est consolidée dans le DCIR.[118] On peut saluer ici la grande prudence administrative puisque la CNAM nous informe également que 99% des feuilles sont remontées à 6 mois. Pour les données issues du PMSI/ATIH, le délai maximal est plus court mais le délai moyen plus important, car il passe par une étape de consolidation annuelle : l'ensemble des données de l'année calendaire *« n »* sont disponibles l'été de l'année *« n +1 »*, par exemple en juillet 2022 pour les données du 1er janvier au 31 décembre 2021[18]. Le processus qualité nécessaire à l'application de la T2A garantit une certaine stabilité de la base.[13] Dans la qualification, fréquemment rencontrée mais invérifiable, de « *possiblement la plus grande base de données de remboursement continue et homogène* » [119], nous pouvons apprécier à plein le caractère continu et homogène, qui permet une exploitation pour l'instant inenvisageable dans notre EDS issus du soin.

Mais gel de base n'implique pas fiabilité des données. L'analyse de DMC illustre bien comme il est facile pour le statisticien, une fois absorbé par ses analyses, d'accepter des limites pourtant majeures. Nous avons déjà abondamment discuté de la qualité du codage, chaque famille de codes utilisés pour définir une pathologie méritant d'être validée. Mais contrairement au cas de la donnée recueillie à fin de recherche, aucune donnée n'est réputée manquante, ce qui signifie en réalité que seules les

[18] Rien n'étant jamais si simple avec le SNDS, l'honnêteté nous oblige à ajouter un bémol : les séjours ne sont remontés dans l'ATIH qu'une fois le patient sorti, et il faudra donc attendre une année supplémentaire pour les patients entrés avant le 31 décembre et sortis après le 1er janvier, ce qui entraîne un vrai biais de mesure pour une analyse qui comparerait les mois de décembre 2020 vs 2021 avec des données disponibles en 2022

informations « positives » (p. ex. hospitalisation pour infarctus du myocarde) seront rapportées, tandis que nous ne pourrons pas différencier le cas d'un patient ayant bénéficié d'un bilan cardiologique complet et normal de celui d'un patient refusant le soin, les traitements et les hospitalisations (ou n'y ayant pas accès), et souffrant d'une insuffisance cardiaque s'exprimant cliniquement mais non encore décompensée et donc invisible dans le SNDS.

Dans le cas des données issues du soin, en particulier celles issues d'un outil automatisant l'extraction des EDS, tel eHOP, la question du gel de base et du contrôle de la qualité des données est beaucoup plus complexe et doit être vue comme un problème à plusieurs facteurs. De façon non limitative, il s'agira de savoir (i) si la donnée est remontée vers l'EDS, (ii) sa qualité brute au temps de la saisie et (iii) sa qualité finale après les étapes de transformation induites par le circuit des données et en lien avec l'objectif d'exploitation.

Premièrement, la remontée effective de la donnée. Pour que cette donnée nous parvienne via l'EDS, il faudra d'abord qu'elle ait été saisie (« produite »), et ensuite qu'elle ait été chargée vers l'EDS. Le chargement dans l'EDS peut être maîtrisé : l'ingénieur en charge du déploiement peut proposer une fréquence de remontée, avec, dans le cas du CHU de Nantes, une fréquence « au pire » mensuelle, comme pour le PMSI. Mais il faudra aussi considérer le caractère discontinu de la saisie des données dans le SI avant de considérer les tables comme définitives. C'est particulièrement vrai pour les CR d'hospitalisation ou de consultation, dont le temps de remontée est très conjoncturel (habitudes du praticien et du service, disponibilité du secrétariat) avec parfois des délais de 6 mois.

Deuxièmement, la qualité originelle de la donnée, avec plusieurs questions pour le *data manager* : quand une donnée doit-elle être considérée comme suspecte (glycémie à 6 g/L) ou aberrante (glycémie nulle ou négative) ? Lorsque deux données sont contradictoires et que le retour à la source n'est pas possible, quelle valeur finale donner à l'information ? Si l'on revient au cas de DMC, le décès du patient pouvait précéder la remontée de l'information sur l'insuffisance cardiaque, et nous avons alors décidé de corriger le temps de l'insuffisance cardiaque en lui attribuant le temps du décès, ce qui est critiquable, et crée un biais dans l'étude du lien entre ces deux événements (qui ne constituait cependant pas la question principale de l'étude, mais modifiait le modèle multi-états).

De façon surprenante, de telles erreurs, même identifiées par les soignants, peuvent persister dans les données issues du soin. A titre d'exemple, un patient était rapporté à tort comme décédé par un SI extérieur à mon hôpital, et sa consultation a été automatiquement annulée. En dépit de l'absurdité et de l'inhumanité de la situation, l'éditeur de la solution logicielle n'a pas corrigé le bug et a clos le

signalement. La variable « décès » pouvant prendre plusieurs valeurs indépendantes dans le SI, le patient est numériquement revenu à la vie lors de la visite suivante.

A l'échelle du SIH nantais, la volonté de produire des indicateurs automatisés crée également des données erronées, aux dépens des soignants et des patients. Ainsi, le format (ou « *template* ») de certaines lettres de liaison contient systématiquement un champ « allergies connues », ce qui permet d'affirmer que ces lettres remplissent à 100% cet objectif d'indicateur. Cependant, ce champ est alimenté par un champ rarement utilisé par les soignants, qui documentent plus souvent les allergies dans le corps du texte principal du CR. Ainsi, si le médecin n'a pas le réflexe de supprimer l'information produite automatiquement, un chercheur sur l'EDS qui voudrait l'exploiter et ne serait pas conscient de cette faille s'appuierait sur des données invalides – sans parler, bien sûr, des conséquences potentielles pour le soin.

Ces exemples caricaturaux sont malheureusement non limitatifs. Les dates automatiques insérées dans les CRC contiennent également un très grand nombre d'erreurs, dont il nous faudra tenir compte pour dater les CR, la date de production étant un indice insuffisant. Je sors ici du cadre de mon manuscrit, mais il apparaît essentiel de souligner qu'il s'agit ici d'effets contre-productifs d'une démarche qualité basée sur des indicateurs (statistiques de clôture de ticket, allergie, date, mais cela a également été vu pour le poids ou les traitements médicamenteux). Cette approche statistique de la qualité du soin nuit potentiellement au soin et, par ricochet, aux travaux de recherche associés.

Troisièmement, pour l'exploitant des données de santé issues du soin, en sus de la qualité de la donnée saisie, un contrôle qualité du circuit des données est nécessaire, chaque étape étant susceptible de l'altérer. Tous les informaticiens connaissent les problèmes de reformatage, fréquents quand un nombre est pris pour une chaîne de caractères ou vice-versa. Or, pour rendre la donnée quantitative directement exploitable dans eHOP, une série d'opérations est appliquée. Un exemple : pour la CRP, la colonne donnant sa valeur quantitative *a priori* en mg/L pourra afficher le texte « non dosable » ou « < 5 mg/L ». Si l'ingénieur a décidé que tout champ textuel était d'emblée exclu de l'analyse quantitative, les valeurs les plus basses ne seront pas récupérables, et nous perdrons l'information des patients dont la CRP est sous le seuil de détection, augmentant artificiellement la proportion de syndromes inflammatoires. Pour résoudre ce cas de figure, il n'y a à ma connaissance pas d'autre solution que connaître l'ensemble des modalités pouvant être prises pour l'ensemble des variables, afin de décider d'une conduite à tenir automatisée (en bref, un *mapping*) propre à chaque situation. Du fait de la taille considérable des bases et de l'hétérogénéité des données et de leurs modalités, je plaide pour un traitement local et au cas par cas sur des données ciblées en lien avec une étude active,

comme ce que nous avons amorcé pour GAVROCHE, par exemple pour discuter les valeurs limites de poids, taille, NT-proBNP ou encore glycémie.

Dans le cas des EDS, ces approches seront encore compliquées par la diversité des SI, plusieurs logiciels du CHU étant souvent impliqués, avec un risque de rupture temporelle si l'étude porte sur plusieurs années d'activité. C'est ce que nous avons illustré dans GAVROCHE en représentant l'évolution des valeurs des données cliniques et biologiques, ce qui nous a permis (i) de nous forcer à nous poser des questions sur l'évolution du nombre de patients et de mesures et (ii) de mettre en lumière une absence de données sur plusieurs dizaines de jours en 2018.

En définitive, l'exploitation des EDS constitue sans doute la situation la plus complexe pour le gel de base, mais avec des possibilités de correction dépassant le cas des données médico-administratives puisqu'un retour direct à la source est possible. Des retombées positives sont également visible pour le soin grâce à l'identification de mauvaises pratiques de recueil, de transfert ou de production automatisée de l'information. Elles nécessitent cependant de formaliser l'approche, en particulier pour la rendre critiquable et reproductible, thème abordé également dans la section 5 « Démarche idéale […] » de la présente partie. S'agissant de données plus figées, comme le SNDS, un avantage pratique est le « réusage » direct de codes génériques à grande échelle, ce que j'ai pu réaliser entre le projet DMC présenté ici et le projet DETECT déjà mentionné, sur deux populations très différentes mais ayant pour point commun un même schéma relationnel.

b. Interopérabilité et enrichissements inter-bases

L'interopérabilité, c'est ici la capacité de nos systèmes d'informations à échanger entre eux. Plus spécifiquement, dans la recherche sur données, c'est notre capacité à croiser des données de différentes sources. Au niveau local, et en particulier pour les EDS, c'est une question essentielle dans l'intégration des flux issus de différents SI. Selon sa définition, elle peut intégrer la notion de qualification, c'est-à-dire tenter de donner un sens plus précis à une donnée externe avant de l'intégrer dans le SI. On « l'harmonise », c'est-à-dire qu'elle est modifiée afin de porter le même sens que d'autres données de notre source.[3]

L'interopérabilité est un point critique tant au niveau local (pour le soin comme pour toutes les composantes de la coordination hospitalière, et bien sûr pour la mise en place des EDS) qu'au niveau inter-régional (ODH) et national (HDH). Avec la création du HDH, l'un des objectifs est la standardisation des SI afin d'améliorer l'interopérabilité à tous les niveaux.[29] Cette approche a été

également critiquée, interprétée comme une vision centralisatrice idéaliste poussant à réunir les données d'abord et à les harmoniser ensuite, reportant la question de la qualité et de l'interopérabilité, sans standardisation préalable, et vouée à l'échec devant l'ampleur de la tâche.[3]

C'est, à mon sens, le nœud du problème : résister à la volonté naïve de vouloir tout réunir d'emblée (« *I want it all, and I want it now* ») en pensant qu'une approche globale absorbant toutes les données est le meilleur point de départ pour leur harmonisation. Dans cette perspective, si toutes les données sont visibles au même endroit, il suffit d'en établir la liste, de les examiner toutes de façon systématique puis, en bon *data manager*, de leur appliquer la transformation congrue qui les rendra exploitables, indépendamment de leur finalité.

Mon retour d'expérience sur DMC et GAVROCHE plaide pour le contraire : nos ressources, tant humaines que structurelles, sont limitées, et la qualification de la donnée nécessite un retour à la source, voire souvent à l'être humain générateur de la source. Le maintien du « lien local » reste absolument nécessaire, et le cas par cas dans la qualification de la donnée est la seule solution raisonnable et raisonnée, d'autant que celui-ci sera à adapter à l'objectif scientifique. Pour revenir à l'interopérabilité, cela suppose donc de traiter les questions d'harmonisation localement, pour éviter des allers-retours épuisants bien connus du trio clinicien/*data manager*/biostatisticien dans les études épidémiologiques classiques.

Une fois la donnée qualifiée localement, nous avons cependant un besoin majeur de cet enrichissement inter-bases rendu possible par l'interopérabilité. L'analyse de la cohorte CONSTANCES appariée aux données SNDS a donné sa crédibilité à la population de DMC, en validant l'algorithme de définition du diabète.[32] J'ai déjà donné des exemples de contrôle de codage PMSI par échantillonnage sur les EDS. Nous n'avons pas encore su l'exploiter ici, mais l'analyse des CR de GAVROCHE, d'abord par lecture humaine puis par TALN, nous permettra de proposer une démarche de validation des cas d'insuffisance cardiaque aiguë selon le codage PMSI, en particulier CIM-10 et/ou GHM.

De façon moins classique, nous observons localement que le travail de standardisation des données de soin pour la recherche permet un vrai « cercle vertueux » soin -> recherche -> soin, *etc.* Effectivement, en raison du changement du principal logiciel de soins du CHU, le format actuel des données de biologie ne permettait pas le chargement rapide vers le nouveau logiciel, mais celui-ci sera permis par la standardisation des données de laboratoires nécessaire à leur usage dans l'EDS, anticipant ainsi sur le besoin d'interopérabilité du soin.

Pour conclure, l'enrichissement inter-bases est également possible à partir de données dites *open source*, certaines données étant disponibles au grand public, comme les fichiers des décès des personnes françaises de 1970 à nos jours[19]. Le statut vital des patients étant très incomplet dans notre SIH, nous avons pu, grâce aux efforts du Dr Vianney Guardiolle et d'Adrien Bazoge, développer un algorithme basé sur l'identité (nom et prénom notamment, avec une « distance » acceptable pour les approximations d'écriture) qui permet d'automatiser la récupération de ce statut vital, approche qui sera utilisée dans GAVROCHE pour Nantes et proposée aux autres centres, et permettra une substantielle économie de moyens par rapport à la solution alternative qui consisterait à solliciter un appariement aux données du CépiDc[20].

c. Justification de l'utilisation des données massives et perspectives françaises

Malgré une complexité d'accès et d'exploitation, l'utilisation des bases de données massives présente des avantages importants. Augmenter la taille d'échantillon permet de connaître plus précisément la prévalence et l'incidence des maladies rares, et d'identifier des « petits » déterminants de santé. La structuration des données massives permet des économies d'échelle (diminution du coût unitaire de la donnée) par le partage des moyens et la mutualisation des ressources, en particulier lorsque celles-ci risquent d'être sous-exploitées (données dormantes ou « froides », serveurs de calcul non sollicités). L'étude des déterminants de santé à grande échelle permet de diminuer la variabilité. De plus, un facteur de risque identifié et validé à l'échelle nationale ou internationale risque moins d'être biaisé qu'à l'échelle locale, par exemple lorsqu'il s'agit d'étudier un toxique environnemental supposé : si les travailleurs de l'amiante présentent davantage de mésothéliomes dans toutes les régions du globe, l'argument de causalité est beaucoup plus fort que si cela n'est observé qu'à l'échelle d'un groupe géographiquement localisé.

En effet, en dépit des grandes réussites françaises que constituent le SNDS et la cohorte CONSTANCES, un risque est la sur-représentation de données « étrangères », au sein de la littérature médicale internationale, sans que soit posée la question de la reproductibilité dans la population nationale. La célèbre base de données *UK Biobank* (UKBB) est un exemple écrasant, qui regroupe les données cliniques, biologiques, et même génomiques de plus de 500 000 individus au Royaume-Uni, avec

[19] Cf. https://arbre.app/insee, les données pouvant être interrogées en ligne ou les bases (annuelles ou mensuelles) téléchargées dans leur entièreté depuis Internet

[20] Article accepté dans *JMIR medical informatics* en janvier 2022, non encore publié au temps de soumission du présent manuscrit

plusieurs milliers de publications associées écrites par des équipes du monde entier.[120] La facilité d'accès à ces données offre une comparaison cruelle au chercheur français intéressé par le SNDS. Or, de par mon expérience personnelle avec l'étude « CORONADO », il me semble que le fait de se savoir patient ou investigateur potentiel pour une étude (appartenir à la population cible, voire source) accroit l'intérêt du patient comme du chercheur pour les résultats – ils se sentent plus « concernés » par ceux-ci. Dès lors, l'enjeu de l'accessibilité aux autres grandes cohortes nationales – qu'il s'agisse de la UKBB, de la cohorte suédoise SNDR sur la population diabétique[121], ou encore des cohortes danoises nationales portant séparément sur les diabètes de type 1 et 2 [122] – est aussi celui du contrôle de la validation en population française. Le partage des données étant rarement acquis (UKBB étant à ma connaissance la principale exception), la formation d'un réseau d'analystes est essentielle, chacun pouvant répliquer les modèles (et parfois les résultats) de l'autre à partir des données dont il dispose. C'est en suivant cette démarche que nous répliquons actuellement des résultats issus de CONSTANCES à partir de UKBB, concernant les comorbidités associées à un LDL-cholestérol bas de cause non pharmacologique (projet HYPOBETA.fr avec le Pr Bertrand Cariou).

