



UNIVERSITE DE LILLE – SECTEUR DROIT ET SANTE
FACULTE DE MEDECINE HENRI WAREMBOURG
Année 2024

THESE POUR LE DIPLOME D'ETAT
DE DOCTEUR EN MEDECINE

Réutilisation de données issues de « lifestyle centers » : étude de faisabilité

Présentée et soutenue publiquement le 18 septembre 2024
à 16:00 au Pôle Recherche

Par Ruben Dursus-Élisée

JURY

Président :

Monsieur le Professeur Philippe AMOUYEL

Assesseurs :

Monsieur le Docteur Aghiles HAMROUN

Madame le Docteur Lysiane LEGUIER

Directeur de thèse :

Monsieur le Professeur Emmanuel CHAZARD

Avertissement

La Faculté n'entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Sigles

ACLM	<i>American College of Lifestyle Medicine.</i>
AHA/ACC	<i>American Heart Association / American College of Cardiology.</i>
AMA	<i>American Medical Association.</i>
ATP-III	<i>Adult Treatment Panel III.</i>
BMI	<i>Body Mass Index.</i>
CCDA	<i>Consolidated Clinical Document Architecture.</i>
CI95%	<i>95% Confidence Interval.</i>
CIM-10	<i>Classification Internationale des Maladies, version 10.</i>
DBP	<i>Diastolic Blood Pressure.</i>
EHR	<i>Electronic Health Records.</i>
HBP	<i>High Blood Pressure.</i>
HDL	<i>High-Density Lipoprotein.</i>
HL7	<i>Health Level Seven.</i>
IC95%	<i>Intervalle de confiance à 95%.</i>
ICD10	<i>International Classification of Diseases, 10th revision (voir CIM10).</i>
IMC	<i>Index de Masse Corporelle.</i>
ITLC	<i>Intensive Therapeutic Lifestyle Change.</i>
JAMA	<i>Journal of the American Medical Association.</i>
KDD	<i>Knowledge Discovery in Databases.</i>
LDL	<i>Low-Density Lipoprotein.</i>
NF	<i>Normalized Form. 3NF: third degree, 4NF: fourth degree.</i>
OMS	<i>Organisation Mondiale de la Santé.</i>
PMSI	<i>Programme de Médicalisation des Systèmes d'Information.</i>
SBP	<i>Systolic Blood Pressure.</i>

WHO *World Health Organization* (voir OMS).

XML *eXtensible Markup Language*.

Sommaire

Avertissement.....	2
Remerciements	3
Sigles.....	8
Sommaire	10
Préambule	13
Introduction.....	14
1 Introduction générale.....	14
2 <i>Intensive Therapeutic Lifestyle Changes</i> (ITLC).....	14
2.1 Définition de la <i>lifestyle medicine</i>	14
2.2 Le concept <i>Intensive Therapeutic Lifestyle Change</i> (ITLC).....	16
2.3 Les <i>lifestyle centers</i>	17
2.4 Contexte socioculturel des <i>lifestyle centers</i>	17
3 Concepts en réutilisation de données.....	18
3.1 <i>Knowledge discovery in databases</i>	19
3.2 Réutilisation de données en santé.....	19
3.2.1 <i>Data preprocessing</i>	21
3.2.2 <i>Feature extraction</i>	21
3.3 Science des données	22
3.4 <i>Big data</i>	22
4 Les données réutilisées en santé	22
4.1 Avantages et inconvénients de la réutilisation de données en santé.....	22
4.2 Les classements des types de données réutilisées.....	24
4.3 Classement des données selon l'intention d'analyse	24
4.3.1 Données issues de la recherche.....	24
4.3.2 Données administratives.....	25
4.3.3 Données collectées au cours des soins.....	25
4.3.4 Données écologiques	26
4.3.5 Données utilisateurs	26
4.4 Classement des données selon leur origine.....	26
4.4.1 Hôpitaux et cliniques.....	26
4.4.2 Structures administratives publiques et privées.....	27
4.4.3 Cabinets de médecine	27

4.4.4	Open data	27
4.4.5	Outils numériques à destination des usagers	28
4.5	Problématiques rencontrées dans la réutilisation de données en santé.....	28
5	Interopérabilité.....	31
5.1	Définitions.....	31
5.2	Standards d'interopérabilité fréquents en santé	31
5.2.1	Standards d'interopérabilité syntaxiques	31
5.2.2	Standards d'interopérabilité sémantiques	33
6	Objectif	34
	Abstract	35
	Article en Anglais.....	36
1	Introduction.....	36
1.1	Context.....	36
1.2	Objective	37
2	Material and Methods	37
2.1	Data Collection	37
2.2	ITLC Intervention	38
2.3	Software	38
2.4	Data Preprocessing	38
2.5	Feature Extraction	38
2.6	Data Analysis.....	39
3	Results	40
3.1	Types of Collected Data	40
3.2	Data Preprocessing and Feature Extraction	41
3.3	Data Analysis.....	42
3.3.1	Lifestyle Center A – EHR Data.....	43
3.3.2	Lifestyle Center B – Semi-structured Letters.....	46
3.3.3	Lifestyle Center C – Research-purposed Database	49
3.3.4	Lifestyle Center D – Research-purposed Database	52
3.3.5	Comparison between Datasets	55
4	Discussion	57
5	Conclusion.....	60
	Discussion en Français.....	61
	Conclusion.....	65
	Liste des tables.....	66

Liste des figures	67
Références	68

Préambule

Le travail scientifique présenté dans cette thèse de médecine fait l'objet d'une publication d'article international en anglais. Il suit le plan suivant :

- Une introduction longue en français, qui poursuit deux objectifs : présenter le contexte médical avec une orientation principalement pédagogique, et présenter le contexte scientifique et l'objectif, comme le fait également l'introduction de l'article en anglais
- L'abstract en anglais tel qu'il sera soumis en complément de l'article reproduit juste après, avec une traduction en français.
- L'article en anglais, tel qu'il sera soumis à une revue scientifique internationale. Cet article suit le plan classique, dans le format imposé par le journal (introduction, matériel et méthodes, résultats, discussion).
- Une discussion et une conclusion en français

Le document est structuré ainsi en application de la circulaire Toubon¹.

Les références présentées en fin de document résultent de la fusion des parties en Anglais et en Français. La numérotation est donc incrémentée dans l'ensemble du document, que les parties soient anglophones ou francophones.

¹ Circulaire du 19 mars 1996 concernant l'application de la loi no 94-665 du 4 août 1994 relative à l'emploi de la langue française. JORF n°68 du 20 mars 1996 page 4258. NOR: PRMX9601403C

Introduction

1 Introduction générale

Les *lifestyle centers*, ou « centres de mode de vie », se développent dans certains pays, mais pas encore en France. À l'instar des cabinets médicaux, la question de la réutilisation des données de ces centres à des fins de recherche médicale se pose. L'objectif de cette thèse est de s'intéresser à la faisabilité de cette réutilisation de données.

Afin d'introduire ce travail, nous présentons la *lifestyle medicine*, la notion d'*intensive therapeutic lifestyle change* qui en découle et les *lifestyle centers* qui pratiquent ces thérapies, dans leur contexte médical et culturel spécifique. Ensuite, nous présentons le concept de réutilisation de données, dans le contexte de la science des données et dans le domaine de la recherche en santé. Cela nous permettra de souligner l'importance de l'interopérabilité, concept clé à ce travail, développé en partie 5.

Cette introduction se terminera par une présentation des objectifs du travail.

2 Intensive Therapeutic Lifestyle Changes (ITLC)

2.1 Définition de la *lifestyle medicine*

La *lifestyle medicine*, terme parfois traduit par « médecine des modes de vie » peut être définie à partir de la littérature comme l'étude et la technique des démarches de soin visant à favoriser des modes de vie sains – en termes de nutrition, activité physique, gestion du sommeil et gestion des émotions [1,2].

On retrouve des pratiques s'apparentant à la *lifestyle medicine* depuis l'Antiquité. Par exemple, Hippocrate souligne l'importance des habitudes alimentaires dans l'origine des maladies [3]. À la fin du 18^e siècle, dans l'ouvrage de Voltaire, Zadig convainc le sultan Ogul « qu'on se porte toujours bien avec de la sobriété et de l'exercice » [4]. Le développement des connaissances sur l'hygiène des 18^e et 19^e siècles permet les premiers programmes modernes de prévention de grande échelle et le développement des approches hygiénistes, cherchant à promouvoir la santé auprès des populations en promouvant des attitudes perçues comme favorables à la santé [5]. En la fin du 20^e siècle, la promotion de la santé est définie de manière académique grâce à la charte d'Ottawa et accélère l'essor de la santé publique [6]. Ces notions sont à distinguer de la *lifestyle medicine*. En effet, la promotion de la santé est constituée généralement d'approches populationnelles et vise principalement la prévention (primaire, secondaire ou tertiaire) des maladies. La santé publique est également caractérisée par une approche populationnelle ; elle concerne aussi bien la promotion de la santé que la gestion administrative et économique des soins, ainsi que l'épidémiologie. Par contraste, la *lifestyle medicine* s'inscrit dans la pratique clinique, souvent dans un colloque singulier, englobant des approches de traitement de la pathologie ainsi que des approches de prévention [7].

La *lifestyle medicine* moderne se caractérise par l'usage de l'enseignement de modes de vie sains comme participant au traitement de fond de certaines maladies chroniques, telles que l'hypertension, le diabète de type 2 et l'athérosclérose et ses complications [2]. Les mécanismes par lesquels un changement de mode de vie conduit à un rétablissement des patients sont un champ d'étude à part entière [8]. Le rétablissement des patients est décrit comme une rémission ou une « réversion » (*reversal* en anglais) de la maladie chronique. Dans l'exemple du diabète de type 2, chez certains patients pris en soin par *lifestyle medicine*, on peut observer un recouvrement de la sensibilité à l'insuline et la perte des signes diagnostiques cliniques et biologiques de la maladie [9,10].

Le changement du mode de vie du patient dans un but thérapeutique nécessite une forte adhésion du patient, adhésion qui est souvent plus difficile à générer et à maintenir que pour des thérapeutiques médicamenteuses ou chirurgicales [11]. Pour obtenir une telle adhésion, les changements de mode de vie nécessitent d'être congruents aux références culturelles, aux activités professionnelles et familiales, au psychisme et aux représentations personnelles et spirituelles du patient. La congruence de tous ces points est possible grâce au travail des professionnels de *lifestyle medicine* avec les patients. Ce travail peut être décrit en deux aspects. D'une part, le professionnel de santé doit réaliser un travail d'adaptation des changements de modes de vie proposés au patient pour qu'ils soient à la fois efficaces pour sa pathologie spécifique et applicables dans le cadre de ses contraintes et réalités individuelles. D'autre part, le professionnel doit réaliser un travail d'accompagnement du patient pour la levée des difficultés d'ordre psychologique, social ou culturel qui l'empêcheraient de bénéficier du changement de mode de vie adéquat pour sa maladie. Ces deux aspects du travail se retrouvent dans les divers protocoles de *lifestyle medicine*. C'est pourquoi les équipes de *lifestyle medicine* sont souvent pluridisciplinaires, intégrant de manière non exhaustive des médecins spécialistes (nutritionnistes et somnologues par exemple), des diététiciens, des psychologues et psychothérapeutes, des professionnels du sport et des coachs spécialisés en mode de vie [12].

La pratique de la *lifestyle medicine* est de plus en plus organisée au niveau international. Les institutions de référence sont les associations professionnelles *American College of Lifestyle Medicine* – ACLM, et *International Board of Lifestyle Medicine* – IBLM. Ces institutions visent ouvertement à développer la *lifestyle medicine* comme une spécialité médicale à part entière, aussi bien au niveau de son enseignement qu'au niveau de sa pratique. De nombreux cours universitaires abordent spécifiquement la *lifestyle medicine*, et des programmes d'internats états-uniens intègrent des enseignements de *lifestyle medicine* dans leurs programmes pédagogiques [13]. L'*American Medical Association* – AMA, a reconnu l'importance des modes de vie sains et recommande aux médecins de pratiquer des interventions de *lifestyle medicine* pour la prévention et le traitement de maladies chroniques [14]. Cette recommandation affirme aussi l'importance pour les médecins d'acquérir un ensemble de 15 compétences de *lifestyle medicine* publiées en 2010 et présentées ci-dessous [15]. Une version étendue de 85 propositions a été publiée en 2022 mais n'a pas encore la portée officielle de la version présentée ici [16].

Leadership	1. Promouvoir les comportements sains en tant qu'élément fondamental du soin médical, la prévention des pathologies et la promotion de la santé. 2. Chercher à adopter des comportements sains et créer un environnement qui soutient ces comportements dans le cadre étudiant, professionnel et familial.
Connaissance	3. Démontrer une connaissance des preuves que des changements spécifiques de mode de vie peuvent avoir un effet positif sur le pronostic des patients. 4. Décrire des moyens par lesquels le médecin peut agir auprès des patients et de leur famille afin d'avoir un effet positif sur les comportements sanitaires des patients.
Compétences d'évaluation	5. Évaluer les prédispositions sociales, psychologiques et biologiques des comportements des patients et les résultats sur leur santé. 6. Évaluer l'état de préparation, de volition, et la capacité du patient à faire des changements de comportements sanitaires. 7. Réaliser un interrogatoire et un examen physique médicaux spécifiques au statut de santé lié au mode de vie, en particulier les « constantes » du mode de vie telles que la consommation de tabac, d'alcool, le régime alimentaire, l'activité physique, l'index de masse corporel, le niveau de stress, et le bien-être émotionnel. Sur la base de cette évaluation, obtenir et interpréter les tests appropriés pour dépister, diagnostiquer et suivre les maladies liées au mode de vie.
Compétences de gestion	8. Utiliser les recommandations reconnues au niveau national (telles que celles sur l'hypertension et sur l'arrêt du tabac) pour aider les patients à gérer eux-mêmes leurs comportements de santé et de leurs modes de vie. 9. Établir des relations efficaces avec les patients et leurs familles pour provoquer et soutenir des changements comportementaux par l'usage de méthodes psychothérapeutiques et par des outils et des méthodes de suivi basées sur des preuves scientifiques. 10. Collaborer avec les patients et leurs familles pour développer des plans d'action basés sur des preuves, atteignables, spécifiques et écrits, tels que des prescriptions de modes de vie. 11. Aider des patients à gérer et à maintenir des habitudes de mode de vie favorables à la santé et référer autant que nécessaire les patients à d'autres professionnels de santé selon que cela est nécessaire au soin de maladies liées au mode de vie.
Usage des outils communautaires et bureautiques	12. Avoir la capacité de pratiquer en tant qu'équipe interdisciplinaire de professionnels de santé et favoriser une approche en équipe. 13. Développer et mettre en place des systèmes et des pratiques de cabinet appropriés aux soins médicaux de mode de vie, telle que des technologies d'aide à la décision. 14. Mesurer les processus et les résultats pour améliorer la qualité des interventions de mode de vie chez des individus et des groupes de patients. 15. Avoir recours aux ressources communautaires appropriées à la promotion de l'implémentation de modes de vie favorables à la santé.

Tableau 1. Compétences de lifestyle medicine recommandées aux médecins par l'AMA [14,15]

2.2 Le concept *Intensive Therapeutic Lifestyle Change* (ITLC)

Dans cette thèse, nous étudions une intervention de *lifestyle medicine* spécifique, nommée *Intensive Therapeutic Lifestyle Change* – ou ITLC. Cette intervention

consiste à faire entrer le patient dans un mode de vie favorable à son rétablissement de manière rapide, d'où l'appellation *intensive* [17]. Les changements de mode de vie sont réalisés en quelques jours. Pour ce faire, l'équipe de professionnels de santé accompagne les patients au quotidien. L'accompagnement offert par les professionnels peut varier, certains programmes sont réalisés à l'aide de temps éducatifs quotidiens ou pluri-hebdomadaires, d'autres programmes sont dits « résidentiels » car les patients sont logés dans un centre de *lifestyle medicine* et sont accompagnés toute la journée par les professionnels [17]. Le contenu de ces programmes est détaillé dans la partie 2.3 ci-dessous.

Ce type de thérapeutique peut donc couvrir une multitude de pratiques différentes. Ces pratiques comportent en commun le changement de mode de vie du patient en quelques jours.

L'ITLC est préconisée [17] par certains auteurs dans plusieurs buts :

- Diminuer rapidement les symptômes de la maladie
- Favoriser la motivation des patients
- Accélérer l'apprentissage du mode de vie sain

2.3 Les *lifestyle centers*

Les *lifestyle centers* sont des centres spécialisés dans l'ITLC. La description qui suit procède de mes visites de tels centres. Les *lifestyle centers* proposent des programmes d'une durée d'une à trois semaines selon les centres et selon les choix des patients, durant lesquels les patients sont accompagnés au quotidien par des équipes pluriprofessionnelles.

Les programmes proposés comportent les éléments suivants :

- Activités physiques supervisées : des activités physiques modérées et des activités intenses sont proposées et sont adaptées aux patients
- Substitution alimentaire : 2 à 3 repas équilibrés servis par jour, à des heures précises, avec une alimentation à large prédominance végétale et crue. Parfois, pratique de périodes de régimes stricts ou de jeûnes supervisés
- Promotion de la santé en groupes : cours vulgarisés de physiologie, de physiopathologie, et sur les modes de vie favorisant la santé
- Accompagnement psychologique : groupes de parole, programmes de relaxation, suivi par psychothérapeute
- Mesures favorisant le sommeil : heures précises de lever et coucher, luminothérapie, musicothérapie
- Coaching : accompagnement personnalisé motivationnel avec mesures régulières des progrès personnels
- Soins de confort : massages, relaxation, etc.

Ces centres sont destinés principalement à une patientèle autonome, compétente sur le plan cognitif et disposée à s'approprier des changements de mode de vie – du fait de pathologies ou d'appétences personnelles.

2.4 Contexte socioculturel des *lifestyle centers*

Si l'histoire de la *lifestyle medicine* remonte à l'Antiquité, celle des *lifestyle centers* est plus récente. On peut aisément retracer les origines des *lifestyle centers* modernes à une tradition chrétienne datant du 19^e siècle, « la réforme sanitaire ». À

titre d'exemple, le *lifestyle center* considéré comme le plus ancien centre encore en fonctionnement et dédié à la *lifestyle medicine* serait le « *Wildwood Lifestyle Center* », fondé en 1942 dans le but de la promotion de la réforme sanitaire [18,19]. Le terme « réforme sanitaire » fait référence aux objectifs de santé de cette tradition et à la Réforme protestante. La réforme sanitaire peut être décrite comme un mouvement hygiéniste du 19^e siècle visant à favoriser la santé des croyants et des non-croyants avec des objectifs humanitaires, mais aussi sur la base d'une théologie eschatologique [20,21]. L'Église Adventiste du Septième Jour est la principale église concernée par ce mouvement, mais quelques-uns de ses éléments constitutifs sont également retrouvés dans certaines traditions évangéliques, baptistes, et dans l'Église des Mormons² [20,21]. La réforme sanitaire associe à une pratique traditionnelle de la *lifestyle medicine* des éléments apparentés à la naturopathie et au christianisme puritain. À ces éléments naturopathiques et puritains, s'ajoute une approche normative stricte de la nutrition, avec une tradition végétarienne. Tous ces éléments sont enseignés dans les *lifestyle centers* adventistes durant la courte période où réside le patient, avec plus ou moins d'insistance, afin de permettre l'ITLC.

Bien qu'ayant largement contribué à l'invention de la *lifestyle medicine*, il est regrettable que les travaux réalisés dans ces centres ne fassent généralement pas l'objet de publications ni de recherches scientifiques [22]. Cela peut être rapproché de ce que leurs pratiques trouvent leur origine dans des écrits considérés comme divinement révélés, les écrits d'Ellen White (leader religieuse adventiste du 19^e siècle), voire dans la Bible [20]. Cependant, aujourd'hui, la pertinence d'un certain nombre de ces pratiques est largement documentée par la *lifestyle medicine*. Malgré cela, les résultats obtenus des ITLC pratiquées dans ces centres ne font pas l'objet de travaux universitaires et ne sont donc pas accessibles à la communauté scientifique. Heureusement, de plus en plus de ces centres numérisent leurs données. Il est donc envisageable, pour pallier ce manque de travaux universitaire, d'organiser un travail de réutilisation des données cliniques générées dans ces centres. L'étude de la faisabilité d'un tel travail de réutilisation de données est l'objet de cette thèse.

3 Concepts en réutilisation de données

Dans cette partie, nous présentons notre discipline de travail. Notre travail s'intéresse à la réutilisation de données, ou *data reuse*, ou *secondary use of data*. Cette discipline s'inscrit dans la démarche de *Knowledge discovery in databases* ou KDD, concept fondamental de la science des données. Du fait de notre sujet spécifique de travail, nous aborderons les concepts clés suivants :

- *Knowledge discovery in databases*
- Réutilisation de données en recherche médicale
- *Feature extraction*
- Science des données

² L'Eglise des Mormons est officiellement nommée « Eglise de Jésus-Christ des Saints des Derniers Jours ».

3.1 Knowledge discovery in databases

En 1990 est présentée une formalisation du processus par lequel on peut obtenir des connaissances scientifiques à partir de bases de données, sans exiger que les données aient été collectées à cette fin précise, nommé *Knowledge Discovery in Databases* (KDD) [23]. Cette notion de connaissance scientifique issue de l'analyse de base de données est ambitieuse. Ce travail a ensuite été étendu et précisé [24,25], en parallèle avec le développement des notions de *data mining* et de *big data*.

La *big data* – anglais, signifiant « données massives » ou « données en gros » – peut être décrite comme un large champ de travail traitant de l'analyse systématisée de grands volumes de données. L'augmentation importante des capacités de collecte et de stockage des données accompagnant les progrès technologiques s'est accompagnée d'une formalisation des opérations réalisées dans ce champ de travail. Les données massives peuvent revêtir différentes caractéristiques [26]. Prosaïquement, on pourra trouver des données massives par le nombre d'individus qu'elles décrivent, notamment dans le champ de la santé publique ou de la réutilisation de données, par le nombre de variables qu'elles formalisent, notamment dans le champ des -omiques (génomique, protéomique, métabolomique, etc.), ou à travers ces deux aspects, notamment dans le champ des données produites en masse par des dispositifs portables (smartphones, etc.).

La *data mining* (ou exploration de données [27]) peut être définie comme un ensemble divers de méthodes et travaux visant à extraire des informations pertinentes à partir de données brutes [25]. Cette notion d'information pertinente est à la fois plus générale et moins ambitieuse que la notion de connaissance scientifique proposée par la KDD. Le terme *data mining* est ainsi utilisé dans le domaine de l'informatique au-delà de la recherche scientifique.

