EDSaNCoh et EDILS2.0 : résultats préliminaires et perspectives

Francesco Monti

FACULTÉ DE MÉDECINE ET PHARMACIE DE ROUEN

ANNÉE 2022

THÉSE POUR LE DOCTORAT EN MÉDECINE

(Diplôme d'État)

PAR

Francesco MONTI

NÉ LE 27 JUIN 1990 À BAGNO A RIPOLI (FI, ITALIE)

PRÉSENTÉE ET SOUTENUE PUBLIQUEMENT LE 11 OCTOBRE 2022

EDSaNCoh et EDILS2.0 : résultats preliminaires et perspectives

Président du jury: *M. le Professeur Pierre DECHELOTTE*

Directeur de thèse: *Mme le Professeur Marie-Pierre TAVOLACCI*

Membre du jury: *M. le Professeur Jacques BENICHOU*

Membre du jury: *M. Julien GROSJEAN, PhD*

ANNEE UNIVERSITAIRE 2021 - 2022

U.F.R. SANTÉ DE ROUEN

DOYEN : **Professeur Benoît VEBER**

ASSESSEURS : **Professeur Loïc FAVENNEC**

Professeur Agnès LIARD

Professeur Guillaume SAVOYE

I - **MEDECINE**

PROFESSEURS DES UNIVERSITES – PRATICIENS HOSPITALIERS

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

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

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

Mme Marie-Laure **WELTER**	HCN	Physiologie

MAITRES DE CONFERENCES DES UNIVERSITES – PRATICIENS HOSPITALIERS

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

PROFESSEUR AGREGE OU CERTIFIE

Mr Thierry **WABLE**	UFR	Communication
Mme Mélanie **AUVRAY-HAMEL**	UFR	Anglais

ATTACHE TEMPORAIRES D'ENSEIGNEMENT ET DE RECHERCHE à MI-TEMPS

Mme Justine **SAULNIER**	UFR	Biologie

II - PHARMACIE

PROFESSEURS DES UNIVERSITES

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

MAITRES DE CONFERENCES DES UNIVERSITES

Mme Margueritta **AL ZALLOUHA**	Toxicologie
Mme Cécile **BARBOT**	Chimie Générale et Minérale
Mr Frédéric **BOUNOURE**	Pharmacie Galénique
Mr Thomas **CASTANHEIRO MATIAS**	Chimie Organique
Mr Abdeslam **CHAGRAOUI**	Physiologie

Mme Camille **CHARBONNIER (LE CLEZIO)**	Statistiques
Mme Elizabeth **CHOSSON**	Botanique
Mme Marie Catherine **CONCE-CHEMTOB**	Législation pharmaceutique et économie de la santé
Mme Cécile **CORBIERE**	Biochimie
Mme Nathalie **DOURMAP**	Pharmacologie
Mme Isabelle **DUBUC**	Pharmacologie
Mme Dominique **DUTERTE- BOUCHER**	Pharmacologie
Mr Gilles **GARGALA** (MCU-PH)	Parasitologie
Mme Nejla EL **GHARBI-HAMZA**	Chimie analytique
Mr Chervin **HASSEL**	Biochimie et Biologie Moléculaire
Mme Maryline **LECOINTRE**	Physiologie
Mme Hong **LU**	Biologie
Mme Marine **MALLETER**	Biologie Cellulaire
M. Jérémie **MARTINET** (MCU-PH)	Immunologie
M. Romy **RAZAKANDRAINIBÉ**	Parasitologie
Mme Tiphaine **ROGEZ-FLORENT**	Chimie analytique
Mr Mohamed **SKIBA**	Pharmacie galénique
Mme Malika **SKIBA**	Pharmacie galénique
Mme Christine **THARASSE**	Chimie thérapeutique

PROFESSEURS ASSOCIES

Mme Cécile **GUERARD-DETUNCQ**	Pharmacie officinale
Mme Caroline **BERTOUX**	Pharmacie

PAU-PH

M. Mikaël **DAOUPHARS**

M. Pierre **BOHN**

PAU

M. Damien **SALAUZE**

PROFESSEUR CERTIFIE

Mme Mathilde **GUERIN**	Anglais

ASSISTANTS HOSPITALO-UNIVERSITAIRES

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

ATTACHES TEMPORAIRES D'ENSEIGNEMENT ET DE RECHERCHE

Mme Chaïma **EZZINE**	Pharmacologie
M. Abdelmounaim **MOUHAJIR**	Parasitologie
M. Olivier **PERRUCHON**	Pharmacognosie

ATTACHE TEMPORAIRE D'ENSEIGNEMENT

M. Maxime **GRAND**	Bactériologie

LISTE DES RESPONSABLES DES DISCIPLINES PHARMACEUTIQUES

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

III – MEDECINE GENERALE

PROFESSEUR MEDECINE GENERALE

Mr Jean-Loup **HERMIL** (PU-MG)	UFR	Médecine générale

MAITRE DE CONFERENCE MEDECINE GENERALE

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

PROFESSEURS ASSOCIES A MI-TEMPS – MEDECINS GENERALISTE

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

MAITRE DE CONFERENCES ASSOCIE A MI-TEMPS – MEDECINS GENERALISTES

Mr Julien **BOUDIER**	UFR	Médecine Générale
Mme Laëtitia **BOURDON**	UFR	Médecine Générale
Mme Elsa **FAGOT-GRIFFIN**	UFR	Médecine Générale
Mr Emmanuel **HAZARD**	UFR	Médecine Générale

ENSEIGNANTS MONO-APPARTENANTS

PROFESSEURS

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

MAITRES DE CONFERENCES

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

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

HCN - Hôpital Charles Nicolle	*HB - Hôpital de BOIS GUILLAUME*
CB - Centre Henri Becquerel	*CHS - Centre Hospitalier Spécialisé du Rouvray*
CRMPR - Centre Régional de Médecine Physique et de Réadaptation	*SJ – Saint Julien Rouen*

Acknowledgements

Words cannot express my gratitude for the members of the jury who have, each in their own way and time, inspired and marked me.

To Professor J.BENICHOU, for guiding me through my residency and giving me the honour of being here today.

To Professor P.DECHELOTTE, for welcoming me into his team as if I were a lifelong member.

To Julien GROSJEAN, for taking part in this jury, your technical support, encouragements and for helping me figure out what I want to do when I grow up.

To Dr. Marie-Pierre TAVOLACCI, for taking part in this jury and for the respect and benevolence she has always shown me.

I could not have embarked on this journey without your input, the knowledge and expertise generously provided.

To André Gillibert, for your kindness, patience, deep knowledge and leaving in me a sense of incompetence I'll never be able to get rid of.

To Professor S.DARMONI, for being very "stimulating" and to all the DIM's team for being so adorable.

Thanks to my colleagues Mikaël DUSENNE, Thibaut SABATIER, Anca VASILIU, Marion PHILIPPE, Karl HERMANN, Louise KIEKEN, Yoann JACOB, Tifeen CLABAUT, Julien RIO and Sorina DANA for their good mood, their support and the moments spent together. I never felt abroad in your company.

To Clément MASSONAUD and Thibaut LAFFOUILLHERE, for being the first to break through my armour since my arrival in France.

To Alessandra ZAGO, for being the person who best understands me in the world and for being an emotionless cynical monster along with me.

To Marion LACASSIN, for relentlessly tolerating my gruffness and your genuine friendship.

To Anaïs REMY, for being my nymphetamine and the most adorable feral-child ever.

To Alexandre Schirrecker, to make me feel like I was socially skilled and being as much a coffee-geek as I am.

I am also grateful to my office mates I encountered along this last semester, Sebastien, Amina, Charlotte et Claire, for their precious moral support, their kindness, sympathy, and always being there to jog my memory about clinical stuff. Thanks should also go to the nurses, dieticians, and patients, who impacted and inspired me.

To all the physicians in the nutrition unit for welcoming and patiently supporting me without ever making me feel inadequate. Special thanks go to Diane CAZAUX, for her humanity.

Lastly, I would be remiss in not mentioning my family.

To my parents, Marco and Maria Stella, for always supporting and encouraging me. I wouldn't have gotten here without you.

To my brother Giovanni, for the countless hours spent conquering virtual worlds together.

To my grand-father Valerio, for always loving and encouraging me. I am sorry I was not there when you left. I miss you.

To Francesca and Avvenente, for being my family 2.0, leaving for the unknown “land of baguettes” with me, and making me feel like life is a latin soap-opera.

Your belief in me has kept my spirits and motivation high during this process.

I would also like to thank my dog Sbino for all the entertainment, emotional support and showing me what LOVE truly is.

And to all the others whose list is too long to mention.

Summary

1. Intro

The term “cohort” refers to a group of subjects sharing common characteristics who are followed individually in a prospective manner. Throughout the follow-up period, data concerning the subjects are collected: notable exposure to risk factors and health events. From a methodological point of view, the clear advantage of prospective cohorts is the possibility of clearly observing the temporal sequence of exposure (or intervention) and the consequent outcome. It is thus possible to model the interactions of different factors relating to living conditions (diet, housing, access to care, social network, etc.), the environment (working conditions, occupational and environmental exposures, etc.), and health status (preclinical states, chronology of pathological phenomena).

The primary ambition of cohorts is therefore to seek associations and, if possible, causal relationships between different risk factors and health events by identifying the sequence of interactions between the health status of the persons included and the factors which may have contributed to it. In principle, therefore, in the collection of information they carry out they favor the search for "internal validity" based on the precise description of the state of health of the persons included ("phenotyping") and the detailed categorization of the types of pathologies and risk factors, rather than the representativeness of the populations ("external validity").

This objective usually implies collecting, generally, voluntarily, a variety of detailed data at different times (clinical examinations, answers to questionnaires or self-questionnaires, biological samples of various kinds, and sometimes imaging...), paying close attention to the conditions and quality of these collections. The collection of these data has for a long time been done manually by clinical research technicians: a very costly and time-consuming solution for studies that last several years and involve hundreds or more patients.

Analysis and text mining tools are a possible solution to this problem, automating data collection thanks to the health data warehouses. Additional data, complementary to medical records (patient questionnaires, scores, etc.), can be obtained through active data collection directly from patients with the patient entering information directly into online applications. Several studies have shown that the collection of data for research via the Internet is feasible and that the quality of responses from data provided directly by patients is comparable to measures adjudicated by physicians or from medical administrative databases.[25]

1.1 E cohorts

With the widespread adoption of Electronic Health Record (EHR) systems, progressively greater amounts of electronic clinical data are being generated, making researchers, healthcare administrators, and clinicians alike increasingly interested in the use of EHR data which offers new opportunities for knowledge advancement covering a wide range of categories. Though the primary motivation for adopting EHRs is improved documentation of patient health and healthcare service provision, with its expected benefits for healthcare quality (medical error reduction, proper reimbursement, and litigation protection), several secondary uses of EHR data show promise, including the assessment of quality improvement initiatives, population health tracking, and epidemiological research.

The exploitation of these data remains nonetheless difficult for several reasons. First, the data are produced and maintained by different systems and health professionals and are consequently spread over multiple databases and even across multiple establishments. Second, the significant amount of data generated results in problematic management of data both in terms of data storage capabilities and data access performances. In 2018 healthcare datasphere was already on par with the media and entertainment ones and, while it was far away from the data size of manufacturing or financial services, it is primed to grow faster than any other sector for the period 2018-2025. This growth is mostly driven by the digitalization of non-digital systems and MRI image creation: the trend is more images with thinner slices and the introduction of 3D capabilities. [30]

Moreover, the health data produced are of different nature; some data are natively structured (eg, diagnosis-related group [DRG] codes and laboratory tests results), but an important part of medical information remains in unstructured free-text clinical narratives (i.e., admission notes, hospitalization reports, discharge summaries, radiology reports and pathology reports). This unstructured information seems to be particularly relevant in the context of cohort selection tasks: Raghavan et al [29], found that not only unstructured data is essential to resolve between 59% and 77% of some clinical trials criteria but also that combining the use of structured and unstructured data enables leverage of patient recruitment.