3. Le consentement du patient et la finalité des données

Pour l'étude SURDIAGENE, le consentement écrit des patients a pu être obtenu lors de l'inclusion dans l'étude, après qu'un investigateur ait expliqué l'objectif et les contraintes associées. Pour DMC et GAVROCHE, la recherche du consentement était jugée trop complexe à mettre en place à l'échelle individuelle, et a donc été réputée acquise. Il y a cependant une différence de traitement entre ces deux projets : pour GAVROCHE, les patients pouvaient être informés, ou plus exactement *s'informer activement*, du déroulement de l'étude par consultation du site Internet de leur CHU d'inclusion ou via le site du GCS HUGO. De plus, les patients déjà opposés localement à l'utilisation de leurs données, pour toute étude en lien avec l'entrepôt, ne pouvaient être inclus. Mais pour l'étude DMC, il leur était impossible de s'opposer directement, bien que l'information soit diffusée sur le site du HDH. En effet il n'existe pas, à ma connaissance, de circuit pour s'opposer à l'utilisation des données nous concernant dans le SNDS.

Cette situation, très « confortable » pour le chercheur, constitue néanmoins une lacune pour l'exercice du droit des patients. Une piste d'amélioration pourrait être l'établissement d'un registre national d'autorisation à voir utiliser ses données de santé à des fins de recherche, comme il existe un « Registre National des Refus » pour le don d'organe, administré par l'Agence Nationale de la Biomédecine. Chacun est réputé consentant s'il n'a pas exprimé activement son opposition en s'inscrivant à ce

registre. Un tel registre nécessite cependant l'appariement par un identifiant national unique, comme le numéro de sécurité sociale.

Cette volonté de donner un statut officiel au consentement du patient, aux motivations d'abord éthiques et réglementaires, peut avoir des conséquences négatives. Pour DMC, on observe que ce choix n'en est pas un puisque le consentement est réputé acquis, sans possibilité d'opposition, tandis que l'information du patient suppose qu'il soit « fort acteur de sa santé » puisqu'il lui faut consulter régulièrement les quelques 5000 projets déclarés sur le site du HDH[21]. Pour GAVROCHE, la capacité à exercer le droit d'opposition est réelle mais suppose là encore une démarche active, avec une opposition spécifique au projet nécessitant une acculturation numérique forte, chez des patients essentiellement âgés et malades et dont l'espérance de vie est réduite. Une information individuelle plus active est très difficile à mettre en place, voire irréaliste. Dans nos discussions avec la CNIL, celle-ci avait d'abord demandé que les patients « *consultant régulièrement dans les établissements de santé pour la pathologie concernée […] se voient remettre une note d'information individuelle sur le projet* ». Avec tout ce qu'une telle démarche suppose de réidentification individuelle et donc de risque pour les données du patient, imaginer mettre en place, pour chaque centre, une information systématique aux cardiologues pour que des patients hospitalisés pour ICA en 2012-2019 et consultant en 2022 pour le suivi de leur insuffisance cardiaque se voient informer qu'ils étaient peut-être susceptibles d'être analysés dans GAVROCHE nous apparaissait déraisonnable et même nuisible à l'exercice du soin. La CNIL a finalement accepté de retirer cette clause.

Enfin, s'agissant du consentement général des patients, l'existence même des EDS les contraint à une alternative qui n'est pas anodine, « être ou ne pas être » dans les entrepôts. En réalité, le patient ne peut en être tout à fait exclu puisque l'on conserve l'information individuelle de leur refus. Dans mon expérience d'investigateur clinique, une personne accepte de participer à une étude parce qu'elle a confiance dans l'investigateur qui la lui propose, qui est souvent son soignant. Par extension, elle accepte d'être incluse dans l'EDS ou dans une cohorte comme CONSTANCES parce qu'elle a confiance dans la mission d'intérêt public du CHU ou de l'INSERM. A l'inverse, le risque de dérive est réel si une société totalitaire ou perçue comme telle peut interpréter ce refus comme une marque de défiance, sociétalement suspecte. Cette interprétation peut apparaître théorique, mais le cas pratique s'est posé au CHU de Nantes lorsqu'il a été question de savoir si l'EDS pouvait être consulté sur réquisition judiciaire.

[21] Consultable sur https://www.health-data-hub.fr/projets - 5047 projets déclarés au 2 octobre 2022

4. Positionnement du chercheur et accessibilité aux données

Ce travail d'exploration des différentes sources nécessite des connaissances théoriques mais aussi pratiques. Pour l'accès aux données, je bénéficie d'une situation très privilégiée en tant que salarié du CHU de Nantes depuis 2017, avec en plus de mon rôle au sein de la C2D une pratique clinique.

Je peux ainsi disposer des différents accès permettant d'étayer ces propos par des exemples issus de mes travaux personnels. **Des données issues de la recherche,** à travers différentes études dont CORONADO, SURDIAGENE, MOTHIF-II, mais aussi la cohorte CONSTANCES associée au SNDS, avec un accès au CASD dans le cadre du projet HYPOBETA.fr ; **des données médico-administratives SNDS,** par un accès d'abord EGB, puis par projet (MOTHIF-II, DMC, DETECT) ; **des données issues du soin via l'EDS du CHU de Nantes,** grâce à notre travail de mise en place et d'exploitation via l'outil eHOP développé par le laboratoire LTSI de Rennes. Ce dernier accès est fondamental et ne peut être dissocié de l'ancrage local puisque comme on l'a vu chaque flux d'information doit pouvoir être remis en question, et provoquer des discussions avec ces « producteurs » de données que sont les soignants (médicaux et paramédicaux), le SIM, les biologistes, et bien sûr les spécialistes du SIH de nos services numériques.

Ce positionnement salarié implique une priorisation des tâches qui a eu un fort impact sur ma capacité à approfondir les différents sujets. L'accès aux données de SURDIAGENE était garanti d'emblée par le statut d'investigateur principal du Pr Hadjadj, et la qualité des données assurée par Elise Gand du CHU de Poitiers. Je leur dois d'avoir pu mener rapidement l'analyse et la publication des résultats avec un investissement personnel raisonnable. Le projet DMC, non financé et avec un accès non acquis aux données, et qui nécessitait un travail de transformation à des fins de recherche des données SNDS beaucoup plus important, n'était acceptable dans l'exercice de mes fonctions que parce qu'il me permettait aussi un gain de compétences directement mis à profit pour d'autres projets institutionnels financés (MOTHIF-II, DETECT, DOXY-COVID). Enfin, le projet GAVROCHE, pour lequel l'investissement est considérable, est légitimé par le financement obtenu par l'AAP GIRCI-GO et la contribution privée du laboratoire AstraZeneca.

Par ailleurs, j'ai été soutenu dans la réalisation du projet DMC par mon appartenance au groupe de travail REDSIAM Endocrinologie, nutrition et métabolisme, dirigé par Sandrine Fosse-Edorh (Santé publique France), et mes échanges fructueux avec les différents membres du groupe, tous utilisateurs réguliers du SNDS.

Au-delà de ma situation individuelle, à l'échelle de l'institution, il est permis d'estimer la solidité d'un projet en fonction des ressources humaines et techniques. Sur ce plan, la cohorte SURDIAGENE

apparaît la plus solide : à partir de quelques documents (protocole, eCRF, schéma relationnel, dictionnaire des variables) et de tables de volume mémoire anecdotique (quelques Mo), il est possible de transférer le savoir nécessaire à l'exploitation des données. Pour le SNDS, ces documents existent également et les tables restent relativement stables dans le temps, mais la dimension plus importante de la base et la complexité du recueil des données nécessitent un personnel stabilisé, tant à la CNAM qu'au CHU, pour garantir la viabilité des projets. Cela peut être rendu très difficile du fait du délai d'accès aux données (volontiers 12 à 24 mois). Pour GAVROCHE et les EDS, l'investigateur doit s'assurer de la stabilité des équipes locales (CDC, DSN), de l'outil (eHOP ou autre système d'interrogation local de la base) et de la plateforme, chacun de ces éléments constituant un point de vulnérabilité critique du projet.

5. Démarche idéale pour les projets RWE – plaidoyer pour de Bonnes Pratiques Epidémiologiques en « vie réelle »

Qu'il s'agisse de données issues de sources médico-administratives ou issues du soin, les projets de recherche sur données de vie réelle (RWE/RWD en anglais – *real-world evidence* ou *real world data*) sont souvent pensés par des chercheurs ayant une expérience de l'analyse de données issues de la recherche. Je propose ici un condensé de mon expérience sur ces sources, sous la forme d'une démarche « idéale » qui se veut fondée sur le bon sens.

Comme pour les projets de recherche, il est impératif de construire cette démarche en considérant le recueil de données comme un moyen, et non comme une fin, la fin étant toujours la question scientifique à laquelle prétend répondre l'étude. En effet, nous avons vu que la récupération des données de vie réelle impliquait de connaître leurs conditions de recueil, et bien souvent de les transformer (*screening*, *mapping*, TALN, contrôle qualité) pour les rendre utiles pour la recherche, et qu'une approche de qualification globale était vouée à l'échec. Or, seul un objectif centré sur une question de recherche précise permet de définir cette transformation de façon raisonnée. Je propose la démarche suivante :

1) **Formulation de l'objectif et de la question principale à laquelle prétend répondre l'étude**

2) **Ecriture complète des critères d'éligibilité des patients** (ou des observations d'autre nature)

3) **Ecriture d'un dictionnaire des variables explicite et à visée exhaustive (*Data statement*),** listant les données, leurs conditions de recueil, les gages de leur qualité *a priori* (recueil clinique contrôlé, standards biologiques, démarche qualité pour les données médico-administratives) et le contrôle *a posteriori* de cette qualité (monitoring basé sur échantillonnage, recherche de valeurs aberrantes, cohérence à l'échelle du jeu de données) et les conditions du gel de base – correspondant donc à une mise à plat du *Data Management*

4) **Ecriture d'un plan d'analyses statistiques complet (*Statistical Analyses Plan* ou SAP),** détaillant les variables créées à partir du gel de base, l'analyse principale et les analyses de sous-groupes /analyses de sensibilité, et, pour les études étiologiques, le graphe dirigé acyclique (*Directed acyclic graph ou DAG*) justifiant les modèles retenus dans une perspective causale

5) **Mise en ligne publique et datée de l'ensemble de ces éléments,** par exemple sur le site https://clinicaltrials.gov, avant la récupération des premières données

Cette proposition simple souligne le caractère indissociable de la question scientifique, de la population d'étude, des données et de leur analyse. La mise en ligne d'un protocole force l'investigateur à la transparence et lui impose d'appliquer un principe de parcimonie au trio population/données/analyse, et le pousse à penser également son travail en terme de moyens. Cela engage l'investigateur devant la communauté scientifique à justifier d'éventuels changements d'approche, et donne la possibilité à cette même communauté de reproduire l'approche choisie dans un autre contexte et donc d'en tester l'universalité. Pour les études observationnelles aux conséquences cliniques immédiates, comme les effets indésirables médicamenteux [123,124], la confiance dans la reproductibilité des résultats est essentielle pour des raisons à la fois médicales, éthiques et juridiques. Toujours dans un esprit de transparence, c'est un argument fort pour valoriser davantage les résultats dits « négatifs ».

Enfin, une telle démarche permet bien sûr d'éviter la redondance de projets de recherche, comme cela était craint aux débuts de l'épidémie à SARS-CoV-2. Cependant, depuis début 2022 et l'extension de l'accès aux données SNDS « France entière » aux équipes de recherche INSERM et des CHU, l'obligation matérielle de déclarer les études de vie réelle sur le site du HDH a disparu, et l'on peut craindre une sous-déclaration et un gaspillage de ressources humaines, dont l'originalité du travail serait perdue

par la publication d'une équipe « concurrente » travaillant à partir des mêmes accès aux mêmes données.

Pour revenir à mes travaux, cette démarche n'a été mise en œuvre dans son intégralité pour aucun des trois projets présentés ici, bien que les protocoles de DMC et de GAVROCHE, non diffusés hors des soumissions CESREES/CNIL, aient été pensés dans cet esprit, avec en particulier un essai de description non ambiguë de la population, des données et de l'analyse. Ce travail était compliqué ici du fait d'un défaut de connaissance des bases SNDS et de la réalité du déploiement des EDS, tant localement qu'à l'échelle de l'inter-région. Dans le cadre d'un projet plus récent, DETECT, portant sur l'efficacité et la sécurité des dispositifs à élution de paclitaxel dans l'angioplastie des lésions des membres inférieurs, nous avons pu exposer notre démarche sur le site *Clinical Trials* avant l'accès effectif aux données. L'approche a été modifiée au cours de l'étude et un SAP actualisé doit encore être mis en ligne, qui fera état des changements opérés depuis le *Data statement* et le SAP initiaux[22].

[22] Consultable sur https://clinicaltrials.gov/ct2/show/NCT05254106

6. Conclusion

Nous avons proposé une perspective épidémiologique fondée sur la finalité du recueil des données, classées selon trois sources : recherche, médico-administratif et soin, ces deux dernières relevant de la grande famille des données de « vie réelle ». Cette finalité du recueil des données a été opposée à la finalité scientifique, soulignant que ce recueil des données est un moyen, et non une fin.

Avec en toile de fond la question clinique du lien entre diabète et insuffisance cardiaque, nous avons illustré l'exploitation de ces sources grâce à trois projets de recherche : (1) Biomarqueurs nutritionnels et insuffisance cardiaque chez les patients diabétiques de type 2 suivis au CHU de Poitiers (cohorte prospective SURDIAGENE), (2) Evénements rétiniens graves et risque d'insuffisance cardiaque chez les personnes diabétiques dans le SNDS (étude sur données médico-administratives DMC), et (3) Variabilité glycémique et pronostic vital chez les patients hospitalisés pour insuffisance cardiaque (étude inter-régionale sur données des entrepôts GAVROCHE).

Sans prétendre à une approche « innovante » explorant de nouvelles méthodes algorithmiques, nous avons pris le parti d'insister sur les données de santé elles-mêmes, leurs conditions de production, leur circuit et leur qualité, avec des approches analytiques que l'on pourrait qualifier de traditionnelles. Pour les données de « vie réelle », SNDS ou EDS, nous nous sommes attachés à expliciter des méthodes, critiquables et reproductibles, de définition de la population et de contrôle qualité des données, à partir notamment de nombreux exemples de contrôles de cohérence interne, d'enrichissements inter-sources et de validations d'algorithmes par échantillonnage.

Laissant donc de côté des méthodes de *machine learning* dépassant nos capacités de modélisation, et qui ont déjà largement fait leur preuve dans les domaines, par exemple, du traitement du signal et de l'image, ce manuscrit est plutôt une défense des fondamentaux : le fondement de l'analyse des données étant les données elles-mêmes, nous ne pouvons pas prétendre à une science de la donnée quand, arrachée de son contexte, elle risque de perdre de son sens et de s'en voir prêter un nouveau et invérifiable.

Dans cet esprit, je fais le pari que c'est par la mise à l'épreuve de ces nouvelles sources, et par l'acculturation des équipes - cliniciens, ingénieurs, chercheurs - que l'on saura définir à quelles questions elles peuvent réellement répondre, et dans quelles conditions leur exploitation est souhaitable et utile.

BIBLIOGRAPHIE

1. WHOCC - ATC/DDD Index [Internet]. [cited 2022 Aug 5]. Available from: https://www.whocc.no/atc_ddd_index/

2. Outil PIA : téléchargez et installez le logiciel de la CNIL | CNIL [Internet]. [cited 2022 Aug 5]. Available from: https://www.cnil.fr/fr/outil-pia-telechargez-et-installez-le-logiciel-de-la-cnil

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

4. Zins M, Cuggia M, Goldberg M. [Health data in France: Abundant but complex]. Med Sci (Paris). 2021;37:179–84.

5. Lee D, Lee J. Testing on the move: South Korea's rapid response to the COVID-19 pandemic. Transp Res Interdiscip Perspect. 2020;5:100111.