3.2 Réutilisation de données en santé

Le domaine de la santé regroupe un certain nombre de travaux de nature très diverse (financiers, pharmaceutiques, médicaux, politiques, etc.). Pour simplifier notre propos, nous utilisons l'expression « réutilisation de données en santé » pour désigner les usages de la réutilisation de données appliqués à la recherche médicale.

La réutilisation de données en santé joue un rôle notable dans un certain nombre de domaines d'études, par exemple l'épidémiologie s'intéressant à des individus, l'épidémiologie des soins (recours au soin, suivi des soins), ou encore les études écologiques s'intéressant au lien entre l'environnement et la santé au niveau populationnel.

La réutilisation, ou utilisation secondaire de données, désigne un usage des données qui soit distinct des utilisations initialement prévues lors de la collecte [24,28]. Cette définition souligne ainsi l'utilisation de données avec une nouvelle finalité, et non pas uniquement son usage dans le but d'accroître la connaissance scientifique.

Dans la recherche médicale, les travaux de réutilisation de données peuvent par exemple être des travaux épidémiologiques basés sur l'étude d'une base de données servant aux remboursements de prestations spécifiques à une pathologie précise [29,30]. Les travaux dans ce domaine sont extrêmement variés, certains

s'intéressant à l'analyse spatiale des phénomènes de santé [31,32], à la détection et la prévention des effets indésirables du médicament [33–35], aux effets des médicaments d'après l'analyse de bases de données extra-hospitalières [36,37], ou encore aux phénomènes de santé mesurables dans les cabinets de médecine générale [38,39].

Afin de clarifier le texte par la suite, nous opposerons la notion de réutilisation à celle « d'utilisation primaire de données » pour désigner les utilisations de données initialement prévues lors de la collecte de ces données. En santé, l'utilisation primaire correspond souvent à l'utilisation des données pour le soin des patients sur lesquelles elles ont été recueillies.

De manière intéressante, les données sont habituellement collectées sous une forme native et peu propice à leur traitement. En effet, le format initial des données dépend des instruments et méthodes de leur collecte, tandis que le format pour leur traitement doit être adapté à leurs usages. Ainsi, les données collectées subissent une ou plusieurs transformations avant d'être stockées et/ou utilisées.

L'usage premier de la collecte et du stockage des données est souvent limité à un usage individuel, par un petit nombre d'utilisateurs. Par exemple, les données de la Sécurité sociale en France sont collectées dans le but premier d'assurer le remboursement des prestations de soins : cadre dans lequel chaque personne enregistrée ne bénéficie que des données la concernant. En plus de cet usage immédiat, les données sont stockées dans un format propice à un premier travail administratif ; dans notre exemple, le format pourrait être conçu de manière à analyser le budget dépensé selon les types de prestations remboursées.

Pour permettre la réutilisation de données, les données originales doivent donc être transformées selon de nouveaux critères, spécifiques aux objectifs de la réutilisation.

Par contraste, lors d'une utilisation primaire de données pour la recherche médicale, le format des données stockées correspond aux analyses programmées – ce format est désigné par certains auteurs comme étant des données de formulaires, ou *questionnaire data* [24,40].

Sur la base du processus de la KDD, il a été proposé [24] de formaliser le travail de réutilisation de données en recherche médicale en cinq phases, que nous avons schématisées en Figure 1.

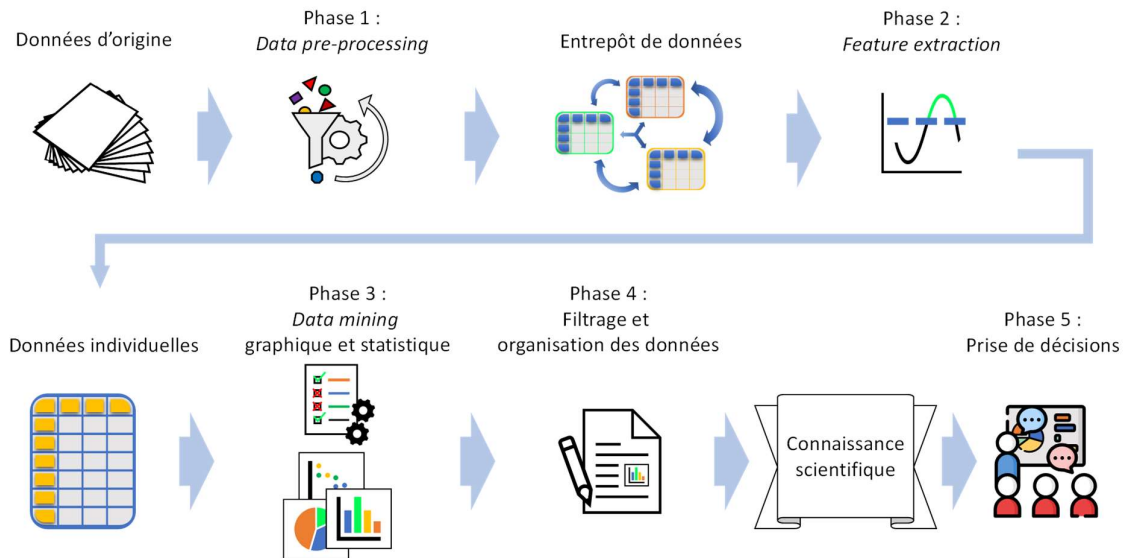


Figure 1. Réutilisation de données en recherche médicale.
 Cette figure est adaptée de [24].

Dans ce processus, les phases 1 et 2 sont spécifiques à la réutilisation de données, et permettent d'obtenir un format de données dites « données individuelles » qui est équivalent au format des données de formulaire. Ainsi, les phases 3 à 5 sont communes aux utilisations primaire et secondaire de données [24]. C'est pourquoi dans le cadre de cette introduction, nous décrivons spécifiquement les phases 1 et 2.

3.2.1 Data preprocessing

L'étape de *data preprocessing*, ou préparation des données, est le traitement informatique consistant à transformer les données initiales en un ensemble de tables reliées, appelé entrepôt de données, répondant à des critères de qualité de formalisation issues de la théorie relationnelle des données [41,42]. Cet ensemble est complexe et tend à coder l'ensemble des données initialement présentes de manière structurée, en évitant les redondances d'une part, et en évitant d'altérer les données par des interprétations ou des regroupements abusifs d'autre part. Cela est permis grâce à des méthodes de formalisation de données comportant des étapes de nettoyage des données, d'alignement terminologique, de correction ou d'alignement d'identité, d'éviction de valeurs aberrantes, de conversion d'unité, de sécurisation, et d'anonymisation des données, etc. L'entrepôt de données est généralement réalisé avec une forme normalisée de degré 3 ou 4 – selon la classification de la théorie relationnelle des données [41,42]. La structure de l'entrepôt de données dépend de la structure informatique des données initiales et non directement de l'utilisation qui en sera faite. Ainsi, les entrepôts de données de santé se veulent pluripotents.

3.2.2 Feature extraction

Le travail de *feature extraction* – extraction de caractéristiques, consiste à interpréter les données de manière systématique dans le but de classer les informations selon les caractéristiques étudiées dans l'analyse subséquente. Ce travail consiste à appliquer des algorithmes afin d'obtenir des données individuelles à partir de

l'entrepôt de données : sur la base de la validation d'un certain nombre de critères, une certaine caractéristique pourra être attribuée à un individu. Par exemple, pour les variables quantitatives, des seuils peuvent être utilisés pour interpréter la donnée pour chaque individu.

Dans le cadre de l'exemple proposé dans le paragraphe 3.2 ci-dessus, afin d'obtenir des données épidémiologiques sur une pathologie, il serait nécessaire de classer les prestations remboursées selon qu'elles sont spécifiques ou non de cette pathologie, à l'aide d'un algorithme défini. Ce classement constitue une extraction de caractéristiques.

L'extraction de caractéristiques, contrairement à l'étape précédente, est spécifique à chaque étude. Malgré les tentatives de documentation et de réutilisation de chaque création de variable dans des *feature stores* [43], à notre connaissance ces mécanismes sont redéfinis pour chaque étude.

3.3 Science des données

La science des données est une discipline scientifique et technique qui s'intéresse à l'ensemble des procédés techniques, concepts scientifiques et moyens mis en œuvre pour générer de la connaissance à partir de données – sans exigence sur le type de données concerné [44].

Cette discipline connaît aujourd'hui un essor important avec le développement de la numérisation, de l'informatique, et de la *big data* [45].

Elle suppose, et reconnaît donc implicitement comme indispensable, la collaboration entre profils issus de l'informatique, de la statistique, et du champ d'application (dans notre cas, la santé). Elle promeut souvent la formation de mêmes personnes aux différents profils, ou au minimum l'acculturation d'ingénieurs polyvalents (informaticiens et statisticiens) au champ d'application.

3.4 Big data

Les données massives sont définies par leur caractère massif uniquement [26]. Elles partagent des caractéristiques communes fréquemment présentées comme les 7 V : Volume, Vitesse, Variété, Vérité, Validité, Volatilité, Valeur [46]. Parce que les données destinées à être réutilisées sont souvent des données de production, elles sont généralement de grande taille. Pour cette raison, le terme de *big data* est habituellement utilisé dans le grand public pour désigner les données qui ont vocation à être réutilisées [26]. Voir aussi partie 3.1 pour une plus ample description de ce domaine.

4 Les données réutilisées en santé

4.1 Avantages et inconvénients de la réutilisation de données en santé

Nous proposons ici de comparer la réutilisation des données à l'utilisation primaire de données en santé, c'est-à-dire les données analysées dans la finalité pour laquelle elles ont été recueillies. Nous mènerons cette comparaison dans le cadre de

la recherche, et non d'autres activités décisionnelles (ex : contrôle de gestion, prévision des ventes, etc.).

Comme dans d'autres domaines, la réutilisation de données permet de limiter les coûts des recherches. En particulier, des économies sont réalisées parce qu'il n'est pas nécessaire de recruter des sujets ni de réaliser une nouvelle collecte de données. Cette considération s'entend lorsqu'on parle de coût marginal. Elle serait moins affirmée si on parlait de coût moyen.

Au niveau éthique, il est intéressant de valoriser, par de telles études, toutes les données disponibles sans nécessité de nouvelles collectes de données sensibles. Cela permet de répondre aux principes éthiques de parcimonie et d'efficacité des études sur l'être humain. À cet égard, rappelons l'emploi du terme « recherche sur les données », qui s'oppose de fait à celui de « recherche impliquant la personne humaine » [47,48].

Ces études de réutilisation de données demandent également moins de temps (marginal) à mettre en œuvre que leurs équivalents prospectifs avec collecte de nouvelles données. En effet, du fait des avantages financiers et éthiques mentionnés ci-dessus, les études s'appuyant sur de la réutilisation de données requièrent souvent moins de démarches administratives. Bon nombre d'entre elles s'inscrivent dans les méthodologies de référence de la Commission nationale de l'informatique et des libertés (CNIL), notamment la MR004 et la MR005 [49].

Ces quelques avantages sont à contrebalancer par d'importantes considérations scientifiques.

Du fait que les études s'appuyant sur la réutilisation de données sont rétrospectives par nature, un certain nombre de problèmes se posent pour l'analyse statistique.

Dans la mesure où les données sont disponibles avant même la définition du protocole de recherche, il est possible de mener une étude « data driven ». Du point de vue du chercheur, cela offre une grande souplesse, et permet une adaptation pragmatique du protocole aux résultats observés. Néanmoins, cela revient finalement à multiplier les variables d'intérêt et les tests statistiques. Multiplier les tests sur un même set de données augmente le risque de première espèce, ou risque alpha. Ce risque n'est généralement pas corrigé dans ces études, contrairement aux études prospectives.

Il y a également une inflation du biais de non-publication : des données réutilisées montrant un résultat négatif ont moins de chances d'être publiées qu'un résultat négatif venant de l'utilisation primaire de données. Ce biais est accru car ces études sont peu coûteuses et souvent non financées, rendant moins crucial la nécessité de publier les résultats négatifs : le temps qui aurait pu être dédié à la publication est avantageusement réutilisé pour réaliser une autre étude. Le déséquilibre entre le temps de travail lié à l'analyse et le temps de travail lié à la publication des résultats explique en partie la majoration du biais de non-publication, en comparaison aux recherches traditionnelles.

On peut également souligner les difficultés de définition liées directement aux caractéristiques des données réutilisées (format, existence, granularité, etc.). En effet, les intentions d'analyse déterminent grandement la structure et le format des données collectées, et il est difficile d'en tirer exactement caractéristiques recherchées. En réutilisation de données, il est fréquent de modifier fortement la question de recherche pour s'adapter à ce que les données permettent d'étudier de

manière scientifiquement acceptable. Ainsi, il n'est pas possible de réaliser toutes les analyses qu'on souhaiterait. Cela constitue une limite à la réutilisation de données qui doit être documentée avec précision pour permettre d'analyser les résultats obtenus.

Enfin, comme toutes les études observationnelles, ces études pâtiennent de biais tels le biais d'indication [50]. Ce biais d'indication concerne non seulement le traitement auquel le patient est soumis (on traite différemment les patients selon la gravité de leur état, mais aussi selon leur potentiel de tolérance d'un traitement, l'option finale curative ou palliative retenue, etc.), et a un impact majeur. Mais il concerne également le recueil de données (on ne réalise pas des examens dont on sait d'avance qu'ils seront négatifs, pour des raisons éthiques et économiques, ou dont on sait que leur tolérance ne sera pas acceptable). Ainsi, prosaïquement, le biais d'indication concerne certes les variables d'exposition, mais également les variables de résultat !

4.2 Les classements des types de données réutilisées

De manière à contextualiser le travail de cette thèse, nous proposons deux manières distinctes de classer les données habituellement réutilisées dans le domaine de la recherche en santé : premièrement, un classement par intention d'analyse et, deuxièmement, un classement par origine des données. Dans un troisième temps, nous discutons également la pertinence de ces classements.

4.3 Classement des données selon l'intention d'analyse

Nous proposons un classement par intention d'analyse, en 5 catégories :

- Les données issues de la recherche, dont la collecte et le stockage sont réalisés dans un but de recherche, mais avec des objectifs parfois différents (ex : étude nichée dans une cohorte, registre, etc.)
- Les données issues de bases administratives, dont la collecte est à objectif administratif (tarification, recensement, imposition, etc.)
- Les données collectées au cours du soin courant, dont la collecte est supervisée par des professionnels principalement dans le but de prodiguer des soins au patient ou d'assurer la continuité des soins
- Les données écologiques, dont la collecte cherche à décrire des environnements et qui ne sont pas structurées de manière à caractériser la santé d'individus
- Les données utilisateurs, dont la collecte est réalisée directement par l'utilisateur (ex : forums, *patient-reported outcomes*, etc.)

Nous présenterons chacune des cinq catégories, en développant particulièrement les catégories concernées par notre travail de thèse – il s'agit des trois premières catégories ci-dessus : données issues de la recherche, données issues de bases administratives, et données collectées au cours du soin.

4.3.1 Données issues de la recherche

Lors de la conception d'une étude médicale, des objectifs scientifiques principaux et secondaires sont déterminés. C'est dans le cadre de la réponse à ces objectifs scientifiques que le format et les méthodes de collecte des données sont choisis. De

la même manière, les méthodes statistiques d'analyse sont choisies pour correspondre aux données collectées et aux objectifs de l'étude.

Ces données peuvent cependant être réutilisées. Cette réutilisation est parfois nommée « analyse secondaire des données » [51]. Lors de la mise en œuvre d'une réutilisation de ces données, les contraintes ainsi posées sont prises en compte et documentées de manière à pouvoir qualitativement interpréter correctement les résultats de l'analyse après réutilisation de données [51]. En particulier, les différentes étapes de changement du format des données doivent être minutieusement suivies pour éviter la perte ou l'altération du sens réel des données étudiées.

À ces considérations conceptuelles s'ajoutent des considérations techniques. Le format des bases de données à visée de recherche peut varier grandement d'une structure à une autre ou d'un scientifique à un autre. Les différences de format peuvent constituer un travail technique important de formatage [52]. Ce type de travaux constitue une forme de *data preprocessing* (cf. partie 3.2.1 ci-dessus). Dans le cadre de la réutilisation de données, ces travaux sont augmentés dans la mesure où les partenaires chercheurs ayant collecté les données ne travaillent pas nécessairement dans le même champ disciplinaire et/ou théorique que les chercheurs menant la réutilisation des données. Ces différences peuvent conduire à des standardisations des bases de données.

4.3.2 Données administratives

Les bases administratives sont généralement d'usage transactionnel, utiles aux procédures de paiement et remboursement des soins. On citera en particulier les bases nationales du Programme de Médicalisation des Systèmes d'Information (PMSI) et du Système National d'Information Inter Régimes de l'Assurance Maladie (SNIIRAM) [53].

Les bases administratives sont une source habituelle de données pour les études en réutilisation de données [54,55]. En effet, de taille importante et souvent exhaustives (en termes de présence d'un séjour par exemple, pas nécessairement en qualité de codage), ces bases permettent des études exploratoires ou avancées en épidémiologie des soins et en épidémiologie. Il est à noter que ces bases de données ne sont plus des échantillons, mais la représentation de la population entière, rendant théoriquement inutile la réalisation de tests statistiques.

Les données administratives de soin détaillent généralement de manière plus détaillée et précise certaines caractéristiques que d'autres. Ainsi, par exemple, le diagnostic d'appendicite est plus fiable et exhaustif que le diagnostic de dépendance au tabac. De même, l'acte de pose de prothèse totale de hanche est plus exhaustif que celui de gazométrie artérielle. Ces différences sont multifactorielles, relevant tantôt de la précision des terminologies utilisées, tantôt de la motivation du codage et de l'incitation financière à utiliser un code. Ainsi, leur exploitation en réutilisation de données pour la recherche médicale nécessite la consultation d'experts du codage et du domaine médical traité.

4.3.3 Données collectées au cours des soins

En cours de soins, l'usage de systèmes informatiques d'aide à la pratique médicale peut être associé à un archivage des informations et ainsi la constitution d'une base de données. L'utilisateur de ces systèmes est un professionnel de santé. Ces bases

ont donc la particularité de contenir des informations individuelles de santé précises et pertinentes pour les professionnels de santé. Elles sont collectées selon les habitudes de pratique, qui peuvent être variables d'un professionnel de santé à l'autre. La réutilisation de ces données nécessite de tenir compte de ces particularités, en particulier pour la *feature extraction* (partie 3.2.2 ci-dessus) [24].

4.3.4 Données écologiques

Les données dites écologiques sont des données recueillies pour un territoire ou pour une condition de vie, et non pour des individus [56]. Il peut s'agir de données de santé, mais également de données environnementales (ex : pollution aux particules fines) ou sociodémographiques (ex : taux de chômage de longue durée). Il est possible de recouper ces données avec des données médicales (épidémiologiques ou individuelles) de manière à conduire des recherches médicales précises. Ce type de réutilisation de données est particulièrement important pour étudier les facteurs de risques environnementaux, tels que la pollution, les milieux de vie, ou les conditions climatiques.

La relation entre les données écologiques et les données médicales nécessite souvent des méthodes spécifiques d'attribution et de test statistiques [56], et soulève des questions d'interopérabilité spatiale [57].

4.3.5 Données utilisateurs

En santé, la réutilisation de données utilisateurs constitue une nouvelle source innovante d'information [58,59]. La nature de ces données est présentée de manière plus détaillée en partie 4.4.5.

Le comportement de l'utilisateur intervient de manière particulière dans la collecte de ces données. Ce facteur est majeur pour toutes les étapes de la réutilisation de ces données.

Les données ainsi collectées sont particulièrement pertinentes pour étudier des aspects épidémiologiques, de recours au soin, ou culturels liés à la santé.

4.4 Classement des données selon leur origine

Nous proposons un classement des données réutilisées en santé selon leur structure d'origine.

- Hôpitaux et cliniques
- Structures administratives publiques et privées
- Cabinets de Médecine Générale
- *Open data*
- Outils numériques à destination des usagers

Nous allons développer ces 5 catégories et expliciter leurs particularités.

4.4.1 Hôpitaux et cliniques

De manière individuelle, les hôpitaux et cliniques ont adopté des systèmes électroniques d'aide à la pratique clinique, souvent regroupés sous l'appellation *Electronic Health Records* (EHR) [24]. Ces systèmes permettent une gestion dématérialisée des données médicales des patients, avec un dossier individuel informatique qui transite d'un service à l'autre en fonction des besoins. Ces dossiers

peuvent contenir des informations médicales, remplies par les professionnels de santé ou les instruments médicaux de mesures ainsi que des informations administratives remplies par le personnel administratif. Certaines données peuvent également être téléversées par des organismes tiers, tels les résultats d'analyses de biologie médicale téléversés par les laboratoires de ville, ou les données de suivi des facturations téléversées par l'Assurance Maladie.

En termes de portée, ces bases ont la particularité d'être limitées à un centre et aux patients de ce centre. En outre, chaque centre dispose potentiellement d'un modèle de données unique et non-interopérable avec d'autres centres. La constitution d'entrepôts de données partagés, ou l'utilisation de logiciels communs peut permettre de réduire la diversité des modèles de données.

4.4.2 Structures administratives publiques et privées

Les assurances de santé ainsi que d'autres dispositifs de services administratifs, publiques ou privées, constituent des bases afin de gérer les prestations offertes à leurs usagers. Ces bases comportent des données importantes sur de larges territoires, et souvent permettent des études exhaustives ou presque exhaustives sur de grandes populations [24,54,55].

Les informations contenues ne sont pas à visée médicale, et souvent ne comportent pas les informations détaillées de manière fiable en ce qui concerne les diagnostics ou les résultats d'examens. Elles permettent au contraire de suivre de manière rapprochée les prestations de soins réalisées, les pratiques de soins et les recours à des prises en charge. À partir de ces données il est parfois possible de faire des études épidémiologiques sur ces territoires, mais cela nécessite des études de validation de la correspondance entre les données de recours au soin recueillies et les situations épidémiologiques.

4.4.3 Cabinets de médecine

Les cabinets de médecine, les maisons de santé et autres lieux de soins dits ambulatoires pourraient être une nouvelle origine de données pour la recherche médicale. En effet, un certain nombre de logiciels d'aide à la pratique médicale et de constitution de dossier électroniques sont utilisés en pratique cliniques. L'usage de ces logiciels se généralise, et la médecine ambulatoire constitue la majorité des prestations de santé ; ces éléments font des données gérées par ces logiciels un vaste champ de recherche exploitable [60]. Un obstacle majeur est la variété des logiciels utilisés, ainsi que la variété des formats de données ainsi créée. Cette variabilité peut constituer un obstacle, mais avec le développement de méthodes appropriées de *feature extraction* permettrait de lever cet obstacle.