Aggregating all these scattered, big, complex, and diversely structured data is, the role of Health Data Warehouses (HDWs). An HDW is defined as a regrouping of data from diverse sources accessible via a single data management system [23]. This kind of data repository centralizes clinical, demographic, and administrative data within a uniform and consistent data model. The appeal of using EHR data for epidemiology is clear: EHRs passively generate large datasets through the routine interactions of patients with healthcare organizations on real-world patient populations in easily retrievable form, allowing the cost-efficient and timely execution of epidemiologic studies on a broad array of topics. When patients repeatedly patronize a particular healthcare organization, a longitudinal data stream develops, whose value increases as time and data accrue.

EDSaNCoh project

The EDSaNCoh project, selected and financed by the ERDF (European Regional Development Fund), aims to develop a platform for creating and automatically feeding active automatic open prospective clinical e-cohorts. The core source of data will be the Rouen University Hospital's health data warehouse (EDSaN) which provides data on hospital stays (biology, CRH, prescriptions, etc.). The EDSaNCoh project also foresees the possibility to link these data with the National Health Data System (SNDS) and to data directly provided by patients themselves.

The end goal of the project is to optimize non-interventional research on epidemiological and clinical data by reducing human errors, the workload, the complexity of data entry, and the time spent on data collection compared to current research methods. In addition, it increases the quality and diversity of data without the often-large number of lost to follow-up that plagues cohort studies which can last for several years. Ultimately, EDSaNCoh will allow for numerous scientific publications throughout the cohort. The platform will have a regional and national vocation by making it possible to create multi-center cohorts, thus promoting scientific cooperation, and innovation and bringing statistical power (high number of patients) to research protocols.

Where such a project particularly shines is in offering the possibility to link local data (hospital data) to the National Health Data System (SNDS - Système National des Données de Santé, in French) and data from self-administered questionnaires sent to patients. To the best of my knowledge, this is the first time in France that such an option is available. Technically this possibility has always been there but no one dared to relentlessly pursue that path to meet compliance with all the GPRD requirements.

With my thesis, I would like to give an overview of the EDSaNCoh infrastructure, what its potentials are, and indirectly what the potentials are for automated electronic cohorts in 2022, their pros and cons, and the difficulties in play. I will present what is its first use case and the first preliminary results. I'll re-examining that 1st evaluation, highlighting current problems and proposing how to correct them. This will hopefully give a more accurate idea of what the real limitations of this technology are and serve as a starting point for a new round of improvements to the EDILS2.0 cohort algorithm. As a first use case, any optimization done to the project will provide new knowledge and insight which will carry-over to future projects built on the EDSaNCoh infrastructure.

EDILS 2.0 – a first use-case

EDILS2.0 (Eating Disorders Inventory Longitudinal Study) will be the first project taking advantage of the EDSaNCoh infrastructure, with patient inclusion beginning at the end of 2022. Eating disorders (ED) are serious pathologies of adolescence and young adulthood, likely to become chronic with a severe impact in terms of morbidity and mortality in the long term [28]. The Clinical Investigation Center (CIC) with the nutrition department of RUH have already conducted a prospective cohort about the determinants and prognostic factors of EDs. It was named EDILS1.0 and it took place from 2015 to 2021 with a 3 years follow-up. EDILS2.0, backed by the EDSaNCoh platform, aims to build upon the success of EDILS1.0 greatly improving recruitment procedures and patient data exhaustivity without sacrificing quality [11].

To the best of my knowledge, there are no cohort studies in France on the prognostic factors and evolution of the EDs combining data from health warehouses, SNDS, and self-administered questionnaires. In literature, no study has provided reliable guidance on the prognostic factors for recovery from EDs, partly due to the lack of representativeness of the samples included in these studies or to methodological limitations. Studies abroad using national databases were retrospective for the period before the diagnosis of the disease and showed the impact of early management [19,34,43]. These studies also report survival data but without investigating the factors influencing these outcomes.

To give some context, eating disorders are defined by the DSM-V classification criteria [6] and include different forms: anorexia nervosa, bulimia nervosa, binge eating disorder, and atypical or unspecified forms (EDNOS, Eating Disorder Not Otherwise Specified). Worldwide prevalence of these diseases reaches 8.4% in women and 2.2% in men [10] and it's raising fast among young people. Adolescence is often marked by various stresses or traumas (emotional, professional, violence) and is therefore conducive to a first peak in the frequency of occurrence of EDs, in particular anorexia nervosa and bulimia. In France, all types of EDs increased significantly in students between 2009 and 2021 with bulimia nervosa being the most common ED (29.8% for women and 15.7% for men). Over this period incidence increased from 30.6% (CI95%=28.0-33.2%) to 51.8% (CI95%=49.8-53.7%) for women and from 11.3% (CI95%=8.7-13.6%) to 31.3% (CI95%=28.1-34.4%) for men [36,37].

The somatic complications of EDs are multiple, both in the case of restriction and hyperphagia, strongly impacting quality of life [31]. Many patients with anorexia and bulimia nervosa feel a strong sense of helplessness in the face of their disorder, frequently associated with severe anxiety and/or depressive disorder: 56.4 % of people with anorexia nervosa [8,39] and 95.4 % of people with bulimia nervosa [13,38] have a comorbidity of mood disorder, anxiety disorder, substance abuse and/or impulse control difficulties [17].

EDs generally evolve in bouts and their evolution can be protracted, with some forms evolving towards chronicity or even death. The major difficulties in reliably estimating the evolutionary and prognostic profile of EDs are related to the size of the series found in scientific literature, the diversity of recruitment according to the centers, and the fluctuations in the clinical presentation, as the same patient may go from one typical or atypical picture to another in a few months or years [12].

Our hope is to be able, at the end of this project, to shed some light on the question.

2. METHODS

2.1 EDILS2.0 protocol

Objectives

The **main objective** of EDILS2.0 is to identify, at 2 and 5 years after the first consultation for an eating disorder, the **prognostic factors** of:

- Recovery
- Remission
- Change in type of eating disorder
- Death (suicide/other cause).

The **secondary objectives** are:

- To characterize the population taken care of by the nutrition service
- To know the annual incidence rates of cure, remission, and death (suicide/other cause) at 5 years follow-up
- To identify the care pathways at 2 years and 5 years follow-up and within 2 years before the first nutrition care
- Identify specificities according to gender, category, and severity of ED for each of the aforementioned objectives.

Evaluation criteria

The evaluation criteria are the following:

- **Recovery** and **remission** will be assessed with each of the Bardone-Cone criteria [2] which allow an evaluation of both physiological, psychological, and also the quality of life [1,40] (composite criterion):
    - No longer meeting the criteria for diagnosis of an ED (absence of consultation/hospitalization for this reason)
    - BMI $\geq$ *18.5* Kg/m^2 for anorexic disorders and $\leq$ 25kg/m^2 for hyperphagic disorders
    - Score within 1 standard deviation of the norms corresponding to age on all the subscales of the EDE-Q
    - WHO QOL-BREF improvement
- **Deaths** and causes of death will be identified via the CépiDC database.
- The **care pathway** will be described by drug consumption, hospitalizations, and consultations (SNDS).

Inclusion criteria

- Inclusion period: 2 years
- Men and women, aged between 18 and 65 years.
- First time consulting/being hospitalized for ED in the Nutrition ward of RUH
- PMSI codes (at least one of them)
    - ICD10 "F50*": (non-organic eating disorders)
    - ICD10 "E66*": obesity associated with compulsion or hyperphagia or night eating syndrome.
- Clinical narrative containing one of the following keywords: compulsion, nibbling, hyperphagia, bulimia, binge eating, tachyphagia, night-time compulsion, night eating syndrome, NED, vesper compulsion.
- The patient has read and understood the information letter and not objecting to participating in the study.

Exclusion criteria

- Patient's objection to participation
- Patient not meeting age criteria.
- The patient has already consulted for ED in the nutrition service at RUH.

Follow-up

Duration of participation of each patient: the active prospective collection will be done by questionnaire for 2 years (at M0, M6, M12, M18, M24)

Passive follow-up with the information systems (SNDS and data warehouse of the Rouen University Hospital): 5 years prospectively and 2 years retrospectively

The total duration of the research (duration between the 1st inclusion and the last visit of the last included patient): 7 years

Table 1: scheduled follow-up via self-administered questionnaires.

		M0	M6	M12	M18	M24
Weight		X	X	X	X	X
Socio-demographic characteristics		X	X	X	X	X
Major events in life		X				X
Eating disorders	EDE-Q	X		X		X
	Food insecurity (HFSSM)	X	X	X	X	X
	Orthorexie (BOT)	X		X		X
Physiology evaluation	ROME IV criteria	X		X		X
	Tobacco (FAGERSTROM)	X		X		X
	Alcool (AUDIT)	X		X		X
Addictions and risky behaviors	Cannabis (CAST)	X		X		X
	Social Networks (BSMAS)	X		X		X
	Physical activity (IPAQ and EAI)	X		X		X
QoL and psychologic evaluation	WHOQOL-BREF	X	X	X	X	X
	HAD	X	X	X	X	X

Estimated sample and preconized statistical analysis

Approximately 550 new patients suffering from an ED are taken care of at RUH's Nutrition Unit every year. It is reasonable to hypothesize that 100 patients could object to the use of their data; that would mean that approximately 1000 patients will be included in the cohort over the two years inclusion period. According to the results of the EDILS1.0 it is expected that around 50% of these individuals will answer and complete self-administered questionnaires.

