Méthode globale de prédiction des durées de séjours hospitalières avec intégration des données incrémentales et évolutives

Vincent Lequertier

THESE de DOCTORAT
DE L'UNIVERSITE CLAUDE BERNARD LYON 1

Ecole Doctorale 205
Ecole Doctorale Interdisciplinaire Science-Santé

Discipline :
Epidémiologie, santé publique, recherche sur les services de santé

Soutenue publiquement le 30/09/2022, par :
Vincent Lequertier

Méthode globale de prédiction des durées de séjours hospitalières avec intégration des données incrémentales et évolutives

---

Devant le jury composé de :

Schott-Pethelaz, Anne-Marie, PU-PH, Université Claude Bernard Lyon 1, Présidente
Boyer, Laurent, PU-PH, Université Aix-Marseille, Rapporteur
Bringay, Sandra, PU, Université Paul Valéry Montpellier III, Rapporteure
Soualmia, Lina, MCF, Université de Rouen, Rapporteure
Fromont, Elisa, PU, Université de Rennes, Examinatrice

Duclos, Antoine, PU-PH, Université Claude Bernard Lyon 1, Directeur de thèse
Wang, Tao, MCF, Université Jean Monnet Saint-Etienne, Co-directeur de thèse
Fondrevelle, Julien, MCF, Institut National des Sciences Appliquées Lyon, Co-directeur de thèse

Cette thèse a été préparée dans les laboratoires suivants.

Laboratoire RESHAPE
INSERM U1290
Université Claude Bernard Lyon 1
Domaine Rockefeller - 2ème étage (couloir CD)
8 Avenue Rockefeller,
69003 Lyon,
France

☎ 33 (0)4 26 68 82 24
33 (0)3 21 46 55 75
✉ hesper@univ-lyon1.fr
Site https://www.reshapelab.fr

Laboratoire DISP
UR 4570
Bâtiment Léonard de Vinci,
21 avenue Jean Capelle,
69100 Villeurbanne,
France

☎ 33 (0)4 72 43 82 19
33 (0)4 72 43 83 14
✉ disp@insa-lyon.fr
Site https://www.disp-lab.fr

Méthode globale de prédiction des durées de séjours hospitalières avec intégration des données incrémentales et évolutives

Résumé

Prédire la durée de séjour des patients est un enjeu important pour l'organisation des activités de soin dans les hôpitaux, notamment en termes de gestion des lits et de préparation de la sortie des patients. Faciliter l'organisation des activités de l'hôpital influence l'accès, la qualité et l'efficience des soins. Dans cette thèse, nous avons cherché à prédire la durée de séjour pour tous les patients de l'hôpital, à toutes les étapes qui composent leurs parcours de soins, à l'aide de données médico-administratives standardisées de Médecine, Chirurgie, Obstétrique qui sont collectées pour le remboursement des soins. Nous avons commencé par faire une revue systématique de la littérature sur les méthodes de prédiction des durées de séjours, afin de mieux comprendre la préparation des données, les différentes approches de prédiction et la façon de rapporter les résultats. Nous avons ensuite travaillé sur une méthode de prétraitement des données et déterminé si les *embeddings* peuvent représenter les concepts médicaux dans le cadre des prédictions de durées de séjours *via* un réseau de neurones. La capacité du réseau de neurones à correctement prédire la durée de séjour a été évaluée et comparée avec celle d'une forêt aléatoire et d'une régression logistique. Nos travaux montrent que la durée de séjour hospitalière peut être prédite au moyen d'un réseau de neurones avec des données médico-administratives standardisées disponibles pour tous les patients.

Mots clés : durée de séjour, prédiction, intelligence artificielle, aide à la décision, dossier médical informatisé

Global method for predicting the hospital length of stay using incremental and evolutionary data

Abstract

Predicting patient length of stay is an important issue for the organization of care activities in hospitals, especially for beds management the and preparation for patients discharge. Facilitating the organization of hospital activities influences access, quality and efficiency of care. In this thesis, we sought to predict length of stay for all patients in the hospital, at all stages that make up their care pathways, using standardized Medical, Surgical, Obstetric medico-administrative data collected for reimbursement of care.
We began by conducting a systematic review of the literature on methods for predicting lengths of stay, in order to better understand data preparation, the different prediction approaches, and how to report the results. We then worked on a data preprocessing method and investigated the ability of embeddings to represent medical concepts in the context of length of stay predictions *via* a neural network. The ability of the neural network to correctly predict length of stay was rigorously evaluated and compared with a random forest and a logistic regression. This work shows that hospital length of stay can be predicted by a neural network using standardized medical-administrative data available for all patients.