4.4.4 Open data

Les *open data*, ou « données libres d'accès », désignent les bases de données d'accès libre. Deux grandes origines de données sont à distinguer : les données issues d'institutions partageant des informations au public, souvent par soucis de transparence [61], et les données scientifiques mises à disposition à la communauté scientifique par des professionnels, chercheurs ou cliniciens [62]. Le caractère sensible des données de santé restreint les champs de l'*open data* en santé [63]. Cependant, des bases avec des données anonymisées ou écologiques sont disponibles.

Il est difficile de référencer les usages des données issues de l'*open data*, du fait de la diversité des types de données et des usages qui en sont possibles. Les données écologiques sont utilisées et transformées pour références démographiques dans certaines études de territoires [64], pour des études d'exposition à des facteurs de risques [65], pour des études épidémiques [66–68], et pour des études de performances de structures ou de politiques publiques [69], entre autres réutilisations [70].

Les données individuelles sont plutôt réutilisées dans des contextes de recherche précis, afin de valider des études préexistantes [62], ou pour poser de nouvelles questions scientifiques [70]. Ces données sont également essentielles aux processus d'apprentissage machine pour le développement d'outils d'intelligence artificiel [63].

4.4.5 Outils numériques à destination des usagers

Un certain nombre d'outils numériques sont développés pour accompagner la santé des usagers. Ces outils conduisent à la réalisation de bases de données qui peuvent être réutilisées. Trois types de sources distinctes sont à considérer.

- **Les applications de santé.** L'utilisateur cherche à suivre sa santé à l'aide de l'application en question, les informations sont recueillies par son initiative [71].
- **Les systèmes de mesure physiques ou numériques associés à des services d'usage courant,** tel que le téléphone portable, la voiture, ou les moteurs de recherche internet. Ces systèmes de mesures requièrent un simple consentement de l'utilisateur, parfois tacite. L'utilisateur ne participe pas activement à la collecte de données, mais son usage de ces services est monitoré d'une telle manière qu'il est possible de réutiliser ces données pour la recherche médicale [58,59].
- **Les réseaux sociaux** et autres moyens de partage d'informations : l'utilisateur participe à une communauté ou un projet, et partage délibérément des informations de santé dans un but précis – généralement pour aider ou être aidé par d'autres usagers [59]

La réutilisation pour la recherche médicale des données ainsi collectées pose des questions de confidentialité, de protection de la vie privée, et d'anonymat spécifiques. Une fois ces interrogations levées, des défis statistiques et d'interprétations sont posés, et font actuellement l'objet de recherches. En effet, ces données comportent une grande quantité de données manquantes car elles ne peuvent pas prétendre être recueillies de manière systématique, mais au contraire leur collecte même est dépendante du comportement de l'utilisateur.

4.5 Problématiques rencontrées dans la réutilisation de données en santé

Nous avons souligné, dans les paragraphes précédents plusieurs problématiques retrouvées pour la réutilisation de données en santé, de manière éparse. Ici nous allons proposer une synthèse non exhaustive de problématiques dans ce travail [28,72].

- Limitation des sources de données

Malgré le développement rapide et important de l'informatique, la numérisation des actes de santé est encore limitée à certains acteurs. La numérisation de ces données nécessite des investissements pouvant constituer un frein à la réutilisation de données, même lorsque les données concernées comporteraient une excellente qualité pour la recherche.

- Respect de la vie privée

Les données de santé sont des données sensibles dont la réutilisation nécessite des précautions particulières de confidentialité. Afin de respecter ces règles, l'extraction de données nécessite souvent des méthodes d'anonymisation, dont la complexité et la fiabilité peuvent constituer des obstacles à la fois à la collecte des données ainsi qu'à leur analyse [72].

Il est généralement assez simple d'éliminer des bases destinées à la réutilisation de données, les informations directes d'identification, telles que les numéros de téléphone, le nom du patient, etc. En revanche, il est parfois difficile d'évaluer la pertinence de données indirectes d'identification, telles que la date du séjour, la date de naissance, ou le lieu de résidence du patient. En effet, ces données peuvent être utiles directement à la recherche médicale ou afin de permettre un alignement des identités. Ce problème n'est pas spécifique à la réutilisation de données ; mais il se pose en des termes particuliers dans la mesure où les données sont déjà collectées, et parfois ainsi accessibles au chercheur – contrairement à la recherche se basant sur une utilisation primaire de données, où la collecte est définie en fonction des besoins.

D'autre part, des problèmes éthiques de consentement à la réutilisation des données sont parfois posés. Il est difficile de permettre de contacter les personnes qui ont contribué aux données. Pour contourner ce problème, les établissements de soins affichent ou informent les patients que les données collectées pourront être réutilisées à des fins de recherche. Mais l'information des patients est parfois difficile à assurer, en particulier lorsque les données ne sont pas issues de centres de soin.

Un certain nombre de solutions techniques ou réglementaires à ces problèmes sont possibles. Elles passent par des techniques d'anonymisation avancées et appliquées avant que ne commence le travail de recherche.

- Interopérabilité

La notion d'interopérabilité, au cœur de notre travail de thèse, est plus complètement décrite dans la partie 5 ci-dessous. L'interopérabilité de bases de données désigne la compatibilité des formats de ces bases entre divers logiciels et méthodes d'analyse. Cette problématique est particulièrement présente en santé du fait de l'existence d'une grande variété de formats et d'usage des données de santé, ayant conduit au fil des ans à la création d'un certain nombre de standards et de codifications qui ne sont pas nécessairement interopérables [73].

- Données manquantes
 - Carence de données accidentelle

Lorsque les données sont manquantes du fait de causes aléatoires, tels que des oublis ou la perte de fichiers, ces données manquantes ne comportent pas de particularités par rapport à des données manquantes dans le cadre d'utilisation primaire de données. On parle alors de données *missing at random* (manquantes aléatoirement), et il existe des méthodes de remplacement des données

manquantes, dites « d'imputation », afin de gérer ces données manquantes [74]. D'autres décisions peuvent être prises pour la gestion de telles données, et cela est soigneusement décrit dans la littérature [75,76].

- Carence inhérente au type de données

Certaines données peuvent être « manquantes » du fait que le recueil des données était de nature à collecter les données lors d'événements spécifiques [24]. Par exemple, très peu de patients bénéficient d'une IRM de l'hypophyse. Les patients en bénéficiant sont ceux qui sont les plus susceptibles d'avoir effectivement un résultat anormal, par exemple une suspicion de tumeur. Mais chez ceux n'ont pas eu un tel examen, les résultats seraient très probablement normaux. Ainsi, la donnée « résultats d'une IRM de l'hypophyse » n'est donc pas manquante aléatoirement. Cela constitue une différence majeure avec la recherche clinique interventionnelle, dans laquelle le recueil des données est protocolisé. Le postulat du *missing at random* est souvent le plus raisonnable en recherche interventionnelle protocolisée, tandis que l'imputation des valeurs manquantes par des valeurs normales est au contraire une méthode plus raisonnable en réutilisation de données en santé. Cela montre que dans le cadre d'une étude en réutilisation de données en santé, il est nécessaire de documenter les causes de données manquantes avec les partenaires impliqués dans la collecte des données pour permettre l'interprétation.

Une fois identifiées les causes de carences de données, la gestion des données manquantes est en soi un domaine de recherche en science des données. Certaines informations manquantes peuvent être simplement exclues de la réutilisation du fait qu'elles ne sont pas pertinentes aux variables étudiées. D'autres situations où la carence de données est apparente, peuvent être codifiées afin de constituer des modalités à part entière. Par exemple, l'absence de recours au soin n'est pas nécessairement une donnée manquante mais peut représenter une suspicion de rétablissement ou un décès en fonction du type de données considéré. Enfin, il est parfois possible dans le processus de *feature extraction* d'inclure des méthodes pour compléter de manière automatisée les données manquantes. Ces quelques modalités présentées, de manière non exhaustive, illustrent les difficultés rencontrées dans ce domaine.

- Interprétation des données

L'interprétation de données réutilisées dans le cadre de la recherche médicale nécessite fréquemment la gestion de divers éléments [24] que nous citons ici de manière non exhaustive :

- Alignement des identités. Il est souvent nécessaire d'assurer l'attribution systématique d'un code unique à une personne donnée afin de documenter pour ce même individu l'ensemble des données disponibles à son sujet.
- Divergences lexicales. Lors de l'utilisation primaire de données les définitions des données collectées sont définies avant l'étude par les investigateurs, tandis que dans le cadre de la réutilisation de données il est important de documenter les définitions des données obtenues, ainsi que la persistance de ces définitions pour toutes les sources de données utilisées.
- Fiabilité de l'information recueillie. Les données peuvent manquer de fiabilité du fait d'erreurs humaines lors de la collecte, d'insincérité des agents de collecte, ou de défaillances du protocole de collecte. Il

convient donc de documenter les conditions de la collecte des données pour soutenir la fiabilité des informations étudiées.

- Cadre académique de référence

La réutilisation des données en santé est un champ de travail encore innovant dont les méthodes et les possibilités ne sont pas codifiées autant que d'autres disciplines plus anciennes [77,78]. Cela constitue une opportunité mais également une limitation dans la mesure où la variabilité du vocabulaire ou des techniques utilisées d'une équipe à l'autre limitent les possibilités de comparaison et de communication nécessaire à la vitalité de la discipline.

5 Interopérabilité

5.1 Définitions

L'interopérabilité d'une base de données désigne la propriété d'une telle base à être utilisée dans de multiples cadres de travail [73,79].

La notion d'interopérabilité regroupe plusieurs concepts. En fonction du concept auquel il est fait référence, l'interopérabilité peut être qualifiée de trois manières :

1. Physique

L'interopérabilité physique décrit l'accessibilité de la base de données par des matériels divers, par exemple par des machines informatiques ou des logiciels différents.

2. Syntaxique

L'interopérabilité syntaxique décrit la compatibilité de l'organisation des données d'un système à l'autre. Par exemple, l'usage d'un modèle similaire d'organisation des variables de la base.

3. Sémantique

L'interopérabilité sémantique est la correspondance des termes codés entre divers systèmes. Ainsi, l'interopérabilité sémantique peut être assurée par l'usage de systèmes de codage de référence.

5.2 Standards d'interopérabilité fréquents en santé

5.2.1 Standards d'interopérabilité syntaxiques

En recherche médicale, nous présentons ici quelques standards d'interopérabilité syntaxiques communs.

Commençons par trois standards définis par le consortium Health Level 7 (HL7). HL7 est un consortium international dont le nom se réfère au septième niveau du modèle Open Systems Interconnection (cadre de travail théorique d'interopérabilité standardisé par l'International Organization for Standardization – ISO), le niveau d'interopérabilité logicielle [80]. Trois standards produits par HL7 sont présentés ici.

- **CCDA : Consolidated Clinical Document Architecture :**

Le standard CCDA, inclus dans les systèmes HL7 v2 et v3, régie le format de documents médicaux, avec une série de normes définissant la syntaxe,

l'organisation et le codage des informations correspondant au dossier d'un patient. Il est le résultat d'une amélioration du standard CDA (Clinical Document Architecture), qui définissait ces mêmes points mais gardait une trop grande liberté pour l'organisation des données dans le fichier, ce qui conduisait à une interopérabilité limitée. Le standard CCDA comporte une grande interopérabilité, mais le format dans lequel les données sont stockées correspond à un classement des informations par dossier médical (généralement correspondant à un patient) et n'est donc pas optimisé pour une réutilisation des données dans un but statistique. Considéré comme une référence, il est habituellement possible d'exporter les informations des logiciels américains de dossiers électroniques sous ce format [81]. Cependant, la norme FHIR est généralement préférée actuellement [82,83].

➤ RIM : Reference Information Model

Tandis que le CCDA peut être approché comme un format de fichier informatique, le RIM est un cadre conceptuel par lequel sont définies les informations codées dans les systèmes informatiques médicaux [84]. Ainsi les concepts de patient, d'acte médical, de consultation ou de soignant sont définis afin de permettre aux concepteurs de systèmes informatiques médicaux de définir des échanges d'information.

➤ FHIR : *Fast Healthcare Interoperability Resources*

Le standard FHIR est un système récent (proposé en 2011) de codage des informations médicales [84]. Ce système inclut à la fois des normes conceptuelles liées au RIM, des références d'implémentation informatiques et des normes de codage. L'interopérabilité de FHIR permet aux logiciels adoptant cette norme une certaine interopérabilité qui a été décrite comme robuste [84,85]. Par comparaison, les normes dites HL7 v2 et v3 créées avant FHIR laissaient une grande place à la customisation, ce qui limitait l'interopérabilité entre systèmes d'informations suivant les mêmes normes [84]. Cependant, l'usage du système FHIR pour la réutilisation de données en vue de recherche est encore un sujet en cours de développement dans la communauté de la recherche médicale informatique [85].

En plus de ces standards développés par HL7, mentionnons les standards OMOP et HPRIM.

- OMOP

Observational Medical Outcomes Partnership est une organisation cherchant à promouvoir des normes pour les bases de données médicales à visée de recherche. C'est dans un but de recherche que cette organisation a créé le standard OMOP CDM (*OMOP Common Data Model*) qui permet de traduire dans un type de données précisément normalisé des informations issues de diverses sources (assurances, hôpitaux, etc.) sous un format commun permettant de conduire des recherches cliniques et plus largement la découverte de connaissance [86,87]. Ainsi, contrairement aux normes HL7 visant principalement les structures de santé et de soins, les normes définies par l'OMOP visent la recherche médicale.

- HPRIM

Le standard HPRIM Santé a été créé initialement par l'association française nommée Harmoniser et Promouvoir l'Informatique Médicale (HPRIM) à partir de la norme américaine ATSM E12-38 de 1991 [88]. Elle concerne l'échange d'informations

médicales entre professionnels de santé (laboratoires de biologie médicale, services cliniques, services d'imagerie, etc.). Elle est maintenant régie par l'association Interop'Santé et figure de référence en France. Présentée parfois comme une norme pour les examens biologiques, elle est cependant polyvalente [88].

5.2.2 Standards d'interopérabilité sémantiques

Les standards d'interopérabilité sémantiques proposent des vocabulaires contrôlés, chacun dans son domaine d'application.

➤ CIM-10

La Classification Internationale des Maladies CIM, version 10 – ou CIM-10 (*International Disease Classification, 10th version* – ICD-10) est un dictionnaire de référence réalisé par l'OMS qui référence les pathologies ou troubles documentés chez l'être humain par catégories. Des modifications mineures sont réalisées chaque année et des réorganisations plus importantes conduisent à l'édition de versions. La version 10 est une version de référence publiée depuis 2008. Actuellement une version 11 est éditée mais n'est pas encore aussi diffusée que la version 10 [89].

➤ CCAM

La Classification Commune des Actes Médicaux est une classification de la Sécurité Sociale française codant les divers actes diagnostiques ou thérapeutiques réalisés essentiellement par des médecins ou des chirurgiens-dentistes, dans le but de leur tarification [90].

➤ ATC

La classification ATC permet un référencement des spécialités pharmacologiques. Elle est éditée et maintenue par l'OMS [91].

➤ LOINC

Le système de classification LOINC est un code permettant de classer certains examens communs, notamment en biologie médicale. Cette classification possède la particularité de documenter les méthodes et les unités utilisées pour les mesures. Elle est donc la méthode de choix pour l'analyse de données de laboratoire médicaux ou le codage de tables de mesures cliniques [92].

➤ UCD et CIP

Tout médicament bénéficiant d'une autorisation de mise sur le marché (AMM) en France se voit attribué par l'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM) une série de numéros d'identification nationaux, nommés codes identifiants de présentation (CIP), chacun correspondant à une présentation de mise sur le marché (une présentation correspond à un conditionnement particulier, par exemple une boîte de 30 comprimés) [93]. Pour qu'une présentation soit admise en établissement de santé, elle doit également recevoir un numéro UCD (unité commune de dispensation), attribué par le club inter-pharmaceutique sous la responsabilité de l'ANSM [93]. Comme il s'agit de codes pour la dispensation des médicaments, ils sont premièrement à usage administratif et logistique, concernant l'achat, la vente et le remboursement des médicaments.

➤ RxNorm

Le RxNorm (nom abrégé venant de l'appellation Rx des prescriptions américaines et du terme « normalized ») est un système de nomenclature des médicaments défini

par la *National Library of Medicine* (NLM) aux États-Unis [94]. Ce système a la particularité d'être lié à plusieurs systèmes de nomenclature de médicaments, ce qui lui confère une certaine interopérabilité [95].

Il existe de nombreux autres standards d'interopérabilité sémantique en santé. Le portail terminologique HeTOP [96] décrit ainsi plus de 3 millions de concepts à travers une centaine de terminologies ou ontologies.

6 Objectif

L'introduction précédente a permis de souligner le caractère unique des thérapies proposées dans des *lifestyle centers*, la rareté des données sur leurs résultats médicaux, et la place en science des données que représenterait la réutilisation de données médicales issues de centres de santé.

Notre objectif principal est d'évaluer la faisabilité de la réutilisation de données issues de *lifestyle centers*. Nous souhaitons tenter une réutilisation de ces données et évaluer quels sont les verrous techniques ou méthodologiques que représente un tel travail.

Nos objectifs secondaires concernent les résultats que nous obtiendrons à travers cette étude : nous souhaitons décrire les caractéristiques démographiques et médicales des patients de ces *lifestyle centers*, et décrire l'impact de la thérapie sur leurs paramètres cliniques et biologiques.

Afin de poursuivre ces objectifs, nos objectifs opérationnels sont 1) de collecter des données issues de plusieurs *lifestyle centers* et d'estimer la quantité de travail nécessaire à l'extraction de ces données, 2) de mesurer des caractéristiques démographiques et des paramètres biologiques de base sur les populations obtenues, 3) de comparer les résultats obtenus en fonction du format des données collectées et des *lifestyle centers*.

À partir de ces travaux, nous comptons compiler des preuves pour une discussion autour, premièrement, de la faisabilité de la réutilisation de données issues de *lifestyle centers*, et, deuxièmement, autour de la robustesse des thérapies proposées dans ces centres.

Abstract

TITLE: Consistency in metabolic changes from intensive therapeutic lifestyle intervention: a multicentric data reuse study of over 3500 records

AUTHORS: Ruben Dursus-Élisée, Nicolea Dan, Emmanuel Chazard

While previous studies have demonstrated the effectiveness of residential intensive therapeutic lifestyle change (ITLC) programs in various small-scale settings, concerns have been raised regarding the comparability and scalability of the methods employed. We hypothesized that reusing routinely collected data from lifestyle centers could provide a feasible approach to investigate the efficacy of residential ITLC programs. We estimated the working time for data preprocessing to be 150.3 and 226.1 hours per 1000 records for the two research-purposed databases, and 157.5 and 197.3 hours per 1000 records for electronic health records and semi-structured letters, respectively; thus, the workloads were surprisingly comparable. The obtained samples exhibited similar age-sex structures but differed in the prevalence of reported common chronic diseases. The metabolic effects observed were highly consistent across samples, with significant reductions in lipidemia among hyperlipidemic patients (-33.5 mg/mL [range: -16.1 to -46.6] in total cholesterol, -26.5 mg/mL [range: -10.9 to -38.5] in LDL cholesterol), body mass index (BMI) among overweight and obese patients (-2.8% [range: -1.8 to -4.0]), and systolic and diastolic blood pressure among hypertensive patients (-13.7 mmHg [range: -4.0 to -29.2] and -7.1 mmHg [range: -3.1 to -18.3], respectively). Overall, our results suggest that data reuse is an effective technique for evaluating the impact of ITLC programs.

Article en Anglais

TITLE: Consistency in metabolic changes from intensive lifestyle intervention: a multicentric data reuse study of over 3500 records

1 Introduction

1.1 Context

Lifestyle medicine is a practice of medicine that involves training and coaching patients into healthy lifestyle habits for the prevention and treatment of chronic diseases [1]. Extensive research has demonstrated the efficacy of healthy habits in preventing diseases, and evidence has been published on the effectiveness of lifestyle interventions in achieving some level of reversal or remission of type-2 diabetes [10], ischemic heart disease [97], hypertension [98], certain cancers [99], and dementia [100].

Intensive Therapeutic Lifestyle Change (ITLC) programs, typically lasting 1-5 weeks, involve comprehensive lifestyle care, including meal replacement therapy, culinary and exercise coaching, stress-relieving therapies, and psychotherapeutic counseling [17]. Historically, ITLC interventions have been practiced in lifestyle centers, also known as "sanitariums", for over a century, with pre- and post-intervention tests performed on patients" [18,101]. While some studies have documented the medical efficiency of ITLC programs, larger studies are lacking, particularly regarding residential programs. Notably, Pritikin et al. [102,103], Ornish et al. [97,99,104], Dan et al. [11], and Sanchez et al. [105] have published findings on the effectiveness of ITLC programs for various conditions. Still, these studies had limited sample sizes, varying from 40 to 120 patients. In contrast, non-residential programs, such as the Pivio program (also known as the Coronary Health Improvement Project (CHIP)), have been studied more extensively [106–108].

Data reuse, the technique of using existing data for research purposes, offers a promising approach to conducting larger studies on residential ITLC programs [24,77]. Data reuse involves accessing, anonymizing, and processing routinely collected clinical data for research purposes. This approach has become increasingly important for evaluating public health interventions and therapeutic practices on a large scale [24,51].

Two phases of data reuse are particularly relevant to our study: data preprocessing and feature extraction [109]. Data preprocessing includes transforming primary data into a data warehouse, typically using third or fourth normal form (3NF/4NF) databases [110]. Feature extraction involves creating databases that only include variables of interest relevant to the study [24]. The quantity and complexity of the work necessary for data preprocessing and feature extraction largely depend on the structure of the data [24,111]. Data structuring facilitates data preprocessing and feature extraction, allowing for systematic mapping of terms instead of more time-consuming methods [112]. For our study, we decided to seek both structured and

semi-structured data to allow for the systematization of the feature extraction. However, we decided not to include unstructured data since we deemed it impractical for a large-scale study.

We differentiated two systems contributing to data structuring: organization and nomenclature. Data organization refers to the nature of the frameworks in which the data is organized, including file formatting that ensures readability by third-party software, regardless of the actual content. National and international entities have published multiple systems of this kind, such as the Health Level Seven models (HL7) [80,113], or the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) [114,115]. This is useful in clinical practice, particularly for interoperability. In our study, we worked with the Consolidated Clinical Document Architecture (CCDA) file format, part of the HL7 version 2 and 3 systems [116]. This is an international standardized medical document format to facilitate the exchange of medical data between software, centered on a one patient-one document format. It is based on the XML file coding system [83,116]. Nomenclatures, also known as coding or naming systems, are necessary to interpret the specific content in the file, such as the International Statistical Classification of Diseases and Related Health Problems (ICD) [117], the Logical Observation Identifiers Names and Codes (LOINC) [118] and the RxNorm [94] coding systems.

Large cohorts of patients allow a more robust and detailed analysis of the efficiency of a therapeutic intervention. Data reuse is a common methodology to allow larger studies to be conducted. To our knowledge, no large data reuse study has been conducted on residential ITLC programs.