Several statistical tests have been preconized in the protocol:

- *Chi2 test* for categorical variables and *T-student test* for quantitative variables
- *Kaplan Meier* survival analysis with *log Rank test* to compare survival curves
- Multivariate analyses (logistic regression and generalized estimation equation) to identify prognostic factors

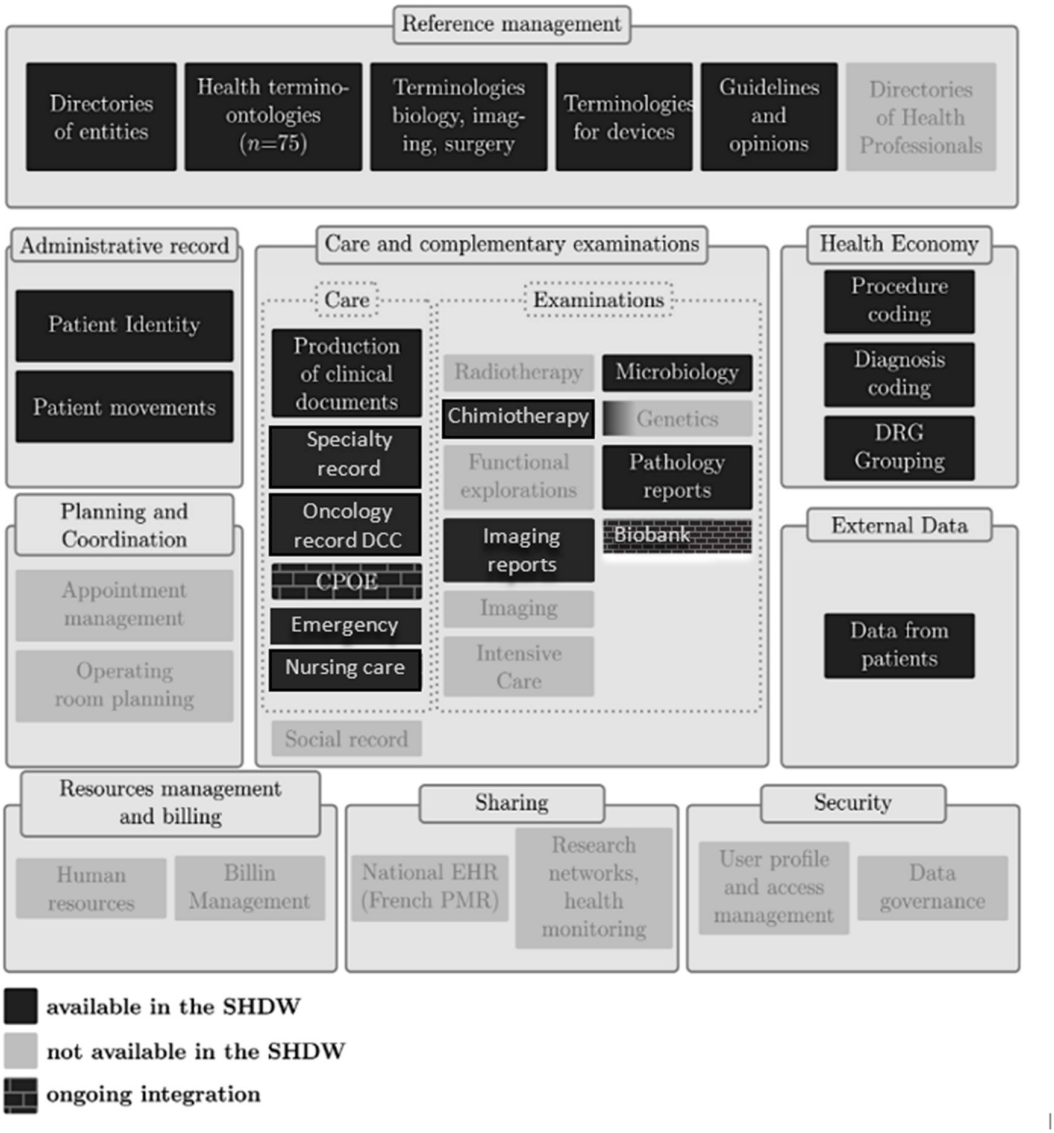
The analysis will then be repeated stratifying by ED category and gender.

2.2 EDSaN – RUH's HDW

EDSaN is the internal Semantic Health Data Warehouse (SHDW) of Rouen University Hospital (RUH). Since 1992, RUH has gathered and kept track of patients by collecting demographic data (such as name, date of birth, and gender), clinical data (such as biological test results, medical procedures, visit records, letters, and discharge summaries), administrative data and, less frequently, omics data. Overall, RUH maintains the data of several millions of patients.

Figure 1 summarizes included data according to their specific domain. A dark grey opaque backdrop represents data that is already present in the SHDW, whereas a light grey background represents data that is neither present nor anticipated to be so in the near or medium term. Bricks covering some or all of the background represent data whose short- or medium-term inclusion in the progressor is planned [23].

Figure 1: Functional coverage of the semantic health data warehouse in terms of data according to each domain. SHDW: semantic health data warehouse; CPOE: computerized physician order entry; DCC: French cancer communication file; PMR: personal medical record; DRG: diagnosis-related group; EHR: electronic health record.



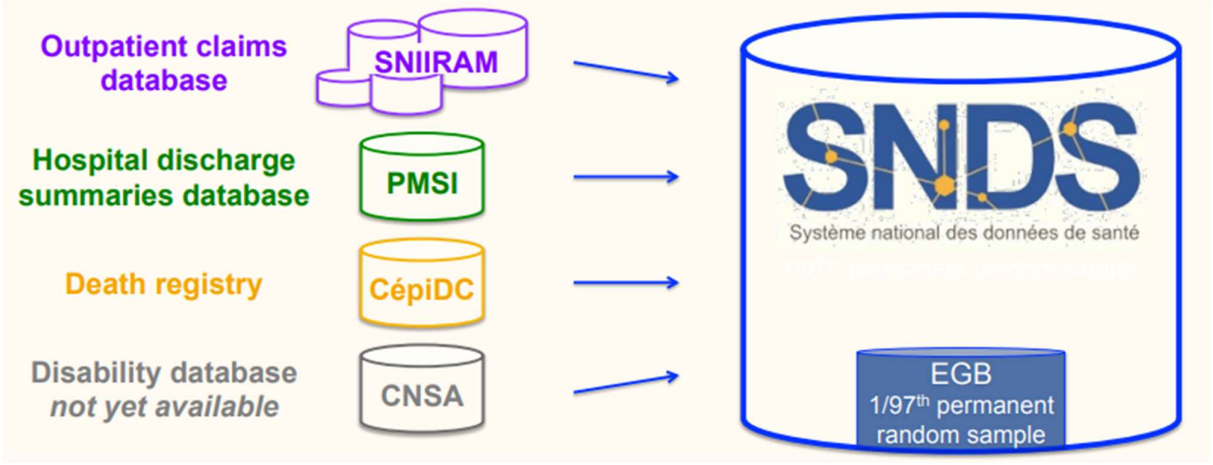
2.3 The "SNDS"

By adopting a pseudonymized version of the unique national identification, the SNDS (Système national des données de santé) database integrates claims with hospital-discharge summaries (PMSI) and the national death register. One of the largest continuous homogeneous claims databases in the world, the French healthcare system today covers 99% of the French population, or over 66 million people, from birth or immigration until death or emigration.

The database includes demographic data, date, and cause of death, Long Term Disease (LTD) registration for full reimbursement (mainly for chronic and costly disease), outpatient reimbursed healthcare encounters such as physician or paramedical visits (e.g. nursing, physiotherapy), medicines prescribed, medical devices, lab tests, all private and public hospitalizations with primary, linked and associated ICD10 diagnoses, procedures, duration, and cost coding system (Diagnostic Related Group), as well as most very expensive drugs and practitioner's information (specialty, sex, age, site and modality of practice). The power of the database is correlatively great, and its representativeness is guaranteed. Every year 450Tb of text-based data is generated, and 3000 variables and 500M of procedures are enregistered, accounting for 11M hospitalizations.

French law regulates SNDS access with a well-defined process through INDS (Institut National Données de Santé), the National Institute of health.

Figure 2: Constitution of French nationwide healthcare data system (SNDS). EGB (Echantillon Généraliste de Bénéficiaires) is the 1/97th random permanent representative sample of SNDS, easier to access in comparison to the full database set.



2.4 EASYMEDSTAT, third-party partner

To collect and analyze the final cohort data, the workgroup has opted to partner with a specialized company among the many in the field. They offer simple, powerful, and ergonomic tools which are also compliant with data protection regulations (HDH, Health Data Host). That said, each tool has its specificities. Three were the French runners-up with solid experience: 1) Doqboard, 2) Arone and 3) EasyMedStat.

After extensive dialogue with the contenders for partnership, their main strengths and weaknesses can be resumed as follows:

Table 2: Comparison between third-party services Doqboard, Arone, EasyMedStat.

	Doqboard	Arone	EasyMedStat
Strengths	• HDH • Reactivity • Autonomy • API adaptable to project needs	• HDH • Reactivity • Ergonomy • e-CRF available	• HDH • Reactivity • Ergonomy • API adaptable to project needs • Autonomy • Price • Statistical analysis tool • Unlimited projects
Weaknesses	• Lack of statistical analysis tool • Price • API needs to be adapted • e-CRF to be developed	• Lack of autonomy concerning variable's definition • Poor coherence controls • Lack of statistical analysis tool	• API needs to be adapted • e-CRF to be developed

The choice was therefore made for EasyMedStat. Several functionalities had to be developed but still less than for the other two tools. As ESM is a third-party platform, it is necessary to define and standardize the data format to be transferred which requires a specific API to be designed and developed by this provider in direct link with the D2IM team. Once everything will be up and running, EasyMedStat will provide a complete and easy-to-use solution to researchers with all the necessary tools to carry out their studies: from data collection to the publication of articles.

Every patient will receive a unique ID to access his data, enabling them to add more information via online surveys proposed by the research team. Opening a direct communication channel with included patients via the platform other than allowing active follow-up at distance, may allow recovery of some information that is not available within the RUH's HDW or SNDS data set (or that varies at a faster rate than the consultations rate or that patients can easily provide more accurately compared to automatic extraction) like:

- Paraclinical data
- Risk factors such as tobacco, alcohol, nutrition, ….
- Drugs delivered during hospital stays
- Social data

Several other interesting features are available such as automatic data pseudonymization, remote collaboration, and data sharing without data leaving the secured environment, ".exif" data stripping of uploaded images, automatic generation of rapports or content for articles like data tables or graphs, and data management tools not requiring programming skills.

To analyze the data collected for each cohort, an **environment** will be designed **for the researchers.** The goal is to be able to propose semi-automatic reports and statistics and to make this model compatible with biostatistical analysis supports to optimize the creation of statistics according to the research questions that arise during the cohort (automatic and dynamic creation of analyses, survival curves, etc.).

Data will be hosted on HDH-certified servers in Berlin (according to ISO 9001:2015 and ISO 27001:2013 norms), RGPD and CNIL compliant: all of the legal aspects are taken care of so that researchers can focus on what they do best.

2.5 Data collection and management

To facilitate collaboration across multiple centers, a generic data model able to feed any type of cohort is needed. The choice came down to the **OMOP CDM** (Observational Medical Outcomes Partnership Common Data Model) because of its focus on interoperability between the different health analysis databases, whether they are clinical or medico-administrative ones.

Moreover, OMOP CDM is promoted in Europe by the European Health Data and Evidence Network (EHDEN); there are currently 61 data center partners in this network across 16 countries. The Lille University Hospital (partner of the Rouen University Hospital in the “G4” with Amiens and Caen) is already an EHDEN member making bi- or multi-center studies more feasible.

More information and what data is collected and how is detailed in Annex 2.

Data flow

See Figure 3 for a visual scheme.

Beginning of follow-up: Every week, for 2 years, the list of all new patients fitting inclusion criteria is extracted from EDSaN to create and feed this data into "BDD1 Extraction EDSaN inclusion". From the EDSaN secured environment, EDSaN IDs and date of inclusion are then shared with a third party of thrust (DIM) allowing pairing with email addresses and creation of T2 (contingency table 2) which is shared weekly with CIC (reading purpose only). At the same time, D2IM asks EMS for EMS-IDs to pair with EDSAN-IDs to create T1 (contingency table 1). Replacing BDD1 EDSaN-IDs with EMS-IDs and consequent sharing of "BDD1-bis" with EMS will make it so that EMS will never get to know the key allowing matching with data from the EDSaN secured environment.

Patients’ e-questionnaires: Having read permission on T1 and T2, the CIC can pair EMS-IDs with patients' respective emails. The procedure will be repeated on a biweekly base for 2 years. This will allow (at M0, M6, M12, M18, M24) automatic sending of online self-administered questionnaires to patients via email leading to the creation on the EMS side of "BDD2 e-autoquestionnaires"

Patient ID management for SNDS data matching and collection: Once a year over the 2 years inclusion period, the DIM will pair EDSAN-IDs with NIR and demographic information to unequivocally identify patients and ask the CNAM to extract data from the SNDS. A raw version of the newly created database (BDD3) hosting patient data for the 2 years before the inclusion will be sent over to the EDSaN secured environment.