Keywords: length of stay, forecasting, artificial intelligence, clinical decision support, electronic health record

Remerciements

Je remercie mes directeurs de thèse Antoine Duclos, Tao Wang et Julien Fondrevelle, ainsi que Vincent Augusto, pour m'avoir fait confiance et m'avoir guidé tout au long de ces trois ans de thèse. Vos précieux conseils et encouragements m'ont été très utiles et m'ont permis d'apprendre à la fois sur le plan scientifique, technique et personnel. Je remercie aussi Stéphanie Polazzi pour notre collaboration sur les données de ce projet.

Je remercie Lina Soualmia, Laurent Boyer et Sandra Bringay d'avoir accepté d'être rapporteurs de ma thèse, et Anne-Marie Schott-Pethelaz ainsi qu'Élisa Fromont d'avoir accepté d'être membres du jury en tant qu'examinatrices.

Je remercie également mes collègues des Hospices Civils de Lyon dans le Service des Données de Santé, et mes collègues des laboratoires RESHAPE et DISP pour m'avoir transmis leurs connaissances et pour tous nos échanges formels et informels.

Enfin, je souhaite remercier ma famille et mes amis, sans qui je ne serais rien.

Préambule

Être en mesure d'anticiper la durée de séjour d'un patient est important pour bien organiser les activités de soins à l'hôpital. Comme les activités de soin sont regroupées par unités médicales possédant leurs thématiques précises, la gestion des ressources à la fois en termes de lits disponibles et de professionnels de santé (e.g. infirmières, médecins, chirurgiens, etc.) ainsi que la préparation de la sortie du patient s'effectue au niveau de chaque unité médicale. Il est donc important d'estimer la durée de séjour des patients lors de leurs passages dans les différentes unités médicales de l'hôpital. Faire une prédiction tout au long du parcours du patient au sein de l'hôpital permet également de prendre en compte les modifications de sa condition médicale. Cette estimation ne peut être conduite par les professionnels de santé, car elle dépend de facteurs nombreux et complexes difficiles à prendre en compte. Comme cette prédiction concerne tous les patients de toutes les unités médicales, les indicateurs sur lesquels baser les prédictions ne peuvent pas être spécifiques à une unité médicale ou une spécialité. De plus, utiliser des données médico-administratives standardisées permet de faciliter l'utilisation d'un système de prédiction en routine à l'hôpital, car ces données sont déjà collectées pour le remboursement des soins. Des prédictions précises et fiables sont requises pour permettre mieux gérer les ressources de l'hôpital. Cela justifie la nécessité d'étudier les méthodes informatiques pour prédire la durée de séjour des patients de manière globale en prenant en compte la nature incrémentale et évolutive des données.

Ce manuscrit de thèse est organisé en 8 chapitres. Le chapitre 1 présente le contexte, afin de comprendre ce qui a motivé ce travail ainsi que les informations qui ont guidé les choix scientifiques. En particulier, ce chapitre introduit la notion de durée de séjour et explique pourquoi la prédire est important, fait une présentation des bases de données utilisées, décrit les principes de fonctionnement de l'intelligence artificielle et plus particulièrement des réseaux de neurones, et liste des usages de l'intelligence artificielle en santé. Le chapitre 2 décrit l'objectif de cette thèse ainsi que ses 3 sous-objectifs. Les chapitres 3, 4 et 5 présentent les contributions scientifiques de la thèse. Plus spécifiquement, le chapitre 3 retranscrit une revue systématique de la littérature sur la méthodologie

autour de la prédiction des durées des séjours, le chapitre 4 présente les résultats préliminaires d'un modèle de prédiction des durées de séjour utilisant un réseau de neurones vers l'avant avec des *embeddings*, et le chapitre 5 propose une évaluation plus poussée des performances prédictives ainsi qu'une comparaison avec d'autres méthodes. Ces trois chapitres comportent chacun une introduction, une copie de l'article présentant la contribution scientifique et une discussion. Le chapitre 6 résume le travail effectué et discute des perspectives scientifiques sur ce travail, afin d'améliorer les performances du modèle de prédiction, et de faciliter son intégration dans le système d'information des hôpitaux. Enfin, le chapitre 7 conclut ce manuscrit de thèse et le chapitre 8 liste l'ensemble des contributions liées à la thèse.