1.2 Objective

Our study aimed to explore the feasibility of reusing clinical data from residential ITLC programs conducted in lifestyle centers. We sought to describe the process and results of reusing pre/post data, analyze the types of datasets that can be collected, and examine the data preprocessing necessary to obtain the data. Additionally, we compared the datasets obtained.

2 Material and Methods

2.1 Data Collection

We anticipated four scenarios for data collection from lifestyle centers: physical medical records, digital administrative records, electronic medical records, and research-purposed digital databases. We focused on centers with digital records, without discrimination based on the original purpose. We requested records containing lab results for comparison. Eight lifestyle centers were contacted, and four provided digital databases for the project: Herghelia Lifestyle Center, Uchee Pines Institute, Weimar Institute, and Wildwood Lifestyle Center. Throughout this study, we refer to the four sources as datasets A to D, noting that this designation does not correspond to a specific alphabetical order of the centers' names.

2.2 ITLC Intervention

Residential ITLC programs have been described in the literature [17,97]. Briefly, volunteer patients participated in a routine program, including daily moderate aerobic exercise, a low-fat high-carbohydrate plant-based diet, daily lifestyle education sessions, daily chaplaincy sessions, weekly psychological counseling sessions, and medical consultations. The residential nature of the program and the rural location of the centers reinforced the patients' immersion in healthy lifestyle habits. Individual patients were prescribed herbs, supplements, massages, and bath treatments based on their specific complaints. The length of stay varied between programs (typically 7-28 days), but we focused on data from the first two weeks: pre-intervention data were collected on day 0 or 1, and post-intervention data were collected between days 6-15.

2.3 Software

Data was preprocessed using Microsoft Excel[®] software, WinRAR[®], R[®], and RStudio[®] software, with multiple R packages [119–127].

2.4 Data Preprocessing

For this study, we define the following terms related to data preprocessing:

- Data retrieval: the process of obtaining data files from the lifestyle center storage system.
- Data reduction: the process of cleansing and selecting features, including removing duplicates and incomplete or corrupted data, aligning terminology, converting units, and removing irrelevant data.
- Data transformation: the process of handling data to obtain 3NF relational databases and setting up a data warehouse with the same structure and coding for all four datasets.

2.5 Feature Extraction

Features were chosen based on literature relevance and availability in the datasets. We extracted 17 features, organized into 5 categories: demographic, diagnostic, medication, clinical, and lab data (Table 1). Demographic, diagnostic, and medication data were considered stable during the lifestyle intervention. Clinical and lab data were extracted as pre/post-intervention values; unpaired data were discarded as irrelevant.

Demographic features, such as age and sex, were directly extracted from the datasets. In cases where age was not readily available, it was estimated from the intervention year and year of birth.

Diagnostic information was obtained from the diagnostic section of the records, which included both previously known diagnoses and those identified during the lifestyle intervention. In instances where the categorization of a diagnosis was unclear, the ICD10 classification was referenced using the codes listed in Table 1.

Medication records were interpreted as medications used before the lifestyle intervention. Typically, medication usage decreases in dosage or ceases altogether during lifestyle interventions, with new medications rarely introduced.

No specific transformations were required for clinical or laboratory data.

Category	Features
Demographic	Age Sex
Diagnostic (ICD-10 codes)	Cancer (C01 to C96) Diabetes (E08 to E14) High blood pressure (I10 to I15) Cardiovascular disease other than high blood pressure (I00 to I09 and I16 to I99)
Medication	Hypolipidemic Antidiabetic Antihypertensive
Clinical data	Height } used as BMI Weight Systolic and diastolic blood pressure
Lab results	Total cholesterol LDL-Cholesterol HDL-Cholesterol Triglycerides

Table 1. Selected features for data extraction

2.6 Data Analysis

When a patient, whether sick or healthy, had already adopted healthy lifestyle habits, it was not expected that any of his/her metabolic parameters would change because of the lifestyle intervention, due to the very nature of the intervention. Therefore, our analysis focused on the impact of the lifestyle intervention depending on the initial patient status. Subgroups were defined (Table 2) for each parameter based on initial patient values, using thresholds from the Adult Treatment Panel III (ATP-III) and American Heart Association/American Cardiology Association (AHA/ACC) guidelines for lipidemia, blood pressure, and BMI [128,129].

	Unit				
Total Cholesterol	mg/ml	Desirable < 200	High ≥ 200 & < 240	Very High ≥ 240	
LDL Cholesterol	mg/ml	Desirable < 130	High ≥ 130 & < 160	Very High ≥ 160	
HDL Cholesterol	mg/ml	Desirable < 40	High ≥ 40 & < 60	Very High ≥ 60	
Triglycerides	mg/ml	Desirable < 150	High ≥ 150 & < 200	Very High ≥ 200	
BMI	kg/m ²	Underweight < 20	Normal ≥ 20 & < 25	Overweight ≥ 25 & < 30	Obese > 30
Blood pressure	mmHg	Normal SBP: < 120 and DBP: < 80	Elevated SBP: ≥ 120 & < 130 and DBP: < 80	High Blood Pressure Stage 1 SBP: ≥ 130 & < 140 or/and DBP: ≥ 80 & < 90	High Blood Pressure Stage 2 SBP: ≥ 140 or/and DBP: ≥ 90

Table 2. Subgroups definitions for lipidemia, BMI, and blood pressure analysis.

3 Results

3.1 Types of Collected Data

We obtained three types of data (Table 3), from four lifestyle centers, labeled A, B, C, and D.

Lifestyle center A provided data collected via electronic health records (EHR) software, AllegianceMD®. Lifestyle center B provided semi-structured letters that were routinely written at the discharge of the patient, to be sent to their primary care provider. Lifestyle centers C and D provided data that they manually collected for the purpose of research.

The data collected from EHR (lifestyle center A) were in a Health Level 7 Consolidated Clinical Document Architecture (CCDA) format.

The semi-structured letters from lifestyle center B were digitally written by the medical secretaries of the lifestyle center, one by one, with the use of a template, from doctors' notes. The files were in a standard Microsoft Word® format, with one file per patient.

The research-type databases were organized in a standard fashion by clinicians. Whereas these databases were made for the purpose of studying the work made at the lifestyle center, they did not suit our data reuse purpose and required extensive data management. The databases were on a .x/sx standard Microsoft Excel® format.

Type of data	Lifestyle center	Standardized format	File format	Primary use	Generation
Software-generated EHR	A	CCDA	.xml One file per patient	Clinical care and billing	EHR Software
Semi-structured letter	B	None	.doc One file per patient	Communication with a family physician	Medical secretary typing, using a template
Research-type table	C and D	None	.xlsx Single database file	In-house research	Manually collected from unstructured medical records

Table 3. Types of data received from lifestyle centers.

3.2 Data Preprocessing and Feature Extraction

We chose 17 features to extract, organized into 5 categories of data: demographic, diagnostic, treatment, clinical, and lab data, as listed in Table 1. The feature extraction was investigated in two aspects: the systematicity of coding and the time necessary to extract the data. The two are interdependent: less systematic coding led to more time to extract the data. Our findings are summarized in Table 4.

Lifestyle center	Original format	Type of feature				
		Demographic	Diagnostic	Medication	Clinical	Labs
A	EHR	Systematic	Incomplete	Incomplete	Systematic	Systematic
B	Semi-structured letters	Systematic	Free text	No data	Inconsistent	Inconsistent
C	Research-purposed 1	Systematic	Systematic	Systematic	Systematic	Systematic
D	Research-purposed 2	Systematic	Inconsistent	Inconsistent	Systematic	Systematic

Table 4. Quality evaluation of feature coding in collected datasets

The working time necessary for data preprocessing is presented in Figure 1.A. Three steps for data preprocessing categories were identified (as described in the Material and Methods section): data retrieval, data reduction, and data transformation. Data retrieval was necessary for EHR and letters preprocessing, but not for the research-purposed database. Data retrieval represented 76 hours of work for EHR data and 212 hours for data from letters. Data reduction represented 101 hours for EHR data and 253 hours for letters, but only 26 and 45 hours for the two research-purposed databases. Finally, data transformation was not dependent on the type of data, since it was estimated at 5 hours for any data set. The total numbers of hours of data preprocessing were found to be 182, 470, 31, and 50 hours for datasets A, B, C, and D respectively, as indicated in Figure 1.A.

Since data retrieval and data reduction were dependent on both the volume and type of data, we reported the number of working hours to the number of finally usable records, to estimate the working hours per 1000 records (Figure 1.B). We obtained 157.5 hours of data preprocessing work for data from EHR records (lifestyle center A), 197.3 hours for data from semi-structured-letters (lifestyle center B), and 150.3 and 226.1 hours for data from research-purposed databases (lifestyle centers C and D respectively), as presented in Figure 1.B. Thus, surprisingly, when reported to the volume of the data, we obtained that the number of hours necessary for preprocessing of data originating from research-purposed database was comparable to the number of hours necessary for data originating from semi-structured letters or EHR.

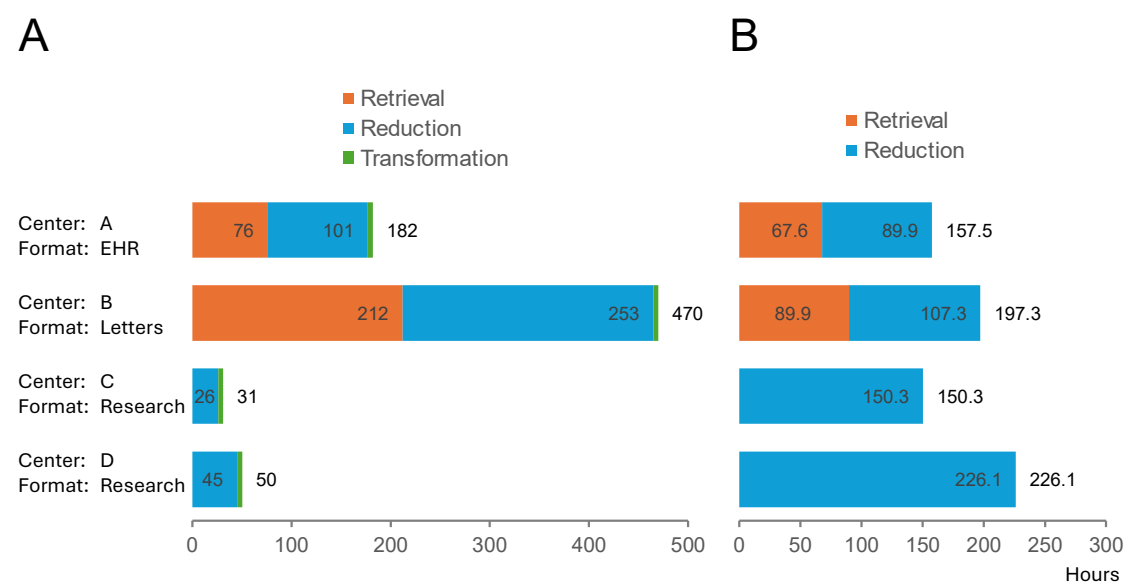


Figure 1. Time requirements for data preprocessing, by dataset and original format, shown in both (A) raw hours and (B) hours per 1000 records.

3.3 Data Analysis

The data extracted from the databases can be described as follows, with three figures and two tables providing an overview. The first figure presents demographic data, with panel A showing the age-sex distribution and panel B displaying the proportion of male/female patients and patients aged 50-69. The two tables provide information on the proportions of patients diagnosed with major chronic conditions (HBP, cancer, cardiovascular disease other than HBP) and the proportions of patients receiving treatment for chronic diseases (hypertension, diabetes, dyslipidemia). The third figure displays the pre- to post-intervention changes in plasma, LDL, and HDL cholesterol levels, triglycerides levels, BMI, and blood pressure, categorized by relevant subgroups.

3.3.1 Lifestyle Center A – EHR Data

3.3.1.1 *Description of the Patients*

From the electronic health records (EHR) from lifestyle center A, we extracted 1,124 patient records. We had no missing data for age or sex. Unsurprisingly, we had a majority of female patients, 69.1% [CI95%: 66.4-71.8] (Figure 2.B). Most patients were middle-aged, with 66.3% [63.5-69.0] being between 50 and 69 years old (Figure 2.C). For the subsequent population analysis, we standardized the results for age and sex, using the age-sex distribution structure obtained from pooling the 4 databases of our study.

We examined the reported diagnoses for four chronic disease categories: cancer, high blood pressure, cardiovascular disease (excluding high blood pressure), and diabetes (Table 5). The standardized rate (SR) of patients diagnosed with high blood pressure was 26.4% [23.3-29.8], while 10.9% [8.9-13.2] had other cardiovascular disease diagnoses, 10.9% [9.0-13.3] had cancer, and 16.5% [14.1-19.4] had diabetes.

We also investigated the number of patients receiving pharmaceutical treatment at the beginning of their stay for three major chronic diseases (Table 6). The SR for patients receiving antidiabetic medication was 18.7% [16.0-21.7], antihypertensive medication was 39.7% [35.8-43.9], and hypolipidemic medication was 15.0% [12.7-17.8].

3.3.1.2 *Lipidemia, BMI, and blood pressure changes*

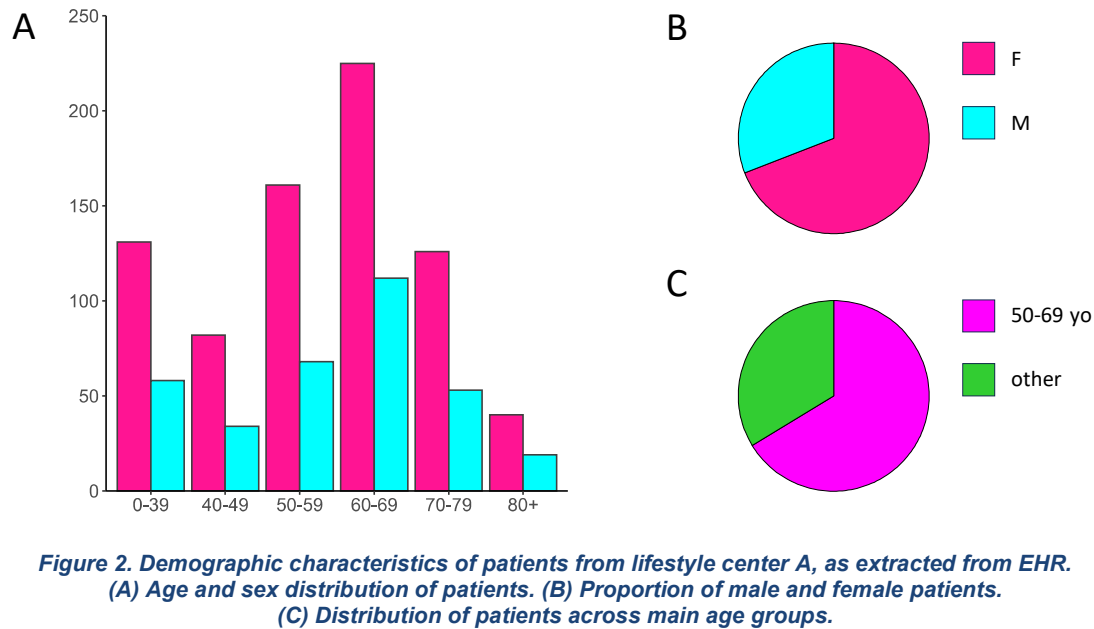
We examined the impact of the lifestyle intervention on lipidemia, BMI, and blood pressure (Figure 3). Patients had these parameters measured on the first day of the intervention and 6 or 7 days after (while the intervention was still in progress, as described in the Material and Methods section). We analyzed the mean difference between the first and second measures. We defined subgroups as described in the Material and Methods section.

Concerning total cholesterol and LDL cholesterol, our analysis revealed a significant reduction across all subgroups, with a clear dose-response relationship, where subgroups with higher initial lipidemia levels exhibited a greater decline (Figure 3). In contrast, HDL cholesterol levels only decreased significantly in subgroups with initially elevated or very high levels, but not in those with normal initial levels.

Regarding BMI, we observed a significant percentage decrease in the normal, overweight, and obese subgroups, whereas underweight patients experienced a non-significant increase.

Regarding blood pressure, patients with normal or elevated blood pressure did not exhibit significant changes in their average values, whereas those with stage 1 or stage 2 HBP showed a significant decrease.

Overall, our findings indicate a significant improvement in blood lipid profiles, blood pressure, and body mass index following the lifestyle intervention.



Condition	% (n)	Standardized % [CI95%]	Average % in the other centers (n) [CI95%]	Test
Cancer	10.4% (115)	10.9% [9.0-13.3]	6.3% (159) [5.4-7.3]	p < 0.0001
Cardiovascular disease	10.0% (111)	10.9% [8.9-13.2]	28.5% (773) [26.9-30.3]	p < 0.0001
Diabetes	15.1% (168)	16.5% [14.1-19.4]	16.7% (453) [15.4-18.2]	ns
High blood pressure	24.6% (273)	26.4% [23.3-29.8]	36.4% (987) [34.6-38.3]	p < 0.0001

Table 5. Prevalence of chronic diseases among patients at lifestyle center A, as extracted from EHR.

Medication	% (n)	Standardized % [CI95%]	Average % in the other centers (n) [CI95%]	Test
Antidiabetic	16.9% (187)	18.5% [15.9-21.5]	23.0% (86) [19.0-27.5]	ns
Antihypertensive	37.1% (411)	39.2% [35.5-43.4]	40.9% (152) [36.0-46.0]	ns
Hypolipidemic	14.0% (155)	14.9% [12.6-17.6]	14.2% (53) [11.1-18.2]	ns

Table 6. Prevalence of chronic disease medications among patients at lifestyle center A, as extracted from EHR.

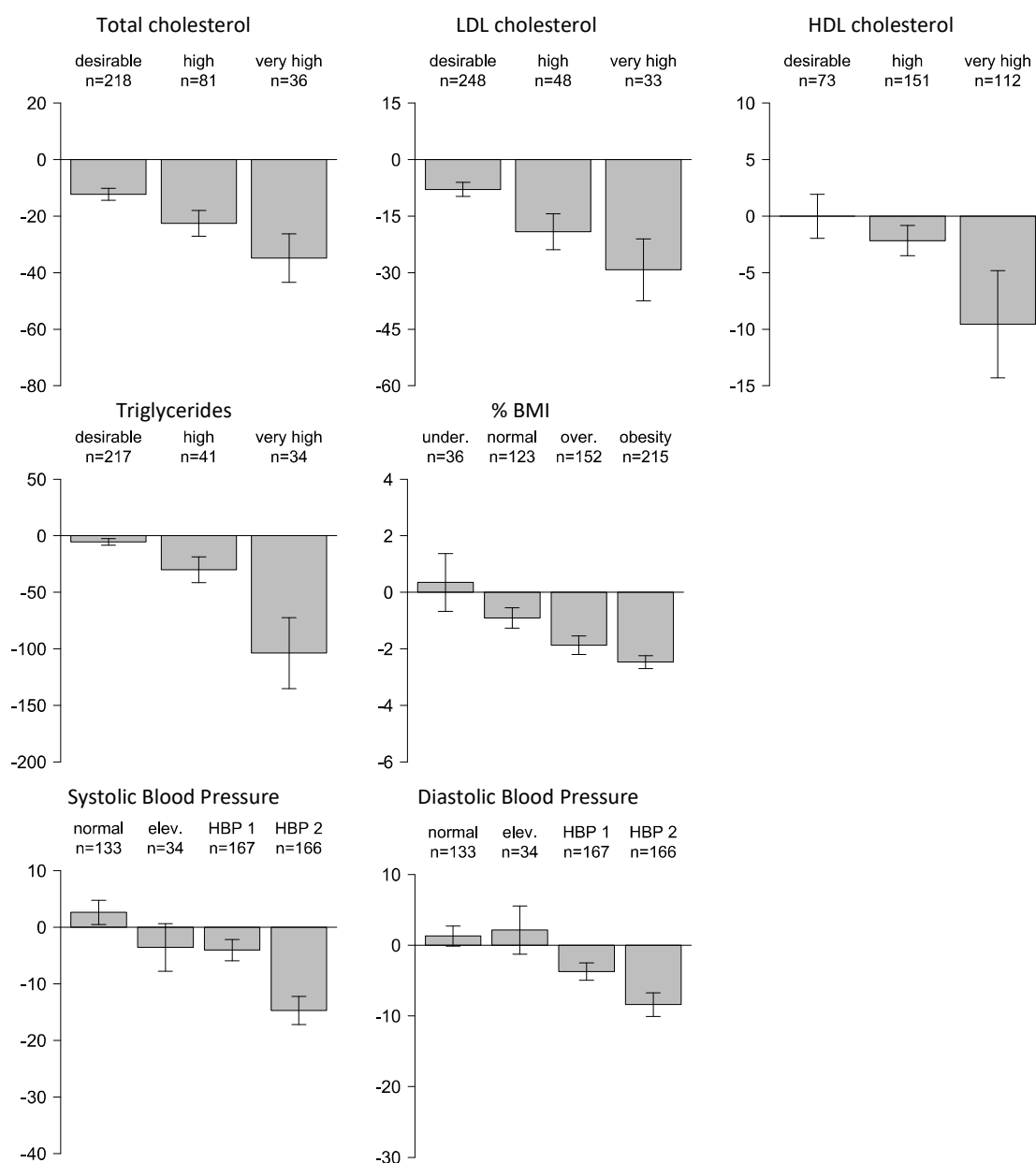


Figure 3. Mean changes in blood lipid levels, BMI, and blood pressure following a 6/7-day intervention, as recorded in EHR from lifestyle center A.

The mean differences were calculated as $\text{mean}(\text{parameter}_{\text{after}} - \text{parameter}_{\text{before}})$. Subgroups were defined as presented in Table 2. Error bars represent 95% confidence intervals, calculated using a t-distribution. The error bar was drawn for a 95% confidence interval with a t law. Units are as follows: Total cholesterol, LDL cholesterol, HDL cholesterol, Triglycerides: mg/ml; Systolic and Diastolic blood pressure: mmHg. BMI: Body Mass Index. Under.: underweight. Over.: overweight. Elev.: Elevated. HBP 1/2: High blood pressure stage 1 or 2.

3.3.2 Lifestyle Center B – Semi-structured Letters

3.3.2.1 Description of the Patients

We extracted 2,357 patient records from the semi-structured letters obtained from lifestyle center B. Notably, only 20 records (< 1% of the total) were missing age and/or sex data. To analyze age and sex, we excluded these records from the study.

The age distribution of both male and female patients followed a bell-curve pattern, peaking at 60-69 years of age (Figure 4.A). The patient population was predominantly female, with 64.1% [62.1-66.0] of patients being women (Figure 4.B). Furthermore, a significant majority (72.4% [70.5-74.1]) of patients were middle-aged, falling within the 50-69 year-old age range (Figure 4.C).