6. Vaidya T, Thomas-Ollivier V, Hug F, Bernady A, Le Blanc C, de Bisschop C, et al. Translation and Cultural Adaptation of PROactive Instruments for COPD in French and Influence of Weather and Pollution on Its Difficulty Score. Int J Chron Obstruct Pulmon Dis. 2020;15:471–8.

7. CUESP - Collège Universitaire des Enseignants de Santé Publique. Item 20 - Interprétation d'une enquête épidémiologique. Santé Publique. 3ème. 2015.

8. Connor Gorber S, Tremblay MS. The bias in self-reported obesity from 1976 to 2005: a Canada-US comparison. Obesity (Silver Spring). 2010;18:354–61.

9. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the Epidemiology of Cardiovascular Diseases: A Historical Perspective. Lancet. 2014;383:999–1008.

10. Zins M, Goldberg M, CONSTANCES team. The French CONSTANCES population-based cohort: design, inclusion and follow-up. Eur J Epidemiol. 2015;30:1317–28.

11. Frey S, Bourgade R, Le May C, Croyal M, Bigot-Corbel E, Renaud-Moreau N, et al. Effect of Parathyroidectomy on Metabolic Homeostasis in Primary Hyperparathyroidism. J Clin Med. 2022;11:1373.

12. Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc J-L, Sailler L. French health insurance databases: What interest for medical research? Rev Med Interne. 2015;36:411–7.

13. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Rev Epidemiol Sante Publique. 2017;65 Suppl 4:S149–67.

14. Polton D, Ricordeau P. Le SNIIRAM et les bases de données de l'Assurance Maladie en 2011. :65.

15. Guittet M, Lamirault G, Connault J, Durant C, Hamidou M, Wargny M, et al. [Evaluation of a woman's care program after pre-eclampsia]. Rev Med Interne. 2021;42:154–61.

16. Portail d'accueil GCS HUGO ET GIRCI GRAND OUEST [Internet]. [cited 2022 Aug 5]. Available from: https://www.chu-hugo.fr/

17. Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny M-L, et al. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Stud Health Technol Inform. 2019;264:1536–7.

18. eHOP, l'entrepôt de données de l'HOPital | Centre de données cliniques [Internet]. 2022 [cited 2022 Aug 5]. Available from: https://centrededonneescliniques.univ-rennes1.fr/ehop-lentrepot-de-donnees-de-lhopital

19. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24:1716–20.

20. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/

21. SQL Developer | Oracle [Internet]. [cited 2022 Aug 5]. Available from: https://www.oracle.com/database/sqldeveloper/

22. Diagnosis | ADA [Internet]. [cited 2022 Sep 15]. Available from: https://diabetes.org/diabetes/a1c/diagnosis

23. Cottin V, Larrieu S, Boussel L, Si-Mohamed S, Bazin F, Marque S, et al. Epidemiology, Mortality and Healthcare Resource Utilization Associated With Systemic Sclerosis-Associated Interstitial Lung Disease in France. Front Med (Lausanne). 2021;8:699532.

24. Cariou B, Hadjadj S, Wargny M, Pichelin M, Al-Salameh A, Allix I, et al. Phenotypic characteristics and prognosis of inpatients with COVID-19 and diabetes: the CORONADO study. Diabetologia. 2020;

25. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed. 2019;181:104804.

26. General Data Protection Regulation (GDPR) – Official Legal Text [Internet]. General Data Protection Regulation (GDPR). [cited 2022 Sep 20]. Available from: https://gdpr-info.eu/

27. HDS [Internet]. [cited 2022 Sep 20]. Available from: https://esante.gouv.fr/produits-services/hds

28. Hourdeaux J. La Cnil demande l'arrêt du stockage de nos données de santé par Microsoft [Internet]. Mediapart. [cited 2022 Sep 20]. Available from: https://www.mediapart.fr/journal/france/091020/la-cnil-demande-l-arret-du-stockage-de-nos-donnees-de-sante-par-microsoft

29. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare. Yearb Med Inform. 2019;28:195–202.

30. Horvais V, Wargny M, Repessé Y, Guillet B, Beurrier P, Ardillon L, et al. rFVIII-Fc in severe haemophilia A: The incentive switch in case of high risk of joint bleedings. Eur J Clin Invest. 2022;52:e13824.

31. Publication du BEH n°4 dédié aux "Hospitalisations pour Covid-19 au 1er semestre 2020 chez les personnes traitées pharmacologiquement pour un diabète en France" (Document) [Internet]. La Veille Acteurs de Santé. 2021 [cited 2022 Sep 15]. Available from: https://toute-la.veille-acteurs-sante.fr/171776/publication-du-beh-n4-dedie-aux-hospitalisations-pour-covid-19-au-1er-semestre-2020-chez-les-personnes-traitees-pharmacologiquement-pour-un-diabete-en-france-communique/

32. Fuentes S, Cosson E, Mandereau-Bruno L, Fagot-Campagna A, Bernillon P, Goldberg M, et al. Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort. Int J Public Health. 2019;64:441–50.

33. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2021;42:3599–726.

34. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol. 2022;79:e263–421.

35. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, et al. 2013 ACCF/AHA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Journal of the American College of Cardiology. 2013;62:e147–239.

36. Savarese G, Uijl A, Lund LH, Anker SD, Asselbergs F, Fitchett D, et al. Empagliflozin in Heart Failure With Predicted Preserved Versus Reduced Ejection Fraction: Data From the EMPA-REG OUTCOME Trial. J Card Fail. 2021;27:888–95.

37. Heidenreich PA, Albert NM, Allen LA, Bluemke DA, Butler J, Fonarow GC, et al. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. Circ Heart Fail. 2013;6:606–19.

38. Huffman MD, Berry JD, Ning H, Dyer AR, Garside DB, Cai X, et al. Lifetime risk for heart failure among white and black Americans: cardiovascular lifetime risk pooling project. J Am Coll Cardiol. 2013;61:1510–7.

39. Sakata Y, Shimokawa H. Epidemiology of heart failure in Asia. Circ J. 2013;77:2209–17.

40. De Peretti C, Pérel C, Tuppin P, Iliou M-C, Juillière Y, Gabet A, et al. Prévalences et statut fonctionnel des cardiopathies ischémiques et de l'insuffisance cardiaque dans la population adulte en France : apports des enquêtes déclaratives « Handicap-Santé ». Bull Epidémiol Hebd. 2014;172–81.

41. Tuppin P, Cuerq A, de Peretti C, Fagot-Campagna A, Danchin N, Juillière Y, et al. Two-year outcome of patients after a first hospitalization for heart failure: A national observational study. Arch Cardiovasc Dis. 2014;107:158–68.

42. Insuffisance cardiaque [Internet]. [cited 2022 Sep 10]. Available from: https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-cardiovasculaires-et-accident-vasculaire-cerebral/insuffisance-cardiaque

43. Chen J, Normand S-LT, Wang Y, Krumholz HM. National and regional trends in heart failure hospitalization and mortality rates for Medicare beneficiaries, 1998-2008. JAMA. 2011;306:1669–78.

44. Loehr LR, Rosamond WD, Chang PP, Folsom AR, Chambless LE. Heart failure incidence and survival (from the Atherosclerosis Risk in Communities study). Am J Cardiol. 2008;101:1016–22.

45. Kosiborod M, Inzucchi SE, Spertus JA, Wang Y, Masoudi FA, Havranek EP, et al. Elevated admission glucose and mortality in elderly patients hospitalized with heart failure. Circulation. 2009;119:1899–907.

46. Mebazaa A, Gayat E, Lassus J, Meas T, Mueller C, Maggioni A, et al. Association between elevated blood glucose and outcome in acute heart failure: results from an international observational cohort. J Am Coll Cardiol. 2013;61:820–9.

47. Emmons-Bell S, Johnson C, Roth G. Prevalence, incidence and survival of heart failure: a systematic review. Heart. 2022;108:1351–60.

48. Lin D-Y, Woodman R, Oberai T, Brown B, Morrison C, Kroon H, et al. Association of anesthesia and analgesia with long-term mortality after hip fracture surgery: an analysis of the Australian and New Zealand hip fracture registry. Reg Anesth Pain Med. 2022;rapm-2022-103550.

49. Home, Resources, diabetes L with, Acknowledgement, FAQs, Contact, et al. IDF Diabetes Atlas 2021 | IDF Diabetes Atlas [Internet]. [cited 2022 Sep 21]. Available from: https://diabetesatlas.org/atlas/tenth-edition/

50. Ricordeau P, Weill A, Vallier N, Bourrel R, Gulhot J, Fender P, et al. Prévalence et coût du diabète en France métropolitaine : quelles évolutions entre 1998 et 2000 ? Revue Médicale de l'Assurance Maladie. 2002;33.

51. Prévalence et incidence du diabète [Internet]. [cited 2022 Sep 21]. Available from: https://www.santepubliquefrance.fr/maladies-et-traumatismes/diabete/prevalence-et-incidence-du-diabete

52. Fuentes S, Mandereau-Bruno L, Regnault N, Bernillon P, Bonaldi C, Cosson E, et al. Is the type 2 diabetes epidemic plateauing in France? A nationwide population-based study. Diabetes Metab. 2020;46:472–9.

53. Piffaretti C, Mandereau-Bruno L, Guilmin-Crepon S, Choleau C, Coutant R, Fosse-Edorh S. Trends in childhood type 1 diabetes incidence in France, 2010-2015. Diabetes Res Clin Pract. 2019;149:200–7.

54. SPF. Évolution de la mortalité et de la surmortalité à 5 ans des personnes diabétiques traitées pharmacologiquement en France métropolitaine : comparaison des cohortes Entred 2001 et Entred 2007. Numéro thématique. Mortalité liée au diabète en France [Internet]. [cited 2022 Sep 21]. Available from: https://www.santepubliquefrance.fr/maladies-et-traumatismes/diabete/evolution-de-la-mortalite-et-de-la-surmortalite-a-5-ans-des-personnes-diabetiques-traitees-pharmacologiquement-en-france-metropolitaine-comparais

55. Harding JL, Shaw JE, Peeters A, Cartensen B, Magliano DJ. Cancer risk among people with type 1 and type 2 diabetes: disentangling true associations, detection bias, and reverse causation. Diabetes Care. 2015;38:264–70.

56. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020;323:1239–42.

57. Barron E, Bakhai C, Kar P, Weaver A, Bradley D, Ismail H, et al. Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study. Lancet Diabetes Endocrinol. 2020;8:813–22.

58. Cariou B, Wargny M, Boureau A-S, Smati S, Tramunt B, Desailloud R, et al. Impact of diabetes on COVID-19 prognosis beyond comorbidity burden: the CORONADO initiative. Diabetologia. 2022;

59. Géodes - Santé publique France - Indicateurs : cartes, données et graphiques [Internet]. [cited 2022 Sep 21]. Available from: https://geodes.santepubliquefrance.fr/#view=map2&c=indicator

60. Fosse S, Hartemann-Heurtier A, Jacqueminet S, Ha Van G, Grimaldi A, Fagot-Campagna A. Incidence and characteristics of lower limb amputations in people with diabetes. Diabet Med. 2009;26:391–6.

61. Fosse Edorh S, Mandereau Bruno L, Hartemann Heurtier A. Les hospitalisations pour complications podologiques chez les personnes diabétiques traitées pharmacologiquement en France en 2013. Bulletin Epidémiologique Hebdomadaire. 2015;638–44.

62. Fosse-Edorh S, Mandereau-Bruno L, Regnault N. Le poids des complications liées au diabète en France en 2013. Synthèse et perspectives. Bull Epidémiol Hebd. 2015;619–25.

63. Assogba GFA, Couchoud C, Roudier C, Pornet C, Fosse S, Romon I, et al. Prevalence, screening and treatment of chronic kidney disease in people with type 2 diabetes in France: the ENTRED surveys (2001 and 2007). Diabetes Metab. 2012;38:558–66.

64. Fagot-Campagna A, Fosse Edorh S, Romon I, Penfornis A, Lecomte P, Bourdel-Marchasson I, et al. Caractéristiques, risque vasculaire et complications chez les personnes diabétiques en France métropolitaine : d'importantes évolutions entre Entred 2001 et Entred 2007. Bull Epidémiol Hebd. 2009;450–5.

65. Cougnard-Grégoire A, Korobelnik J-F, Delyfer M-N, Rigalleau V, Daien V, Creuzot-Garcher C, et al. Trends in the Use of Eye Care Services in Adults Treated for Diabetes between 2008 and 2017 in France: A Nationwide Study. Ophthalmic Res. 2020;63:452–9.

66. Yau JWY, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. 2012;35:556–64.

67. Song P, Yu J, Chan KY, Theodoratou E, Rudan I. Prevalence, risk factors and burden of diabetic retinopathy in China: a systematic review and meta-analysis. J Glob Health. 2018;8:010803.

68. Heintz E, Wiréhn A-B, Peebo BB, Rosenqvist U, Levin L-A. Prevalence and healthcare costs of diabetic retinopathy: a population-based register study in Sweden. Diabetologia. 2010;53:2147–54.

69. Avisar R, Friling R, Snir M, Avisar I, Weinberger D. Estimation of prevalence and incidence rates and causes of blindness in Israel, 1998-2003. Isr Med Assoc J. 2006;8:880–1.

70. Report of the expert committee on the diagnosis and classification of diabetes mellitus - PubMed [Internet]. [cited 2022 Oct 6]. Available from: https://pubmed.ncbi.nlm.nih.gov/12502614/

71. American Diabetes Association Professional Practice Committee, Draznin B, Aroda VR, Bakris G, Benson G, Brown FM, et al. 6. Glycemic Targets: Standards of Medical Care in Diabetes-2022. Diabetes Care. 2022;45:S83–96.

72. Pop-Busui R, Januzzi JL, Bruemmer D, Butalia S, Green JB, Horton WB, et al. Heart Failure: An Underappreciated Complication of Diabetes. A Consensus Report of the American Diabetes Association. Diabetes Care. 2022;45:1670–90.

73. Ohkuma T, Komorita Y, Peters SAE, Woodward M. Diabetes as a risk factor for heart failure in women and men: a systematic review and meta-analysis of 47 cohorts including 12 million individuals. Diabetologia. 2019;62:1550–60.

74. Rosengren A, Edqvist J, Rawshani A, Sattar N, Franzén S, Adiels M, et al. Excess risk of hospitalisation for heart failure among people with type 2 diabetes. Diabetologia. 2018;61:2300–9.

75. Wamil M, Coleman RL, Adler AI, McMurray JJV, Holman RR. Increased Risk of Incident Heart Failure and Death Is Associated With Insulin Resistance in People With Newly Diagnosed Type 2 Diabetes: UKPDS 89. Diabetes Care. 2021;44:1877–84.

76. Leung AA, Eurich DT, Lamb DA, Majumdar SR, Johnson JA, Blackburn DF, et al. Risk of heart failure in patients with recent-onset type 2 diabetes: population-based cohort study. J Card Fail. 2009;15:152–7.

77. Hallgren Elfgren I-M, Grodzinsky E, Törnvall E. The Swedish National Diabetes Register in clinical practice and evaluation in primary health care. Prim Health Care Res Dev. 2016;17:549–58.

78. Nationella Diabetesregistret [Internet]. [cited 2022 Sep 22]. Available from: https://www.ndr.nu/#/variabler

79. Rawshani A, Rawshani A, Franzén S, Sattar N, Eliasson B, Svensson A-M, et al. Risk Factors, Mortality, and Cardiovascular Outcomes in Patients with Type 2 Diabetes. N Engl J Med. 2018;379:633–44.

80. Gabet A, Juillière Y, Lamarche-Vadel A, Vernay M, Olié V. National trends in rate of patients hospitalized for heart failure and heart failure mortality in France, 2000-2012. Eur J Heart Fail. 2015;17:583–90.