Acronymes

API *Application Programming Interface*, Interface de programmation d'applications. 102, 109

CCAM Classification Commune des Actes Médicaux. 10

CEREES Comité d'Expertise pour les Recherches, les Études et les Études dans le domaine de la Santé. 13

CNIL Commission Nationale de l'Informatique et des Libertés. 13

CNN *Convolutional Neural Network*, Réseau de neurones convolutif. 17, 22

DMP Dossier Médical Partagé. 96

DPI Dossier Patient Informatisé. 96

GHM Groupement Homogène de Malades. 7

HCL Hospices Civils de Lyon. xvii, 5, 6, 13

IA Intelligence Artificielle. 14, 18, 25, 115

MCO Médecine, Chirurgie, Obstétrique. xvii, 2, 5, 7

PMSI Programe de Médicalisation du Système d'Information,. 5, 7

RNN *Recurrent Neural Network*, Réseau de neurones réccurent. 20, 25, 100

RPU Résumé de Passages aux Urgences. xv, 5, 11

RSS Résumé de Sortie Standardisé. 7

RUM Résumé d'Unité Médicale. xv, 7

SSR Soin de Suite et de Réadaptation. 97

UM Unité Médicale. 7

Sommaire

Liste des tableaux

Table des figures

Chapitre 1

Contexte

1.1 Durée de séjour

La durée de séjour s'exprime comme le nombre de jours entre l'admission d'un patient à l'hôpital et sa sortie.

Avoir une estimation de la durée de séjour permet d'anticiper le taux d'occupation d'une unité médicale, et d'estimer les besoins en termes de lits et de personnels qui sont des ressources critiques. Une prédiction correcte de la durée de séjour permet d'anticiper la sortie du patient (de l'hôpital ou de l'unité médicale), et une bonne anticipation contribue à fluidifier ce processus. Par conséquent, une bonne prédiction participe à améliorer l'organisation des services hospitaliers. Les estimations de durées de séjour peuvent donc être prises en compte par des programmes servant à la planification des ressources de l'hôpital. Par exemple, elles sont utilisées dans des modèles de simulation [1]. De plus, une bonne organisation réduit les complications [2] (morbi-mortalité) et joue un rôle prépondérant dans la satisfaction des patients [3]. Il convient donc de faire des prédictions précises pour améliorer la gestion des ressources des hôpitaux. Cependant, une estimation de la durée de séjour uniquement basée sur les moyennes par diagnostics principaux, bien que pratique et simple, peut ne pas être assez précise. L'influence de l'incertitude sur la durée de séjour a été mesurée et montre que prendre en compte les variations de durées de séjour liées aux caractéristiques des patients et leurs prises en charge améliore la

planification et la gestion des lits [4]. Les prédictions de la durée de séjour par le personnel médical ne sont pas toujours exactes [5, 6], car la durée de séjour est associée à de nombreux facteurs historiques, médicaux et environnementaux dont les professionnels de santé n'ont pas forcément connaissance au moment de son estimation. De tout cela découle la nécessité de s'appuyer sur des bases de données de tailles importantes et des algorithmes informatiques pour prédire la durée de séjour.

La figure 1.1 montre l'évolution annuelle de la durée de séjour en France en MCO. On peut constater une baisse de la durée de séjour annuelle, qui peut être liée à des changements organisationnels et de gestion économique des hôpitaux ainsi que des progrès dans les pratiques médicales. Cela peut être corroboré par une comparaison avec d'autres pays. En 2019, la durée de séjour moyenne en MCO en Turquie était de 4,1 jours, aux États-Unis de 5,5 jours, de 7 jours en Italie et de 16 jours au Japon. La durée de séjour moyenne dépend donc des pays, ce qui permet de faire l'hypothèse qu'elle dépend de choix organisationnels et politiques [7].