We also examined the reported diagnoses for four chronic diseases (Table 7). Interestingly, the SR for cancer diagnosis was relatively low at 4.1% [3.3-5.1]. In contrast, the SR for high blood pressure diagnosis was substantially higher at 35.8% [33.4-38.4], while 30.0% [27.7-32.4] of patients had other cardiovascular diseases. The SR for diabetes was 13.3% [11.8-14.9]. Unfortunately, we were unable to extract data on patient medication regimens from the letters collected (Table 4).

3.3.2.2 Lipidemia, BMI, and Blood Pressure Changes

We investigated the impact of lifestyle interventions on lipidemia, BMI, and blood pressure using the same approach as for the lifestyle center A dataset (Figure 5).

Regarding total cholesterol and LDL-cholesterol, our analysis revealed a significant decrease across all subgroups. Notably, the magnitude of this decrease was directly related to the initial level of lipidemia, with larger drops observed in subgroups with higher initial levels. In contrast, the results for HDL-cholesterol were more complex, with significant increases observed in patients with initially normal blood levels, significant decreases in those with initially high levels, and non-significant decreases in those with initially very high levels.

Regarding BMI, we observed significant decreases in the normal, overweight, and obese subgroups, while the underweight subgroup showed a non-significant increase.

Overall, our study demonstrated a significant decrease in blood lipid levels, blood pressure, and BMI for patients with high initial values, except for the complex pattern observed for HDL-cholesterol.

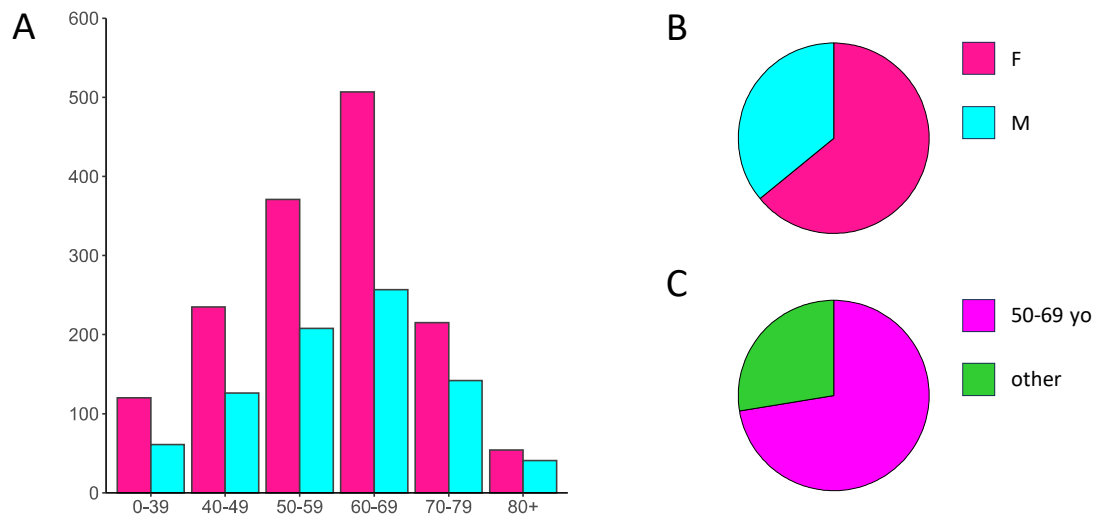


Figure 4. Demographic characteristics of patients from lifestyle center B, as extracted from semi-structured letters. (A) Age and sex distribution of patients. (B) Proportion of male and female patients. (C) Distribution of patients across main age groups.

Condition	% (n)	Standardized % [CI95%]	Average % in the other centers (n) [CI95%]	Test
Cancer	4.1% (96)	4.1% [3.3-5.1]	14.1% (184) [12.3-16.1]	p < 0.0001
Cardiovascular disease	29.4% (688)	30.0% [27.7-32.4]	13.1% (194) [11.5-14.9]	p < 0.0001
Diabetes	13.5% (316)	13.3% [11.8-14.9]	22.0% (326) [20.0-24.2]	p < 0.0001
High blood pressure	35.8% (836)	35.8% [33.4-38.4]	29.9% (443) [27.7-32.3]	p < 0.0005

Table 7. Prevalence of chronic diseases among patients at lifestyle center B, as extracted from semi-structured letters.

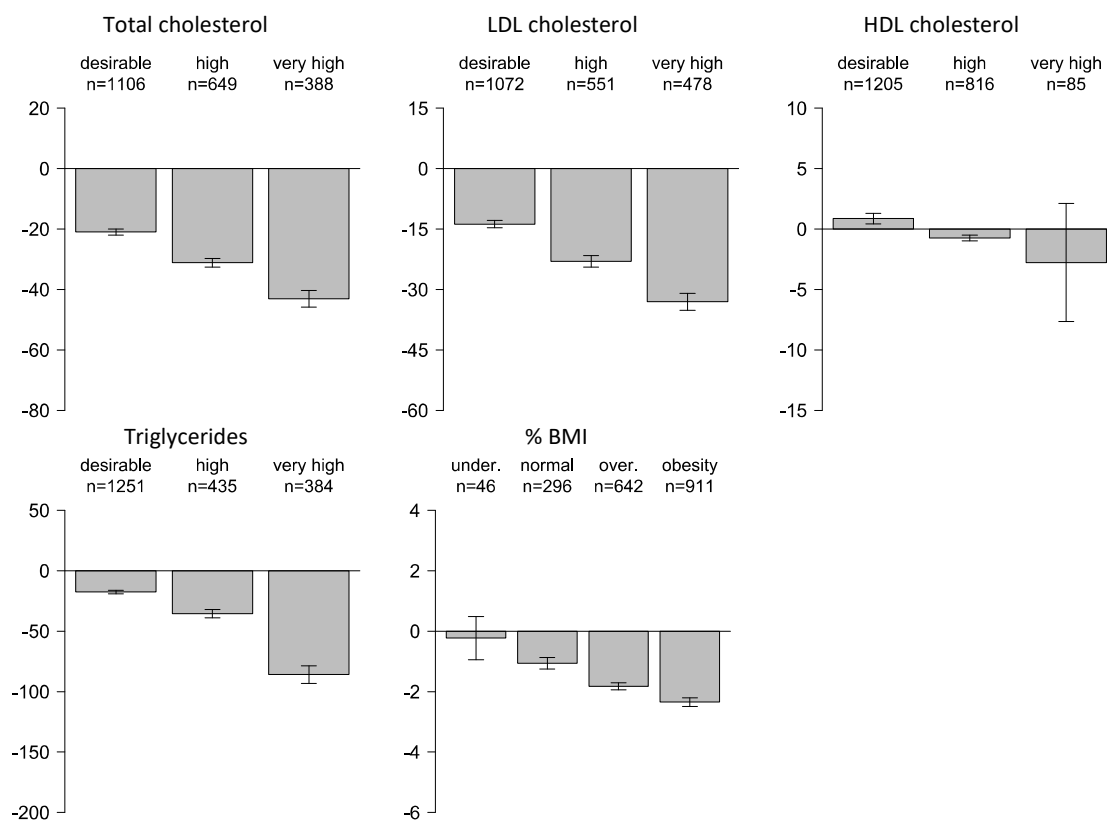


Figure 5. Mean changes in blood lipid levels and BMI following a 7-day intervention, as recorded in in semi-structured letters from lifestyle center B.

The mean differences were calculated as $\text{mean}(\text{parameter}_{\text{after}} - \text{parameter}_{\text{before}})$. Subgroups were defined as presented in Table 2. Error bars represent 95% confidence intervals, calculated using a t-distribution. The error bar was drawn for a 95% confidence interval with a t law. Units are as follows: Total cholesterol, LDL cholesterol, HDL cholesterol, Triglycerides: mg/ml. BMI: Body Mass Index. Under.: underweight. Over.: overweight.

3.3.3 Lifestyle Center C – Research-purposed Database

3.3.3.1 Description of the Patients

We extracted 173 patient records from the research database obtained from lifestyle center C. Notably, there were no missing data points for age or sex. As expected, our analysis revealed a majority of female patients, comprising 65.3% [58.0-72.0] of the sample (Figure 6.B). Most patients were middle-aged, with 77.5% [70.7-83.0] falling between 50 and 69 years old (Figure 6.C).

We examined the prevalence of three chronic diseases: cardiovascular disease (excluding high blood pressure), high blood pressure, and diabetes (Table 8). Interestingly, the data from this center did not include cancer patients, unlike the other three centers. This is likely due to the center's policy of not recommending this lifestyle intervention to cancer patients. The prevalence rates for these diseases were substantial, with 36.9% [27.7-50.8] of patients having diabetes, 50.6% [40.0-65.7] having high blood pressure, and 20.4% [13.8-31.9] having cardiovascular disease other than high blood pressure.

We also analyzed the number of patients receiving pharmaceutical treatment for these three chronic illnesses at the beginning of their stay (Table 9). The results showed that 28.4% [21.2-40.3] of patients were treated with antidiabetics, 50.5% [40.4-65.2] with antihypertensive medication, and 17.9% [12.5-28.3] with hypolipidemic medication.

3.3.3.2 Lipidemia, BMI, and Blood Pressure Changes

We investigated the effects of the lifestyle intervention on lipidemia, BMI, and blood pressure (Figure 6).

Regarding plasma cholesterol, LDL cholesterol, and HDL-cholesterol, our analysis revealed a significant decrease only for patients with high or very high levels at the beginning of the intervention. In contrast, there was a non-significant increase in mean blood levels for patients with normal levels. Although the mean drop was slightly larger for patients with very high levels compared to high levels, this difference was not statistically significant.

Regarding triglycerides, we observed a significant small increase in mean blood levels for patients with normal levels at the start of the program. There was a small non-significant mean decrease for patients with high levels and a significant decrease for those with very high levels.

Regarding BMI, we found a significant decrease in patients who were overweight or obese, while the decrease was non-significant for underweight and normal-weight patients. It is worth noting that the small sample size of underweight (n=4) and normal-weight (n=20) patients limited the statistical power of these findings.

Regarding blood pressure, our analysis showed that patients with normal or elevated blood pressure levels did not experience significant changes in their average values, whereas those with stage 1 or stage 2 high blood pressure had a significant decrease in both systolic and diastolic blood pressure values.

Overall, our study demonstrated a significant decrease in blood lipid levels, blood pressure, and body mass index for patients who started the intervention with high values, with a more complex pattern observed for triglycerides.

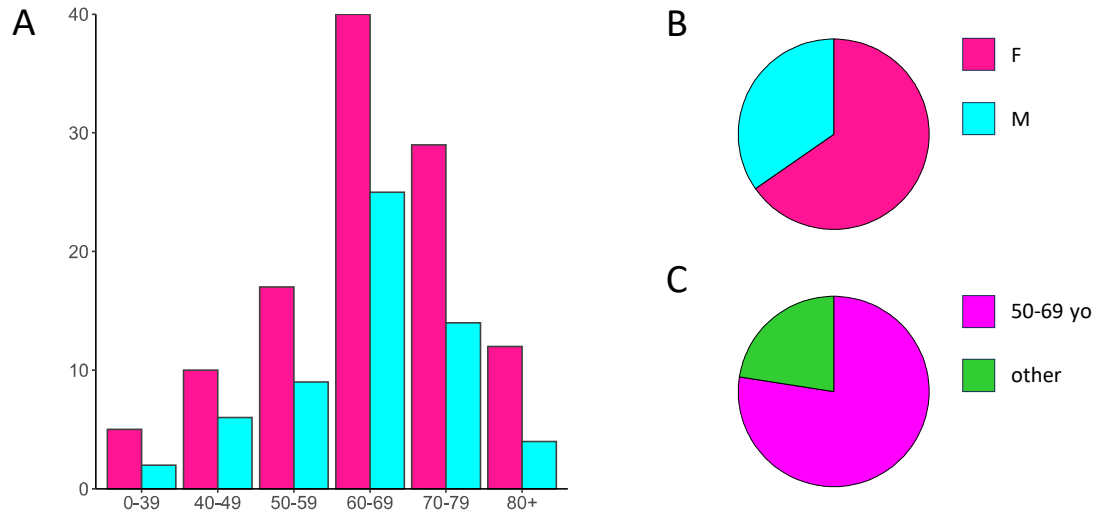


Figure 6. Demographic characteristics of patients from lifestyle center C, as extracted from semi-structured letters. (A) Age and sex distribution of patients. (B) Proportion of male and female patients. (C) Distribution of patients across main age groups.

Condition	% (n)	Standardized % [CI95%]	Average % in the other centers (n) [CI95%]	Test
Cardiovascular disease	23.7% (41)	20.4% [13.7-31.9]	23.6% (859) [22.2-25.0]	ns
Diabetes	41.6% (72)	36.9% [27.3-50.8]	15.7% (573) [14.6-16.9]	p < 0.0001
High blood pressure	56.6% (98)	50.6% [39.7-65.5]	32.7% (1193) [31.2-34.3]	p < 0.0001

Table 8. Prevalence of chronic diseases among patients at lifestyle center C, as extracted from a research-purposed database.

Treatment for chronic illness	% (n)	Standardized % [CI95%]	Average % in the other centers (n) [CI95%]	Test
Antidiabetic	33.5% (58)	20.1% [13.6-31.9]	18.5% (242) [16.5-20.7]	p < 0.005
Antihypertensive	60.7% (105)	36.6% [27.4-50.9]	38.3% (500) [35.7-40.9]	p < 0.005
Hypolipidemic	23.1% (40)	50.2% [39.7-65.5]	14.3% (187) [12.5-16.3]	ns

Table 9. Prevalence of chronic disease medications among patients at lifestyle center C, as extracted from a research-purposed database.

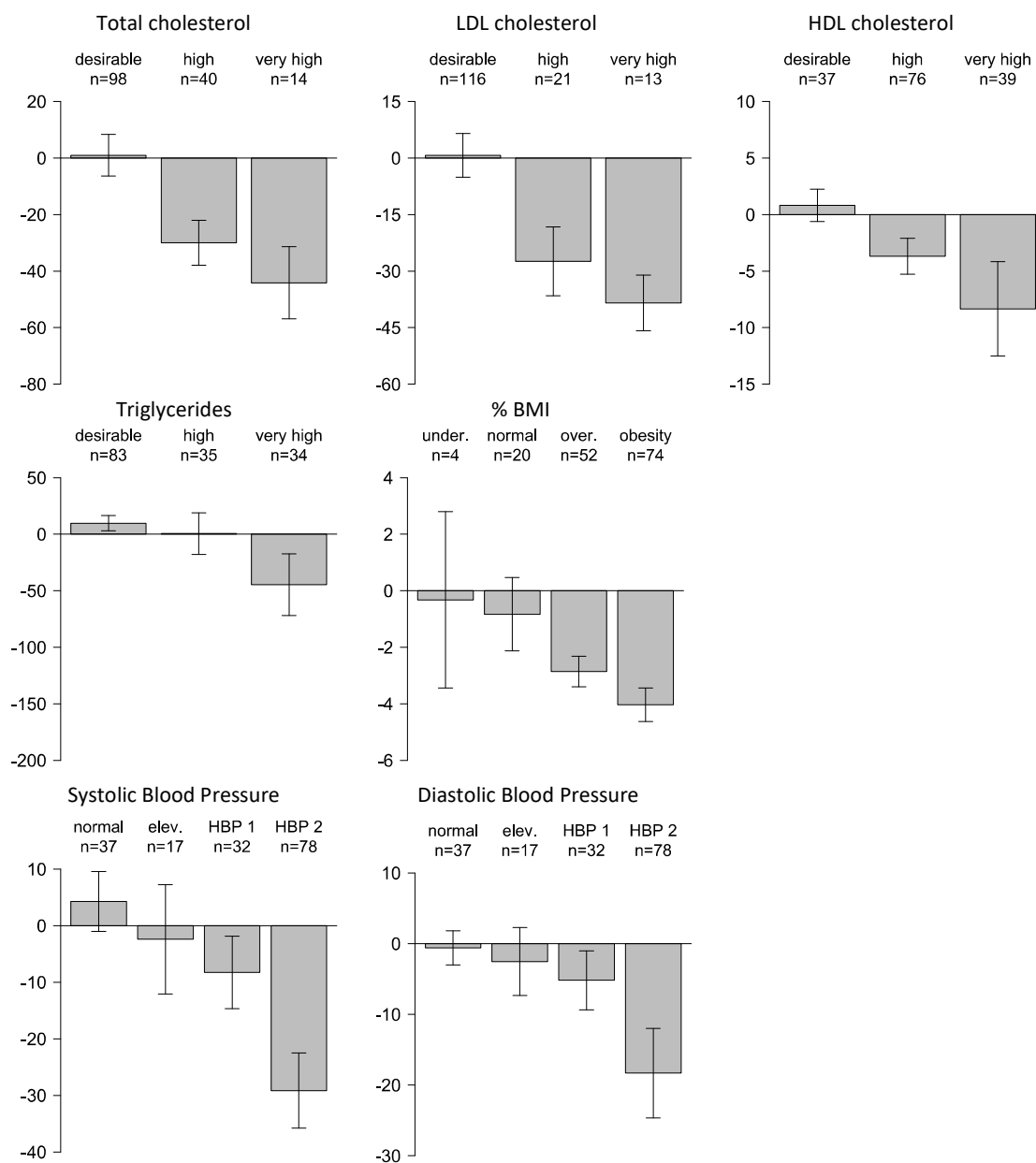


Figure 7. Mean changes in blood lipid levels and BMI following a 14-day intervention, as recorded in the research-purposed database from lifestyle center C.

The mean differences were calculated as $\text{mean}(\text{parameter}_{\text{after}} - \text{parameter}_{\text{before}})$. Subgroups were defined as presented in Table 2. Error bars represent 95% confidence intervals, calculated using a t-distribution. The error bar was drawn for a 95% confidence interval with a t law. Units are as follows: Total cholesterol, LDL cholesterol, HDL cholesterol, Triglycerides: mg/ml; Systolic and Diastolic blood pressure: mmHg. BMI: Body Mass Index. Under.: underweight. Over.: overweight. Elev.: Elevated. HBP 1/2: High blood pressure stage 1 or 2.

3.3.4 Lifestyle Center D – Research-purposed Database

3.3.4.1 Description of the Patients

We extracted 199 patient records from the research database obtained from lifestyle center D, with no missing data for age or sex. The patient population was predominantly female (70.9% [64.2-76.7]; Figure 7.B) and middle-aged, with 72.4% [65.8-78.1] of patients falling between 50 and 69 years old (Figure 7.C).

We examined the prevalence of four chronic diseases: cancer, cardiovascular disease (excluding high blood pressure), high blood pressure, and diabetes (Table 10). Notably, the prevalence of diabetes was high, at 40.0% [31.6-51.1]. The prevalence of cancer, high blood pressure, and other cardiovascular diseases was 31.6% [23.8-42.2], 32.3% [24.9-42.2], and 19.3% [13.8-27.5], respectively.

We also investigated the number of patients receiving treatment for three major chronic illnesses at the beginning of their stay (Table 11). The treatment rates were 18.7% [13.2-26.9] for diabetes, 32.7% [25.4-42.7] for high blood pressure, and 11.1% [7.0-18.1] for dyslipidemia.

3.3.4.2 Lipidemia, BMI, and Blood Pressure Changes

We analyzed the effects of the lifestyle intervention on lipidemia, BMI, and blood pressure (Figure 7).

Regarding plasma cholesterol and LDL-cholesterol, for patients with initially normal levels, there was a non-significant increase in mean blood levels. In contrast, patients with high or very high initial levels showed a significant decrease. Triglyceride levels increased significantly in patients with normal initial values but decreased significantly in those with high or very high initial values. HDL cholesterol levels decreased non-significantly in patients with normal or high initial values but showed a significant drop in those with very high initial values.

Regarding BMI, a significant reduction was observed in patients with normal, overweight, or obese BMI, whereas no such decrease was noted in underweight individuals.

Regarding blood pressure, patients having normal levels experienced a small but significant increase, while those with elevated levels showed a non-significant decrease. Patients with stage 1 or stage 2 hypertension had a significant decrease in blood pressure.

Overall, our study demonstrated significant improvements in blood lipid levels, blood pressure, and BMI among patients with high initial values, while patients with normal initial values showed no significant changes, except for an unexpected increase in triglyceride levels and blood pressure.

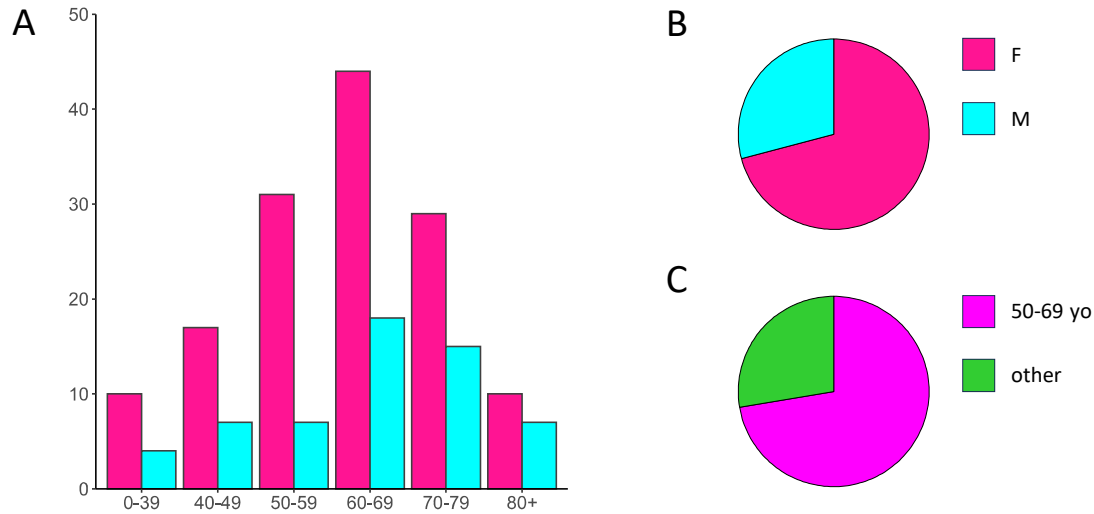


Figure 8. Demographic characteristics of patients from lifestyle center D, as extracted from semi-structured letters. (A) Age and sex distribution of patients. (B) Proportion of male and female patients. (C) Distribution of patients across main age groups.

Condition	% (n)	Standardized % [CI95%]	Average % in the other centers (n) [CI95%]	Test
Cancer	30.7% (61)	31.6% [23.8-42.2]	6.3% (218) [5.6-7.2]	p < 0.0001
Cardiovascular disease	22.1% (44)	19.3% [13.8-27.5]	23.6% (855) [22.3-25.0]	ns
Diabetes	43.2% (86)	40.1% [31.6-51.1]	15.4% (557) [14.2-16.6]	p < 0.0001
High blood pressure	37.2% (74)	32.3% [24.9-42.2]	33.6% (1215) [32.1-35.1]	ns

Table 10. Prevalence of chronic diseases among patients at lifestyle center D, as extracted from a research-purposed database.