Merging all the sources: Once a year, for 7 years, EDSaN will perform extraction of all the local health data concerning included patients' visits and/or stays (BDD4) and merge those data with those coming from the SNDS finally recreating an exhaustive picture of patient's clinical history. This new table (BDD5) is then transferred over to EMS and merged with BDD2 (answers from self-administered questionnaires).

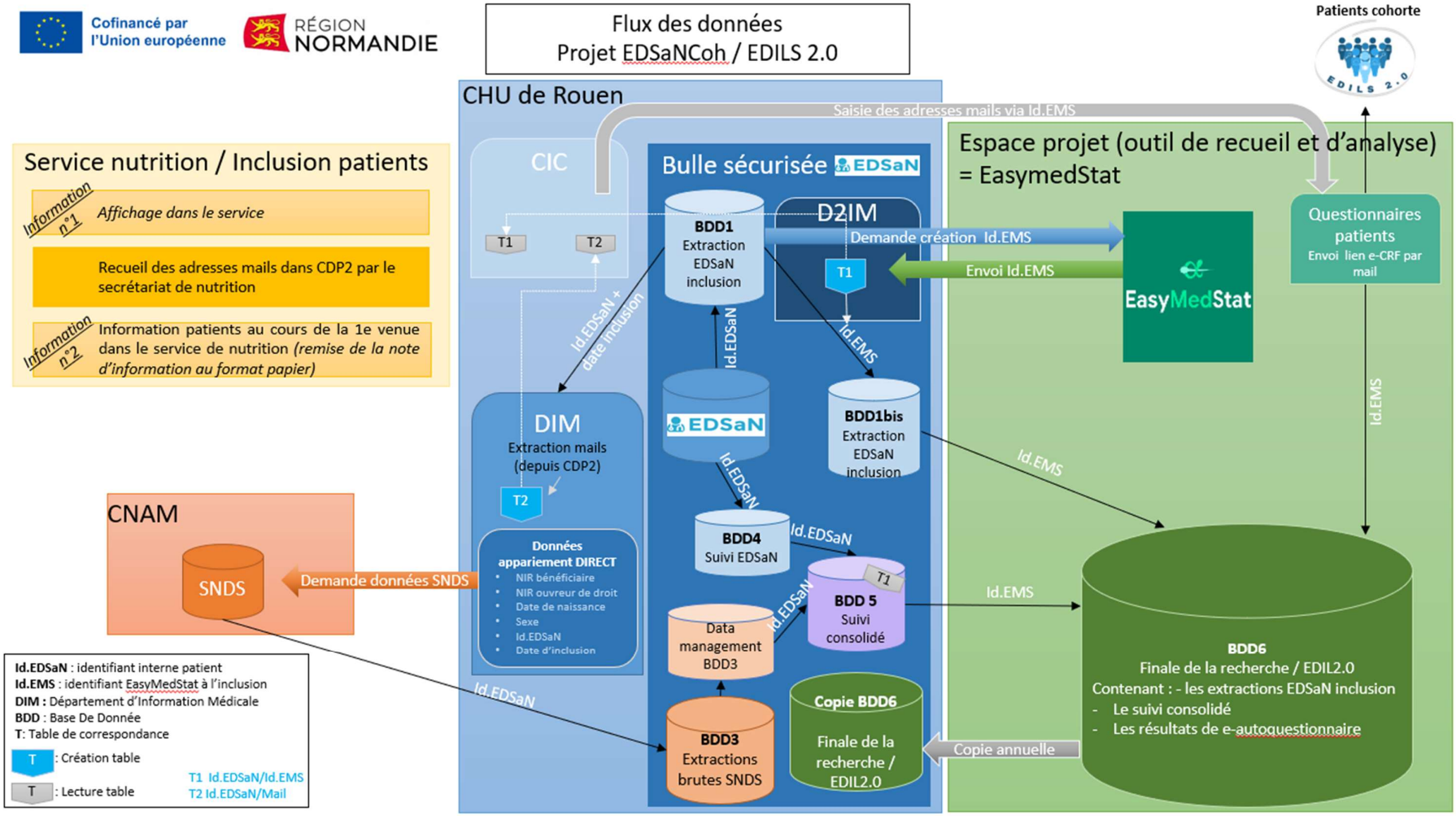
For patients whose email addresses are not entered in CDP2 or who don’t have one, the search is done only on the EDSaN and SNDS data without the e-questionnaire part.

The table below represents who has access to what kind of information. It is important to know that no actor has access to the whole set of information.

Table 3: data access and permissions.

	EDSaN-ID	NIR	EMS-ID	Mail	Tables edit/creation permission	Table reading permission
D2IM	X		X		T1	/
CNAM	X	X			/	/
DIM	X	X		X	T2	/
CIC	X		X	X	/	T1, T2
EasyMedStat			X	X	/	/

Figure 3: data flow and EDSaNCoh architecture.



2.6 Evaluation

Patient inclusion

The patient's inclusion will be evaluated by calculating *precision* and *recall*. In information retrieval and classification (machine learning), these two performance metrics are used to evaluate how well a machine is capable of retrieving or classifying data (from a collection, corpus, or sample space). **Precision** is the fraction of relevant instances among the retrieved ones and it's the equivalent of PPV *(positive predictive value)*. **Recall** is the fraction of relevant instances that were retrieved and it's the equivalent of *sensitivity*.

$$Precision = \frac{TP}{TP + FP} = 1 - FDR \qquad Recall = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

Equation 1: TP = true positives; FP = false positives; FN = false negatives; FDR = false discovery rate; TPR = true positive rate; FNR = false negative rate

Precision and recall are not particularly useful metrics when taken out of context. For instance, it is possible to have perfect *recall* by simply retrieving every single item sacrificing *specificity*. Likewise, it is possible to have near-perfect *precision* by selecting only a very small number of extremely likely items. Either value for one measure is also generally compared for a fixed level at the other measure or they are combined into a single measure. On this occasion, an **F_1 score** is calculated (harmonic mean of precision and recall).

The machine will be compared against the human on 2 random sets of 100 hospital stays: 100 included and 100 non-included patients. The human control will respectively inspect the two sets for false positives and false negatives.

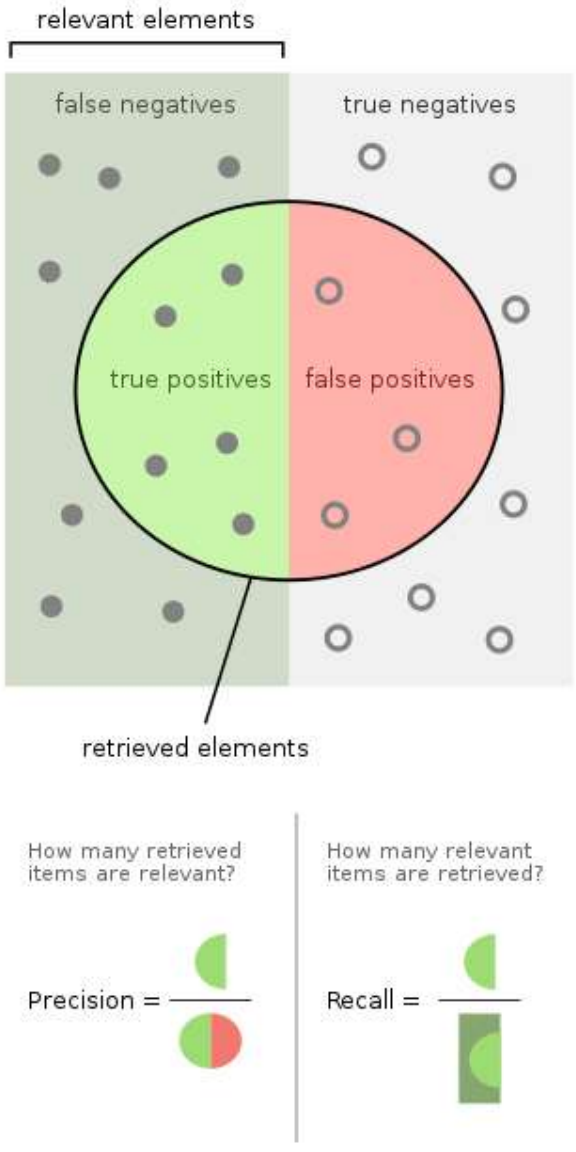


Figure 4: Precision and recall

Feature extraction

Feature extraction is evaluated on a set of 100 random documents. The variables targeted are 24 and according to their nature, they are extracted via EDSaN Docs queries or regular expressions. For each document two queries are run, one "positive" (looking for keywords match) and one "negative" (looking for an explicit negation of the same set of words/concepts). Out of positive and negative query results, four outcomes are possible (see Table 4). In the case of numerical features, if the regular expression matches a value is extracted, else, a "null" value is returned. The measure used for evaluation will be "*accuracy*", representing the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined.

Table 4: possible combinations of query results and consequent behavior.

Positive query	Negative query	Value
TRUE	TRUE	?
TRUE	FALSE	Yes
FALSE	TRUE	No
FALSE	FALSE	Don't know

2.7 Legal and regulatory aspects

The sponsor is responsible for obtaining the agreement of all parties involved in the research to guarantee direct access to source data, source documents, and reports for quality control and auditing purposes at all locations where the research is conducted. Under the legislative provisions in force (articles L.1121-3 and R.5121-13 of the public health code), people with direct access to source data will take all necessary precautions to ensure the confidentiality of information relating to the research, the persons involved in the research and in particular to their identity.

These people, in the same way as the persons who direct and supervise the research, are subject to professional secrecy. Only the health professional who directs the conduct of the research in a center can keep the link between the coded identity of the persons who lend themselves to the research and their first and last names. Each person who takes part in the research has been informed of his or her rights and how to exercise them. Similarly, each person who takes part in the research has given his or her consent to access personal data.

Within the framework of the research, computer processing of certain personal data of health professionals will also be implemented to allow the implementation and the progress of the research. To this end, these personal data will be transmitted to the sponsor or persons or companies acting on its behalf, in France. This data may also be transmitted, under conditions ensuring their confidentiality, to French or foreign health authorities and other entities of the sponsor. Following the provisions of the French law on data processing, files and freedoms, healthcare professionals have the right to access and rectify computerized data concerning them at any time. They also have the right to object to the transmission of data covered by professional secrecy that may be used in the context of this research and be processed.

Archiving of research data

The sponsor and the person(s) responsible for the research archive the essential documents and data related to this research for 20 years following the end of the research (non-interventional research). The retention and archiving of all these respective documents is the responsibility of the person conducting and supervising the research during the archiving period.

All data, documents, and reports may be subject to audit or inspection. These data, unless the person involved in the research objects, may be transmitted to the French and international health authorities and any other partner of the sponsor in France, in the European Union, and outside the European Union, under conditions that ensure their confidentiality and a sufficient level of security. This data, unless the person who is to be involved in the research objects at any time, may also be used in other subsequent research projects exclusively for scientific purposes.

Ethical considerations

The sponsor and the persons directing and supervising the research undertake that this research will be conducted following the Declaration of Helsinki as amended by the 64th General Assembly in Fortaleza

(Brazil) in October 2013, the law n° 2004-806 of August 9, 2004, and its consolidated versions. The research is conducted under this protocol.