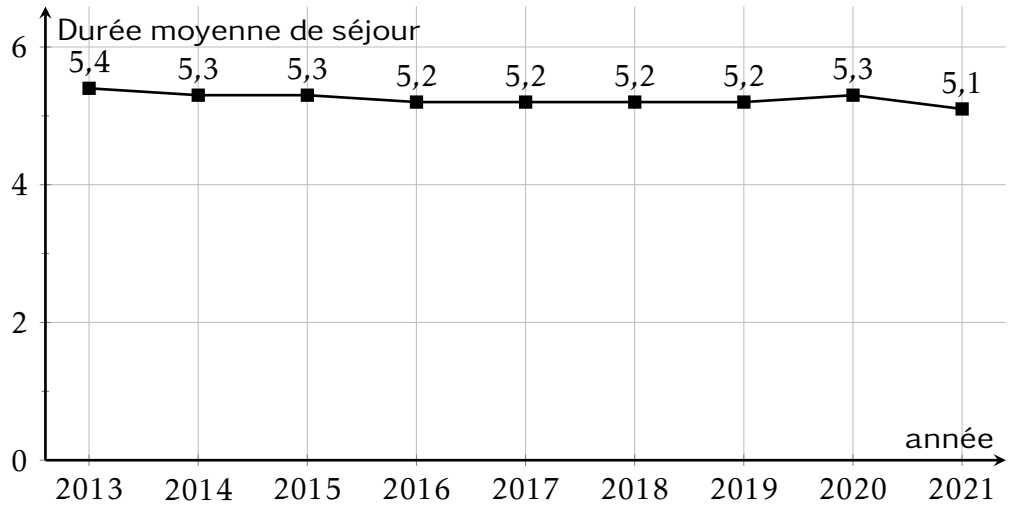


Figure 1.1 – Durée moyenne de séjour annuelle en MCO en France entre 2013 et 2021. Données de l'Agence Technique de l'Information sur l'Hospitalisation [8].

La figure 1.2 montre l'évolution du nombre de lits en hospitalisation complète, avec une diminution de 34 000 lits entre 2013 et 2021 en MCO. Comme le nombre de lits en hospitalisation complète diminue, et qu'aucun indicateur

ne montre une baisse des besoins en termes de prises en charges, il en découle la nécessité d'améliorer la qualité des hospitalisations et leur gestion, de façon à réduire la durée de séjour.

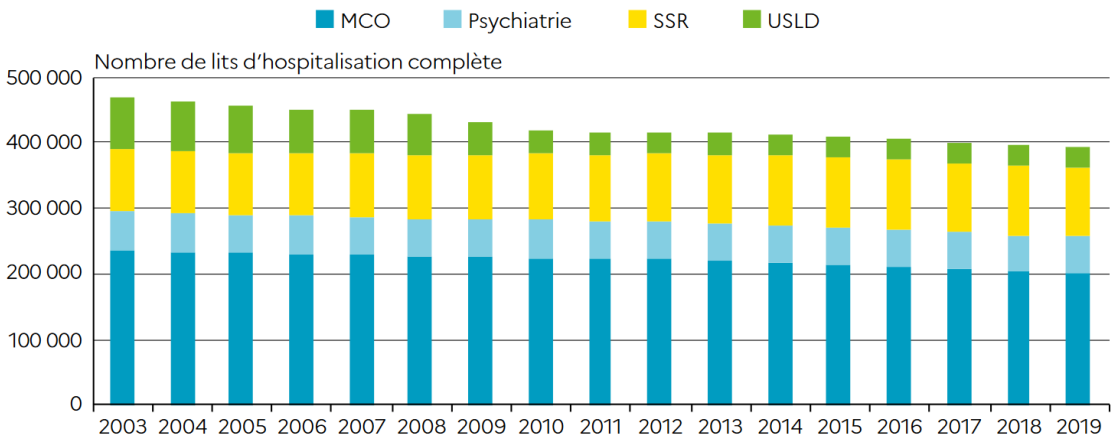


FIGURE 1.2 – Nombre de lits à l'hôpital en France entre 2003 et 2019 [9].

La durée de séjour est donc importante pour la gestion des établissements hospitaliers. Améliorer les performances d'algorithmes qui la prédisent a ainsi des implications plus larges qu'une meilleure organisation des services hospitaliers.

Prédire la durée de séjour peut servir des objectifs différents, induisant différentes cibles de prédiction et différentes façons de les utiliser. Par exemple prédire la durée de séjour est utilisé pour identifier les patients à risque d'avoir une durée de séjour prolongée. Cette identification permet une meilleure allocation des ressources de l'hôpital pour mieux accommoder les besoins des patients. Dans ce cas, la prédiction de la durée de séjour est le plus souvent binaire, car cela suffit pour déterminer si la durée de séjour sera supérieure à une valeur déterminée préalablement, ou non [10-12]. La durée de séjour peut également être prédite pour identifier les facteurs qui sont associés à une réduction ou une augmentation de la durée de séjour, afin d'éclairer les prises de décision en matière de politiques sanitaires. Par exemple, une revue systématique a suggéré que les données liées aux patients dans les services de néonatalogie comme le poids à la naissance, l'âge de la mère et le sexe du nouveau-né suffisent à prédire la durée de séjour [13]. Il a également été trouvé que trois facteurs suffisent à pré-