81. Bugger H, Abel ED. Molecular mechanisms of diabetic cardiomyopathy. Diabetologia. 2014;57:660–71.

82. Giamouzis G, Schelbert EB, Butler J. Growing Evidence Linking Microvascular Dysfunction With Heart Failure With Preserved Ejection Fraction. J Am Heart Assoc. 2016;5:e003259.

83. Boudina S, Sena S, Theobald H, Sheng X, Wright JJ, Hu XX, et al. Mitochondrial energetics in the heart in obesity-related diabetes: direct evidence for increased uncoupled respiration and activation of uncoupling proteins. Diabetes. 2007;56:2457–66.

84. Ishikawa T, Kajiwara H, Kurihara S. Alterations in contractile properties and Ca2+ handling in streptozotocin-induced diabetic rat myocardium. Am J Physiol. 1999;277:H2185-2194.

85. Choi KM, Zhong Y, Hoit BD, Grupp IL, Hahn H, Dilly KW, et al. Defective intracellular Ca(2+) signaling contributes to cardiomyopathy in Type 1 diabetic rats. Am J Physiol Heart Circ Physiol. 2002;283:H1398-1408.

86. Kenny HC, Abel ED. Heart Failure in Type 2 Diabetes Mellitus. Circ Res. 2019;124:121–41.

87. Pop-Busui R, Cleary PA, Braffett BH, Martin CL, Herman WH, Low PA, et al. Association between cardiovascular autonomic neuropathy and left ventricular dysfunction: DCCT/EDIC study (Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications). J Am Coll Cardiol. 2013;61:447–54.

88. Gerstein HC, Nair V, Chaube R, Stoute H, Werstuck G. Dysglycemia and the Density of the Coronary Vasa Vasorum. Diabetes Care. 2019;42:980–2.

89. Wargny M, Gallini A, Hanaire H, Nourhashemi F, Andrieu S, Gardette V. Diabetes Care and Dementia Among Older Adults: A Nationwide 3-Year Longitudinal Study. J Am Med Dir Assoc. 2018;19:601-606.e2.

90. Henny J, Nadif R, Got SL, Lemonnier S, Ozguler A, Ruiz F, et al. The CONSTANCES Cohort Biobank: An Open Tool for Research in Epidemiology and Prevention of Diseases. Front Public Health. 2020;8:605133.

91. Caisse Nationale de l'Assurance Maladie (CNAM). Méthodologie médicale de la cartographie des pathologies et des dépenses, version G8 (années 2015 à 2019, Tous Régimes) [Internet]. 2021 [cited 2022 Jan 20]. Available from: https://assurance-maladie.ameli.fr/sites/default/files/2021_methode-reperage-pathologies_cartographie_1.pdf

92. ALD n°13 - Maladie coronarienne [Internet]. Haute Autorité de Santé. [cited 2022 Jan 23]. Available from: https://www.has-sante.fr/jcms/c_534304/fr/ald-n13-maladie-coronarienne

93. ALD n°1 - Accident vasculaire cérébral [Internet]. Haute Autorité de Santé. [cited 2022 Jan 23]. Available from: https://www.has-sante.fr/jcms/c_534745/fr/ald-n1-accident-vasculaire-cerebral

94. ALD n°3 - Artériopathie oblitérante des membres inférieurs [Internet]. Haute Autorité de Santé. [cited 2022 Jan 23]. Available from: https://www.has-sante.fr/jcms/c_534760/fr/ald-n3-arteriopathie-obliterante-des-membres-inferieurs

95. Grave C. HOSPITALISATIONS POUR VALVULOPATHIE EN FRANCE : CARACTÉRISTIQUES DES PATIENTS ET ÉVOLUTION 2006-2016 / HOSPITALIZATIONS FOR VALVULAR HEART DISEASE IN FRANCE: PATIENTS CHARACTERISTICS AND TRENDS 2006-2016. :10.

96. Giral P, Neumann A, Weill A, Coste J. Cardiovascular effect of discontinuing statins for primary prevention at the age of 75 years: a nationwide population-based cohort study in France. Eur Heart J. 2019;40:3516–25.

97. SNDS : Système National des Données de Santé | CNIL [Internet]. [cited 2022 Sep 6]. Available from: https://www.cnil.fr/fr/snds-systeme-national-des-donnees-de-sante#:~:text=Cr%C3%A9%C3%A9%20par%20la%20loi%20de,donn%C3%A9es%20de%20sant%C3%A9%20publiques%20existantes.

98. Identifiants des bénéficiaires | Documentation du SNDS [Internet]. [cited 2022 Sep 6]. Available from: https://documentation-snds.health-data-hub.fr/fiches/fiche_beneficiaire.html#utilisation-des-identifiants

99. Caisse Nationale de l'Assurance Maladie (CNAM). SNDS Fiche pratique : Thème Bénéficiaires / Notions de bénéficiaires [Internet]. 2020 [cited 2022 Aug 9]. Available from: https://documentation-snds.health-data-hub.fr/files/Cnam/2019-06_CNAM-INDS_SNDS_Fiches_Thematiques_BENEF_MAJ-2020-09_MPL-2.0.pdf

100. Equipe SNDS de la Direction Appui, Traitements et Analyses des données. SNDS - Ce qu'il faut savoir [Internet]. 2021 [cited 2022 Aug 9]. Available from: https://documentation-snds.health-data-hub.fr/files/Sante_publique_France/2021-10-SpF-SNDS-ce-quil-faut-savoir-v3-MPL-2.0.pdf

101. SAS : logiciels et solutions Analytiques, IA & Data Management [Internet]. [cited 2022 Sep 7]. Available from: https://www.sas.com/fr_fr/home.html

102. Visualisation de la structure du SNDS [Internet]. [cited 2022 Sep 7]. Available from: https://health-data-hub.shinyapps.io/dico-snds/

103. Requête type de sélection des affections de longue durée (ALD) | Documentation du SNDS [Internet]. [cited 2022 Sep 7]. Available from: https://documentation-snds.health-data-hub.fr/fiches/requete_type_ald.html#le-referentiel-medicalise-ir-imb-r

104. Médicaments [Internet]. [cited 2022 Jan 23]. Available from: https://www.ameli.fr/etablissement/exercice-professionnel/nomenclatures-codage/medicaments

105. Therneau T, Crowson C, Atkinson E. Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model. :31.

106. Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Oudshoorn CGM. Time-varying covariates and coefficients in Cox regression models. Ann Transl Med. 2018;6:121.

107. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics--2015 update: a report from the American Heart Association. Circulation. 2015;131:e29-322.

108. Wiener RS, Wiener DC, Larson RJ. Benefits and risks of tight glucose control in critically ill adults: a meta-analysis. JAMA. 2008;300:933–44.

109. Krinsley JS. Glycemic variability: a strong independent predictor of mortality in critically ill patients. Crit Care Med. 2008;36:3008–13.

110. Gerbaud E, Darier R, Montaudon M, Beauvieux M-C, Coffin-Boutreux C, Coste P, et al. Glycemic Variability Is a Powerful Independent Predictive Factor of Midterm Major Adverse Cardiac Events in Patients With Diabetes With Acute Coronary Syndrome. Diabetes Care. 2019;42:674–81.

111. Lazzeri C, Valente S, Chiostri M, D'Alfonso MG, Gensini GF. Prognostic impact of early glucose variability in acute heart failure patients: a pilot study. Int J Cardiol. 2014;177:693–5.

112. Danne T, Nimri R, Battelino T, Bergenstal RM, Close KL, DeVries JH, et al. International Consensus on Use of Continuous Glucose Monitoring. Diabetes Care. 2017;40:1631–40.

113. Granger CB, Goldberg RJ, Dabbous O, Pieper KS, Eagle KA, Cannon CP, et al. Predictors of hospital mortality in the global registry of acute coronary events. Arch Intern Med. 2003;163:2345–53.

114. Serruys PW, Onuma Y, Garg S, Sarno G, van den Brand M, Kappetein A-P, et al. Assessment of the SYNTAX score in the Syntax study. EuroIntervention. 2009;5:50–6.

115. Prodigy · An annotation tool for AI, Machine Learning & NLP [Internet]. Prodigy. [cited 2022 Aug 8]. Available from: https://prodi.gy

116. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J. 2018;60:431–49.

117. Serruys PW, Onuma Y, Garg S, Sarno G, van den Brand M, Kappetein A-P, et al. Assessment of the SYNTAX score in the Syntax study. EuroIntervention. 2009;5:50–6.

118. L'Essentiel du SNDS | Documentation du SNDS [Internet]. [cited 2022 Sep 17]. Available from: https://documentation-snds.health-data-hub.fr/formation_snds/documents_cnam/essentiel_snds.html#le-circuit-d-alimentation-du-snds

119. Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, et al. The national healthcare system claims databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2017;26:954–62.

120. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12:e1001779.

121. Eliasson B, Gudbjörnsdottir S. Diabetes care--improvement through measurement. Diabetes Res Clin Pract. 2014;106 Suppl 2:S291-294.

122. Carstensen B, Rønn PF, Jørgensen ME. Prevalence, incidence and mortality of type 1 and type 2 diabetes in Denmark 1996-2016. BMJ Open Diabetes Res Care. 2020;8:e001071.

123. Weill A, Païta M, Tuppin P, Fagot J-P, Neumann A, Simon D, et al. Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. Pharmacoepidemiol Drug Saf. 2010;19:1256–62.

124. Hoisnard L, Laanani M, Passeri T, Duranteau L, Coste J, Zureik M, et al. Risk of intracranial meningioma with three potent progestogens: A population-based case-control study. Eur J Neurol. 2022;29:2801–9.

125. Ammirati E, Lupi L, Palazzini M, Hendren NS, Grodin JL, Cannistraci CV, et al. Prevalence, Characteristics, and Outcomes of COVID-19-Associated Acute Myocarditis. Circulation. 2022;145:1123–39.

126. Martin-Blondel G, Lescure F-X, Assoumou L, Charpentier C, Chapplain J-M, Perpoint T, et al. Increased risk of severe COVID-19 in hospitalized patients with SARS-CoV-2 Alpha variant infection: a multicentre matched cohort study. BMC Infectious Diseases. 2022;22:540.

ANNEXES

Annexe 1. Autorisation CNIL de l'EDS nantais, conditions d'accès et gouvernance

L'outil eHOP a été mis à disposition du CHU de Nantes depuis 2017 et son exploitation est autorisée par la CNIL depuis juillet 2018 (demande n°2129203, délibération n°2018-295) suite aux efforts de la Direction de la Recherche et de l'Innovation (DRI), efforts coordonnés par Marie Lebigre, chargée de mission pour le SI recherche. Cela fait de l'EDS nantais le premier du groupe HUGO à avoir été autorisé par la CNIL. Cette autorisation définit en particulier les obligations d'information aux patients, la gouvernance de l'outil, et ses finalités.

L'autorisation CNIL de l'EDS du CHU de Nantes est conditionnée à certains obligations d'information, individuelles ou collectives, générales ou propres à un projet spécifique. Pour résumer, l'exploitation des données à visée de recherche est possible pour tous les patients ayant consulté ou ayant été hospitalisés au CHU de Nantes depuis le mois d'août 2018, date à partir de laquelle ils sont réputés avoir reçu la note d'information relative à la constitution de l'EDS. L'information est à présent mentionnée de façon systématique dans les CR de consultations (avec impossibilité pour le rédacteur de supprimer l'information) et une note d'information doit être transmise de façon automatique lors de la prise de rendez-vous. La note d'information est relative à l'usage potentiel des données concernant le patient et aux modalités d'exercice de ses droits, en particulier son droit d'opposition. Par ailleurs, afin d'informer les patients n'ayant pas été admis au CHU de Nantes après le mois d'août 2018, une campagne de communication par les médias locaux a été mise en place[23,24] et une information est donnée sur le site Internet du CHU et par affichage interne à l'établissement (hall principal des différents bâtiments)[25]. Pour ces derniers patients, l'information individuelle reste cependant requise pour les projets de recherche multicentriques (inter-services) nécessitant des informations potentiellement identifiantes.

Ainsi les personnes suivies au CHU de Nantes peuvent, par voie de courrier postal ou électronique, nous informer de leur refus de voir leurs données utilisées soit pour un projet de recherche précis, soit pour l'ensemble des finalités de l'EDS. Dans ce dernier cas, leur opposition est enregistrée dans une

[23] Vidéo accessible sur Télénantes : https://telenantes.ouest-france.fr/societe/info-soir/article/info-soir-du-lundi-24-juin

[24] https://www.ouest-france.fr/pays-de-la-loire/chu-de-nantes-ces-millions-de-donnees-si-precieuses-6409068

[25] https://www.chu-nantes.fr/le-traitement-de-vos-donnees-personnelles-au-chu-de-nantes

table d'eHOP, et les requêtes via l'interface graphique indiqueront le nombre et la proportion de patients opposés sans donner accès à leur identité.

La gouvernance de l'outil est, elle aussi, liée à l'autorisation donnée par la CNIL. Elle pose en particulier les principes de formulation des demandes d'exploitation de l'outil et du circuit de ses demandes. Elle est résumée dans la **Figure 15**, et dépend notamment du besoin (dénombrement, *screening*, étude ou autre) et de la période d'intérêt (couvrant ou non la période antérieure à l'autorisation CNIL et donc à la campagne d'information). Le circuit de la demande est le suivant :

- Le demandeur doit travailler au CHU de Nantes. Il effectue une demande par un portail informatique accessible via l'Intranet du CHU, une procédure qui ne prend que quelques minutes
- Si la demande concerne un usage d'eHOP, elle est communiquée à l'équipe de la Clinique des Données, sous la responsabilité du Pr Pierre-Antoine Gourraud. L'équipe se réunit en staff hebdomadaire, staff où le SIM est également représenté (Dr Christophe Leux). Les demandes sont discutées collectivement et attribuées à l'un des agents à l'issue du staff. En cas de difficultés réglementaires, l'agent sollicite d'autres membres de l'équipe ou le personnel compétent

Enfin, les finalités autorisées définies par la CNIL sont distribuées en deux catégories :

- Finalités liées à la recherche : screening, dénombrement, études observationnelles dont médico-économiques…
- Finalités hors recherche : appui à la prise de décision médicale, maîtrise des vigilances et des risques, optimisation de l'organisation des soins…

L'exploitation des EDS n'est donc pas limitée à la recherche et son développement est appelé à appuyer d'autres besoins des CHU, comme l'organisation du système de soin ou l'optimisation de la tarification à l'activité.