The data recorded during this research are subject to computerized processing in compliance with the law n°78-17 of January 6, 1978, relating to data processing, files, and freedoms, modified by the law 2004-801 of August 6, 2004, and its consolidated versions. This research will be submitted to the opinion of a Committee for the Protection of Persons before its implementation and will be notified to the ANSM.

The sponsor will obtain the authorization of the CNIL for the realization of a data processing of the research data. Any substantial modification of the protocol is the subject of a written amendment that is submitted to the manager and biostatisticians, if applicable, and to the sponsor for validation; the latter must obtain a favorable opinion from the CPP before its implementation and notify it to the ANSM. All amendments to the protocol must be brought to the attention of all persons participating in the research and they must undertake to respect its content.

Audit and inspection

An audit may be carried out at any time by persons mandated by the sponsor and independent of those responsible for the research. The purpose of the audit is to ensure the quality of the research, the validity of its results, and compliance with applicable laws and regulations. Those conducting and monitoring the research agree to comply with the requirements of the sponsor and the health authorities regarding an audit or inspection of the trial. The audit may apply to all stages of the study, from the development of the protocol to the publication of results and the filing of data used or generated in the study.

3. Results

3.1 Inclusion

Evaluation of patient inclusion is resumed in the table beside.

This version of the inclusion algorithm reached 96% ***precision*** and 88.1% ***recall*** for patients to be included.

The two measures can be combined into an ***F_1-Score*** yielding a result of 91.9%.

	Human control	
Algorithm	+	-
+	96	4
-	13	87

Table 5: outcome of the inclusion algorithm

3.2 Feature extraction

With 100 documents and 24 variables, there is a total of 2400 matches/mismatches to be evaluated. Feature extraction yielded good results with 2226/2400 matches (92.75% **accuracy**), according to the automatic comparison. The outcome of the automatic comparison between the machine and the human is resumed in the table below.

Table 6: FP=false positive; FN=false negative; IDK="i don't know" (no match neither for the positive or negative query); "?"=unintelligible (match for both the positive and negative query), ■ = extracted via EDSaN queries; □ = extracted via regular expressions.

	MATCHES	MISMATCHES	FP	FN	"IDKs" OR "?"
Physical activity	55	45	12	2	31
Depression	80	20			20
Anxiety	80	20	1		19
Smoking/vaping	86	14	3	1	10
History of anxiety	89	11			11
Diabetes	90	10			10
Death of a beloved one	92	8			8
Suicidal thoughts/self-harming attempts	94	6		1	5
History of depression	94	6			6
Free-fat mass (kg)	95	5			5
Fat mass (kg)	96	4			4
Bone tissue mass (kg)	96	4			4
BDM Femur Z score (z)	96	4			4
Lean mass (kg)	97	3			3
ED (infancy or adolescence)	97	3			3
Trauma (infancy or adolescence)	97	3			3
BDM Spine Z score (z)	98	2			2
Overweight/obesity (infancy or adolescence)	98	2			2
BMI (kg/m²)	99	1			1
Body weight (kg)	99	1			1
Body height (m)	99	1			1
History of physical/verbal abuse	99	1			1
Endometriosis	100	0			0
Autistic disorder	100	0			0
TOTAL	**2226 (92.75%)**	**174 (7.25%)**			
	2400				

4. Discussion

Before going into the technical details of the results, I'd like to present an updated version of the results concerning feature extraction after reviewing previous work. I feel like this is necessary as in my opinion these results, and, taking also into account what could be improved further, have a serious impact on perspectives.

I will share a few reflections and propose a series of improvements limited to the easiest and risk-free to implement, leaving out difficult cases and *unicorns*. The assembly of baseline characteristics derived from the vast array of elements available within EHRs should involve creating rules for the 99%—an informal expression implying that imperfect rules must be implemented that work well for the majority but rarely universally. The counterpoint is that it is often possible to find scenarios where strict application of a proposed rule provides incorrect information; however, changing the rule to accommodate the scenario improperly changes correct information for many more study subjects and could thus be counterproductive.

Table 7: FP=false positive; FN=false negative; IDK="I don't know" (no match neither for the positive or negative query); "?"=unintelligible (match for both the positive and negative query), ■ = extracted via EDSaN queries; □ = extracted via regular expressions. Between brackets, the improvements respect to the previous automatic evaluation.

	MATCHs	MISMATCHEs	FP	FN	"IDKs" OR "?"
Physical activity	57(+2)	43	12	2	29
Depression	83(+3)	17			17
Anxiety	85(+5)	15	1		14
Smoking/vaping	89(+3)	11	3		8
Diabetes	90	10			10
Death of a beloved one	92	8			8
History of anxiety	93(+4)	7			7
History of depression	94	6			6
Suicidal thoughts/self-harm attempts	95(+1)	5			5
Fat-free mass (kg)	97(+2)	3			3
Fat mass (kg)	97(+1)	3			3
ED (infancy or adolescence)	97	3			3
Trauma (infancy or adolescence)	97	3			3
Overweight/obesity (infancy or adolescence)	98	2			2
Bone tissue mass (kg)	99(+3)	1			1
BMI (kg/m²)	99	1			1
Body weight (kg)	99	1			1
Body height (m)	99	1			1
History of physical/verbal abuse	99	1			1
BMD Femur Z score (z)	99(+3)	1			1
BMD Spine Z score (z)	100(+2)	0			0
Endometriosis	100	0			0
Lean mass (kg)	100(+4)	0			0
Autistic disorder	100	0			0
TOTAL	**2258 (94.08%)**	**142 (5.92%)**			
	2400				

Ultimately, imperfect rules must be implemented, and researchers must accept tolerance for this noise. Indeed, a reliance on rules is mandatory in the case of EHR data. Fortunately, direct observation of electronic

medical charts permits scrutiny of EHR-based rules, allowing researchers to "pull back the curtain" to uncover and correct suboptimal rules, a feature not available in most insurance-claims-based studies.

4.1 Inclusion

The algorithm for patient inclusion yielded good results with high accuracy (96%) and a low number of false positives. According to the results, we can estimate for the EDILS2.0 project, 13% of false negatives which is unfortunate but considered still acceptable given that the priority is accuracy and that the number of forecasted inclusions is solid enough to withstand a hit.

The four false positives are due to the inclusion-query matching with some keywords in the anamnesis section of the document. The 13 false negatives are also in good part easily rectifiable. Four of them have been missed because there are no EHRs on them. This is so because, currently, physicians are not obligated to create an EHR after a consultation so there are a few of them that still completely rely on paper clinical narratives. In contrast, participation in working groups, organised to teach patients how to better deal with their condition, is systematically reported in digital form which can allow us to recover these patients with a few tweaks to the query. Currently, exploiting those documents would be the only way to be able to include these patients.

EIP1 (EDILS Improvement Proposal 1):

Query:

- add the following words/phrases:
  - "TAP: Gestion de l'impulsivité alimentaire"
  - "addiction au sucre"
  - "trouble du comportement alimentaire"
  - "Crises compulsives"
  - Replace "grignotage" with *Grignote**

Such changes would impose the following updates to the performance metrics:

- Accuracy for patient inclusion: 96% → **100%.**
- Recall for patient inclusion: 88.1% → **95%.**
- F_1-Score: 91.9% → **97.5%**

4.2 Feature extraction

I will refer to the Table 7 as the "true" set of results for feature extraction. The first thing we can note is that all features extracted via regular expressions (in white in the table above) performed really well scoring at or above 97% accuracy. Those are all numeric variables and their extraction proved to be reliable, as expected: numbers are easier to target and extract than concepts and the clinical narrative template explicitly demands the physician to type them in.

Physical activity, depression, anxiety, and smoking stand out as the most difficult patient features to extract automatically, all of them scoring below the 90% mark in accuracy. It is interesting to note that these are the only variables for which we have false positive or false negatives which indicates that, overall, either the machine is certain about the variable or it refrains from expressing his judgment, which is a good thing: the machine will never assume "*missing information equals X"* unless specifically told so. But why are these variables more problematic than others?

Physical activity

By far the variable that poses more problems.

- Semantics is extremely variable.

- Difficulty to quantify physical activity (recall bias, desirability bias, intensity) and the practice itself is easily subject to frequent variations because of external factors (work/injuries/sickness/motivation/etc…)
- The query is too sensitive: “marche” and “sport” are too sensitive and the “discharge recommendations and follow-up” section is responsible of several false positives and “?” as well (contemporary matches for positive and negative query)
- a few keywords are missing

EIP2 (EDILS Improvement Proposal 2):

- Clearly define what is to be considered “physically active” within the context of EDs. This may vary according to the underlying pathology. If it isn’t precisely defined for the physician, it will never be for the machine either. A good starting point could be the WHO guidelines [5,44] which take into account the global health status of the patient or the American College of Sports Medicine and American Heart Association guidelines [41,46] and/or the utilisation of metabolic equivalent of task (MET)[18,21]. Precisely assessing the level of physical activity would require physicians to adapt their practice and take extra time with the patient which could cause resistance.
- Semantics and temporality: standardize as much as possible, be concise et precise
- Query:
  - replace “marche” with “marche X [time unit]”;
  - add the following to the positive query: “activité physique néant”~3, “Zumba”, “yoga”, “body balance”, “HIIT”, “chasse” ;
  - add the following to the negative query: “reprendre activité physique”~3, “sedentarité”, “absence d’activité physique”~3, “activité physique aucune”~3;
  - do not look for matches in “discharge recommendations and follow-up” section.

EDSaN DOCs can automatically detect negations but it’s not flawless: explicitly adding negations to the negative query is a form of backup plan. Correct identification of negations is closely tied to document formatting. Links between document layout and EDSaN processing need to be better understood.

The last two points are fairly simple to implement and they shouldn’t cause any unwanted side effects. Such implementations would allow recovery of 21 mismatches, boosting accuracy for physical activity to 78%.

Depression and anxiety

- Document layout & temporality: actually the query for *depression* or *anxiety* looks for matches in the whole document, antecedents included, while the query for “history of depression/anxiety” correctly searches only the anamnesis portion of the document. To distinguish “past depression/anxiety” from “current depression/anxiety”, the query should ignore antecedents. This issue is currently responsible for 8 mismatches for *depression* (which, if correct, would bring accuracy much closer to the good results of “history of depression”) and 4 mismatches in the case of *anxiety*.
- Semantics: The patient’s state of mind can be articulated and difficult to interpret and describe. In a few lines, the physician is supposed to synthesize all the relevant nuances of the patient’s state of mind. Let’s take as an exemple the following patient. Forty years old female, with an history of Obsessive Compulsive Disorder and Bulimia during adolescence is consulting for a relapse: *“Good thymia under FLUOXETINE, declares that she can't stand herself anymore, very low self-esteem and dysmorphophobia”*.
  A human being can understand that what the patient is likely trying to be saying is something along the lines of “I’m mostly fine but I hate myself when I look at the mirror”. Such ambiguity will

occasionally cause contemporary matches of the positive and negative query making it difficult to come down to a binary answer.

EIP3 (EDILS Improvement Proposal):

- Query:
  - Make sure to exclude patient's history if the target is current condition
  - Add: "*anxious-depressive context*", "*aboulia*", "*agoraphobia*", "*apragmatism*".
- Cure the document layout to minimize the risk of missing explicit negations (responsible of 7 mismatches for anxiety). Explicitly add negations to the negative query as a backup solution.

Clearly, these proposals are not a panacea and reaching a very high accuracy score for depression and anxiety will remain challenging, nevertheless, they can be implemented easily and quickly and should boost the accuracy score for depression to 92% (+9%) and the accuracy score for anxiety to 96% (+11%) without occasioning an increase in false positives.