dire la durée de séjour après une fracture de la hanche : (i) le score de la société américaine des anesthésistes, (ii) le test mental abrégé et (iii) la mobilité [14]. En cardiologie, il a été déterminé que les facteurs les plus importants pour prédire la durée de séjour sont la fréquence cardiaque, la pression sanguine et l'âge [15]. Des facteurs spécifiques à des types d'hospitalisation ou d'opération chirurgicale pourraient donc permettre de prédire la durée de séjour. Cela suggère qu'une prédiction de la durée de séjour doit utiliser des facteurs spécifiques aux caractéristiques des différentes unités médicales et diagnostics, ce qui rendrait difficile l'utilisation du même modèle de prédiction pour tous les patients d'un hôpital. Des facteurs organisationnels ont également été identifiés, comme le type d'hôpital, le ratio entre le nombre de docteurs et d'infirmières, les dépenses et la taille de l'hôpital [16, 17]. Cela montre que la prédiction de la durée de séjour repose sur des données de différents types. Ces données sont diverses et variées comme des données médicales, sociales, démographiques, organisationnelles et administratives, les rendant difficiles à utiliser. Tout cela montre que la prédiction de la durée de séjour a des implications sur l'identification des facteurs permettant d'améliorer les techniques médicales et l'organisation de l'hôpital.

Prédire la durée de séjour est donc un enjeu important pour les hôpitaux, et constitue par conséquent un champ de recherche essentiel en santé publique. De plus, de par la nécessité pour les hôpitaux d'être toujours plus efficaces, ce besoin d'une estimation précise de la durée de séjour ne peut que devenir plus important.

1.2 Bases de données de santé

Les données sur la durée de séjour des patients sont stockées dans une base de données standardisée avec le Programme de Médicalisation du Système d'Information (PMSI). Comme il est connu que les séjours aux urgences influent sur la durée d'une hospitalisation [18], il est également important d'avoir des informations sur les séjours aux urgences précédant une hospitalisation. C'est pourquoi des Résumés de Passages aux Urgences (RPU) sont utilisés conjointement aux données du PMSI. Cette section décrit les Hospices Civils de Lyon (HCL), les bases de données hospitalières en Médecine, Chirurgie et Obstétrique (MCO) utilisées au long de cette thèse, et en détaille notre utilisation.

1.2.1 Les Hospices Civils de Lyon

Les données utilisées dans le cadre de cette thèse proviennent des Hospices Civils de Lyon (HCL) un Centre Hospitalier Universitaire situé dans la ville de Lyon, son agglomération et le Var. La figure 1.3 montre une carte des HCL. Les données utilisées proviennent de 4 groupements dans Lyon et son agglomération. Les établissements des HCL sont entre autre spécialisés en Gériatrie, traitements dentaires, neurologie et neurochirurgie, hématologie, cardiologie, oncologie, pneumologie. Même si les données proviennent de quatre groupements hospitaliers, ceux-ci dont donc constitués de plusieurs centres de spécialités diverses.

En 2020, 1 200 000 journées d'hospitalisation ont été réalisées, 228 710 passages aux urgences et 74 441 opérations chirurgicales. Cette même année, 2496 articles ont été publiés, et 28 projets européens ont été menés [19]. Les HCLs sont le 2ème hôpital de France.