Charte à signer par tous les personnes impliquées par le projet et sous la responsabilité d'un investigateur du CHU de Nantes

ANNEXE : circuit des demandes d'exploitation entrepôt de données

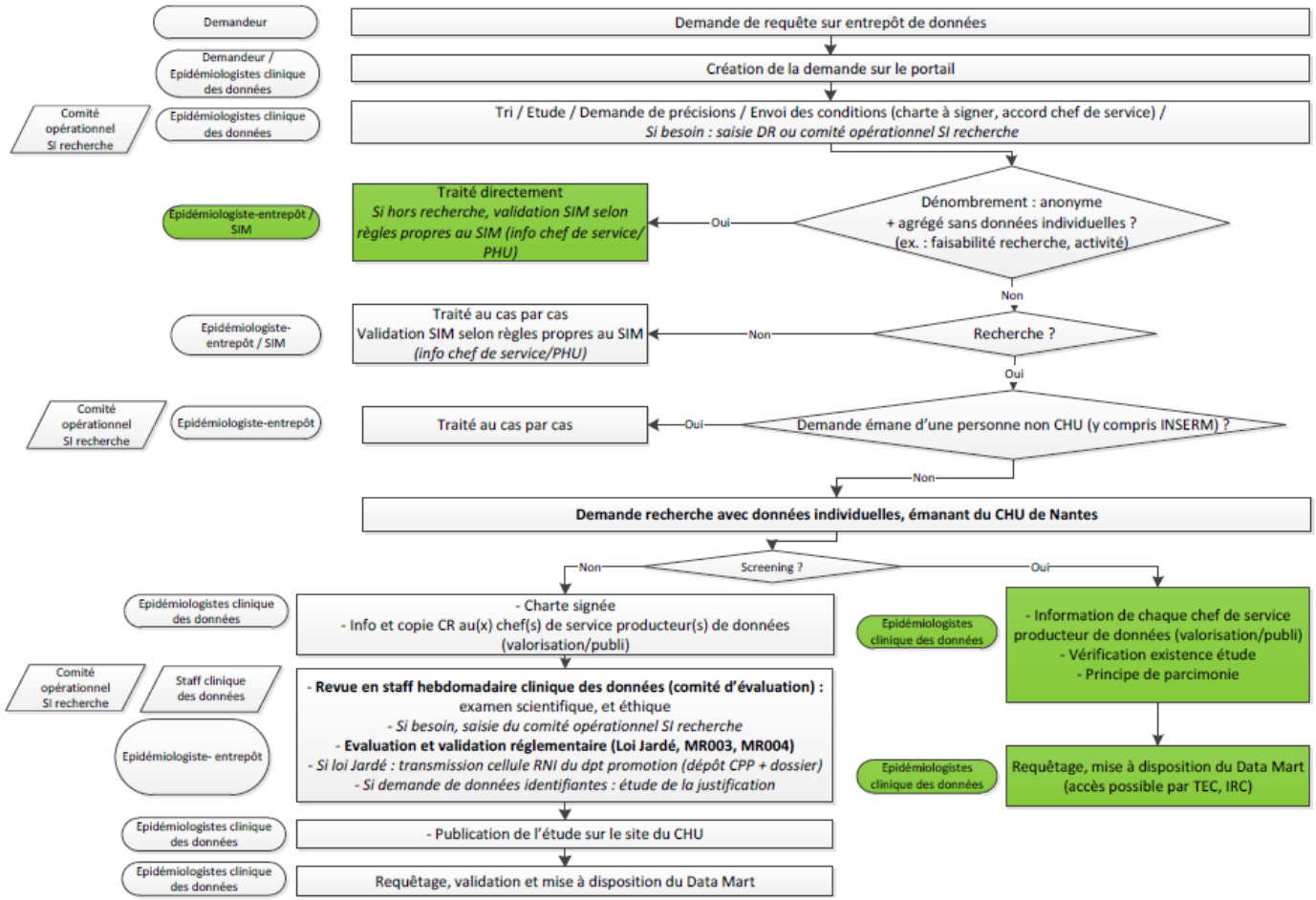


Figure 15. EDS nantais - Circuit des demandes d'accès à eHOP au CHU de Nantes

Annexe 2. Cas d'usage eHOP (1) : dénombrement dans l'EDS nantais

Comme son nom l'indique, le dénombrement consiste à obtenir le nombre de patients correspondant à un profil précis dans la base.

Dans le cadre d'une étude observationnelle, un laboratoire souhaite identifier le nombre de patients présentant une maladie relativement rare et suivis au CHU de Nantes. Prenons l'exemple de la maladie de Verneuil, aussi appelé hidrosadénite aiguë. Bien que théoriquement identifiable par un code CIM-10 (L73.2 – Hidrosadénite suppurée), cette affection sera rarement codée dans le PMSI car ne donnant pas lieu à hospitalisation, et ne contribuant pas à la valorisation d'une hospitalisation pour une autre cause. La réunion des termes « Verneuil » ou « hidrosadénite » est cependant à la fois sensible et spécifique : ils pourront donc être utilisés pour une recherche « texte entier » et permettront l'identification de plusieurs centaines de patients par an.

L'interface ergonomique d'eHOP permet de présenter sur le même écran la liste de patients et le contexte d'utilisation du mot-clef pour un patient donné, ce qui évite la démultiplication des clics. Le segment de phrase où apparaît le mot-clef d'intérêt est présenté sans ouvrir le dossier du patient (cf. **Figure 16**), une approche très confortable pour l'utilisateur qui lui permet de contrôler la validité d'une centaine de cas en quelques dizaines de minutes. Le nombre de cas est ensuite transmis à l'équipe en lien avec le laboratoire. Dans le cadre d'une étude inter-régionale, les modalités de la recherche, ici élémentaires, peuvent être transmises aux autres CDC participants. Le dénombrement obtenu nous renseignera sur la faisabilité de l'étude. S'agissant d'un simple dénombrement sans nécessité d'identification du patient, aucune procédure réglementaire n'aura été nécessaire, bien que l'on se situe dans le cadre de la recherche.

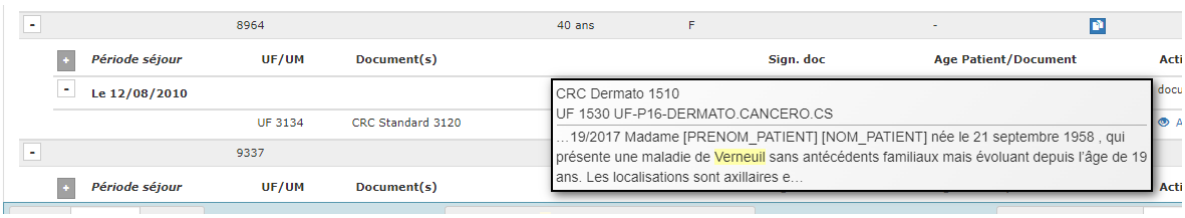


Figure 16. eHOP - Capture d'écran du résultat d'un screening *à visée de dénombrement sur l'ensemble des patients du CHU de Nantes, réalisé le 11/08/2022, à partir du seul mot-clef "Verneuil".*

La recherche textuelle a demandé un peu moins de 5 secondes. Le passage du pointeur de la souris sur le dossier d'un patient affiche un certain nombre de mots avant et après le mot-clef d'intérêt. Comme cela est visible sur la capture ci-dessus, la procédure de désidentification automatique a fonctionné

partiellement en cachant nom et prénom de la patiente mais la date de naissance, exprimée littéralement et non sous un format JJ/MM/AAAA, n'a pas été désidentifiée.

Annexe 3. Cas d'usage eHOP (2) : *screening* dans l'EDS nantais

Le *screening* est un cas plus évolué de dénombrement, où la liste des patients identifiés est mise à disposition d'un investigateur.

Cela pose en particulier la question de la protection de l'identité du patient puisque, contrairement au dénombrement, des données individuelles non anonymes sont susceptibles d'être enregistrées. La levée de l'anonymat n'est cependant pas toujours nécessaire : si l'investigateur peut récupérer toutes les données d'intérêt via eHOP sans requérir à un autre logiciel hospitalier, il peut se voir mettre à disposition des sources désidentifiées sans accès direct au nom des patients. Par contre, la levée de l'anonymat sera nécessaire s'il doit contacter le patient ou compléter sa recherche par d'autres sources, numériques ou papier.

Prenons le cas réel d'un soutien à une étude internationale. L'équipe de cardiologie du CHU de Nantes (Dr Nicolas Piriou) souhaitait identifier tous les patients admis au CHU pour une infection à COVID-19 compliquée de myocardite.[125] Dans le cadre du suivi de l'épidémie, un travail d'identification des patients admis pour COVID-19 a déjà été réalisé par le service d'information médicale (SIM). La liste des identifiants personnels des patients est transmise à l'équipe de la Clinique des Données, qui la charge dans eHOP et constitue un « *datamart* », c'est-à-dire un sous-groupe de patients directement interrogeable par eHOP. Le terme « myocardite » est jugé suffisamment sensible et spécifique pour la question, nous lançons sur ce *datamart* une recherche qui identifie 8 individus parmi les 493 ayant été admis pour COVID-19. Sur ces petits effectifs, les cas de formule négative (« pas de myocardite ») ou hors sujet (« antécédent de myocardite ») sont aisément exclus un par un (**cf. Figure 17**). La liste de 8 patients sera transmise aux cardiologues qui retiendront deux cas confirmés. Comme il s'agit d'une étude sur des hospitalisations postérieures à l'autorisation CNIL de juillet 2018, aucune procédure réglementaire n'est requise dès lors que les patients déclarés comme opposés sont automatiquement exclus de la recherche. Dans le cas de l'épidémie de COVID-19 cette approche a bien sûr été répliquée pour d'autres sous-groupes, comme l'identification en février 2021 des cas de variant anglais avec l'équipe de Maladie Infectieuse (Dr Paul Le Turnier pour Nantes, crédité dans le *CoCliCo study group* [126]), ou les cas de pneumopathie organisée (Dr Stéphanie Piriou et Baptiste Artignan, thèse d'exercice de pneumologie, article en cours de rédaction).

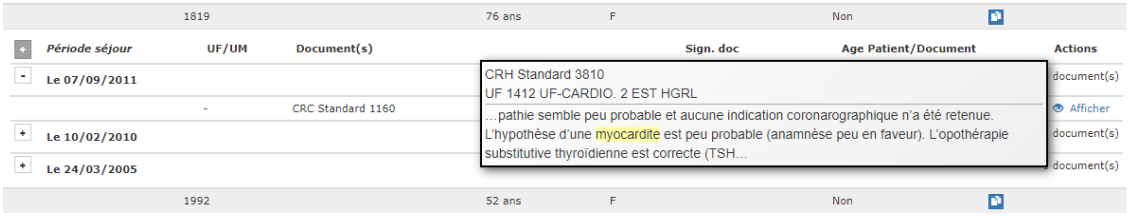


Figure 17. eHOP - Capture d'écran du résultat d'un screening *sur l'ensemble des patients du CHU de Nantes, réalisé le 13/08/2022, à partir du seul mot-clef "myocardite%".*

Lorsque le pointeur survole l'icône de l'œil (« Afficher », tout à droite) pour un document d'un patient donné, le texte associé au mot-clef apparaît en surbrillance, nous permettant d'affiner rapidement le screening. Ici, le terme n'apparaît que dans la phrase « myocardite peu probable », et n'est plus utilisé dans le reste des CR du séjour du patient, il peut donc être considéré comme « *screené* » à tort.

Annexe 4. Cas d'usage eHOP (3) : enrichissement d'une base existante

Comme on l'aura compris, un intérêt majeur de la mise en place d'un EDS est la facilitation de l'accès aux données et le croisement multi-source. Sans EDS, un investigateur souhaitant accéder aux données des patients inclus dans l'une de ses études devra s'adresser au personnel du SIM pour les données PMSI, aux ingénieurs du pôle de biologie pour les données de biologie, aux ingénieurs en charge de tel logiciel de spécialité pour accéder aux constantes numérisées…

Avec les EDS, le clinicien dispose d'un interlocuteur unique qui a accès à ces différentes sources d'information. C'est ce que le Pr Hadjadj a sollicité pour l'étude EDIT (de Nantes). EDIT est une étude de cohorte prospective visant à caractériser précisément les patients présentant un diabète de type 2. Les informations d'EDIT sont recueillies par différentes voies, incluant des questionnaires informatisés (eCRF). Mais EDIT est aussi enrichie directement à partir de la liste des IPP des patients, en y appariant les données de biologie et certains codes associés aux hospitalisations, notamment CIM-10. Cette action était tout à fait possible avant la mise en place des EDS, mais elle est à présent facilitée et accélérée par la centralisation des données et des acteurs.

Annexe 5. DMC - Problèmes rencontrés pour la correspondance entre classes ATC et doses délivrées

Dans DMC, plusieurs éléments vont compliquer la récupération automatisée des informations sur les médicaments lors de la correspondance entre les tables PHARMACIE et le thésaurus

- **Les associations de médicaments.** Non seulement il sera nécessaire de déduire les deux (exceptionnellement trois) médicaments associés à une classe ATC relative à une association, mais la dose ne sera pas disponible sous forme structurée dans le thésaurus. Il sera donc nécessaire de la récupérer « manuellement » à partir du libellé.
    - P. ex. : classe ATC A10BD10 pour l'association de la metformine et de la saxagliptine

- **Le caractère incomplet du thésaurus fourni par la CNAM,** le nombre d'unités délivrés pouvant être vide (rarement) ou nul (très fréquent). Ce n'est pas anecdotique puisque cela représentait 1186 (14%) des codes CIP d'intérêt sur les 8432 extraits. Le thésaurus comprenait également quelques erreurs de saisie.

- **Pour certaines solutions injectables, comme l'insuline**, il est nécessaire de connaître le nombre de stylos, le nombre de mL par stylo, et le nombre d'unités du produit par mL. Cette information n'était pas toujours disponible dans les informations structurées, et jamais de façon directement analysable (« 5/3 », par exemple, pour signifier 5 stylos de 3 mL).

- **Pour les insulines mixtes,** qui combinent une insuline rapide et une insuline lente (ex. NOVOMIX 30), la fraction de l'insuline rapide n'est pas disponible de façon structurée, et il a été nécessaire de modifier le thésaurus en interprétant au cas par cas les libellés.

- **Enfin, d'autres unités n'étaient pas directement analysables, pour différentes raisons**
    - Dose correspondant à la masse du comprimé et non à celle du principe actif
        - Ex. CIP 3971247 pour le clopidogrel : notée 111.86 pour 75 mg
    - Conditionnement contenant différentes doses d'une même molécule
        - Ex. CIP 3007885 du rivaroxaban (XARELTO) : dose notée 15+20 pour 42 cp de 15 mg et 7 cp de 20 mg

- Pour la metformine sous forme STAGID, la dénomination « STAGID 700 » correspond en réalité à 280 mg du principe actif metformine

Par ailleurs, certaines difficultés n'étaient pas liées au thésaurus ou au format des données SNDS mais à la classe ATC elle-même. Trois médicaments n'ont pas de dose définie quotidienne (DDD) disponible dans le thésaurus de l'OMS – une DDD a été proposée, basée sur la prescription habituelle, pour la ciclétanine, le pirétanide et l'altizide.

Annexe 6. GAVROCHE - Avis favorable du CSE HUGO

Groupement de Coopération Sanitaire
des Hôpitaux Universitaires du Grand Ouest

Comité Scientifique et Éthique de la plateforme Ouest Data Hub

Avis d'évaluation
Session du 2 décembre 2020

Date de réception : le 25/11/2020	**Établissement : CHU de Nantes** **Responsable scientifique :** Pr Samy Hadjadj Chef du service EDN, Endocrinologie, Diabétologie, Nutrition du CHU de Nantes
Titre : Glycemia And its Variability with regard tO Congestive Heart failurE GAVROCHE	
AVIS FAVORABLE	

Recommandations et suggestions :

- Les contextes scientifique, médical, épidémiologique sont clairement exposés.
- L'intérêt du projet est justifié.
- Sur le plan de la forme, les objectifs de l'étude pourraient être mieux mis en perspectives dans les différentes parties du protocole
- La période de recrutement doit être en phase avec les dates d'hospitalisations.
- Le mode de constitution de l'échantillon de compte-rendus et le traitement local sur les compte-rendus sont à décrire.
- Les critères de sélection des patients en Insuffisance cardiaque aigüe sont à revoir et à affiner avec les médecins DIM.
- Afin d'assurer le chaînage des données entre les différents centres pour des patients ayant fréquenté plusieurs établissements la procédure de hachage du NIR fournie par le GCS HUGO pourra être utilisée et est à intégrer au protocole.
- L'absence de risque pour le patient n'est pas un justificatif à la demande de dérogation à l'information individuelle.

Tours, le 19/01/2021
Pour le Comité Scientifique et Éthique,

F. Ossant
secrétaire du CSE.

1/1

Annexe 7. GAVROCHE - Avis CESREES : favorable avec recommandation

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE DES SOLIDARITÉS ET DE LA SANTÉ
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION

Comité Éthique et Scientifique pour les Recherches, les Études et les Évaluations dans le domaine de la Santé (CESREES)

Avis du Comité en date du 11 février 2021

Numéro de dossier :	3360056
Titre du projet :	GAVROCHE "Glycemia And its Variability with Regard tO Congestive Heart failurE"
Responsable de Traitement :	CHU de Nantes

	OUI	NON
L'étude est conforme à l'éthique	X	
L'étude présente un intérêt scientifique et/ou social	X	
L'étude présente un caractère d'intérêt public	X	

▫ Avis favorable

▫ Avis réservé

X Avis favorable avec recommandations

▫ Avis défavorable

Observations du Comité :

Il convient de préciser les modalités d'information collective avant de soumettre le dossier à la CNIL. Un point d'attention doit être porté sur la circulation du NIR.