Smoking/vaping

The major problem that plagues this variable is semantic which is currently too variable and it's responsible for 8/11 mismatches. There are two possible solutions to tackle the problem: 1) tweak the query to include all the possible expressions a physician could use, and 2) agree on a precisely-defined expression. The first one is immediate, the second one will need to communicate with physicians and medical secretaries, and them to agree upon adapting their practices. Communication will

Both solutions should be implemented as neither of them will be sufficiently effective alone. The query must be able to find a specific line of text and extract it hence it needs something clearer than "tobaco" or "smoking" as an anchor point.

EIP4 (EDILS Improvement Proposal 4)

Utilise the following expression "*Active smoking: [yes][non], ……".* The use of text fields that the physician can easily click to select or erase should gently nudge him to respect the defined structure and he'll be able then to add all the details he wants concerning the consumption. There will be a clear way to recover the variable and physicians should naturally comply to this solution since it doesn't alter their practice at all.

If successful, that would allow recovery of 8 more matches for what concerns the sample test, boosting accuracy for smoking/vaping to 97%.

Diabetes

By the nature of clinical documentation processes, EHR-based rules for binary characteristics are largely restricted to positive affirmations for defining disease "presence" (i.e., observing a documented code), and the absence of positive affirmations (i.e., not observing a code) for defining disease "absence". That is, structured EHR data rarely contain negative affirmations—documentation that a certain disease was sought but not found, which would lend greater credence to its true absence. As such, EHR-based studies have an inherent inability to differentiate "no disease" from "missing disease status", the former defined by the clinical situation where a specific disease was sought but not found, and the latter defined by a disease not sought in any clinical context. As a consequence, EHR studies will ultimately possess a degree of hidden missingness—qualitative diagnoses labeled "not present" according to rule criteria that were simply not investigated under usual clinical circumstances. The extent of hidden missingness will be related to information quality and creates a misclassification problem.

Diabetes is one of those variables. Out of 100 documents, only in one case did we find an explicit deny of the pathology. There are then 5 mismatches because of incorrect identification of the "Family anamnesis" section and a few false positives because of diabetes about to be confirmed but not yet so.

EIP5 (EDILS Improvement Proposal 5)

- Query: add a negative query for the variable "diabetes"
- Document layout: clearly define key sections of the document

These improvements would boost accuracy to 96% (+6%).

Other non-numerical variables

Concerning the other non-numerical variables I'll go over them quickly as they do not pose as much of a challenge and I'll touch on what can be easily and reliably improved.

- **History of depression** and **history of anxiety**: these variables are much easier to deal with compared to their twins' *current depression* and *current anxiety*, all it takes is a well-defined "Personal antecedents" section which would allow them to recover respectively 3 and 5 mismatches.
- **Death of a beloved one**: Four mismatches due to EDSaN incorrectly identifying the phrase containing "death of…." as an antecedent resulting in it being not querable for unknown reasons and one mismatch due to a missing keyword
- **Suicidal thoughts/self-harm attempts**: 2 mismatches due to missed negations (they should be recoverable explicitly adding such expressions to the negative query) and 3 mismatches due to missing keywords.
- **Overweight/ED/AD in infancy or adolescence**: a total of 5 mismatches would be rectifiable by updating the queries.

Numerical variables

The quantitative data typically available within EHRs (e.g., vital signs, laboratory test results) are a commonly cited strength of EHR data. Inevitably, measurement errors and missing data will abound. Random variation of quantitative measurements from an EHR is often greater than analogous measurements taken under a standardized, prospective research protocol, and is largely uncorrectable [42]. Furthermore, missing data in an EHR are seldom missing at random [42]. Oftentimes, missing data imply better health in ways that are undocumented [42]. The tenuousness of the missing-at-random assumption complicates the application of popular imputation techniques

In this particular use-cases, values of interest are issues of DEXA-scans and osteodensitometry's results. As they are not extracted directly from the Radiology's database but from clinical narratives, they are subject to an unknown degree of human error. Let aside this issue, non-quantifiable in the absence of an access to radiology reports in digital form, the extraction yielded good results. Mismatches are due to missing information or to the presence of the outcome of a previous DEXA-scan copy-pasted into the document *before* the new DEXA-scan results. The query scans the document and stops "too early" matching with old results and missing the new ones. All in all the algorithm did well for these variables and actually outperformed the human in a few of them.

If the whole of my proposals turns out to be correct, the following would be the resulting scenario.

Table 8: Best-case scenario following EIPs implementation.

	MATCHes	MISMATCHes
Physical activity	78	22
Depression	92	8
Anxiety	96	4
Smoking/vaping	97	4
Diabetes	96	4
Death of a beloved one	96	4
History of anxiety	98	2
History of depression	97	3
Suicidal thoughts/self-harm attempts	100	0
Fat-free mass (kg)	97	3
Fat mass (kg)	97	3
ED (infancy or adolescence)	100	0
Trauma (infancy or adolescence)	99	1
Overweight/obesity (infancy or adolescence)	98	2
Bone tissue mass (kg)	99	1
BMI (kg/m²)	99	1
Body weight (kg)	99	1
Body height (m)	99	1
History of physical/verbal abuse	99	1
BMD Femur Z score (z)	99	1
BMD Spine Z score (z)	100	0
Endometriosis	100	0
Lean mass (kg)	100	0
Autistic disorder	100	0
TOTAL	**2335 (97.29%)**	**65 (2.71%)**
	2400	

4.3 Web surveys

Since my ambition was to be able to have, and give the reader, a clearer idea of what can be done with this technology, what are its pros and cons etc, I would like to discuss what is, in my opinion, one of the strengths of EDSaNCoh's infrastructure i.e. the possibility of asking included subjects to fill in questionnaires whose data will go directly into the ESM database.

Fully self-administered questionnaires (SAQs) are not as common as interviewer-administered surveys, however, more and more researchers are considering them, including mail surveys, Web surveys, and interactive voice response (IVR) surveys conducted via telephone primarily as a way to reduce costs. With technology becoming almost ubiquitarian in wealthy countries, survey participants are increasingly capable and willing to respond to web surveys via their smartphone or their personal computers. The three most important advantages are that mobile web surveys are faster, simpler, and cheaper. Concerning the **time** required to conduct a mobile web survey, the following observations can be made:

- The time it takes to get in contact with the respondent can be considerably reduced if the invitation is sent by e-mail or text message (SMS).
- Follow-ups can be carried out very quickly by e-mail. The timing of reminders can be tailored to the respondents. Follow-up via questionnaires can be carried on even if the patient moves to another region.

- The time it takes to deliver a complete questionnaire is also very short. As soon as it is completed, the questionnaire is submitted and delivered. Thus, there is no time lag between the moment the respondent returns the questionnaire and the moment it is received.
- The time it takes to store the collected data is dropped. Responses are instantly recorded into a database and prepared for analysis also avoiding any potential error caused by retyping data into the database by hand.

Secondly, web and mobile web surveys can be tailored to the situation. Therefore, they may make life simpler for the respondents and the researcher.
Here are some examples:

- Respondents may be allowed to save a partially completed form. This allows the subjects to complete the questionnaire at their own pace and reduces the cognitive burden.
- The questionnaire may be filled with already available information.
- There can be a facility that automatically generates an e-mail message to the survey agency if the respondent indicates he has complaints about the questionnaire. Such information can help to improve surveys and avoid future problems.
- Response rates can be monitored over time. If the response is lower than expected, action can be undertaken. For example, customized e-mail reminders can be sent without overloading the patient. The literature shows that this may lead to irritation and break-off or lower data quality [35].
- The proper survey software can check that no respondent can complete the questionnaire more than once.
- Like in computer-assisted interviewing, web questionnaires may contain route instructions. These instructions monitor that respondents only answer relevant questions and skip irrelevant questions.
- *Para-data (*data concerning the actual web questionnaire completion process) can be collected and analysed. For example, information on the characteristics of the respondent's technical environment, response time, errors made, and navigation behaviour can help to detect and correct problems in the questionnaire. This is especially important in mobile web surveys.

An important element of the self-administered mode is that there is, by definition, no interviewer involved. As known, **interviewer error** can contribute significantly to total survey error. Hence, by removing the interviewer from the equation, *survey quality can actually improve*. This may be particularly true when survey topics or specific questions are of sensitive nature.

According to a meta-analysis [32] about studies comparing the Web as a method to collect sensitive information to "traditional" ways, we can say that:

- relative to interviewer-administered telephone surveys, Web data collection appears to yield more reports of sensitive information;
- online administration is at least as good as paper self-administration.

There is also evidence that the increase in reporting represents an increase in accuracy as well [20]. A downside of the SaQs approach is that depending on the population of interest, coverage problems may arise. Despite the huge growth of the Internet, there are still many people who do not have access to or choose not to use, the Internet. There are also wide disparities in Internet access among ethnic, socioeconomic, and demographic groups. In the case of EDILS2.0, the target population is for the most part in his young adulthood and very likely to be comfortable with basic internet tasks.

Concerning **non-response rates**, two meta-analyses [25,33] compared nonresponse rates in Web surveys with nonresponse with other methods of data collection, and both studies agree that the difference in response rates averages roughly 11%, with Web surveys getting lower response rates than surveys using mail or the telephone to collect the data. A few things to note:

- the response rate difference is larger with one-time surveys and reduces over-time in the case of multiple checkpoints
- for college populations, the difference in response rates is 3% in favour of web surveys
- the response rate advantage for mail surveys over Web surveys is larger when one or more reminders are sent to members of the sample. Multiple contacts don't raise Web response rates as much as they do with other types of surveys.

Of course, a web survey requires an **initial time investment** since the absence of an interviewer demands a carefully designed survey instrument that is easy for the respondent to complete on his own. Differences in literacy levels among responders should also be considered in the questionnaire design phase. Because of the lack of interviewer-respondent interaction, nonresponse is more difficult to assess and it is a challenge to disentangle the effects of noncontact, refusal, and a poor sampling frame. All of the questionnaires planned to be administered via the EMS platform in the scope of EDILS2.0 are validated for self-administration.

Once these survey steps are over, there are no further data collection obstacles other than the costs of help-desk personnel. Field data collection is relatively cost-free and not dependent on the number of questionnaires administered and completed. Data input costs are irrelevant as well and no time nor effort is required relative to data entry and verification.

It is important to note though that for any e-cohort built and managed via the EDSaNCoh framework, patients do actively choose to participate and it's in their interest to answer questionnaires diffused via EasyMedStat as it is part of their follow-up hence, expect excellent response rates are likely and expected. Once operational, it will be necessary to **train and monitor users**.

Indeed, these tools will imply changes in the practices of the investigators/researchers and support will be essential for their proper use. It is therefore planned to create documents for investigators/researchers (user manual, video clips, tutorials, etc.), face-to-face training for potential users (with demonstrations) and the creation of educational content and tutorials for patients (for non-opposition and questionnaire collection). Given the previous considerations and evidence, despite having its weaknesses, direct administration of online questionnaires will likely be at least as good as "old school" methods but with the advantage of being way more practical, cost-effective, and less time-consuming.

5. CONCLUSION

Traditionally, data from EHRs have been used to assess adverse effects of treatment, especially unexpected effects. Improvements in the availability and quality of data and advances in study designs and analytical methods have broadened the value of this approach. This enables researchers to answer questions of both regulatory and epidemiological importance more quickly than with traditional study designs where data are collected in real-time after the conception of the study.