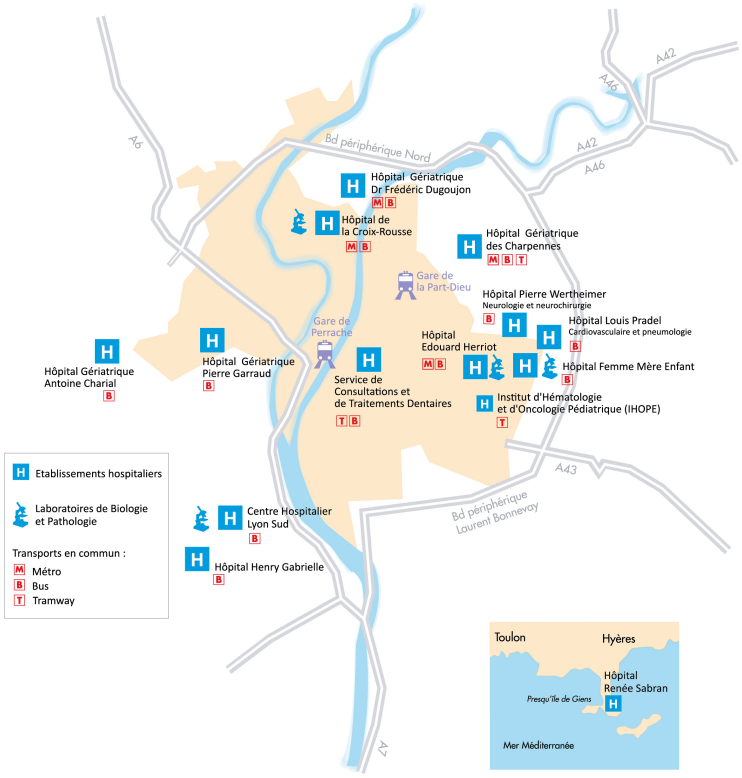


FIGURE 1.3 – Carte des HCL. DirComHCL, CC BY-SA 3.0 *via Wikimedia Commons*

1.2.2 Le Programme de Médicalisation du Système d'Information

Le Programme de Médicalisation du Système d'Information (PMSI) est une base de données standardisée d'informations administratives et médicales sur les séjours des patients [20]. Il a été mis en place dans les années 90 et est inspiré du modèle américain des *Diagnosis Related Groups* (DRG). Dès 1991, tous les hôpitaux français doivent évaluer leurs activités de soin et à partir de 1994, ils doivent transmettre les informations [21] à la Direction Régionale des Affaires Sanitaires et Sociales (DRASS) puis aux Agences Régionales de Santé (ARS). Le PMSI a d'abord été utilisé à des fins épidémiologiques, puis est utilisé à des fins administratives pour le remboursement des actes médicaux, depuis 2005 et l'arrivée de la tarification à l'activité [22].

Le PMSI est utilisé pour plusieurs types d'hospitalisation :

— Les hospitalisations courtes durées en Médecine, Chirurgie et Obstétrique (MCO), qui concernent la majorité des séjours hospitaliers ;

— Les Soins de Suite et Réadaptation (SSR), qui suivent éventuellement une hospitalisation ;

— Les Hospitalisations A Domicile (HAD) ;

— les hospitalisations en psychiatrie (PSY) ;

Comme nous nous intéressons uniquement aux soins aigus à l'hôpital, nous nous sommes focalisés sur les soins en MCO.

Dans le PMSI, chaque hospitalisation est l'objet d'un Résumé de Sortie Standardisé (RSS). Une hospitalisation pouvant être découpée en plusieurs parties selon les mutations du patient dans différentes unités médicales au sein de l'hôpital, le RSS est découpé en Résumés d'Unités Médicales (RUM). Si le patient n'est passé que par une seule unité médicale (UM), on parle de séjour mono-RUM. Une fois le RSS constitué, le séjour est inscrit dans un Groupement Homogène de Malades (GHM) qui contient des prises en charge appartenant aux mêmes catégories médicales. Le regroupement dans un GHM dépend du diagnostic principal et des éventuels diagnostics associés ainsi que de certains actes médicaux qualifiés de « classant ». Le GHM est ensuite converti en Groupement Homogène

de Séjour, auquel correspond un tarif de remboursement. La figure 1.4 illustre ce principe.

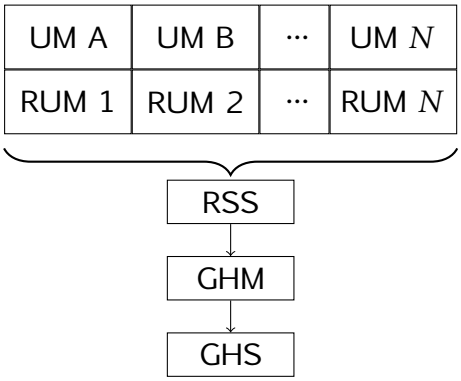


FIGURE 1.4 – Relation entre le Résumé de Sortie Standardisé et les Résumés d'Unités Médicales

La table 1.1 montre le contenu d'un RUM standardisé. La date d'entrée et la date de sortie permettent de calculer la durée de séjour. Chaque séjour est identifié par un identifiant unique `rssnuano`, chaque RUM est identifié par le `rumnu` et chaque patient est identifié par l'identifiant `ippano`.

Les informations sur le mode de sortie et la date de sortie sont censurées car non disponibles au moment de la prédiction de la durée de séjour à l'admission dans une unité médicale. De plus, comme la durée de séjour est calculée en prenant les informations disponibles au début du RUM, de manière à être le plus proche de la réalité, les données ont été organisées de manière à ce que la liste des actes médicaux pour un RUM soit la liste des actes médicaux des RUMs précédents dans le RSS.