Treatment for chronic illness	% (n)	Standardized % [CI95%]	Average % in the other centers (n) [CI95%]	Test
Antidiabetic	21.1% (42)	18.7% [13.2-26.9]	19.8% (253) [17.7-22.0]	ns
Antihypertensive	37.7% (75)	32.7% [25.4-42.7]	40.7% (522) [38.1-43.4]	p < 0.05
Hypolipidemic	13.1% (26)	11.1% [7.0-18.1]	15.3% (196) [13.4-17.4]	ns

Table 11. Prevalence of chronic disease medications among patients at lifestyle center D, as extracted from a research-purposed database.

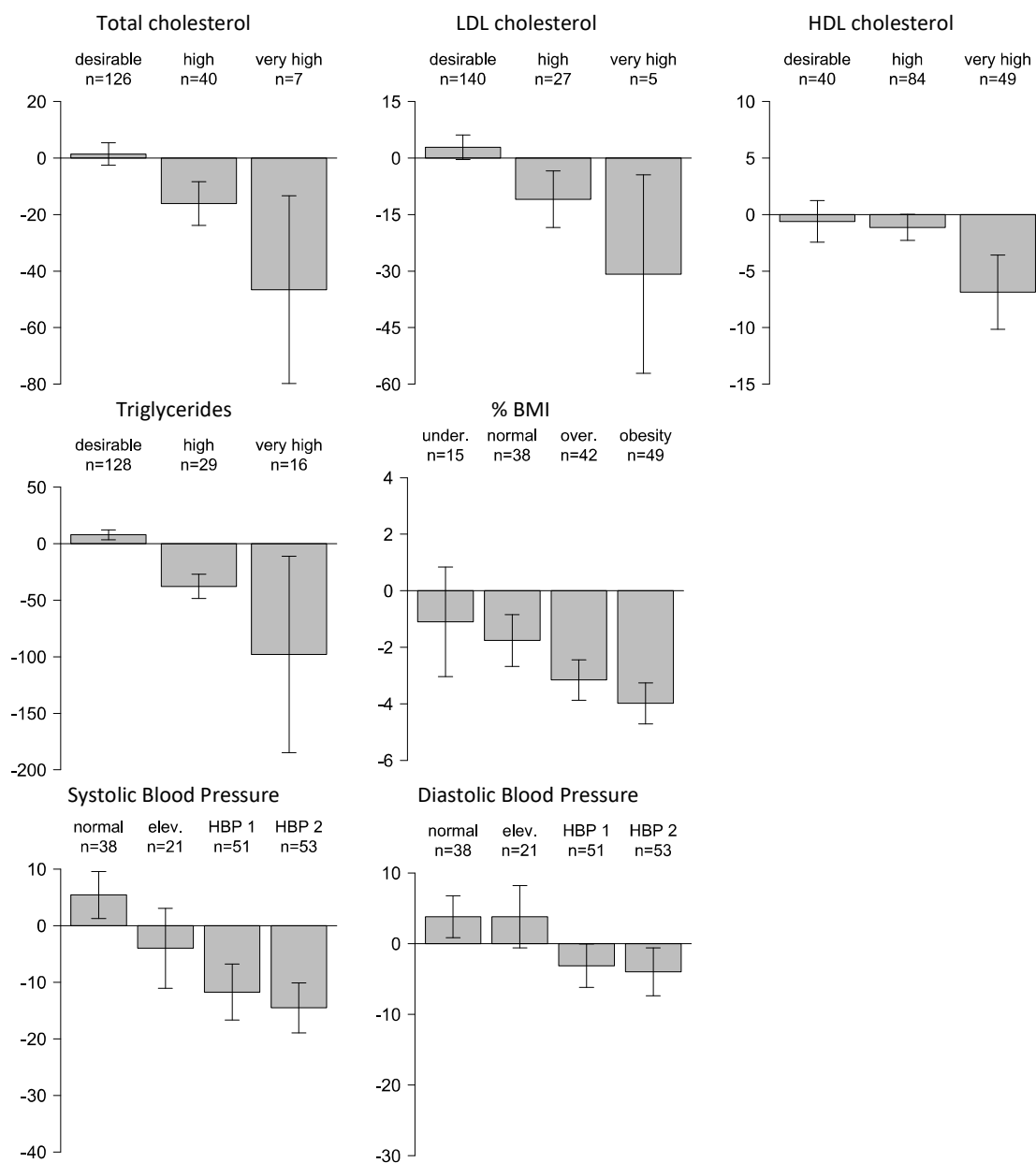


Figure 9. Mean changes in blood lipid levels and BMI following a 14-day intervention, as recorded in the research-purposed database from lifestyle center D.

The mean differences were calculated as $\text{mean}(\text{parameter}_{\text{after}} - \text{parameter}_{\text{before}})$. Subgroups were defined as presented in Table 2. Error bars represent 95% confidence intervals, calculated using a t-distribution. The error bar was drawn for a 95% confidence interval with a t law. Units are as follows: Total cholesterol, LDL cholesterol, HDL cholesterol, Triglycerides: mg/ml; Systolic and Diastolic blood pressure: mmHg. BMI: Body Mass Index. Under.: underweight. Over.: overweight. Elev.: Elevated. HBP 1/2: High blood pressure stage 1 or 2.

3.3.5 Comparison between Datasets

3.3.5.1 Comparison between Populations

A comparative analysis of the population structures across the four datasets was conducted.

The proportions of patients with reported diagnoses of four chronic diseases (cancer, diabetes, high blood pressure, and cardiovascular disease other than high blood pressure) were compared between each center and the average of the other centers. Due to the absence of cancer reports in center C's records, no test was performed for this condition. The results, presented in Tables 5, 7, 8, and 10, revealed that the differences in reported diagnosis proportions between centers were predominantly statistically significant. Out of 15 tests, only 4 did not yield statistically significant differences ($p > 0.05$), indicating that the proportion in an individual dataset was not proportionally different from the average of the other datasets in only 4 cases. These exceptions were observed for diabetes diagnosis in center A's records (Table 5), cardiovascular disease in center C's records (Table 8), and cardiovascular disease and high blood pressure in center D's records (Table 10). Thus, the comparison between centers suggested that the proportions of reported diagnoses for each center were mostly statistically distinct from their average across all centers.

A similar analysis was performed for the proportions of patients reported to be receiving treatment for three chronic diseases (high blood pressure, diabetes, and dyslipidemia) in the three centers with available treatment reports (center B's reports were not reusable, as mentioned earlier). The results are presented in Tables 6, 9, and 11 for centers A, C, and D, respectively. Most tests yielded non-significant results, indicating similar proportions of reported treatments across most datasets. Out of 9 tests, only 3 were statistically significant, specifically: the proportion of patients with reported treatment for diabetes and high blood pressure in center C's records (Table 9), and the proportion of patients with reported treatment for high blood pressure in center D's records (Table 11).

Overall, our population comparison revealed dissimilarities between the four datasets regarding reported diagnoses of chronic diseases, but similarities between centers A, C, and D regarding reported treatments for these diseases.

3.3.5.2 Comparison between intervention impacts

We examined the consistency of changes in BMI, blood pressure, and lipidemia associated with the ITLC intervention.

Table 12 presents the statistically significant changes observed in three datasets or more, with a 2-color set used to represent consistency in the changes. Our analysis revealed that five parameters consistently decreased in specific subgroups across all datasets: total cholesterol and LDL-cholesterol levels in patients with initial high or very high values, BMI in overweight and obese patients, and both systolic and diastolic blood pressure in patients with initial blood pressure consistent with stage 1 or stage 2 high blood pressure. Additionally, three parameters consistently decreased in specific subgroups in 3 out of 4 datasets: HDL cholesterol in patients with initial high or very high levels, triglycerides in patients with initial high blood levels, and BMI in patients with initial normal weight. Notably, systolic blood pressure decreased in 2 out of 3 datasets for patients with initially normal blood pressure, although a significant increase was observed in the third dataset.

Parameter	Subgroup	Mean change	min max	Datasets with significant results
Total Cholesterol	High & Very High	-33.5	-46.6 -16.1	4/4
LDL-Cholesterol	High & Very High	-26.5	-38.5 -10.9	4/4
HDL-Cholesterol	High & Very High	-4.4	-9.6 -0.7	3/4
Triglycerides	Very High	-83.3	-103.8 -44.6	4/4
Triglycerides	High	-25.7	-37.7 +0.5	3/4
BMI	Overweight & Obese	-2.8%	-4.0 -1.8	4/4
BMI	Normal weight	-1.1%	-1.8 -0.8	3/4
SBP	HBP stage 1 & 2	-13.7	-29.2 -4.0	4/4
DBP	HBP stage 1 & 2	-7.1	-18.3 -3.1	3/3

Legend
All datasets have significant results
Three datasets out of four have significant results

Table 12. Summary of results with significant changes observed in three or four datasets for lipidemia, blood pressure, and BMI.

Mean changes are expressed in mg/mL for lipid concentrations and mmHg for blood pressure measurements.

Table 13 presents the statistically significant changes observed in one or two datasets. Furthermore, we found that total cholesterol, LDL-cholesterol, and HDL-cholesterol levels decreased in 2 out of 4 datasets, with no significant changes in the remaining two datasets. Interestingly, triglycerides increased in patients with initial normal blood levels in 2 out of 4 datasets, while an increase was also observed in the same subgroups in the two other datasets, highlighting an inconsistency in this result.

Parameter	Subgroup	Mean change	min max	Datasets with significant results
Total Cholesterol	Desirable	-7.7	-21.0 +1.4	2/4
LDL-Cholesterol	Desirable	-4.5	-13.8 +2.9	2/4
HDL-Cholesterol	Desirable	+0.3	-0.6 +0.9	1/4
Triglycerides*	Desirable	-11.6	-17.6 -5.5	2/4
Triglycerides*	Desirable	+8.8	+7.9 +9.7	2/4
SBP	Normal	+4.1	+2.6 +5.4	2/3
DBP	Normal	+1.5	-0.6 +3.8	1/3

Legend
Two datasets with significant results
One dataset with significant results

Table 13. Results with significant changes observed in one or two datasets for lipidemia, blood pressure, and BMI.

Mean changes are expressed in mg/mL for lipid concentrations and mmHg for blood pressure measurements. *: Triglyceride results are presented separately to highlight the heterogeneous nature of the significant changes detected.

4 Discussion

This study aimed to explore the feasibility of repurposing data from lifestyle centers to examine the consistency of metabolic changes induced by intensive lifestyle interventions. Following initial contact with eight centers, we successfully obtained data from four of them. Notably, the received data were heterogeneous in format, comprising semi-structured letters, EHR, and research-specific databases.

Through this study, we were able to delineate the essential steps and obstacles involved in feature extraction from these diverse data sources. Notably, our findings revealed a surprising similarity in the time required to extract relevant data from semi-structured letters and EHR, compared to research-purposed databases, when controlling for the volume of records. This highlights the potential efficiency of leveraging data from non-traditional sources, such as clinical records, for research purposes.

Our analysis revealed a consistent pattern of beneficial effects of residential ITLC on patients across all four datasets. Specifically, we observed significant reductions in total cholesterol, LDL-cholesterol, triglycerides, BMI, SBP, and DBP among

subgroups of patients. Notably, these reductions were consistent across datasets, with mean decreases of 33.5 mg/dL in total cholesterol, 26.5 mg/dL in LDL-cholesterol, 83.3 mg/dL in triglycerides, 2.8% in BMI, 13.7 mmHg in SBP, and 7.1 mmHg in DBP among patients with corresponding high or elevated levels. However, we also identified inconsistencies between datasets, with some datasets showing increases while others showed decreases in certain parameters for the same subgroups. For instance, among patients with desirable LDL-cholesterol levels, two datasets exhibited significant decreases, whereas the other two showed non-significant increases.

Our findings indicate that the systematic reuse of data from residential ITLC programs in lifestyle centers is a viable approach. The effort required to facilitate data reuse is manageable, and the results obtained are largely consistent across different lifestyle centers, suggesting that data reuse can be a valuable strategy for extracting meaningful insights from these programs.

Our study revealed that the time required for data preprocessing was comparable for semi-structured letters and structured data, such as EHRs and research-purposed databases. This finding challenges the common assumption that structured data is inherently easier to work with than unstructured data. In reality, preprocessing structured data can be a time-consuming task, often requiring data cleaning, transformation, and integration, which can be complex and labor-intensive [112]. While some studies have suggested that preprocessing structured data may require similar time and effort as preprocessing unstructured data [112,130], our work highlights the need for further investigation into this question. In the context of lifestyle center data, our findings raise an important question: could data preprocessing be simplified by adopting data storage practices that are designed to facilitate data reuse for research purposes?

Our study revealed that the labor-intensive nature of EHR data preprocessing was largely attributed to the limitations of the EHR software, which only allowed file-by-file extraction of CCDA documents. This data retrieval step accounted for nearly half of the preprocessing time for EHR data. In contrast, having a complete database of all patients with lab results, clinical data, and other relevant information would have significantly streamlined the process. Moreover, the research-purposed structured databases we utilized lacked international naming codes for lab results and diagnostics, necessitating time-consuming recoding efforts. These constraints highlight the potential for structured data to be made more preprocess-friendly. The feasibility of data reuse is thus heavily influenced by the data storing practices and tools employed in lifestyle centers, a notion supported by existing literature [83,131,132].

Our findings also demonstrate the viability of reusing semi-structured data, which had initially been a concern due to the potential labor intensity involved. However, our study revealed that the effort required to reuse semi-structured data is comparable to that of structured data. Moreover, semi-structured data reuse offers a convenient means of accessing large datasets. Notably, the structured research-purposed datasets we obtained were significantly smaller than the routinely produced datasets (EHR and letters) we received. This disparity highlights the labor-intensive nature of data cleaning and structuring for research purposes, which may not be feasible for all lifestyle centers to undertake before data sharing.

A notable trend emerged across all our datasets, with approximately two-thirds of the patient population being female and two-thirds falling within the 50-69 age range. This demographic profile is consistent with findings from studies on the CHIP program, a non-residential ITLC [106,133,134], where participants voluntarily enrolled. The similarity in demographics between our datasets and the CHIP program suggests that individuals who opt for ITLC interventions tend to share these characteristics.

While the reported diagnosis proportions varied across lifestyle centers, we found more consistency in the proportions of patients with medication for common chronic diseases. This disparity can be attributed to the known variability in missing data rates depending on the type of data and its intended use [135–137]. Although we couldn't precisely quantify the missing data rate in this study, our analysis of diagnostic and medication data, combined with the authors' expertise, suggests that medication data were more complete and less prone to missing values. This is likely due to the meticulous collection of medication data in lifestyle centers, which is crucial for supporting patients and adjusting medication regimens during their stay [105]. This careful recording of medication data may explain the similarities observed in medication proportions across datasets.

Our findings on the hypolipidemic effects of residential ITLC programs on hyperlipidemic patients align with existing literature [11,97,138]. The low-fat diet employed in these programs likely contributed to the improvements in patients' lipid blood levels [139,140]. In contrast, we did not observe consistent improvements in lipidemia among normolipidemic patients, which is unexpected given that the non-residential CHIP ITLC program has been shown to improve cholesterol levels in normolipidemic patients [106,133,134]. This discrepancy may be attributed to the relatively short duration of the intervention (one to two weeks) compared to the 4-week program described in the literature. Notably, we did observe significant reductions in cholesterol and triglycerides in normolipidemic patients in two out of four datasets, suggesting that differences in initial population characteristics may have contributed to the inconsistent results for normolipidemic patients.

To identify patient subgroups for our analysis of ITLC effects, we relied on international clinical recommendations [128,129]. However, given the limited reporting of residential ITLC interventions in the literature, it is likely that these recommendations do not optimally capture the specific population and effects of such interventions. A more effective approach may be to employ a data-driven method to identify patient subgroups based on the effects of ITLC. This would enable a more accurate characterization of the program's effects, including the amplitude of efficiency, predictive factors, and potential adverse effects. Furthermore, such a data-driven approach could help explain the inconsistencies observed in normolipidemic patients' results by identifying factors that predict variations in blood lipid levels. By leveraging data-driven insights, we may uncover more nuanced and relevant subgroupings that better reflect the complexities of ITLC interventions.

Notably, our study revealed significant improvements in BMI. The multifaceted approach, comprising a whole food, plant-based, low-fat diet, moderate exercise, and counseling, is likely to have contributed to this outcome, as supported by existing literature on ITLC programs [97,106]. However, it is important to acknowledge that some medical providers in the lifestyle centers we studied occasionally recommended caloric restrictions and short-term fasting (2-3 days) for overweight and obese patients seeking weight loss. Although these practices are not standard

protocols and are only employed in specific circumstances, their impact on our results is likely to be minimal. A more in-depth comparative analysis of the treatment approaches used in these centers and their effects on BMI would be necessary to fully elucidate the influence of residential ITLC programs on weight management.

Our findings in hypertensive patients align with existing literature, which has consistently demonstrated the efficacy of residential ITLC programs in improving blood pressure [105,106]. The combination of moderate exercise [98] and a low-fat plant-based diet [141] may contribute to the positive effects of ITLC on hypertension. In contrast, our results show no significant blood pressure reduction in normotensive patients, with a surprising increase in systolic blood pressure (SBP) in two out of three datasets and diastolic blood pressure (DBP) in one out of three datasets. This outcome is unexpected, as it diverges from the literature on ITLC [105,106] and our findings in hypertensive patients. The magnitude of this increase (+4.1 mmHg SBP and +1.5 mmHg DBP) is relatively small compared to the significant hypotensive effects observed in hypertensive patients (-13.7 mmHg SBP and -7.1 mmHg DBP). Further investigation into the management and measurement of blood pressure in normotensive patients is necessary to provide a comprehensive explanation for this phenomenon. Notably, the lifestyle centers we studied often taper blood pressure medication in normotensive patients, but our medication records lacked sufficient detail to quantify this practice or its potential effect.

One of the significant strengths of our study lies in its utilization of routinely generated data, which enables the examination of real-world data and minimizes selection bias [55]. However, a notable limitation is the absence of control groups for comparison, which may introduce confounding bias that could partially explain our findings [142]. In the context of our study, relevant confounders include the participants' motivation to improve their health and the lack of data from patients who do not complete their stay in the lifestyle centers. While the volunteer nature of the program and participant attrition may contribute to our results, we believe that the consistency and magnitude of our findings cannot be fully attributed to these factors. In our experience, only a small proportion of patients typically do not complete their stay in lifestyle centers, and departures are usually due to non-compliance with intervention requirements rather than health status. Nevertheless, further studies are needed to quantify this confounder. Importantly, our results are supported by smaller randomized controlled trials on residential ITLC [11,97,138] and larger studies on non-residential ITLC [106,133,134] which confirm the robust nature of our results.

5 Conclusion

In this study, we successfully demonstrated the feasibility of leveraging data reuse techniques to investigate the effectiveness of residential ITLC programs. Our findings consistently indicate that these programs can lead to significant improvements in lipidemia among hyperlipidemic patients, BMI among overweight and obese patients, and blood pressure among hypertensive patients. To further elucidate the scope and underlying mechanisms of these effects, more in-depth and comparative studies are warranted.

Discussion en Français

Cette étude visait à explorer la faisabilité de la réutilisation des données provenant des *lifestyle centers* pour examiner la constance des changements métaboliques induits par des ITLC. Après un premier contact avec huit centres, nous avons réussi à obtenir des données de quatre d'entre eux. Les données reçues étaient hétérogènes en format, comprenant des lettres semi-structurées, des dossiers médicaux électroniques (EHR), et des bases de données spécifiques à la recherche.

À travers cette étude, nous avons pu délimiter les étapes essentielles et les obstacles impliqués dans l'extraction des caractéristiques à partir de ces sources de données diverses. Notre étude a révélé une similarité surprenante dans le temps requis pour la *data preprocessing* des lettres semi-structurées et des EHR, comparé aux bases de données à vocation de recherche, en tenant compte du volume des données travaillées. Cela permet de mettre en avant l'efficacité potentielle, pour ce domaine, de la réutilisation de données à des fins de recherche.

Notre analyse a révélé une certaine constance des effets bénéfiques sur les patients des ITLC résidentiels à travers les quatre ensembles de données obtenus. Plus précisément, nous avons observé des réductions significatives du cholestérol total, du cholestérol LDL, des triglycérides, de l'IMC, de la SBP et de la DBP parmi les sous-groupes de patients. Notamment, ces réductions étaient similaires à travers les ensembles de données, avec des diminutions moyennes de 33,5 mg/dL pour le cholestérol total, de 26,5 mg/dL pour le cholestérol LDL, de 83,3 mg/dL pour les triglycérides, de 2,8 % pour l'IMC, de 13,7 mmHg pour la SBP, et de 7,1 mmHg pour la DBP parmi les patients présentant des niveaux élevés ou très élevés. Cependant, nous avons également identifié des incohérences entre les ensembles de données, certains montrant des augmentations tandis que d'autres montraient des diminutions pour les mêmes sous-groupes dans certains paramètres. Par exemple, parmi les patients présentant des niveaux de cholestérol LDL souhaitables, deux ensembles de données ont montré des diminutions significatives, tandis que les deux autres ont montré des augmentations non significatives.

Nos résultats indiquent que la réutilisation systématique des données des programmes ITLC résidentiels dans les *lifestyle centers* est une approche viable. L'effort nécessaire pour faciliter la réutilisation des données est gérable, et les résultats obtenus sont largement cohérents entre différents *lifestyle centers*, suggérant que la réutilisation des données peut être une stratégie précieuse pour améliorer notre connaissance de ces programmes.

Notre étude a révélé que le temps nécessaire pour la *data preprocessing* était comparable pour les lettres semi-structurées et les données structurées (les EHR et les bases de données à vocation de recherche). Cette constatation remet en question l'hypothèse commune selon laquelle les données structurées sont intrinsèquement plus faciles à manipuler que les données non structurées. En réalité, le *data preprocessing* structurées peut être une tâche longue, nécessitant souvent un nettoyage, une transformation et une intégration des données, qui peuvent être complexes et exigeants en temps et en main-d'œuvre [112]. Alors que certaines études ont suggéré que la *data preprocessing* de données structurées pourrait nécessiter un temps et un effort similaires à celui des données non structurées

[112,130], notre travail souligne la nécessité de recherches plus approfondies sur cette question. Dans le contexte des données des *lifestyle centers*, nos résultats soulèvent une question importante : la data preprocessing pourrait-elle être simplifiée en adoptant des pratiques de stockage des données conçues pour faciliter la réutilisation des données à des fins de recherche ?