Drawing correct inference and estimating minimally biased effects from EHR-based retrospective studies greatly depends on identifying the most informative patients from an EHR database without compromising the generalizability. The various ways and intensities by which individuals utilize healthcare services suggest a substantial fraction of patients in an EHR system will have data shortcomings (i.e., do not meet a minimal data-completeness standard) and should not be included in research studies. Focusing on patient selection with an eye toward maximizing data completeness is a logical strategy but defies a completely objective definition and will tend to over-select a less-healthy patient population.

Hidden missingness is unavoidable and impossible to quantify, yet precautions can be taken to minimize its impact through proper patient selection and rule creation. Unfortunately, it is easy to perform an EHR study that generates believable results yet is fraught with preventable misclassification and missing data. Results generated from EHR studies can have an aura of credibility because of highly precise results derived from large sample sizes, yet they can be severely biased. Despite these limitations, EHR-based retrospective studies will likely become more prominent as EHR databases proliferate. Epidemiologists must continue to improve the methods of EHR-based epidemiology, given the relevance of EHRs in today's healthcare environment.

Pros of EHR in epidemiological research can be resumed as follows:

- Studies are more cost-effective to conduct as most data are already collected for other purposes
- Data are less prone to recall bias
- Data can be collected prospectively and are available in near-real time (relevant in fast-changing fields)
- Large sample sizes allow for increased power to conduct granular comparisons between population subgroups and to investigate rare outcomes
- High validity of coded data for many diagnoses
- Detailed prescribing and dispensing information often available for medications
- Potential for linkage across a range of healthcare settings
- Samples are often (not always) representative of the source population

They also come with unique challenges:

- patient phenotyping
- accounting for changes in patient composition and data quality
- reaching high accuracy
- identifying the same patient across different health systems
- accounting for the hardware resources needed to hand big data sets
- a potential barrier to data access as informatics skills are required

According to preliminary results, EDILS2.0 shows a promising data quality, a quality that will, I believe, further improve thanks to the cues given by the latest analysis. The inclusion is about to begin and the project can still fully benefit from the proposed corrections. Optimization suggested are of conservative nature, they are easy to implement and do not require adaptations on the physician's side which should translate into a frictionless implementation. Nevertheless, they should, and will be tested. If successful, they would bring the

accuracy of the feature extraction algorithm into a “zone of confidence”, beyond the 95% mark, crucial for the trustworthiness of future analysis. It will also serve as a manifesto of the good results achievable with e-cohorts and the quality of the EDSaNCoh project and foster the interest in this technology.

BIBLIOGRAPHY

[1] Ackard DM, Richter SA, Egan AM, Cronemeyer CL. What does remission tell us about women with eating disorders? Investigating applications of various remission definitions and their associations with quality of life. J. Psychosom. Res. 2014 Jan;76(1):12–18.

[2] Bardone-Cone AM, Harney MB, Maldonado CR, Lawson MA, Robinson DP, Smith R, et al. Defining recovery from an eating disorder: Conceptualization, validation, and examination of psychosocial functioning and psychiatric comorbidity. Behav. Res. Ther. 2010 Mar;48(3):194–202.

[3] Baumann C, Erpelding M-L, Régat S, Collin J-F, Briançon S. The WHOQOL-BREF questionnaire: French adult population norms for the physical health, psychological health and social relationship dimensions. Rev. Epidemiol. Sante Publique. 2010 Feb;58(1):33–39.

[4] Berg KC, Peterson CB, Frazier P, Crow SJ. Psychometric evaluation of the eating disorder examination and eating disorder examination-questionnaire: a systematic review of the literature. Int. J. Eat. Disord. 2012 Apr;45(3):428–438.

[5] Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br. J. Sports Med. 2020 Dec;54(24):1451–1462.

[6] Call C, Walsh BT, Attia E. From DSM-IV to DSM-5: changes to eating disorder diagnoses. Curr. Opin. Psychiatry. 2013 Nov;26(6):532–536.

[7] Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L. Promises and pitfalls of electronic health record analysis. Diabetologia. 2018 Jun;61(6):1241–1248.

[8] Fichter MM, Quadflieg N, Hedlund S. Twelve-year course and outcome predictors of anorexia nervosa. Int. J. Eat. Disord. 2006 Mar;39(2):87–100.

[9] Gache P, Michaud P, Landry U, Accietto C, Arfaoui S, Wenger O, et al. The Alcohol Use Disorders Identification Test (AUDIT) as a screening tool for excessive drinking in primary care: reliability and validity of a French version. Alcohol. Clin. Exp. Res. 2005 Nov;29(11):2001–2007.

[10] Galmiche M, Déchelotte P, Lambert G, Tavolacci MP. Prevalence of eating disorders over the 2000–2018 period: a systematic literature review. Am. J. Clin. Nutr. 2019 May 1;109(5):1402–1413.

[11] Galmiche M, Lucas N, Déchelotte P, Deroissart C, Solliec M-AL, Rondeaux J, et al. Plasma Peptide Concentrations and Peptide-Reactive Immunoglobulins in Patients with Eating Disorders at Inclusion in the French EDILS Cohort (Eating Disorders Inventory and Longitudinal Survey). Nutrients. 2020 Feb 18;12(2):E522.

[12] Garcia F, Delavenne H, Déchelotte P. Atypical eating disorders: a review. Nutr. Diet. Suppl. 2011 Apr 1;3:67.

[13] Godart N-T, Curt F, Perdereau F, Lang F, Venisse J-L, Halfon O, et al. [Are anxiety or depressive disorders more frequent among one of the anorexia or bulimia nervosa subtype?]. L'Encephale. 2005 Jun;31(3):279–288.

[14] Griffiths MD, Szabo A, Terry A. The exercise addiction inventory: a quick and easy screening tool for health practitioners. Br. J. Sports Med. 2005 Jun;39(6):e30.

[15] Hazzard VM, Loth KA, Hooper L, Becker CB. Food Insecurity and Eating Disorders: a Review of Emerging Evidence. Curr. Psychiatry Rep. 2020 Oct 30;22(12):74.

[16] Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance Questionnaire. Br. J. Addict. 1991 Sep;86(9):1119–1127.

[17] Hudson JI, Hiripi E, Pope HG, Kessler RC. The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication. Biol. Psychiatry. 2007 Feb 1;61(3):348–358.

[18] Jetté M, Sidney K, Blümchen G. Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. Clin. Cardiol. 1990 Aug;13(8):555–565.

[19] John A, Marchant A, Demmler J, Tan J, DelPozo-Banos M. Clinical management and mortality risk in those with eating disorders and self-harm: e-cohort study using the SAIL databank. BJPsych Open. 2021 Mar 19;7(2):e67.

[20] Kreuter F, Presser S, Tourangeau R. Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. Public Opin. Q. 2008;72(5,):847–865.

[21] Larson-Meyer DE. A Systematic Review of the Energy Cost and Metabolic Intensity of Yoga. Med. Sci. Sports Exerc. 2016 Aug 1;48(8):1558–1569.

[22] Legleye S, Guignard R, Richard J-B, Ludwig K, Pabst A, Beck F. Properties of the Cannabis Abuse Screening Test (CAST) in the general population. Int. J. Methods Psychiatr. Res. 2015 Jun;24(2):170–183.

[23] Lelong R, Soualmia LF, Grosjean J, Taalba M, Darmoni SJ. Building a Semantic Health Data Warehouse in the Context of Clinical Trials: Development and Usability Study. JMIR Med. Inform. 2019 Dec 20;7(4):e13917.

[24] Lin C-Y, Broström A, Nilsen P, Griffiths MD, Pakpour AH. Psychometric validation of the Persian Bergen Social Media Addiction Scale using classic test theory and Rasch models. J. Behav. Addict. 2017 Dec 1;6(4):620–629.

[25] Manfreda KL, Bosnjak M, Berzelak J, Haas I, Vehovar V. Web Surveys versus other Survey Modes: A Meta-Analysis Comparing Response Rates. Int. J. Mark. Res. 2008 Jan;50(1):79–104.

[26] Meule A, Holzapfel C, Brandl B, Greetfeld M, Hessler-Kaufmann JB, Skurk T, et al. Measuring orthorexia nervosa: A comparison of four self-report questionnaires. Appetite. 2020 Mar 1;146:104512.

[27] Monacis L, de Palo V, Griffiths MD, Sinatra M. Social networking addiction, attachment style, and validation of the Italian version of the Bergen Social Media Addiction Scale. J. Behav. Addict. 2017 Jun 1;6(2):178–186.

[28] Pedram P, Patten SB, Bulloch AGM, Williams JVA, Dimitropoulos G. Self-Reported Lifetime History of Eating Disorders and Mortality in the General Population: A Canadian Population Survey with Record Linkage. Nutrients. 2021 Sep 23;13(10):3333.

[29] Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci. 2014;2014:218–223.

[30] Reinsel D, Gantz J, Rydning J. The Digitization of the World from Edge to Core. 2018;28.

[31] de la Rie S, Noordenbos G, Donker M, van Furth E. The patient's view on quality of life and eating disorders. Int. J. Eat. Disord. 2007 Jan;40(1):13–20.

[32] Roger Tourangeau, Frederick Conrad, Mick Couper. The Science of Web Surveys. OUP USA; 2013.

[33] Shih T-H, Xitao Fan. Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis. Field Methods. 2008 Aug;20(3):249–271.

[34] Silén Y, Sipilä PN, Raevuori A, Mustelin L, Marttunen M, Kaprio J, et al. Detection, treatment, and course of eating disorders in Finland: A population-based study of adolescent and young adult females and males. Eur. Eat. Disord. Rev. J. Eat. Disord. Assoc. 2021 Sep;29(5):720–732.

[35] Silvia Biffignandi, Jelke Bethlehem. Handbook of Web Surveys. John Wiley & Sons; 2011.

[36] Taquet M, Geddes JR, Luciano S, Harrison PJ. Incidence and outcomes of eating disorders during the COVID-19 pandemic. Br. J. Psychiatry. 2022 May;220(5):262–264.

[37] Tavolacci M-P, Ladner J, Déchelotte P. Sharp Increase in Eating Disorders among University Students since the COVID-19 Pandemic. Nutrients. 2021 Sep 28;13(10):3415.

[38] Tozzi F, Thornton LM, Klump KL, Fichter MM, Halmi KA, Kaplan AS, et al. Symptom fluctuation in eating disorders: correlates of diagnostic crossover. Am. J. Psychiatry. 2005 Apr;162(4):732–740.

[39] Viricel J, Bossu C, Galusca B, Kadem M, Germain N, Nicolau A, et al. [Retrospective study of anorexia nervosa: reduced mortality and stable recovery rates]. Presse Medicale Paris Fr. 1983. 2005 Nov 19;34(20 Pt 1):1505–1510.

[40] de Vos JA, LaMarre A, Radstaak M, Bijkerk CA, Bohlmeijer ET, Westerhof GJ. Identifying fundamental criteria for eating disorder recovery: a systematic review and qualitative meta-analysis. J. Eat. Disord. 2017;5:34.

[41] Wahid A, Manek N, Nichols M, Kelly P, Foster C, Webster P, et al. Quantifying the Association Between Physical Activity and Cardiovascular Disease and Diabetes: A Systematic Review and Meta-Analysis. J. Am. Heart Assoc. 2016 Sep;5(9):e002495.

[42] Williams BA. Constructing Epidemiologic Cohorts from Electronic Health Record Data. Int. J. Environ. Res. Public. Health. 2021 Dec 14;18(24):13193.