Il convient par ailleurs dans le résumé d'enlever la mention en gras ci-après car non conforme au GPRD : "une demande de dérogation à l'information individuelle est soumise auprès de la CNIL, ce qui nous apparaît justifié **du fait du caractère à la fois inutile** et disproportionné de l'information individuelle."

Le projet n'appelle pas d'observations quant à son caractère d'intérêt public ni sur sa conformité à l'éthique.

Dans le cas où l'avis est réservé, le responsable de traitement ou, par délégation, le responsable scientifique est invité, dans les meilleurs délais, à signifier au Health Data Hub s'il souhaite procéder à une modification de son dossier pour un nouvel examen par le CESREES ou s'il demande que le Health Data Hub dépose en l'état son étude auprès de la CNIL pour autorisation.
Si la première option est retenue, un nouveau délai d'examen suivra la réception par le CESREES, du dossier modifié.
L'ensemble des modifications apportées devront être obligatoirement apparentes.

Pour le CESREES, le Président
Bernard Nordlinger

Le 11 février 2021

DocuSigned by:
Bernard Nordlinger
2DACCE28E904424...

Annexe 8. GAVROCHE - Autorisation CNIL

CNIL.
COMMISSION NATIONALE INFORMATIQUE & LIBERTÉS

La Présidente

Monsieur Philippe EL SAÏR
DIRECTEUR GENERAL
CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES
5 ALLEE DE L'ILE GLORIETTE - IMMEUBLE CAP OUEST
44093 - NANTES CEDEX 1

Paris, le 3 décembre 2021

N/Réf. : MLD/CBO/AVL/AR2110432

Objet : AUTORISATION

Décision DR-2021-328 autorisant le CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES à mettre en œuvre un traitement de données ayant pour finalité une étude portant sur l'association entre la variabilité glycémique et la mortalité hospitalière chez des patients hospitalisés pour insuffisance cardiaque aiguë, réalisée au sein de la plateforme « Ouest Data Hub », intitulée « GAVROCHE ». (Demande d'autorisation n° 921208)

La Commission a été saisie d'une demande d'autorisation relative à un traitement de données à caractère personnel. Ce traitement, dont la finalité présente un caractère d'intérêt public, relève de la procédure prévue aux articles 66, 72 et suivants de la loi du 6 janvier 1978 modifiée.

Responsable de traitement	Le Centre Hospitalier Universitaire de Nantes
Avis du comité	Avis favorable avec recommandations du Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé du 11 février 2021.
Finalité	Etude portant sur l'association entre la variabilité glycémique et la mortalité hospitalière chez des patients hospitalisés pour insuffisance cardiaque aiguë, intitulée « GAVROCHE ».
Observations liminaires sur le déploiement de la plateforme « HUGO », la responsabilité de traitement et la sous-traitance	Les établissements du Groupement de Coopération Sanitaire des Hôpitaux Universitaires du Grand Ouest (GCS HUGO) ont développé leur entrepôt de données de santé (EDS) à partir d'une plateforme technologique identique : le système eHOP (« entrepôt-Hôpital »). A l'échelle locale, chaque établissement a créé une structure dédiée, le Centre de données cliniques (CDC). Cette structure permet l'intégration des données des patients issues de différents flux du système d'information hospitaliers (SIH), et apporte l'expertise et les services nécessaires à leur exploitation.

	Le GCS HUGO s'est doté d'une plateforme technologique pour le traitement des données massives en santé, dénommée plateforme « Ouest Data Hub » (« ODH »). Il est responsable de l'intégration des données sur la plateforme technologique, de la pseudonymisation des données de la base projet (deuxième pseudonymisation), de la création de la base projet, de la gestion des comptes utilisateurs projet, de l'attribution des droits sur la base projet, du suivi des traces d'activité des gestionnaires HUGO et des utilisateurs sur l'espace projet. La Commission relève que dans le cadre du déploiement de la plateforme ODH, le Centre hospitalier universitaire de Nantes : - est responsable de l'hébergement des données et de l'exploitation technique de la plateforme ; - met à disposition un espace projet dédié, des outils de gestion des comptes et de gestion des traces d'activité sur la plateforme ; - est responsable de la gestion des comptes gestionnaires. S'agissant plus spécifiquement du projet de recherche « GAVROCHE », la Commission relève que : - le GCS Hugo intervient en qualité de sous-traitant dans le cadre de la mise en œuvre de ce projet de recherche. Le traitement des données par le sous-traitant devra être régi par un contrat ou un acte juridique conformément à l'article 28 du RGPD ; - le Centre hospitalier universitaire de Nantes, qui intervient en qualité de responsable de traitement dans le cadre de ce projet, sera chargé de l'hébergement des données et qu'il est certifié hébergeur de données de santé (certificat n° 725810 du 11 décembre 2020).
Points de non-conformité à la méthodologie de référence concernée	La Commission prend acte que le dossier de demande mentionne que le traitement envisagé est conforme aux dispositions de la méthodologie de référence MR-004, à l'exception de la nature des données collectées et des modalités d'information des personnes. En dehors de ces exceptions, ce traitement devra respecter le cadre prévu par la méthodologie de référence MR-004.

Réutilisation de bases de données existantes	Les données issues des entrepôts de données suivants seront réutilisées dans le cadre de ce projet de recherche : - l'entrepôt du Centre hospitalier universitaire d'Angers dénommé « eHop Angers » (demande d'autorisation n° 2215237 - décision DT-2020-003 du 24 août 2020) ; - l'entrepôt du Centre hospitalier universitaire de Brest (demande d'autorisation n° 2217775) ; - l'entrepôt du Centre hospitalier universitaire de Nantes dénommé « eHop » (demande d'autorisation n° 2129203 – délibération n° 2018-295 du 19 juillet 2018) ; - l'entrepôt du Centre hospitalier universitaire de Rennes, dénommé « eHop Rennes » (demande d'autorisation n° 2212496) ; - l'entrepôt du Centre hospitalier universitaire de Tours (demande d'autorisation n° 2212853 - Délibération n° 2020-028 du 27 février 2020).
Catégories particulières de données traitées (autres que données de santé)	*S'agissant de la collecte du numéro d'inscription au répertoire national d'identification des personnes physiques* : Les données des Centres de données cliniques seront appariées entre elles grâce au NIR des personnes concernées. Afin d'assurer la pseudonymisation et le chaînage des données du projet, un traitement du NIR est réalisé localement dans chaque établissement, par application d'une fonction de hachage à clé secrète à l'aide d'un logiciel « boîte noire » disposant des secrets spécifiques au projet. Il est traité au sein de la DSI de chaque établissement par une personne en charge de la gestion de l'application administrative du système d'information de l'établissement, soumise au secret professionnel, dans un environnement sécurisé. Cette personne est habilitée à y accéder, dans le cadre de ses missions ; elle sera la seule à accéder à cette donnée pour le projet et ses actions seront tracées. Les données identifiantes et le NIR sont manipulés dans des fichiers temporaires qui ne portent pas le nom de l'étude, sont stockés dans un espace d'accès restreint et sont effacés dès qu'ils ne sont plus utiles. Le NIR n'est pas exporté et le pseudonyme issu de son traitement fait l'objet d'une deuxième pseudonymisation lors de sa réception sur la plateforme ODH.

Information et droits des personnes	*S'agissant des patients ayant reçu une note d'information individuelle lors de leur admission aux centres hospitaliers universitaires participants prévoyant un dispositif spécifique d'information auquel ils peuvent se reporter préalablement à la mise en œuvre de chaque nouvelle étude réalisée à partir de leurs données :* Une note d'information sera diffusée sur le site web du responsable de traitement de l'étude, de la plateforme « HUGO » et des centres participants. Elle devra comporter l'ensemble des mentions prévues par le Règlement général sur la protection des données. *S'agissant des patients n'ayant pas reçu une note d'information individuelle renvoyant vers un tel dispositif spécifique d'information :* En application de l'article 69 de la loi « Informatique et Libertés » et de l'article 14, 5, b), du Règlement général sur la protection des données, l'obligation d'information individuelle de la personne concernée peut faire l'objet d'exceptions dans l'hypothèse où la fourniture d'une telle information se révélerait impossible, exigerait des efforts disproportionnés ou compromettrait gravement la réalisation des objectifs du traitement. En pareils cas, le responsable de traitement prend des mesures appropriées pour protéger les droits et libertés, ainsi que les intérêts légitimes de la personne concernée, y compris en rendant les informations publiquement disponibles. En l'espèce, il sera fait exception au principe d'information individuelle de ces personnes et des mesures appropriées seront mises en œuvre, notamment par la diffusion d'une note d'information relative au projet de recherche qui devra comporter l'ensemble des mentions prévues par le Règlement général sur la protection des données sur les sites web suivants : • celui du Centre hospitalier universitaire de Nantes ; • celui des centres hospitaliers universitaires participant au projet ; • celui de la plateforme HUGO. *S'agissant des modalités d'exercice des droits des personnes :* Conformément aux principes de transparence et de loyauté de l'article 5 du RGPD et comme précisé dans les lignes directrices sur la transparence adoptées par le groupe de travail « Article 29 » le 29 novembre 2017, le responsable de traitement doit informer les personnes concernées de toute restriction spécifique applicable à ces droits afin de s'assurer que leurs attentes raisonnables n'ont pas été trompées. Si les personnes concernées fournissent des informations complémentaires permettant leur ré-identification, un tel exercice devra être rendu possible conformément à l'article 11 du RGPD.

	La Commission demande au responsable de traitement d'informer les personnes concernées du fait qu'elles disposeront d'un délai de trois mois minimum pour exercer leurs droits, et notamment leur droit d'opposition, à compter de la publication de l'information relative à l'étude. Lors du gel de la base de données en vue de son analyse, les secrets de pseudonymisation seront effacés pour des raisons de sécurité. Par conséquent, les personnes souhaitant exercer leurs droits après cette étape devront fournir des informations permettant de les ré-identifier, conformément à l'article 11.2 du RGPD.
Mesures de sécurité	Une analyse d'impact relative à la protection des données spécifique au projet « GAVROCHE » a été réalisée et fournie à l'appui du dossier. Celle-ci intègre également les éléments liés à la solution technique de la plateforme « ODH ». La Commission relève que la sécurité des données de l'espace projet dédié au projet « GAVROCHE » dépend essentiellement de la solution technique de la Plateforme « ODH ». A cet égard, la Commission relève que les utilisateurs accèdent à leur poste de travail avec une carte individuelle et un code personnel et qu'ils se connectent à la plateforme ODH à l'aide d'un certificat et d'un mot de passe ; elle prend acte que l'authentification forte pour l'accès aux espaces projet sera mise en place en 2022. Elle relève également que les traces sont centralisées dans un système mis à disposition par l'hébergeur et qu'un outil est à l'étude pour leur supervision en vue de détecter des comportements anormaux et de lever des alertes ; elle recommande de le mettre en œuvre dans les meilleurs délais. Concernant la mise au point du traitement automatique du langage afin d'exploiter les comptes rendus d'hospitalisation, la Commission relève qu'elle sera tout d'abord effectuée au CHU de Nantes, à partir des données nantaises préalablement désidentifiées, et que le modèle sera ensuite partagé avec les autres centres. Le modèle sera purgé d'éventuelles traces de données nominatives en effectuant plusieurs itérations de sondage et de suppression par des règles adaptées. Enfin, le modèle sera appliqué localement dans chaque centre participant lors de l'extraction de ses données. La Commission relève que les exports de données depuis l'espace projet ne pourront être constitués que de données anonymes. À cet égard, la Commission rappelle que le responsable de traitement doit réaliser une analyse permettant de démontrer que ses processus d'anonymisation respectent les trois critères définis par l'avis n°05/2014 sur les techniques d'anonymisation adoptés par le groupe de l'Article 29 (G29) le 10 avril 2014. À défaut, si ces trois critères ne peuvent être réunis, une étude des risques de ré-identification doit être menée.

Durée d'accès aux données	Trois ans à compter à compter de la mise à disposition des données au sein de l'espace projet de la plateforme « HUGO ».
Transparence du traitement	Ce traitement devra être enregistré dans le répertoire public mis à disposition par la Plateforme des données de santé.

AUTORISE, dans ces conditions, le CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES à mettre en œuvre le traitement, en application de l'article 13 de la loi précitée et de la délibération n° 2019-021 du 28 février 2019 portant délégation d'attributions de la Commission nationale de l'informatique et des libertés à son président et à son vice-président délégué.

Marie – Laure DENIS

Annexe 9. Liste des documents annexes externes à cette thèse

Pour aller plus loin dans la compréhension du travail associé à cette thèse, les documents suivants sont disponibles sur demande. Du fait de leur volume, ils n'ont pas été joints en annexe.

En lien avec le projet DMC

- Supplementary material associé à l'article publié

En lien avec le projet DMC

- Dossiers CESREES puis CNIL de soumission du projet. En particulier, le protocole DMC faisant état de l'ensemble des codes d'abord retenus pour le projet (ATC, CCAM, CIM-10 et GHM, ainsi que certains thésaurus propres au SNDS comme les codes spécialités médicales), même si le *Supplemental File 1* de l'article finalement soumis propose une synthèse suffisante en première lecture.

En lien avec le projet GAVROCHE

- Autorisation CNIL 2018 de l'EDS lié à eHOP au CHU de Nantes
- Dossiers de soumission CSE, CESREES puis CNIL, dont le PIA et la réponse point à point aux demandes de précisions de la CNIL