Notre étude a révélé que la grande quantité de travail que représente la *data preprocessing* des EHR était largement attribuable aux limitations du logiciel utilisé par le *lifestyle center*. Ce logiciel permettait seulement une extraction des documents CCDA patient par patient. Cette étape de récupération des données (*data retrieval*) représentait près de la moitié du temps de *data preprocessing* des EHR. En revanche, disposer d'une base de données complète de tous les patients avec les résultats de laboratoire, les données cliniques et d'autres informations pertinentes aurait considérablement simplifié le processus. De plus, les bases de données structurées à vocation de recherche que nous avons utilisées manquaient de codes de dénomination internationaux pour les résultats de laboratoire et les diagnostics, nécessitant des efforts de recodage chronophages. Ces contraintes soulignent le potentiel d'amélioration des données structurées pour qu'elles soient mieux préparées pour la *data preprocessing*. La faisabilité de la réutilisation des données est donc fortement influencée par les pratiques et les outils de stockage des données utilisés dans les *lifestyle centers*, notion soutenue par la littérature [83,131,132].

Nos résultats démontrent également la viabilité de la réutilisation des données semi-structurées, alors même que cela nous avait initialement semblé difficile en raison de la quantité de travail envisagée. Cependant, notre étude a révélé que l'effort requis pour réutiliser des données semi-structurées est comparable quantitativement à celui nécessaire pour la réutilisation de données structurées. De plus, la réutilisation des données semi-structurées offre un moyen pratique d'accéder à de grands ensembles de données. Notamment, les bases de données structurées à vocation de recherche que nous avons obtenues étaient de bien plus petite taille que les ensembles de données collectés en routine (EHR et lettres). La petite taille des bases de données que nous avons reçues provient vraisemblablement de la grande quantité de travail nécessaire au nettoyage et à la structuration des données à des fins de recherche, travail qui ne peut vraisemblablement pas être réalisé sur de grandes quantités de données par tous les *lifestyle centers*, du fait de leurs contraintes en termes de budget et de ressources humaines.

Une tendance notable est apparue dans tous nos ensembles de données : environ deux tiers de la population de patients étaient des femmes et environ deux tiers se situaient dans la tranche d'âge de 50 à 69 ans. Ce profil démographique est cohérent avec les résultats des études sur le programme CHIP [106,133,134], un ITLC ambulatoire dans lequel les personnes incluses participent sur la base du volontariat. La similarité démographique entre nos ensembles de données et le programme CHIP suggère que les individus qui optent pour des interventions ITLC tendent à partager ces caractéristiques.

Bien que les proportions de diagnostics rapportées aient varié d'un *lifestyle center* à l'autre, nous avons constaté de plus grandes similitudes en ce qui concerne les proportions de patients traités. Cette disparité peut être attribuée à la variabilité connue des taux de données manquantes en fonction du type de données et de leur utilisation prévue [135–137]. Bien que nous n'ayons pas pu quantifier précisément le taux de données manquantes dans cette étude, notre analyse des données de diagnostic et de traitement médicamenteux, combinée à l'expertise des auteurs,

suggère que les données de traitement étaient plus complètes et moins sujettes à des carences de données. En effet, la collecte des données de traitement dans les *lifestyle centers* est généralement méticuleuse, dans la mesure où cela est crucial pour accompagner les patients et ajuster les traitements médicamenteux pendant leur séjour. Cette consigne minutieuse des données de traitement médicamenteux peut expliquer les similarités observées dans les taux de patients traités dans les trois ensembles de données concernés.

Nos résultats sur les effets hypolipémiants des programmes ITLC résidentiels sur les patients hyperlipémiés sont en accord avec la littérature [11,97,138]. L'alimentation pauvre en lipides offerte dans ces programmes a probablement contribué aux améliorations des taux lipidiques sanguins des patients [139,140]. En revanche, nous n'avons pas observé d'améliorations homogènes de la lipémie chez les patients normolipémiés, ce qui est inattendu étant donné qu'il a été démontré que le programme ITLC ambulatoire CHIP améliore les taux de cholestérol chez les patients normolipémiés [106,133,134]. Cette divergence peut être attribuée à la durée relativement courte des interventions étudiées ici (une à deux semaines) par rapport au programme de 4 semaines décrit dans la littérature. On peut noter aussi que nous avons observé des réductions significatives du cholestérol et des triglycérides chez les patients normolipémiés dans deux des quatre ensembles de données, suggérant que des différences dans les caractéristiques initiales des populations de patients ont pu contribuer à l'hétérogénéité des résultats obtenus pour les patients normolipémiés.

Pour identifier les sous-groupes de patients pour notre analyse des effets des ITLC, nous nous sommes appuyés sur des recommandations cliniques internationales [128,129]. Cependant, étant donné le peu de données, dans la littérature, sur les interventions ITLC résidentielles, il est probable que ces recommandations ne sont pas adaptées, de manière optimale, aux populations spécifiques et aux effets de telles ITLC. Une approche plus efficace pourrait consister à utiliser une méthode basée sur les données (*data-driven*) pour identifier les sous-groupes de patients en fonction des effets des ITLC. Cela permettrait une caractérisation plus précise des effets du programme, notamment l'amplitude de l'efficacité, les facteurs prédictifs et les éventuels effets indésirables. De plus, une telle approche basée sur les données pourrait aider à expliquer les incohérences observées dans les résultats des patients normolipémiés en identifiant les facteurs prédisant les variations des taux de lipides sanguins. En exploitant les résultats d'une approche basée sur les données, nous pourrions mettre en lumière des sous-groupes plus nuancés et plus pertinents qui correspondraient mieux aux complexités des interventions ITLC.

Il est à noter que notre étude a montré des améliorations significatives de l'IMC. L'approche multifacette, comprenant un régime alimentaire complet à base de plantes et pauvre en lipides, de l'exercice modéré et de l'accompagnement psychothérapeutique, a probablement contribué à ce résultat, comme le soutient la littérature sur les programmes ITLC [97,106]. Cependant, il est important de reconnaître qu'une partie du personnel médical dans les centres de mode de vie que nous avons étudiés recommandait occasionnellement des restrictions caloriques et des jeûnes de courte durée (2 à 3 jours) pour les patients en surpoids ou obèses cherchant à perdre du poids. Comme ces pratiques ne faisaient pas partie du protocole standard et n'étaient utilisées que dans des circonstances spécifiques, leur impact sur nos résultats est susceptible d'être minime. Une analyse comparative plus approfondie des approches de traitement utilisées dans ces centres et de leurs effets

sur l'IMC serait nécessaire pour élucider pleinement l'influence des programmes ITLC résidentiels sur la gestion du poids.

Nos résultats chez les patients hypertendus sont conformes à la littérature, qui a démontré l'efficacité des programmes ITLC dans l'amélioration de la pression artérielle [105,106]. La combinaison d'exercices modérés [98] et d'un régime végétal pauvre en lipides [141] pourrait contribuer aux effets positifs des ITLC sur l'hypertension. En revanche, nos résultats ne montrent aucune réduction significative de la pression artérielle chez les patients normotendus, avec une augmentation surprenante de la SBP dans deux des trois ensembles de données et de la DBP dans un des trois ensembles de données. Ce résultat est inattendu, car il diverge de la littérature sur les ITLC [105,106] et de nos propres constatations chez les patients hypertendus. L'ampleur de cette augmentation (+4,1 mmHg de SBP et +1,5 mmHg de DBP) est relativement faible par rapport aux effets hypotenseurs significatifs observés chez les patients hypertendus (-13,7 mmHg de SBP et -7,1 mmHg de DBP). Une enquête plus approfondie sur la prise en charge et la mesure de la pression artérielle chez les patients normotendus est nécessaire pour fournir une explication complète de ce phénomène. Il est à noter que les *lifestyle centers* que nous avons étudiés diminuent souvent les médicaments contre l'hypertension chez les patients normotendus, mais nos dossiers médicaux manquaient des informations nécessaires pour quantifier cette pratique et ainsi évaluer son impact.

Une des forces significatives de notre étude réside dans son utilisation de données générées en routine, ce qui permet l'examen de données du monde réel et limite ainsi le biais de sélection [55]. Cependant, une limitation notable est l'absence de groupes témoins pour la comparaison, ce qui peut introduire un biais de confusion qui pourrait partiellement expliquer nos résultats [142]. Dans le contexte de notre étude, les biais de confusion pertinents incluent la motivation des participants à améliorer leur santé et le manque de données provenant des patients qui ne terminent pas leur séjour dans les centres de mode de vie. Bien qu'il soit possible que la nature volontaire du programme et l'attrition des participants puissent contribuer à nos résultats, nous croyons que la cohérence et l'ampleur de nos résultats ne peuvent pas être entièrement attribuées à ces facteurs. Selon notre expérience, seule une petite proportion de patients ne termine généralement pas leur séjour dans les *lifestyle centers*, et ces départs sont généralement dus à un non-respect des exigences de l'intervention plutôt qu'à l'état de santé. Néanmoins, des études supplémentaires sont nécessaires pour quantifier ce biais de confusion. Enfin, nos résultats sont soutenus par des études randomisées contrôlées sur les ITLC résidentiels [11,97,138], certes plus petites, et par de grandes études sur les ITLC ambulatoires [106,133,134], ce qui renforce la crédibilité de nos résultats.

Conclusion

Dans cette étude, nous avons réussi à démontrer avec succès la faisabilité de l'utilisation de techniques de réutilisation des données pour enquêter sur l'efficacité des programmes ITLC résidentiels. Nos résultats indiquent que ces programmes peuvent entraîner des améliorations significatives de la lipidémie chez les patients hyperlipidémiques, de l'IMC chez les patients en surpoids ou obèses, et de la pression artérielle chez les patients hypertendus. Pour élucider davantage la portée et les mécanismes sous-jacents de ces effets, des études plus approfondies et comparatives sont nécessaires.

Liste des tables

Introduction

Tableau 1. Compétences de lifestyle medicine recommandées aux médecins par l'AMA [14,15]	16
--	----

Article scientifique

Table 1. Selected features for data extraction	39
Table 2. Subgroups definitions for lipidemia, BMI, and blood pressure analysis.	40
Table 3. Types of data received from lifestyle centers.....	41
Table 4. Quality evaluation of feature coding in collected datasets	41
Table 5. Prevalence of chronic diseases among patients at lifestyle center A, as extracted from EHR.	44
Table 6. Prevalence of chronic disease medications among patients at lifestyle center A, as extracted from EHR.	44
Table 7. Prevalence of chronic diseases among patients at lifestyle center B, as extracted from semi-structured letters.	47
Table 8. Prevalence of chronic diseases among patients at lifestyle center C, as extracted from a research-purposed database.	50
Table 9. Prevalence of chronic disease medications among patients at lifestyle center C, as extracted from a research-purposed database.....	50
Table 10. Prevalence of chronic diseases among patients at lifestyle center D, as extracted from a research-purposed database.	53
Table 11. Prevalence of chronic disease medications among patients at lifestyle center D, as extracted from a research-purposed database.....	53
Table 12. Summary of results with significant changes observed in three or four datasets for lipidemia, blood pressure, and BMI.....	56
Table 13. Results with significant changes observed in one or two datasets for lipidemia, blood pressure, and BMI.....	57

Liste des figures

Introduction

Figure 1. Réutilisation de données en recherche médicale. Cette figure est adaptée de [24].	21
---	----

Article scientifique

Figure 1. Time requirements for data preprocessing, by dataset and original format, shown in both (A) raw hours and (B) hours per 1000 records.	42
Figure 2. Demographic characteristics of patients from lifestyle center A, as extracted from EHR. (A) Age and sex distribution of patients. (B) Proportion of male and female patients. (C) Distribution of patients across main age groups.	44
Figure 3. Mean changes in blood lipid levels, BMI, and blood pressure following a 6/7-day intervention, as recorded in EHR from lifestyle center A.	45
Figure 4. Demographic characteristics of patients from lifestyle center B, as extracted from semi-structured letters. (A) Age and sex distribution of patients. (B) Proportion of male and female patients. (C) Distribution of patients across main age groups.	47
Figure 5. Mean changes in blood lipid levels and BMI following a 7-day intervention, as recorded in in semi-structured letters from lifestyle center B.	48
Figure 6. Demographic characteristics of patients from lifestyle center C, as extracted from semi-structured letters. (A) Age and sex distribution of patients. (B) Proportion of male and female patients. (C) Distribution of patients across main age groups.	50
Figure 7. Mean changes in blood lipid levels and BMI following a 14-day intervention, as recorded in the research-purposed database from lifestyle center C.	51
Figure 8. Demographic characteristics of patients from lifestyle center D, as extracted from semi-structured letters. (A) Age and sex distribution of patients. (B) Proportion of male and female patients. (C) Distribution of patients across main age groups.	53
Figure 9. Mean changes in blood lipid levels and BMI following a 14-day intervention, as recorded in the research-purposed database from lifestyle center D.	54

Références

- [1] Rippe JM. The Academic Basis of Lifestyle Medicine. *Am J Lifestyle Med* 2023;15598276231212825. <https://doi.org/10.1177/15598276231212825>.
- [2] LIFESTYLE MEDICINE – EVIDENCE REVIEW. American College of Preventive Medicine; 2009.
- [3] Tsiompanou E, Marketos SG. Hippocrates: timeless still. *J R Soc Med* 2013;106:288–92. <https://doi.org/10.1177/0141076813492945>.
- [4] Voltaire. Œuvres complètes de Voltaire. Paris: Garnier; 1883.
- [5] Tulchinsky TH, Varavikova EA. A History of Public Health. *New Public Health* 2014;1–42. <https://doi.org/10.1016/B978-0-12-415766-8.00001-X>.
- [6] World Health Organization. Ottawa Charter for Health Promotion, 1986, p. 5.
- [7] Kelly J, Shull J. A Comprehensive Clinical Lifestyle Medicine Specialty Fellowship Program: What Intensive Lifestyle Treatment Can Do. *Am J Lifestyle Med* 2017;11:414–8. <https://doi.org/10.1177/1559827617696293>.
- [8] Hyman MA, Ornish D, Roizen M. Lifestyle medicine: treating the causes of disease. *Altern Ther Health Med* 2009;15:12–4.
- [9] Gregg EW, Chen H, Wagenknecht LE, Clark JM, Delahanty LM, Bantle J, et al. Association of an Intensive Lifestyle Intervention With Remission of Type 2 Diabetes. *JAMA* 2012;308:2489–96. <https://doi.org/10.1001/jama.2012.67929>.
- [10] Taylor R, Ramachandran A, Yancy WS, Forouhi NG. Nutritional basis of type 2 diabetes remission. *BMJ* 2021;374:n1449. <https://doi.org/10.1136/bmj.n1449>.
- [11] Dan N, Morosan S, Gheorghita V, Dan V, Petersen F, Hanspreet S, et al. Herghelia lifestyle study - Efficiency of a residential Lifestyle change program. *Rev Ig Și Sănăt Publică* 2009;59:27–41.
- [12] Braman M, Edison M. How to Create a Successful Lifestyle Medicine Practice: *Am J Lifestyle Med* 2017. <https://doi.org/10.1177/1559827617696296>.
- [13] Polak R, Pojednic RM, Phillips EM. Lifestyle Medicine Education. *Am J Lifestyle Med* 2015;9:361–7. <https://doi.org/10.1177/1559827615580307>.
- [14] American Medical Association. Healthy Lifestyles H-425.972 2010.
- [15] Lianov L. Physician Competencies for Prescribing Lifestyle Medicine. *JAMA* 2010;304:202. <https://doi.org/10.1001/jama.2010.903>.
- [16] Lianov LS, Adamson K, Kelly JH, Matthews S, Palma M, Rea BL. Lifestyle Medicine Core Competencies: 2022 Update. *Am J Lifestyle Med* 2022;16:734–9. <https://doi.org/10.1177/15598276221121580>.
- [17] Mechley AR, Dysinger W. Intensive Therapeutic Lifestyle Change Programs: A Progressive Way to Successfully Manage Health Care. *Am J Lifestyle Med* 2015;9:7.

- [18] Retired Senior Volunteer Program. History of Dade County, Georgia. First Edition. Espy Publishing Company; 1981.
- [19] Pierson RH. Miracles Happen Every Day. Pacific Press Publishing Association; 2012.
- [20] Jackson W. The Nineteenth-Century Context of the Seventh-day Adventist Health Message. In: Rogers L, editor. Chang. Attitudes Sci. Advent. Health Med. 1865 - 2015 Collect. Pap. Conf. Held Avondale Coll. High. Educ. March 13 14 2015, Cooranbong, N.S.W.: Avondale Academic Press; 2015, p. 135–59.
- [21] Calvert SJ. A Taste of Eden: Modern Christianity and Vegetarianism. *J Eccles Hist* 2007;58:461–81. <https://doi.org/10.1017/S0022046906008906>.
- [22] Nieto-Martínez R, González-Rivas JP, Mechanick JI. Survey of International Centers that Incorporate Lifestyle Medicine. In: Mechanick JI, Kushner RF, editors. *Creat. Lifestyle Med. Cent. Concept Clin. Pract.*, Cham: Springer International Publishing; 2020, p. 345–53. https://doi.org/10.1007/978-3-030-48088-2_30.
- [23] Piatetsky-Shapiro G. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. *AI Mag* 1990;11:68.
- [24] Chazard E, Ficheur G, Caron A, Lamer A, Labreuche J, Cuggia M, et al. Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features. *Stud Health Technol Inform* 2018;255:15–9.
- [25] Fayyad U, Gregory P-S, Smyth P. From Data Mining to Knowledge Discovery in Databases. *AI Mag* 1996;17:18.
- [26] Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. *BioMed Res Int* 2015;2015. <https://doi.org/10.1155/2015/639021>.
- [27] République Française, editor. Vocabulaire de l'informatique (liste de termes, expressions et définitions adoptés). *J Off Répub Fr* 2003;3531–2.
- [28] Safran C. Update on Data Reuse in Health Care. *Yearb Med Inform* 2017;26:24–7. <https://doi.org/10.15265/IY-2017-013>.
- [29] Filipovic-Pierucci A, Samson S, Fagot-Campagna A. Mesure du recours aux soins pour troubles dépressifs à partir du Sniiram. *Rev D'Épidémiologie Santé Publique* 2016;64:S16. <https://doi.org/10.1016/j.respe.2016.01.054>.
- [30] Safran C. Reuse of clinical data. *Yearb Med Inform* 2014;9:52–4. <https://doi.org/10.15265/IY-2014-0013>.
- [31] Genin M. Discrete scan statistics and generalized likelihood ratio test. *Romanian J Pure Appl Math* 2015;60:83–92.
- [32] Genin M, Duhamel A, Preda C, Fumery M, Savoye G, Peyrin-Biroulet L, et al. Space-time clusters of Crohn's disease in northern France. *J Public Health* 2013;21:497–504. <https://doi.org/10.1007/s10389-013-0580-9>.
- [33] Chazard E, Boudry A, Beeler PE, Dalleur O, Hubert H, Tréhou E, et al. Towards The Automated, Empirical Filtering of Drug-Drug Interaction Alerts in

- Clinical Decision Support Systems: Historical Cohort Study of Vitamin K Antagonists. *JMIR Med Inform* 2021;9:e20862. <https://doi.org/10.2196/20862>.
- [34] Chazard E, Beuscart J-B, Rochoy M, Dalleur O, Decaudin B, Odou P, et al. Statistically Prioritized and Contextualized Clinical Decision Support Systems, the Future of Adverse Drug Events Prevention? *Stud Health Technol Inform* 2020;270:683–7. <https://doi.org/10.3233/SHTI200247>.
- [35] Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. *IEEE Trans Inf Technol Biomed Publ IEEE Eng Med Biol Soc* 2011;15:823–30. <https://doi.org/10.1109/TITB.2011.2165727>.
- [36] Thillard E-M, Gautier S, Babykina E, Carton L, Amad A, Bouzillé G, et al. Psychiatric Adverse Events Associated With Infliximab: A Cohort Study From the French Nationwide Discharge Abstract Database. *Front Pharmacol* 2020;11:513. <https://doi.org/10.3389/fphar.2020.00513>.
- [37] Cren P-Y, Bertrand N, Deley M-CL, Génin M, Mortier L, Odou P, et al. Is the survival of patients treated with ipilimumab affected by antibiotics? An analysis of 1585 patients from the French National hospital discharge summary database (PMSI). *Oncolmunology* 2020;9:1846914. <https://doi.org/10.1080/2162402X.2020.1846914>.
- [38] Boullenger L, Balcaen T, Calafiore M, Legrand B, Rochoy M, Chazard E. Diabétiques de type 2 suivis en médecine générale : séquences de traitements et évolution. *Rev D'Épidémiologie Santé Publique* 2020;68:S23. <https://doi.org/10.1016/j.respe.2020.01.050>.
- [39] Grujon A. Nouveau profil épidémique de la gale humaine dans un cabinet de médecine générale de la métropole Lilloise. Thèse d'exercice. Université du droit et de la santé, 2017.
- [40] Falissard B. Analysis of Questionnaire Data with R. CRC Press; 2011.
- [41] Codd EF. A Relational Model of Data for Large Shared Data Banks 1970;13:11.
- [42] Codd EF. Derivability, redundancy and consistency of relations stored in large data banks. New York, NY: IBM; 1969.
- [43] Lamer A, Saint-Dizier C, Paris N, Chazard E. « Data Lake, Data Warehouse, Datamart and Feature Store » - Contributions au Pipeline de réutilisation des données. *J Epidemiol Popul Health* 2024;72:202245. <https://doi.org/10.1016/j.jep.2024.202245>.
- [44] Donoho D. 50 Years of Data Science. *J Comput Graph Stat* 2017;26:745–66. <https://doi.org/10.1080/10618600.2017.1384734>.
- [45] De Mauro A, Greco M, Grimaldi M. A formal definition of Big Data based on its essential features. *Libr Rev* 2016;65:122–35. <https://doi.org/10.1108/LR-06-2015-0061>.
- [46] Khan MA, Uddin MF, Gupta N. Seven V's of Big Data understanding Big Data to extract value. *Proc. 2014 Zone 1 Conf. Am. Soc. Eng. Educ.*, 2014, p. 1–5. <https://doi.org/10.1109/ASEEZone1.2014.6820689>.