TABLEAU 1.1 – Contenu d'un RUM, ainsi qu'une description et la modalité des variables

nom	description	modalité
ippano	identifiant anonymisé du patient	$\mathbb{Z}^+$
rssnuano	identifiant anonymisé du séjour	$\mathbb{Z}^+$

Suite page suivante

nom	description	modalité
rumnu	Identifiant du RUM au sein du séjour	$\mathbb{Z}^+$
hopital	Établissement où s'est déroulé le séjour	A, …,F
umano	Identifiant anonymisé de l'UM	$\mathbb{Z}^+$
typeauto	Type d'unité médicale	réanimation, SI, SC, urgences, palliatif, médecine, chirurgie, gynecobs
sexe	Sexe du patient	H, F
age	Age du patient	$\mathbb{Z}^+$
modent	Mode d'entrée du RUM	Domicile, Urgences, Tsft Autre, Tsft MCO, Tsft SSR, Tsft SLD, Mutation Autre, Mutation MCO, Mutation SSR, Mutation SLD
modsor	Mode de sortie du RUM	Domicile, Urgences, Tsft Autre, Tsft MCO, Tsft SSR, Tsft SLD, Mutation Autre, Mutation MCO, Mutation SSR, Mutation SLD
dentj	Jour de la date d'entrée	1, …, 31
dentmo	Mois de la date d'entrée	1, …, 12
denta	Année de la date d'entrée	2011, …, 2020
denth	Heure de la date d'entrée	0, …, 23
dentmi	Minute de la date d'entrée	0, …, 59
dsorj	Jour de la date de sortie	1, 31
dsormo	Mois de la date de sortie	1, …, 12

Suite page suivante

nom	description	modalité
dosra	Année de la date de sortie	2011, …, 2020
dsormi	Minute de la date de sortie	0, …, 53
dsorh	Heure de la date de sortie	0, …, 23
dp_dr	Diagnostic principal	Classification Internationale des Maladies, $10^{ème}$ édition (CIM-10)
da_dr	Diagnostic associés	CIM-10, Nombre de diagnostiques de 0 à $+\infty$
actes_dr	Actes médicaux	Classification Commune des Actes Médicaux (CCAM) Nombre d'actes de 0 à $+\infty$
igs2	Score de gravité	$\mathbb{Z}^+ \leq 137$

Dans le PMSI, les diagnostics principaux et associés sont représentés sous forme de codes de la Classification Internationale des Maladies, $10^{ème}$ édition (CIM-10). Cette classification représente les maladies de manière hiérarchique, du plus général jusqu'au plus précis. Plus spécifiquement, le premier caractère du code informe sur le chapitre, et les deux suivants permettent de former la catégorie. Les trois lettres d'après informent sur la localisation de la maladie, la sévérité ou d'autres informations supplémentaires. Le code peut se terminer par une extension. Par exemple, dans le code « J45.1 », la lettre J indique le chapitre des maladies de l'appareil respiratoire, le nombre 45 correspond à l'asthme, et le nombre 1 précise que l'asthme est non allergique.

Les actes médicaux sont codées avec une autre classification, la Classification Commune des Actes Médicaux (CCAM) [23]. Cette classification est également

hiérarchique. Les deux premières lettres renseignent sur la localisation de l'acte, tandis que la troisième et quatrième décrivent la technique utilisée et les chiffres apportent des précisions supplémentaires. Chaque RUM a un diagnostic principal et peut avoir un ou plusieurs diagnostiques associés et actes médicaux.

1.2.3 Résumé de Passage aux Urgences

Le résumé de Passage aux Urgences (RPU) contient les données sur le passage aux urgences d'un patient. La table 1.2 montre le contenu d'un RPU quand joint avec les données du PMSI avec la clé unique `rssnuano`.

Tableau 1.2 – Contenu d'un RPU, ainsi qu'une description et la modalité des variables

nom	description	modalité
ippano	identifiant anonymisé du patient	
rssnuano	identifiant anonymisé du séjour	
dp	Identifiant anonymisé de l'UM	
dureeurg	Durée du passage aux urgences	En minutes
da_dr	Diagnostic associés	CIM-10, Nombre de diagnostiques de 0 à $+\infty$
actes_dr	Actes médicaux	Classification Commune des Actes Médicaux (CCAM) Nombre d'actes de 0 à $+\infty$

1.2.4 Autorisations réglementaires

Accéder à ces données a demandé des autorisations institutionnelles. L'accès aux données a été validé par le Comité d'Expertise pour les Recherches, les Études et les Évaluations dans le domaine de la Santé (CEREES) lors de la session du 16 janvier 2020, avec le numéro de dossier TPS 1171550. Le traitement des

données a été également été approuvé par la Commission Nationale de l'Informatique et des Libertés (CNIL), numéro de dossier DR_2020-196, sous réserve d'information des patients sur leurs droits d'opposition par le biais d'affichages. Ces données sont stockées et traitées sur un serveur GNU/Linux des Hospices Civils de Lyon (HCL) avec 128Gio de mémoire vive, 16 cœurs de processeur. L'accès au serveur se fait de manière distante par un protocole assurant confidentialité, intégrité et authentification.