Annexe 10. GAVROCHE - Procédure SQL pour transmission aux centres, version du 12/10/2022 (1) : sélection des séjours d'intérêt

```
/* Auteur : Matthieu Wargny, CHU de Nantes - matthieu.wargny@chu-nantes.fr */
/* Dernière lecture et exécution : 1er août 2022
/* Temps d'exécution total en local pour Nantes le 1er août 2022 : minutes */

/*********************************************************/
/*********************************************************/
/* PHASE 0 : déclaration des variables */
/*********************************************************/
/*********************************************************/

/* PERIODE D'INTERET */
define debut_inclusion = TO_DATE('2010.12.31', 'YYYY.MM.DD');
define fin_inclusion   = TO_DATE('2020.01.01', 'YYYY.MM.DD');

/* CODES DIAGNOSTICS */
define code_GHM       = TO_CHAR('05M09%');

define code_CIM_3     = TO_CHAR('I50');
define code_CIM_4     = TO_CHAR('I110'), TO_CHAR('I130'), TO_CHAR('I132'), TO_CHAR('I139'),
TO_CHAR('R570');

/* CODES LOINC DES CRH*/
define code_LN_CRH    =  TO_CHAR('LN:11493-4'), TO_CHAR('LN:15507-7'), TO_CHAR('LN:34112-3');

/* CODES LOINC DE LA GLYCEMIE (sans "LN:" pour Nantes) */
define code_LN_GLYCEMIE    =  TO_CHAR('14749-6'), TO_CHAR('39481-7'), TO_CHAR('51596-5');


/*********************************************************/
/* PHASE 1 : sélection de tous les séjours avec une date d'entrée 2011-2019  */
/*********************************************************/

/* Variable associées récupérées : date et mode d'entrée et de sortie, type de séjour */
CREATE TABLE T1_SEJ_ALL AS (
SELECT DISTINCT ID_PAT, ID_SEJ, DATE_ENTREE, DATE_SORTIE, MODE_ENTREE, MODE_SORTIE, TYPE_SEJ
FROM EDBM_EDS.ehop_sejour
WHERE((DATE_ENTREE > &debut_inclusion)  AND (DATE_ENTREE < &fin_inclusion)));


/*********************************************************/
/* PHASE 2 : sélection sur GHM et/ou CIM */
/*********************************************************/

/* 2.1 : sélection de tous les séjours avec codes séjours */
CREATE TABLE T2_SEJ_CODE_DIAG AS (
SELECT DISTINCT ID_PAT, ID_SEJ
FROM EDBM_EDS.ehop_entrepot_structure
WHERE CODE like &code_GHM OR (
(SUBSTR(CODE,1,3) IN (&code_CIM_3) OR
SUBSTR(CODE,1,4) IN (&code_CIM_4))
AND (TEXTE IN ('Principal','Relié'))
));
```

```
/* 2.2 : croisement avec les séjours de la période d'intérêt */
CREATE TABLE T3_SEJ AS (
SELECT DISTINCT A.*
FROM T1_SEJ_ALL A
INNER JOIN T2_SEJ_CODE_DIAG B
ON A.ID_PAT = B.ID_PAT AND A.ID_SEJ = B.ID_SEJ
);

DROP TABLE T1_SEJ_ALL; DROP TABLE T2_SEJ_CODE_DIAG;


/*********************************************************/
/* PHASE 3 : exclusion des opposés                       */
/*********************************************************/

CREATE TABLE T4_SEJ AS (
SELECT DISTINCT A.*
FROM T3_SEJ A
LEFT JOIN EDBM_OPPOSITION.OPPOSITION B
ON (A.ID_PAT = B.ID_PAT)
WHERE B.ID_PAT IS NULL
);

DROP TABLE T3_SEJ;

/*********************************************************/
/* PHASE 4 : exclusion des patients avec 0 nuitée et non décédés (MOD_SORTIE = 9)
/*********************************************************/

/* 4.1 Calcul de durées d'intérêt */

CREATE TABLE T5_SEJ AS (
SELECT A.*, (A.DATE_SORTIE - A.DATE_ENTREE) as duree_sejour,
(to_date(A.DATE_SORTIE, 'dd/mm/yyyy') - to_date(A.DATE_ENTREE, 'dd/mm/yyyy')) as sej_dur_calendaire
FROM T4_SEJ A
);

/* 4.2 Sélection sur les critères de nuitée et de décès */
CREATE TABLE T6_SEJ AS (
SELECT * FROM T5_SEJ WHERE (MODE_SORTIE = 9 OR sej_dur_calendaire > 0)
);

DROP TABLE T4_SEJ; DROP TABLE  T5_SEJ;


/*********************************************************/
/* PHASE 5 : sélection sur l'existence d'au moins 1 CRH et récupération de l'âge */
/*********************************************************/

/* 5.1 Sélection des séjours avec CRH et récupération de l'âge à l'entrée*/
CREATE TABLE T7_SEJ_AVEC_CRH AS (
SELECT DISTINCT ID_PAT, ID_SEJ, MIN(AGE_PAT) as age_entree
FROM EDBM_EDS.ehop_entrepot
WHERE TYPE_DOC IN (&code_LN_CRH)
GROUP BY ID_PAT, ID_SEJ
);
```

```
/* 5.2 croisement avec les séjours déjà identifiés */
CREATE TABLE T8_SEJ AS (
SELECT DISTINCT A.*,B.age_entree
FROM T6_SEJ A
INNER JOIN T7_SEJ_AVEC_CRH B
ON A.ID_PAT = B.ID_PAT AND A.ID_SEJ = B.ID_SEJ
);

DROP TABLE T6_SEJ; DROP TABLE T7_SEJ_AVEC_CRH;


/*********************************************************/
/* PHASE 6 : Simple exclusion des patients mineurs */
/*********************************************************/

CREATE TABLE T9_SEJ AS (SELECT DISTINCT * FROM T8_SEJ
WHERE age_entree >=18);

DROP TABLE T8_SEJ;


/*********************************************************/
/* PHASE 7 : sélection sur l'existence d'au moins 1 glycémie -6h / +24h après l'admission */
/*********************************************************/

/* Récupération des temps d'exécution de glycémie : approche en 4 temps      */
/* Basé sur le mapping LOINC, donc codes communs à l'échelle inter-régionale */

/*  (étape 1/5) : identification des lignes du MAPPING GLOBAL */
CREATE TABLE A1_COD_GLY AS (
SELECT DISTINCT CODE_A
FROM EDBM_EDS.ehop_thesaurus_mapping
WHERE CODE_B IN (&code_LN_GLYCEMIE)
);

/*  (étape 2/5) : on ne conserve que les lignes d'intérêt dans ENTREPOT_STRUCTURE */
CREATE TABLE A2_ALL_GLY AS (
SELECT DISTINCT A.ID_ENTREPOT_STRUCTURE, A.ID_PAT, A.ID_SEJ, A.CODE, A.NOMBRE, A.DATE_DATA
FROM EDBM_EDS.ehop_entrepot_structure A
INNER JOIN A1_COD_GLY B
ON A.CODE = B.CODE_A
);

/*  (étape 3/5) : on ajoute la date d'entrée issue d'EHOP_ENTREPOT */
CREATE TABLE A3_ALL_GLY_DATE AS (
SELECT DISTINCT A.*, B.DATE_ENTREE, (A.DATE_DATA-B.DATE_ENTREE) as delai_gly
FROM A2_ALL_GLY A
LEFT JOIN EDBM_EDS.ehop_sejour B
ON (A.ID_PAT = B.ID_PAT AND A.ID_SEJ = B.ID_SEJ)
);

/*  (étape 4/5) : on ne garde que ceux avec un délai entre -0.25 (-6 heures) et +1 jour */
CREATE TABLE A4_SELECT_GLY_DATE AS (
SELECT DISTINCT ID_PAT, ID_SEJ, DATE_ENTREE
FROM A3_ALL_GLY_DATE
WHERE (delai_gly>(-0.25) AND delai_gly<1)
);
```

```
/* (étape 5/5) croisement avec les séjours déjà identifiés */
CREATE TABLE GAV_POP_FINALE AS (
SELECT DISTINCT A.*
FROM T9_SEJ A
INNER JOIN A4_SELECT_GLY_DATE B
ON (A.ID_PAT = B.ID_PAT AND A.ID_SEJ = B.ID_SEJ)
);

DROP TABLE T9_SEJ;
DROP TABLE A1_COD_GLY; DROP TABLE A2_ALL_GLY; DROP TABLE A3_ALL_GLY_DATE; DROP TABLE
A4_SELECT_GLY_DATE;

/* FIN DE LA SELECTION DE LA POPULATION DE GAVROCHE  */
/* Exécution 12 octobre 2022 : temps total < 1 minute, 7 252 lignes obtenues */
```

Annexe 11. GAVROCHE - Procédure SQL pour transmission aux centres, version du 12/10/2022 (2) : récupération des données d'intérêt associées aux séjours

```
/* Auteur : Matthieu Wargny, CHU de Nantes - matthieu.wargny@chu-nantes.fr */
/* Temps d'exécution total en local pour Nantes le 12 octobre 2022 : environ 1 minute */

/*******************************************************/
/*******************************************************/

/* Ce code considère comme acquise les séjours d'intrérêt pour GAVROCHE, table GAV_POP   */
/* On s'attachera ici à récupérer les informations suivantes associées aux séjours :     */
/* Codes diagnostics associés à l'IEP, biologie, un peu de clinique                      */

/* Base temporaire créée : GAV_STRUCT_TMP, correspondant aux données structurées liées aux séjours de GAV_POP */

/* Puis récupération de différentes tables ciblées */

/* Table 1 des codes CIM-10 / GHM /CCAM                                  */
/* Table 2 des données biologiques d'intérêt                             */
/* Table 3 des "constantes" (peut-être propre à Nantes pour l'instant)   */
/* Table 4 des CRH : approche séparée, non vue ici (pour usage TALN)     */

/* En sus, enrichissement des données GAV_POP (décès, sexe, délais)      */
/* Dans la version finale, aucune date JJ/MM/AAAA ne sera conservée      */


/*************************************************************************************/
/******* TABLE 0 : RECUPERATION DES DONNEES STRUCTUREES D'INTERET (table temporaire)  ********/
/*************************************************************************************/

/** En pratique : on reprend la table structurée qu'on restreint
(i) à la pop d'étude et
(ii) aux données structurées utiles dans GAVROCHE :
ID_ENTREPOT (clef), CODE_THESAURUS (pour distinguer GHM, etc.)
*/

/* Temps habituel : 3 à 5 minutes */
CREATE TABLE GAV_STRUCT_TMP AS (
SELECT DISTINCT A.ID_ENTREPOT, A.ID_PAT, A.ID_SEJ, A.CODE_THESAURUS, A.CODE, A.NOMBRE,
A.TEXTE, A.UNITE_VAL, ROUND(24*(A.DATE_DATA-B.DAT_ENT)) as DELAI_DATA_HEURES
FROM EDBM_EDS.ehop_entrepot_structure  A
INNER JOIN GAV_POP_FINALE B
ON (A.ID_PAT = B.ID_PAT AND A.ID_SEJ = B.ID_SEJ)
);


/*************************************************************************/
/******* BASE 1 : RECUPERATION DES CODES CIM-10 / CCAM / GHM D'INTERET ********/
/*************************************************************************/

/* Temps habituel < 5 secondes */
CREATE TABLE GAV_PMSI AS (
SELECT DISTINCT ID_PAT, ID_SEJ, CODE_THESAURUS, CODE, TEXTE
FROM GAV_STRUCT_TMP
```

```
WHERE
  CODE_THESAURUS IN ('GHM','cim10','ccam')
);


/*****************************************************************************/
/******* BASE 2 : RECUPERATION DES BIOLOGIES d'INTERET                ********/
/*****************************************************************************/

/* Temps total : <5 secondes */

/*  (étape 1/2) : identification des lignes du MAPPING GLOBAL (récupération CODE_A)*/
CREATE TABLE BIO_T1 AS (
SELECT DISTINCT CODE_A
FROM EDBM_EDS.ehop_thesaurus_mapping
WHERE CODE_B IN ('14933-6', '14682-9','1988-5','67151-1',
'2324-2','6690-2','14749-6','39481-7','51596-5','1920-8',
'1742-6','789-8','718-7','4549-2','2823-3','26478-8',
'2951-2','26474-7','26484-6','26449-9','26499-4','6768-6',
'71425-3','777-3')
);

/*  (étape 2/2) : identification des lignes du MAPPING GLOBAL (récupération CODE_A)*/
CREATE TABLE GAV_BIO AS (
SELECT DISTINCT A.ID_PAT, A.ID_SEJ, A.CODE, A.NOMBRE, A.UNITE_VAL, A.DELAI_DATA_HEURES
FROM GAV_STRUCT_TMP A
INNER JOIN BIO_T1 B
ON A.CODE = B.CODE_A
WHERE CODE_THESAURUS = 'labo'
);

DROP TABLE BIO_T1;

/*****************************************************************************/
/******* BASE 3 : RECUPERATION DES CONSTANTES d'INTERET                    */
/*******        Attention : Probablement propre à Nantes                   */
/*******        à voir si adapation locale ou abandon (centre-dépendant)   */
/*****************************************************************************/

/* Temps habituel < 2 secondes */
CREATE TABLE GAV_DPI_CONSTANTES AS (
SELECT DISTINCT ID_ENTREPOT, ID_PAT, ID_SEJ, CODE_THESAURUS, CODE, NOMBRE, UNITE_VAL,
DELAI_DATA_HEURES
FROM GAV_STRUCT_TMP
WHERE (
  CODE_THESAURUS = 'DPI-CST'
  AND
(CODE IN ('FCard_Autre', 'FResp_Autre', 'GlyCapil_Autre', 'PADia_Autre', 'PASys_Autre',
'SpO2_Autre','DebitO2_Autre','ModeO2_Autre','Poids_Autre','Taille_Autre','IMC_Autre',
'IMC_EDBM_Autre','Temp_Autre','Glasgow_Autre')
)));




/*****************************************************************************/
```

```
/******* BASE 4 : RECUPERATION DES CR D'INTERET                    ********/
/*****************************************************************************/

/* Identification des CR d'intérêt, ici les CRH */
/* Temps habituel : 15 secondes                 */
CREATE TABLE GAV_ENTREPOT_preTXT AS (
SELECT DISTINCT A.ID_ENTREPOT, A.ID_PAT, A.ID_SEJ, A.TITRE, A.TYPE_DOC
FROM EDBM_EDS.ehop_entrepot  A
INNER JOIN GAV_POP_FINALE B
ON (A.ID_PAT = B.ID_PAT AND A.ID_SEJ = B.ID_SEJ)
WHERE TYPE_DOC IN ('LN:11493-4',
'LN:15507-7','LN:34112-3')
);

/* Récupération du texte associé */
/* Temps habituel : 26 secondes */
CREATE TABLE GAV_ENTREPOT_TXT AS (
SELECT A.*
FROM EDBM_EDS.EHOP_TEXTE A
INNER JOIN GAV_ENTREPOT_PRETXT B
ON (A.ID_PAT = B.ID_PAT AND A.ID_SEJ = B.ID_SEJ AND A.ID_ENTREPOT = B.ID_ENTREPOT)
WHERE
A.CERTITUDE = '1' AND A.CONTEXTE = 'texte' /* AND A.RANG_CONTEXTE = 1 */
);

DROP TABLE GAV_ENTREPOT_preTXT;

/*****************************************************************************/
/******* EVOLUTION POP : (1) AJOUT DES DONNEES DE DECES + SEXE                ********/
/*******              (2) SUPPRESSION DES DATES                                ********/
/*****************************************************************************/

/* Temps habituel : <10 secondes */

CREATE TABLE GAV_POP_ENRICH AS (
SELECT A.*, B.SEXE, B.DATENAIS, B.CODE_DECES, B.DATE_DECES
FROM  GAV_POP A
LEFT JOIN  EDBM_EDS.ehop_patient B
ON A.ID_PAT = B.ID_PAT
);

/* table GAV_POP_ENRICH_NO_DATE : correspond à la table finale des séjour (IEP) pour GAVROCHE */
CREATE TABLE GAV_POP_ENRICH_NO_DATE AS (
SELECT ID_PAT, ID_SEJ,
MOD_ENT, MOD_SOR, TYPE_SEJ, URGENCES,
ROUND((DAT_ENT-DATENAIS)/365.25) as AGE_P_ENT,
CODE_DECES, ROUND((DATE_DECES-DATENAIS)/365.25) as AGE_P_DECES,
ROUND(DATE_DECES-DAT_ENT) as DELAI_ENT_DECES_JOURS,
EXTRACT(MONTH FROM DAT_ENT) as ENT_MOIS,
EXTRACT(YEAR FROM DAT_ENT) as ENT_ANNEE,
sej_dur_calendaire
FROM  GAV_POP_ENRICH
);

DROP TABLE GAV_POP_ENRICH;
```

```
/* Eléments manquants au 12/10/2022 (MW) */
/* (optionnel)CP de l'habitat et croisement déprivation LOINC */
/* (optionnel)Croisement données décès INSEE_HOP */

/* Quelques DROP si besoin */
/*
DROP TABLE GAV_POP;
DROP TABLE GAV_STRUCT_TMP;
DROP TABLE GAV_BIO;
DROP TABLE GAV_PMSI;
DROP TABLE GAV_DPI_CONSTANTES;
DROP TABLE GAV_POP_ENRICH;
DROP TABLE GAV_POP_ENRICH_NO_DATE;
DROP TABLE STRUCT_THESAURUS;
DROP TABLE TAB_THESAURUS;
DROP TABLE GAV_ENTREPOT_PRETXT;
DROP TABLE GAV_ENTREPOT_TXT;
*/
```

Annexe 12. GAVROCHE - Résumé pas-à-pas de l'approche SQL

(1) Identification de la population d'intérêt, avec comme unité le séjour (cf. Figure 10 partie V)

a. Sélection selon la présence des codes GHM et/ou la présence de codes CIM-10 d'insuffisance cardiaque en diagnostic principal ou relié
b. Puis sélection selon l'âge ≥ 18 ans au temps du séjour, la présence d'au moins un CRH associé au séjour et un début de séjour entre 2011 et 2019 (années entières)
c. Puis sélection selon la présence ≥ 1 glycémie dans les 6h précédant ou les 24h suivant l'admission
d. Récupération des dates de début et de fin du séjour, des mode d'entrée et de sortie, et de la notion d'hospitalisation en urgence oui/non
e. Exclusion des séjours sans nuitée et non soldés par le décès du patient, afin d'exclure les hospitalisations de jour

(2) Récupération des données associés, à partir des séjours obtenus

a. Récupération des codes CIM-10, GHM et CCAM associés au séjour
b. Récupération des données de biologie
c. Récupération des « constantes » associées au séjour (fréquence cardiaque, température, etc.)
d. Récupération du sexe et de la date de décès du patient (si enregistrée dans le SI, et donc possiblement non associée au séjour)
e. Création des délais d'intérêt et suppression des dates

Annexe 13. GAVROCHE - Ensemble des variables d'intérêt, au 4 août 2022

Pour mémoire, l'unité statistique élémentaire d'intérêt est le séjour, et non le patient. Se référer également au schéma relationnel (**cf. Figure 10**) pour mieux comprendre les relations entre les différentes tables de données dans GAVROCHE.