- [47] Gao Y, Janssen M, Zhang C. Comprendre l'évolution de la recherche sur les données gouvernementales ouvertes : vers la durabilité et l'intelligence des données ouvertes. *Rev Int Sci Adm* 2023;89:41–57. <https://doi.org/10.3917/risa.891.0041>.
- [48] Roriz S, Toulouse E, Bacle M, Granier S, Elotmani L, Ambert A, et al. Réglementation pour la recherche clinique en France en 2023. *Anesth Réanimation* 2023;9:414–27. <https://doi.org/10.1016/j.anrea.2023.09.002>.
- [49] Souche R, Mas S, Scatton O, Fabre J-M, Gimeno L, Herrero A, et al. Législation encadrant la recherche clinique rétrospective sur dossiers médicaux : ce qu'il faut savoir et ce qu'il faut faire. *J Chir Viscérale* 2022;159:236–42. <https://doi.org/10.1016/j.jchirv.2021.10.009>.
- [50] Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC Med Res Methodol* 2019;19:53. <https://doi.org/10.1186/s12874-019-0695-y>.
- [51] Johnston MP. Secondary Data Analysis: A Method of which the Time Has Come. *Qual Quant Methods Libr* 2014.
- [52] Logan T. A practical, iterative framework for secondary data analysis in educational research. *Aust Educ Res* 2020;47:129–48. <https://doi.org/10.1007/s13384-019-00329-z>.
- [53] Système National des Données de Santé. Composantes du SNDS | SNDS n.d. <https://www.snds.gouv.fr/SNDS/Composantes-du-SNDS> (accessed March 27, 2024).
- [54] Pour une meilleure utilisation des bases de données administratives et medico-administratives nationales pour la santé publique et la recherche. Haut Conseil de la santé publique; 2012.
- [55] Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. *J Epidemiol Community Health* 2014;68:283–7. <https://doi.org/10.1136/jech-2013-202744>.
- [56] Greenland S, Robins J. Invited Commentary: Ecologic Studies—Biases, Misconceptions, and Counterexamples. *Am J Epidemiol* 1994;139:747–60. <https://doi.org/10.1093/oxfordjournals.aje.a117069>.
- [57] Ghenassia A, Beuscart JB, Ficheur G, Occelli F, Babykina E, Chazard E, et al. A generic method for improving the spatial interoperability of medical and ecological databases. *Int J Health Geogr* 2017;16:36. <https://doi.org/10.1186/s12942-017-0109-5>.
- [58] Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin Infect Dis Off Publ Infect Dis Soc Am* 2009;49:1557–64. <https://doi.org/10.1086/630200>.
- [59] Shakeri Hossein Abad Z, Kline A, Sultana M, Noaen M, Nurmambetova E, Lucini F, et al. Digital public health surveillance: a systematic scoping review. *NPJ Digit Med* 2021;4:41. <https://doi.org/10.1038/s41746-021-00407-6>.
- [60] Calafiore M, Chazard E, Averlant L, Ramez C, Sarrazin F, Leveque N, et al. GPs' interest in integrated care for frail older adults and corresponding consulting and prescribing data: qualitative and quantitative analyses of the

- PAERPA integrated care project. *Br J Gen Pract* 2022;72:e809–15. <https://doi.org/10.3399/BJGP.2021.0626>.
- [61] Attard J, Orlandi F, Scerri S, Auer S. A systematic review of open government data initiatives. *Gov Inf Q* 2015;32:399–418. <https://doi.org/10.1016/j.giq.2015.07.006>.
 - [62] Hardwicke TE, Mathur MB, MacDonald K, Nilsson G, Banks GC, Kidwell MC, et al. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*. *R Soc Open Sci* 2018;5:180448. <https://doi.org/10.1098/rsos.180448>.
 - [63] Seastedt KP, Schwab P, O'Brien Z, Wakida E, Herrera K, Marcelo PGF, et al. Global healthcare fairness: We should be sharing more, not less, data. *PLOS Digit Health* 2022;1:e0000102. <https://doi.org/10.1371/journal.pdig.0000102>.
 - [64] Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* 2015;347:78–81. <https://doi.org/10.1126/science.1260825>.
 - [65] Ventura-Cots M, Watts AE, Cruz-Lemini M, Shah ND, Ndugga N, McCann P, et al. Colder Weather and Fewer Sunlight Hours Increase Alcohol Consumption and Alcoholic Cirrhosis Worldwide. *Hepatology* 2019;69:1916–30. <https://doi.org/10.1002/hep.30315>.
 - [66] Gao F, Tao L, Huang Y, Shu Z. Management and Data Sharing of COVID-19 Pandemic Information. *BIOPRESERVATION BIOBANKING* 2020;18.
 - [67] Harrison PW, Lopez R, Rahman N, Allen SG, Aslam R, Buso N, et al. The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Res* 2021;49:W619–23. <https://doi.org/10.1093/nar/gkab417>.
 - [68] Shover CL, Falasinnu TO, Dwyer CL, Santos NB, Cunningham NJ, Freedman RB, et al. Steep increases in fentanyl-related mortality west of the Mississippi River: Recent evidence from county and state surveillance. *Drug Alcohol Depend* 2020;216:108314. <https://doi.org/10.1016/j.drugalcdep.2020.108314>.
 - [69] Goldberg D, Ross-Driscoll K, Lynch R. County Differences in Liver Mortality in the United States: Impact of Sociodemographics, Disease Risk Factors, and Access to Care. *Gastroenterology* 2021;160:1140–1150.e1. <https://doi.org/10.1053/j.gastro.2020.11.016>.
 - [70] Sielemann K, Hafner A, Pucker B. The reuse of public datasets in the life sciences: potential risks and rewards. *PeerJ* 2020;8:e9954. <https://doi.org/10.7717/peerj.9954>.
 - [71] Kim J, Park M. A Study on ML-Based Sleep Score Model Using Lifelog Data. *Appl Sci* 2023;13:1043. <https://doi.org/10.3390/app13021043>.
 - [72] Priou S, Lame G, Jankovic M, Kempf E. “In conferences, everyone goes ‘health data is the future’”: an interview study on challenges in re-using EHR data for research in Clinical Data Warehouses. *AMIA Annu Symp Proc* 2024;2023:579–88.
 - [73] Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. *Yearb Med Inform* 2017;26:38–52. <https://doi.org/10.15265/IY-2017-007>.

- [74] Little RJ. Missing Data Analysis 2024. <https://doi.org/10.1146/annurev-clinpsy-080822-051727>.
- [75] Ferri P, Romero-Garcia N, Badenes R, Lora-Pablos D, Morales TG, Gómez de la Cámara A, et al. Extremely missing numerical data in Electronic Health Records for machine learning can be managed through simple imputation methods considering informative missingness: A comparative of solutions in a COVID-19 mortality case study. *Comput Methods Programs Biomed* 2023;242:107803. <https://doi.org/10.1016/j.cmpb.2023.107803>.
- [76] Luo Y. Evaluating the state of the art in missing data imputation for clinical data. *Brief Bioinform* 2022;23:bbab489. <https://doi.org/10.1093/bib/bbab489>.
- [77] Zimmerman AS. Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists. Thesis. 2003.
- [78] Sandt S van de, Dallmeier-Tiessen S, Lavasa A, Petras V. The Definition of Reuse. *Data Sci J* 2019;18:22–22. <https://doi.org/10.5334/dsj-2019-022>.
- [79] Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. *BMC Med Inform Decis Mak* 2023;23:18. <https://doi.org/10.1186/s12911-023-02115-5>.
- [80] Kabachinski J. What is Health Level 7? *Biomed Instrum Technol* 2006;40:375–9. <https://doi.org/10.2345/i0899-8205-40-5-375.1>.
- [81] Particle Health Team. What is the C-CDA Healthcare Data Format? 2022. <https://www.particlehealth.com/blog/what-is-ccda-consolidated-clinical-document-architecture> (accessed April 12, 2024).
- [82] Jung S, Bae S, Seong D, Yi B-K. Standard Document Development for Health Information Exchange in Korea. *Appl Clin Inform* 2022;13:592–601. <https://doi.org/10.1055/s-0042-1749331>.
- [83] Bossenko I, Linna K, Piho G, Ross P. Migration from HL7 Clinical Document Architecture (CDA) to Fast Health Interoperability Resources (FHIR) in Infectious Disease Information System of Estonia. *Proc. 38th ACM SIGAPP Symp. Appl. Comput.*, New York, NY, USA: Association for Computing Machinery; 2023, p. 882–5. <https://doi.org/10.1145/3555776.3577836>.
- [84] Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. *Proc. 26th IEEE Int. Symp. Comput.-Based Med. Syst.*, Porto, Portugal: IEEE; 2013, p. 326–31. <https://doi.org/10.1109/CBMS.2013.6627810>.
- [85] Duda SN, Kennedy N, Conway D, Cheng AC, Nguyen V, Zayas-Cabán T, et al. HL7 FHIR-based tools and initiatives to support clinical research: a scoping review. *J Am Med Inform Assoc* 2022;29:1642–53. <https://doi.org/10.1093/jamia/ocac105>.
- [86] Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre M-M, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. *Appl Clin Inform* 2020;11:13–22. <https://doi.org/10.1055/s-0039-3402754>.
- [87] FitzHenry F, Resnic FS, Robbins SL, Denton J, Nookala L, Meeker D, et al. Creating a Common Data Model for Comparative Effectiveness with the

- Observational Medical Outcomes Partnership. *Appl Clin Inform* 2015;06:536–47. <https://doi.org/10.4338/ACI-2014-12-CR-0121>.
- [88] Référentiel du standard HPRIM Santé 2021.
 - [89] French Technical Agency for Hospital Information (ATIH). ICD-10 FR 2017 for PMSI usage 2022. <https://www.atih.sante.fr/cim-10-fr-2022-usage-pmsi> (accessed January 12, 2023).
 - [90] Santé M des S et de la, Santé M des S et de la. Classification commune des actes médicaux (CCAM). Ministère Solidar Santé 2021. <https://solidarites-sante.gouv.fr/professionnels/gerer-un-etablissement-de-sante-medico-social/financement/financement-des-etablissements-de-sante-10795/financement-des-etablissements-de-sante-glossaire/article/classification-commune-des-actes-medicaux-ccam> (accessed June 7, 2021).
 - [91] Anatomical Therapeutic Chemical (ATC) Classification n.d. <https://www.who.int/tools/atc-ddd-toolkit/atc-classification> (accessed June 7, 2021).
 - [92] Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem* 1996;42:81–90.
 - [93] Agence Nationale de Sécurité du Médicament et des produits de santé. Actualité - Codification des médicaments à usage humain : l'ANSM et le Club interpharmaceutique, acteurs de la délivrance des codes CIP et UCD. ANSM 2022. <https://ansm.sante.fr/actualites/codification-des-medicaments-a-usage-humain-lansm-et-le-club-interpharmaceutique-acteurs-de-la-delivrance-des-codes-cip-et-ucd> (accessed April 12, 2024).
 - [94] National Library of Medicine. RxNorm n.d. <https://www.nlm.nih.gov/research/umls/rxnorm/index.html> (accessed April 22, 2024).
 - [95] Liu S, Wei Ma, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. *IT Prof* 2005;7:17–23. <https://doi.org/10.1109/MITP.2005.122>.
 - [96] CISMef. HeTOP, Health Terminology/Ontology Portal n.d. <https://www.hetop.eu/hetop/> (accessed June 7, 2024).
 - [97] Ornish D, Brown SE, Billings JH, Scherwitz LW, Armstrong WT, Ports TA, et al. Can lifestyle changes reverse coronary heart disease? The Lifestyle Heart Trial. *The Lancet* 1990;336:129–33. [https://doi.org/10.1016/0140-6736\(90\)91656-U](https://doi.org/10.1016/0140-6736(90)91656-U).
 - [98] Naci H, Salcher-Konrad M, Dias S, Blum MR, Sahoo SA, Nunan D, et al. How does exercise treatment compare with antihypertensive medications? A network meta-analysis of 391 randomised controlled trials assessing exercise and medication effects on systolic blood pressure. *Br J Sports Med* 2018;bjsports-2018-099921. <https://doi.org/10.1136/bjsports-2018-099921>.
 - [99] Ornish D, Magbanua MJM, Weidner G, Weinberg V, Kemp C, Green C, et al. Changes in prostate gene expression in men undergoing an intensive nutrition

- and lifestyle intervention. *Proc Natl Acad Sci U S A* 2008;105:8369–74. <https://doi.org/10.1073/pnas.0803080105>.
- [100] Rao RV, Subramaniam KG, Gregory J, Bredesen AL, Coward C, Okada S, et al. Rationale for a Multi-Factorial Approach for the Reversal of Cognitive Decline in Alzheimer's Disease and MCI: A Review. *Int J Mol Sci* 2023;24:1659. <https://doi.org/10.3390/ijms24021659>.
 - [101] Banta JE, Lee JW, Hodgkin G, Yi Z, Fanica A, Sabate J. The Global Influence of the Seventh-Day Adventist Church on Diet. *Religions* 2018;9:251. <https://doi.org/10.3390/rel9090251>.
 - [102] Barnard RJ, Lattimore L, Holly RG, Cherny S, Pritikin N. Response of non-insulin-dependent diabetic patients to an intensive program of diet and exercise. *Diabetes Care* 1982;5:370–4. <https://doi.org/10.2337/diacare.5.4.370>.
 - [103] Barnard RJ, Massey MR, Cherny S, O'Brien LT, Pritikin N. Long-term use of a high-complex-carbohydrate, high-fiber, low-fat diet and exercise in the treatment of NIDDM patients. *Diabetes Care* 1983;6:268–73. <https://doi.org/10.2337/diacare.6.3.268>.
 - [104] Ornish D, Weidner G, Fair WR, Marlin R, Pettengill EB, Raisin CJ, et al. Intensive lifestyle changes may affect the progression of prostate cancer. *J Urol* 2005;174:1065–70.
 - [105] Sanchez A, Chung S, Mejia A, Ramirez F, Shavlik G, Bivens R, et al. Multiple lifestyle intervention reverses hypertension. *Cogent Med* 2019. <https://doi.org/10.1080/2331205X.2019.1636534>.
 - [106] Rankin P, Morton DP, Diehl H, Gobble J, Morey P, Chang E. Effectiveness of a Volunteer-Delivered Lifestyle Modification Program for Reducing Cardiovascular Disease Risk Factors. *Am J Cardiol* 2012;109:82–6. <https://doi.org/10.1016/j.amjcard.2011.07.069>.
 - [107] Lifestyle Medicine Institute. Clinical Research. Pivio n.d. <https://piviohealth.com/knowledge-bank/clinical-research/> (accessed April 22, 2024).
 - [108] Kent L, Morton D, Hurlow T, Rankin P, Hanna A, Diehl H. Long-term effectiveness of the community-based Complete Health Improvement Program (CHIP) lifestyle intervention: a cohort study. *BMJ Open* 2013;3:e003751. <https://doi.org/10.1136/bmjopen-2013-003751>.
 - [109] Chazard E, Ficheur G, Caron A, Lamer A, Labreuche J, Cuggia M, et al. Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features. *Stud Health Technol Inform* 2018;255:15–9.
 - [110] Gardner SR. Building the data warehouse. *Commun ACM* 1998;41:52–60. <https://doi.org/10.1145/285070.285080>.
 - [111] García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F. Big data preprocessing: methods and prospects. *Big Data Anal* 2016;1:9. <https://doi.org/10.1186/s41044-016-0014-0>.
 - [112] Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic

- p health records. WIREs Comput Stat 2021;13:e1549.
-
- <https://doi.org/10.1002/wics.1549>
- .
-
- [113] Health Level Seven International. Homepage n.d.
- <https://www.hl7.org/>
- (accessed April 22, 2024).
-
- [114] Observational Health Data Sciences and Informatics. OMOP Common Data Model n.d.
- <https://ohdsi.github.io/CommonDataModel/>
- (accessed April 22, 2024).
-
- [115] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Suchard MA, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers n.d.
-
- [116] HL7 Standards - Section 1a: Clinical Document Architecture (CDA®) | HL7 International n.d.
- https://www.hl7.org/implement/standards/product_section.cfm?section=10
- (accessed April 22, 2024).
-
- [117] World Health Organization. International Classification of Diseases (ICD) n.d.
- <https://www.who.int/standards/classifications/classification-of-diseases>
- (accessed April 22, 2024).
-
- [118] Regenstrief Institute. Home. LOINC n.d.
- <https://loinc.org/>
- (accessed April 22, 2024).
-
- [119] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.
-
- [120] Dowle M, Srinivasan A. data.table: Extension of `data.frame`. 2023.
-
- [121] Aragon TJ. epitools: Epidemiology Tools. 2020.
-
- [122] Signorell A. DescTools: Tools for Descriptive Statistics. 2023.
-
- [123] Posit team. RStudio: Integrated Development Environment for R. Boston, MA: Posit Software, PBC; 2024.
-
- [124] Lang DT. XML: Tools for Parsing and Generating XML Within R and S-Plus. 2023.
-
- [125] Wickham H, Hester J, Ooms J. xml2: Parse XML. 2023.
-
- [126] Wickham H. The Split-Apply-Combine Strategy for Data Analysis. J Stat Softw 2011;40:1–29.
-
- [127] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.
-
- [128] Expert Panel On Detection, Evaluation, And Treatment Of High Blood Cholesterol In Adults. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). JAMA J Am Med Assoc 2001;285:2486–97.
- <https://doi.org/10.1001/jama.285.19.2486>
- .
-
- [129] Jensen MD, Ryan DH, Apovian CM, Ard JD, Comuzzie AG, Donato KA, et al. 2013 AHA/ACC/TOS Guideline for the Management of Overweight and Obesity in Adults. Circulation 2014;129:S102–38.
- <https://doi.org/10.1161/01.cir.0000437739.71477.ee>
- .

- [130] Vuokko R, Mäkelä-Bengs P, Hyppönen H, Lindqvist M, Doupi P. Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data. *Int J Med Inf* 2017;97:293–303. <https://doi.org/10.1016/j.ijmedinf.2016.10.004>.
- [131] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;20:144–51. <https://doi.org/10.1136/amiainl-2011-000681>.
- [132] Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *J Med Internet Res* 2018;20:e9134. <https://doi.org/10.2196/jmir.9134>.
- [133] Drozek D, Diehl H, Nakazawa M, Kostohryz T, Morton D, Shubrook JH. Short-Term Effectiveness of a Lifestyle Intervention Program for Reducing Selected Chronic Disease Risk Factors in Individuals Living in Rural Appalachia: A Pilot Cohort Study. *Adv Prev Med* 2014;2014:1–7. <https://doi.org/10.1155/2014/798184>.
- [134] Leibold C, Shubrook JH, Nakazawa M, Drozek D. Effectiveness of the Complete Health Improvement Program in Reducing Risk Factors for Cardiovascular Disease in an Appalachian Population. *J Osteopath Med* 2016;116:84–91. <https://doi.org/10.7556/jaoa.2016.020>.
- [135] Evans L, London JW, Palchuk MB. Assessing real-world medication data completeness. *J Biomed Inform* 2021;119:103847. <https://doi.org/10.1016/j.jbi.2021.103847>.
- [136] Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* 2016;23:e20–7. <https://doi.org/10.1093/jamia/ocv130>.
- [137] Madden JM, Lakoma MD, Rusinak D, Lu CY, Soumerai SB. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *J Am Med Inform Assoc* 2016;23:1143–9. <https://doi.org/10.1093/jamia/ocw021>.
- [138] Rosenthal MB, Barnard RJ, Rose DP, Inkeles S, Hall J, Pritikin N. Effects of a high-complex-carbohydrate, low-fat, low-cholesterol diet on levels of serum lipids and estradiol. *Am J Med* 1985;78:23–7. [https://doi.org/10.1016/0002-9343\(85\)90456-5](https://doi.org/10.1016/0002-9343(85)90456-5).
- [139] Trepanowski JF, Varady KA. Veganism Is a Viable Alternative to Conventional Diet Therapy for Improving Blood Lipids and Glycemic Control. *Crit Rev Food Sci Nutr* 2015;55:2004–13. <https://doi.org/10.1080/10408398.2012.736093>.
- [140] Freedman MR, King J, Kennedy E. Popular Diets: A Scientific Review, Executive Summary. *Obes Res* 2012;9:1S-5S. <https://doi.org/10.1038/oby.2001.113>.
- [141] Tomé-Carneiro J, Visioli F. Plant-Based Diets Reduce Blood Pressure: A Systematic Review of Recent Evidence. *Curr Hypertens Rep* 2023;25:127–50. <https://doi.org/10.1007/s11906-023-01243-7>.
- [142] Arah OA. Bias Analysis for Uncontrolled Confounding in the Health Sciences. *Annu Rev Public Health* 2017;38:23–38. <https://doi.org/10.1146/annurev-publhealth-032315-021644>.

AUTEUR : Nom : DURSUS-ÉLISÉE

Prénom : Ruben

Date de Soutenance : 18/09/2024

Titre de la Thèse : Réutilisation de données issues de « lifestyle centers » : étude de faisabilité

Thèse - Médecine - Lille 2024

Cadre de classement : Santé publique et médecine sociale

DES + spécialité : Santé publique et médecine sociale

Mots-clés : Réutilisation de données, Lifestyle Medicine, Cholestérol

Résumé :

CONTEXTE : Bien que des études antérieures aient montré à petite échelle l'efficacité des programmes résidentiels avec *Intensive Therapeutic Lifestyle Change* (ITLC), la comparabilité et la pertinence à grande échelle de ces résultats ont été questionnées. Nous avons émis l'hypothèse que la réutilisation des données collectées en routine par les *lifestyle centers* pourrait fournir une approche réalisable pour étudier l'efficacité des programmes résidentiels de ITLC.

METHODE : Nous avons obtenu des données de quatre *lifestyle centers*, comprenant trois types différents de bases de données : dossiers de santé électroniques, lettres semi-structurées et bases de données collectées à des fins de recherche.

RESULTATS : Nous avons estimé le travail de data preprocessing à 150,3 et à 226,1 heures par 1000 dossiers pour les deux bases de données collectées à des fins de recherche, et à 157,5 et à 197,3 heures par 1000 dossiers pour les dossiers de santé électroniques et les lettres semi-structurées, respectivement ; ainsi, ces quantités s'avèrent comparables. Les échantillons obtenus présentaient des structures d'âge et de sexe similaires mais différaient par la prévalence des maladies chroniques courantes rapportées. Les effets métaboliques observés étaient homogènes entre les échantillons, avec des réductions significatives de la lipidémie chez les patients hyperlipidémiques (-33,5 mg/mL [min.|max. : -16,1|-46,6] en cholestérol total, -26,5 mg/mL [-10,9|-38,5] en cholestérol LDL), de l'indice de masse corporelle (IMC) chez les patients en surpoids ou obèses (-2,8 % [-1,8|-4,0]), et de la pression artérielle systolique et diastolique chez les patients hypertendus (-13,7 [-4,0|-29,2] et -7,1 mmHg [-3,1|-18,3], respectivement).

CONCLUSION : Dans l'ensemble, nos résultats suggèrent que la réutilisation des données des *lifestyle centers* est une technique efficace pour évaluer l'impact des programmes de ITLC.

Composition du Jury :

Président : Monsieur le Professeur Philippe AMOUYEL

Assesseurs : Monsieur le Docteur Aghiles HAMROUN

Madame le Docteur Lysiane LEGUIER

Directeur de thèse : Monsieur le Professeur Emmanuel CHAZARD