[43] Wood S, Marchant A, Allsopp M, Wilkinson K, Bethel J, Jones H, et al. Epidemiology of eating disorders in primary care in children and young people: a Clinical Practice Research Datalink study in England. BMJ Open. 2019 Jul;9(8):e026691.

[44] World Health Organization. Global recommendations on physical activity for health. Recomm. Mond. Sur Act. Phys. Pour Santé. 2010;58.

[45] Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr. Scand. 1983 Jun;67(6):361–370.

[46] Physical Activity and Public Health: Updated Recommendation for Adults From the American College of Sports Medicine and the American Heart Association. Circulation. 2007 Aug 28;116(9):1081–1093.

Appendix

Annex 1 - Acronymes

ADD – Attention disorder disease

ANSM – Agence nationale de sécurité du médicament et des produits de santé

API - application programming interface

AUDIT - Alcohol Use Disorders Identification Test

BMD - Bone mineral density

BSMAS - Bergen Social Media Addiction Scale

CAST - Cannabis Abuse Screening Test

CépiDC - Centre d'épidémiologie sur les causes médicales de décès

CIC – Centre d'investigation clinique (Clinical Investigation Center)

CL - Consultation Letters

CL – Consultation letters

CN – Clinical narrative

CPOE - Computerized Physician Order Entry

CPP - Comité de protection des personnes

DIM – Département d'information médicale (Medical Information Departement)

DRCI - Département Recherche clinique et innovation

DRG - Diagnosis-Related Group

DSM-V - Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

EAI - Exercise Addiction Inventory

EAV – Entity-attribute-value

ECMT - Extractor of Concepts from Multiple Terminologies

eCRF - Electronic Case Report Form

ED - Eating disorder

EDE-Q - Eating Disorder Examination Questionnaire

EDILS - Eating Disorders Inventory Longitudinal Study

EDNOS - Eating Disorder Not Otherwise Specified

EDSaN - Entrepôt de Données de Santé de Normandie

EHDEN - European Health Data and Evidence Network

EHR – Electronic Health Records

EMS - EasyMedStat

ERDF - European Regional Development Fund

GDPR - General Data Protection Regulation

HAD - Hospital Anxiety and Depression scale

HDS – Hébergeur données de santé

HDW - Health Data Warehouses

HeTOP - Health T&O Portal

HR - Hospitalization Reports

IMDG - In-Memory Data Grid

INDS - Institut National Données de Santé

IPAQ - International Physical Activity Questionnaire

IR - information retrieval

IS – Information System

IVR – Interactive voice response

LTD – Long-term disease

MET - metabolic equivalent of task

NLP – Natural language processing

NoSQL – Not-only-Structured Query Language

OCD – Obsessive compulsive disorder

OMOP CDM - Observational medical outcomes partnership Common Data Model

PIN – Personal identification number

PMR - Personal Medical Record

PMSI - Programme de médicalisation des systèmes d'information

QOL – Quality of life

RUH – Rouen's University Hospital

SAQ – Self-administered questionnaire

SHDW – Semantic Health Data Warehouses

SNDS - Système National des Données de Santé (National health data system)

SSE – Semantic Search Engine

WHO QOL-BREF – World Health Organisation Quality of Life (short version)

Annex 2 – collected data

Data collected via the EDSaN

Data collected thanks to the EDSaN is derived from text documents, PMSI (ICD-10 codes), and biological results.

- **Patient selection**: based on PMSI code F50* (Eating disorders), code E66* (obesity), Hospitalization Reports (HR), and Consultation Letters (CL):

Weight, height, and BMI: found in Hospitalization Reports and Consultation Letters

Clinical signs: retrieved from Hospitalization Reports and Consultation Letters

History: reconstructed via Hospitalization Reports and Consultation Letters

Treatments: identified Hospitalization Reports and Consultation Letters

Biology: Potassium, glycemia, and liver function

Imaging: Dexa-scan (fat and non-fat masses) and BMD (bone mineral density) results

Comorbidities: determined via PMSI codes

Data gathered via the "SNDS"

Data extracted from the SNDS (2 years retrospectively and 5 years prospectively) are the following :

Ambulatory care:

  - Consulted specialty, date of consultation (month/year)
  - Delivery of drugs (reimbursed): prescriber, date, type of drug (antidepressant, antibiotic, antiviral)
  - LTD (Long-term disease)
  - Procedures: date and code
  - Pregnancy and childbirth: date
- **Hospitalizations**:
  - Home hospitalization, hospital stays and follow-up cares: mode of entry, date, unit of stay, length of stay, main diagnosis and associated codes, realized procedures, and mode of discharge.
- **Death**:
  - Date and cause

Data gathered via e-questionnaire

Data collected via self-administered e-questionnaires are intended to complement the other sources of information, in particular, to assess the patient's lifestyle by collecting information that only the patients themselves can provide.

Socio-demographic characteristics: marital status, number of children, diploma, employment status.

Anamnesis: major life events

Characterization of the eating disorder:

- Weight
- **EDE-Q (Eating Disorder Examination Questionnaire)**: a self-administered questionnaire consisting of 28 questions. It is used to evaluate the extent and severity of the characteristics associated with a

diagnosis of ED. Four subscales are used (restriction, eating problem, body shape problem, and weight problem) as well as a global score [4].

- **Food insecurity**: can lead to eating disorders, particularly bulimia [15]. It will be measured thanks to the Household Food Security Survey Module (HFSSM**)** in 6 items giving a score from 0 to 6 (0-1 = high or marginal food security, 2-4 = low food security, and 5-6 = very low food security).
- **Orthorexia**: Orthorexia Nervosa is characterized by an obsession with eating healthy foods. This atypical eating behavior is not currently characterized as OCD according to the DSM-V. The BOT (Bratman Orthorexia Test) [26] is a self-questionnaire to assess Orthorexia Nervosa in 10 dichotomous questions (0 = no and 1 = yes). A score ≥ 5 indicates a greater risk of orthorexia.

Physiology evaluation

- Functional gastrointestinal disorders: ROME IV criteria.

Addictions and risky behaviors

- **Tobacco:**
  - Self-reported tobacco use
  - Tobacco dependence with the Fagerström questionnaire, which assesses the degree of tobacco dependence [16]. The score goes from 0 and 10, establishing several degrees of dependence, from the lowest to the highest.
  - Use of electronic cigarettes (vaping).
- **Alcohol:**
  - Self-reported alcohol use.
  - Alcohol-related risk behaviors are assessed by the AUDIT (Alcohol Use Disorders Identification Test) score [9]. It consists of 10 questions, a score between 2 and 3 indicates a moderate risk of harmful use. A score greater than or equal to 8 in men and 7 in women is suggestive of current alcohol misuse, whereas a score greater than 12 in men and greater than 11 in women is indicative of alcohol dependence.
- **Cannabis:**
  - Self-reported cannabis use.
  - Cannabis dependence is assessed thanks to the CAST (Cannabis Abuse Screening Test), a 6-items. The scores obtained vary between 0 and 24 (highest level of dependence). A score above 7 indicates problematic cannabis use. scale [22].
- **Social media:**
  - Social network use: which networks and duration of use.
  - Social network addiction is assessed with the Bergen Social Media Addiction Scale (BSMAS) which is a self-administered questionnaire that measures the risk of social network addiction in 6 items measured by a 5-point Likert scale (1= very rarely and 5= very often) [24,27].

Physical activity

- **IPAQ (International Physical Activity Questionnaire):** This self-questionnaire is composed of 7 questions (short version). It evaluates the global physical activity and the level of sedentary behavior during the last 7 days by taking into account the intense, moderate, or walking practices but also the time spent sitting. 3 levels of physical activity are established: low, moderate, or high according to the score obtained
- **EAI** (**Exercise Addiction Inventory):** a 6-item self-questionnaire that assesses exercise dependence based on DSM-IV criteria for addiction [14]. Scores range from 0 to 30 (0-12 = asymptomatic individual, 13-23 = potentially symptomatic individual, ≥24 = individual at risk for exercise dependence).

Quality of life and psychological evaluation

- **Quality of life (QoL): p**atients' QoL is measured by the WHOQOL-BREF self-questionnaire composed of 26 questions divided into 4 domains (physical, psychological, social relationships, and environment). It is presented in the form of a Likert scale going from 1 to 5, then transformed into a score from 0 (worst quality of life) to 100 (the best quality of life) [3].
- **Depression and anxiety:** HAD (Hospital Anxiety and Depression scale), a questionnaire developed and validated to provide non-psychiatric physicians with a screening test for the most common psychological disorders (anxiety and depression) [45]. It identifies the existence of symptomatology and assesses its severity. The HAD calculates a score for anxiety and another score for depression (optimal score < 8, between 8 and 10 = mild, between 11 and 21 = moderate to severe)

ABSTRACT

INTRODUCTION : Avec l'adoption généralisée des dossiers médicaux électroniques (DME), des quantités de plus en plus importantes de données cliniques électroniques sont générées, ce qui fait que les chercheurs, les administrateurs de soins de santé et les cliniciens s'intéressent de plus en plus à l'utilisation de telles données. Le projet EDSaNCoh, sélectionné et financé par le FEDER (Fonds européen de développement régional), vise à développer une plateforme pour créer et alimenter automatiquement des e-cohortes prospectives. L'objectif final du projet est d'optimiser la recherche non interventionnelle sur les données épidémiologiques et cliniques en réduisant les erreurs humaines, la charge de travail, la complexité de la saisie des données et le temps consacré à la collecte des données par rapport aux méthodes de recherche actuelles, ce qui se traduit finalement par une réduction des coûts. Le premier projet tirant parti de l'infrastructure EDSaNCoh est EDILS2.0 (Eating Disorders Inventory Longitudinal Study) dont l'objectif principal est d'identifier, 2 et 5 ans après une première consultation pour Trouble du Comportement Alimentaire (TCA), les facteurs pronostiques de guérison, de rémission, de changement de type de trouble alimentaire et de décès ou suicide.

METHODOLOGIE : Trois sont les sources de données combinées par l'infrastructure EDSaNCoh : l'entrepôt de données de santé du CHU de Rouen, le SNDS (système national des données de santé), et des questionnaires auto-administrables diréctement envoyés aux patients. Afin d'évaluer les capacités de l'algorithme construit pour EDILS2.0 à identifier correctement les patients répondant aux critères d'inclusion et à récupérer correctement les variables ciblées, ses performances ont été comparées automatiquement, sur un ensemble de documents aléatoires, à un *gold standard* humain.

RESULTATS : L'algorithme a donné de bons résultats, atteignant 96 % de *precision* et 88,1 % de *recall* pour l'inclusion des patients. En ce qui concerne ses capacités d'extraction de caractéristiques, il a obtenu, sur un ensemble de 24 variables, une *accuracy* moyenne de 94,08%.

CONCLUSION : Selon les résultats préliminaires, EDILS2.0 montre une qualité de données très prometteuse, une qualité qui est, je crois, encore améliorable. Les optimisations suggérées sont de nature conservatrice, elles sont faciles à mettre en œuvre et ne nécessitent pas d'adaptations des pratiques de travail, ce qui devrait se traduire par une mise en œuvre sans friction. En cas de succès, elles feront passer la précision de l'algorithme d'extraction des caractéristiques dans une "zone de confiance", au-delà de la barre des 95 %, ce qui est crucial pour la fiabilité des analyses futures. Tel résultat servira également de manifeste des bons résultats que l'on peut obtenir avec les e-cohortes et de la qualité du projet EDSaNCoh et favorisera l'intérêt pour cette technologie.