Tableau A1. Unité statistique = 1 séjour - 10 variables						
	Codage	**Source**	**TALN ?**	**LOINC**	**Qualité**	**Clef ?**
GENERAUX						
Identifiant unique/séjour	AlphaNum	DPI	Non	N/A	N/A	Oui
Identifiant unique/patient au niveau du centre	AlphaNum	DPI	Non	N/A	N/A	
Identifiant unique/patient au niveau de l'ODH	AlphaNum	ODH	Non	N/A	N/A	
Age à l'admission, en années	Entier	DPI	Non	30525-0	DPI	
Sexe à l'admission	Binaire	DPI	Non	46098-0	DPI	
Code INSEE de la commune de résidence	5 chiffres	DPI	Non	45401-7	DPI	
Date d'entrée, au mois près	MMAAAA	DPI	Mixte		DPI	
Entrée par les urgences	Binaire	DPI	Non		DPI	
Mode de sortie	Catégorielle	DPI	Non		DPI	
Nombre de jours avant le décès dans l'année suivant l'admission, s'il y a lieu	Entier	DPI	Non		DPI + INSEE	

Tableau A2. Unité statistique = 1 code diagnostic CIM-10 ou GHM – 3 variables						
	Codage	**Source**	**TALN ?**	**LOINC**	**Qualité**	**Clef ?**
Identifiant unique/séjour	Entier	DPI	Non	N/A	N/A	Oui
Diagnostic CIM-10 : type, parmi lesquels…	Quadrimodale	Clinicom	Non		DIM/SIM	
Diagnostic principal	DP			81892-2		
Diagnostic relié	DR			81892-4		
Diagnostics associés (30 maximum par séjour)	DA			NK		
Groupe Homogène de Malades	GHM			NK		
Diagnostic CIM-10 : code international	AlphaNum	Clinicom	Non		DIM/SIM	

Tableau A3. Unité statistique = 1 dosage biologique d'intérêt – 4 variables						
	Codage	**Source**	**TALN ?**	**LOINC**	**Qualité**	**Clef ?**
Identifiant unique/séjour	Entier	DPI	Non	N/A	N/A	Oui
Biologie : délai, en heures depuis l'admission	Réel	DXLAB	Non		DXLAB	
Biologie : valeur	Réel	DXLAB	Non		DXLAB	
Biologie : type, parmi 21 modalités	Catégoriel	DXLAB	Non		DXLAB	
Glycémie, en mmol/L	glycemie	DXLAB	Non	2345-7	DXLAB	
Natrémie, en mmol/L	natremie	DXLAB	Non	2951-2	DXLAB	
Kaliémie, en mmol/L	kaliemie	DXLAB	Non	2823-3	DXLAB	
Créatininémie, en µmol/L	creatininemie	DXLAB	Non	14682-9	DXLAB	
eGFR, formule de Cockroft & Gault, en mL/min	eGFR_CG	DXLAB	Non	35592-5	DXLAB	
eGFR, formule MDRD, en mL/min	eGFR_MDRD	DXLAB	Non	48642-3	DXLAB	
eGFR, formule CKD-EPI, en mL/min	eGFR_CKDEPI	DXLAB	Non	62238-1	DXLAB	
Brain Natriuretic Peptide, en ng/L	BNP	DXLAB	Non	30934-4	DXLAB	
NTpro-Brain Natriuretic Peptide, en ng/L	proBNP	DXLAB	Non	33762-6	DXLAB	
C-Reactive Protein, en mg/L	CRP	DXLAB	Non	1988-5	DXLAB	
HbA_{1c}, en %	HBA1C	DXLAB	Non	4548-4	DXLAB	
Hémoglobinémie, en g/dL	Hb	DXLAB	Non	718-7	DXLAB	
Plaquettes, en G/L	PLT	DXLAB	Non	777-3	DXLAB	
Polynucléaires neutrophiles, en G/L	PNN	DXLAB	Non	26499-4	DXLAB	
Polynucléaires éosinophiles, en G/L	PNE	DXLAB	Non	26449-9	DXLAB	
Polynucléaires basophiles, en G/L	PNB	DXLAB	Non	26444-0	DXLAB	
Lymphocytes, en G/L	LY	DXLAB	Non	26474-7	DXLAB	
Monocytes, en G/L	MO	DXLAB	Non	26484-6	DXLAB	
Acide urique, en µmol/L	ac_urique	DXLAB	Non	14933-6	DXLAB	
Troponinémie I, en µg/L	tropo_I	DXLAB	Non	10839-9	DXLAB	
Troponinémie T, en µg/L	tropo_T	DXLAB	Non	67151-1	DXLAB	

Tableau A4. Unité statistique = 1 constante – 3 variables						
	Codage	**Source**	**TALN ?**	**LOINC**	**Qualité**	**Clef ?**
Identifiant unique/séjour	Entier	DPI	Non	N/A	N/A	Oui
Constante : délai, en heures depuis l'admission	Réel	DXLAB	Non		DXLAB	
Constante : valeur	Réel	DXLAB	Non		DXLAB	
Constante : type, parmi 7 modalités	Catégoriel	DXLAB	Non		DXLAB	
Fréquence cardiaque, en battements par minute	FC	CRH	Non	N/A	N/A	Oui
Pression artérielle systolique, en mmHg	PAS	CRH	Non	N/A	N/A	
Pression artérielle diastolique, en mmHg	PAD	CRH	Non	N/A	N/A	
Poids, en kg	poids	CRH	Non	N/A	N/A	
Taille, en cm	Taille	CRH	Non	N/A	N/A	
IMC, en kg/m²	IMC	CRH	Non	N/A	N/A	
Glycémie, en mmol/L	glycemie	CRH	Non	N/A	N/A	

Tableau A5.a Unité statistique = 1 compte-rendu associé au séjour (CR hospitalier, lettre de sortie ou lettre de synthèse, soit LOINC …) – 1/3 – 22 variables						
	Codage	**Source**	**TALN ?**	**LOINC**	**Qualité**	**Clef ?**
Identifiant unique/séjour	AlphaNum	DPI	Non	N/A	N/A	Oui
Identifiant unique/compte-rendu	AlphaNum	DPI	Non	N/A	N/A	
Bloc « 0 »						
Séjour correspondant à l'admission pour insuffisance cardiaque aiguë	Trimodale : oui/non/NK	CRH	Oui	N/A	TALN	
Délai entre la date d'admission DPI et celle du CR, en jours	Entiers	CRH/DPI	Mixte	N/A	TALN	
Bloc « 1 »						
Facteur déclenchant de l'ICA (1) : trouble du rythme	Booléen	CRH	Oui	N/A	TALN	
Facteur déclenchant de l'ICA (2) : ischémique	Booléen	CRH	Oui	N/A	TALN	
Facteur déclenchant de l'ICA (3) : poussée hypertensive	Booléen	CRH	Oui	N/A	TALN	
Facteur déclenchant de l'ICA (4) : infectieux	Booléen	CRH	Oui	N/A	TALN	
Facteur déclenchant de l'ICA (5) : régime ou traitement	Booléen	CRH	Oui	N/A	TALN	
Facteur déclenchant de l'ICA (6) : autre ou non connu	Booléen	CRH	Oui	N/A	TALN	
Premier épisode d'ICA	Booléen	CRH	Oui	N/A	TALN	
Arrêt cardiaque à l'admission	Booléen	CRH	Oui	N/A	TALN	
Bloc « 2 »						
Cardiopathie causale (1) : ischémique	Booléen	CRH	Oui	N/A	TALN	
Cardiopathie causale (2) : valvulaire	Booléen	CRH	Oui	N/A	TALN	
Cardiopathie causale (3) : rythmique	Booléen	CRH	Oui	N/A	TALN	
Cardiopathie causale (4) : hypertensive	Booléen	CRH	Oui	N/A	TALN	
Cardiopathie causale (5) : autre	Booléen	CRH	Oui	N/A	TALN	
Cardiopathie causale (6) : non connue	Booléen	CRH	Oui	N/A	TALN	
Type d'ICA* (1) : droite isolée	Booléen	CRH	Oui	N/A	TALN	
Type d'ICA* (2): insuffisance cardiaque gauche décompensée	Booléen	CRH	Oui	N/A	TALN	
Type d'ICA* (3) : OAP	Booléen	CRH	Oui	N/A	TALN	
Type d'ICA* (4) : choc cardiogénique	Booléen	CRH	Oui	N/A	TALN	

* Théoriquement, les 4 types d'insuffisance cardiaque aiguë proposée sont mutuellement exclusives. Cependant, nous n'excluons pas de nous intéresser ici au recouvrement des types identifiés par TALN.

Tableau A5.b Unité statistique = 1 compte-rendu associé au séjour (CR hospitalier, lettre de sortie ou lettre de synthèse, soit LOINC …) – 2/3 – 20 variables						
	Codage	**Source**	**TALN ?**	**LOINC**	**Qualité**	**Clef ?**
Bloc « 3 »						
Antécédent d'insuffisance respiratoire chronique	Booléen	CRH	Oui	N/A	TALN	
Antécédent de bronchopneumopathie chronique obstructive	Booléen	CRH	Oui	N/A	TALN	
Antécédent de syndrome d'apnée obstructive du sommeil	Booléen	CRH	Oui	N/A	TALN	
Antécédent d'AVC ou d'AIT	Booléen	CRH	Oui	N/A	TALN	
Type de diabète : néant, type 1, type 2, inconnu ou autre	4 modalités	CRH	Oui	N/A	TALN	
Bloc « 4 »						
A l'admission : Fréquence cardiaque, en battements par minute	Entier	CRH	Oui	N/A	TALN	
A l'admission : Pression artérielle systolique, en mmHg	Entier	CRH	Oui	N/A	TALN	
A l'admission : Pression artérielle diastolique, en mmHg	Entier	CRH	Oui	N/A	TALN	
A l'admission : Poids, en kg	Réel positif	CRH	Oui	N/A	TALN	
A l'admission : Taille, en cm	Réel positif	CRH	Oui	N/A	TALN	
A l'admission : IMC, en kg/m²	Réel positif	CRH	Oui	N/A	TALN	
A l'admission : tabagisme, jamais/sevré/actif	3 modalités	CRH	Oui	N/A	TALN	
Bloc « 5 »						
Antécédent de trouble du rythme	Booléen	CRH	Oui	N/A	TALN	
Antécédent de dépression	Booléen	CRH	Oui	N/A	TALN	
Antécédent de troubles cognitifs	Booléen	CRH	Oui	N/A	TALN	
Antécédent de cancer (tumeur solide ou hémato)	Booléen	CRH	Oui	N/A	TALN	
A l'admission : arythmie cardiaque/fibrillation auriculaire	Booléen	CRH	Oui	N/A	TALN	
FEVG réduite (<41%)	Booléen	CRH	Oui	N/A	TALN	
FEVG modérément réduite (41-49 %)	Booléen	CRH	Oui	N/A	TALN	
FEVG préservée (>49 %)	Booléen	CRH	Oui	N/A	TALN	

Tableau A5.c Unité statistique = 1 compte-rendu associé au séjour (CR hospitalier, lettre de sortie ou lettre de synthèse, soit LOINC …) – 3/3 – 19 variables						
	Codage	**Source**	**TALN ?**	**LOINC**	**Qualité**	**Clef ?**
Interprétation du Bloc « 1 » sur les traitements médicamenteux						
Médicaments du diabète						
Metformine	Booléen	CRH	Oui	N/A	TALN	
Sulfamide hypoglycémiant	Booléen	CRH	Oui	N/A	TALN	
Inhibiteur des DPP4	Booléen	CRH	Oui	N/A	TALN	
Analogue du récepteur au GLP1	Booléen	CRH	Oui	N/A	TALN	
Insuline	Booléen	CRH	Oui	N/A	TALN	
Médicaments anti-hypertenseurs						
Bêta-bloquants	Booléen	CRH	Oui	N/A	TALN	
Inhibiteurs calciques	Booléen	CRH	Oui	N/A	TALN	
Inhibiteurs de l'enzyme de conversion	Booléen	CRH	Oui	N/A	TALN	
Antagoniste du récepteur à l'angiotensine 2	Booléen	CRH	Oui	N/A	TALN	
Diurétiques thiazidiques	Booléen	CRH	Oui	N/A	TALN	
Diurétiques de l'anse	Booléen	CRH	Oui	N/A	TALN	
Autres médicaments à visée cardiovasculaire						
Anti-agrégants plaquettaires	Booléen	CRH	Oui	N/A	TALN	
Anticoagulants	Booléen	CRH	Oui	N/A	TALN	
Dérivés nitrés	Booléen	CRH	Oui	N/A	TALN	
Statines	Booléen	CRH	Oui	N/A	TALN	
Ezétimibe	Booléen	CRH	Oui	N/A	TALN	
Sacubitril	Booléen	CRH	Oui	N/A	TALN	
Autres médicaments						
Antidépresseurs	Booléen	CRH	Oui	N/A	TALN	
Anxiolytiques	Booléen	CRH	Oui	N/A	TALN	

Titre : Diabète et insuffisance cardiaque - Approche épidémiologique par l'analyse croisée des différentes sources de données de santé

Mots clés : diabète, insuffisance cardiaque, épidémiologie, données massives, données médico-administratives, entrepôts de données de santé

Résumé : Les données de santé peuvent être catégorisées en données de recherche et en données de « vie réelle », parmi lesquelles les données médico-administratives et celles issues du soin. Avec pour fil rouge le lien entre deux maladies fréquentes, le diabète et l'insuffisance cardiaque (IC), nous proposons trois analyses épidémiologiques fondées sur ces sources. L'étude SURDIAGENE d'abord, une cohorte prospective « classique » de 1349 personnes diabétiques suivies au CHU de Poitiers, chez qui nous analysons le lien entre des biomarqueurs nutritionnels et le risque d'hospitalisation ou de décès pour IC. L'étude DMC ensuite, sur les données médico-administratives de l'Assurance Maladie de plus de 3 millions de personnes identifiées comme diabétiques entre 2012 et 2018, chez qui nous étudions le risque d'IC après un événement rétinien grave. Enfin, le projet GAVROCHE, une analyse des entrepôts de données de santé hospitaliers à l'échelle inter-régionale, sur le lien entre la variabilité glycémique à l'admission et le pronostic des personnes hospitalisées pour IC aiguë. Cela nécessitera en particulier l'extraction d'informations issues de comptes rendus par traitement automatisé du langage naturel.
Ces trois projets nous permettent d'illustrer la gestion des données selon leur source et les enjeux liés à l'information et au consentement des patients, avant de conclure sur des propositions de bonnes pratiques épidémiologiques dans le traitement des données de vie réelle.

Title: Diabetes mellitus & Heart Failure – Epidemiological study by cross-analysis of different sources of health data

Keywords: diabetes mellitus, heart failure, epidemiology, big data, administrative health data, health data warehouse

Abstract: Health data can be categorized into research data and "real-life" data, the latter including medico-administrative and health care setting. Using the link between diabetes and heart failure (HF) as a common theme, we propose three epidemiological analyses based on these sources.
First, the SURDIAGENE study, a "classic" prospective cohort of 1349 diabetics patients followed at the University Hospital of Poitiers. We analyzed the link between nutritional biomarkers and the risk of HF requiring hospitalization. Second, the DMC study, based on medico-administrative data from the French National Health System, for more than 3 million people with diabetes between 2012 and 2018, in whom we studied the risk of HF after a serious retinal event. Finally, the GAVROCHE project, an inter-regional analysis of hospital health data warehouses, on the link between glycemic variability on admission and the prognosis of individuals hospitalized for acute HF. This will involve natural language processing methods to extract data from hospital reports.

These three projects are an opportunity to illustrate the data management according to data source, the issues related to the information and the consent of patients, and to conclude with proposals for a code of good epidemiological practices in real-life data processing.