1.2.5 Organisation des données utilisées

La figure 1.5 montre l'organisation chronologique du jeu de donnée et sa complexité, en considérant à la fois les données des hospitalisations (PMSI) et des urgences. La durée de séjour d'un patient à prédire dépend de trois éléments : (i) les étapes précédentes du séjour du patient ; (ii) les éventuels autres séjours du patient ; (iii) les séjours des autres patients concomitants et antérieures.

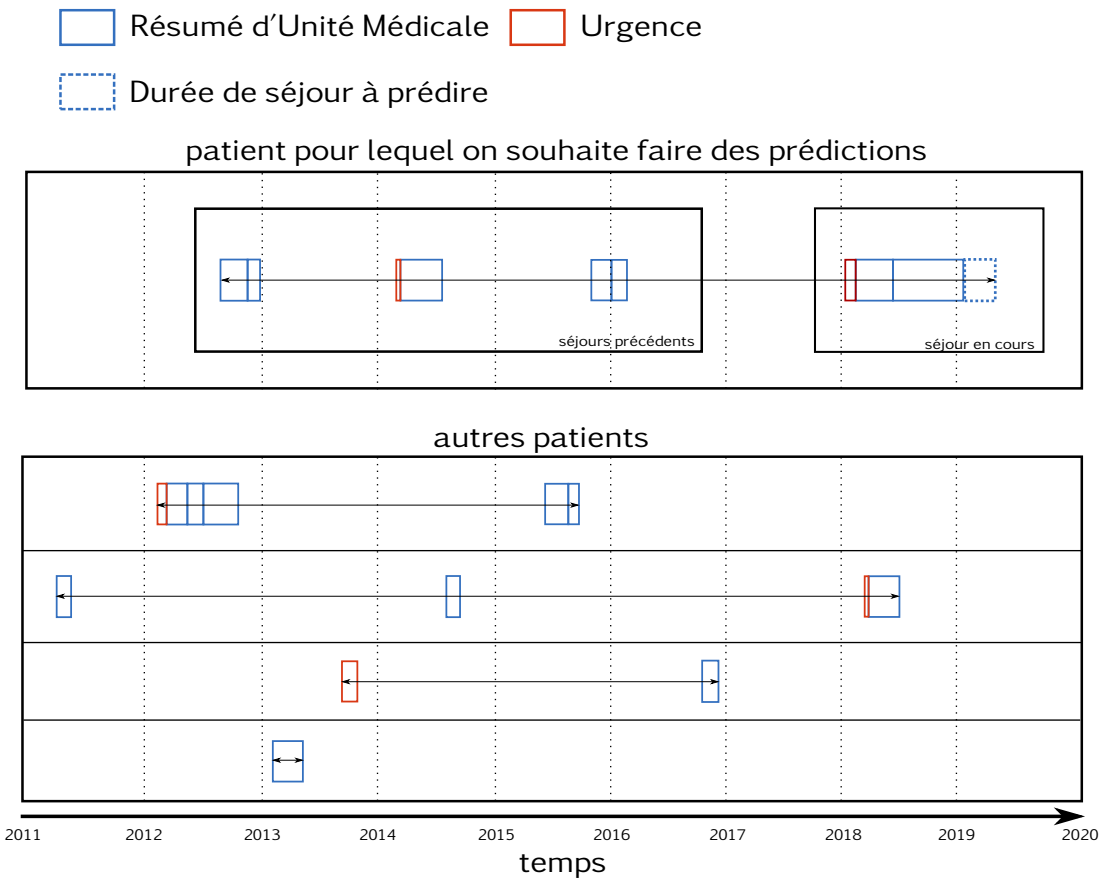


FIGURE 1.5 – Schéma des données utilisées. La durée de séjour prédite dépend des précédentes étapes dans le parcours de soin du patient, de ses hospitalisations antérieures, ainsi que des hospitalisations des autres patients, concomitantes et antérieures.

Le jeu de données étant riche dans sa structure et son contenu, s'appuyer sur des méthodes de prédiction du domaine de l'intelligence artificielle a semblé nécessaire.

1.3 Intelligence artificielle

1.3.1 Une brève histoire de l'intelligence artificielle

Les débuts de l'intelligence artificielle (IA) sont reliés aux débuts de l'informatique. La première machine de traitement mécanique, la machine analytique ainsi que le premier programme pour cette machine sont inventés entre 1837 et 1843 [24]. La machine de Turing, proposée en 1936 [25], décrit un modèle théorique de traitement. Dans ce modèle, un ruban de taille infinie divisé en cellules stockant chacune un symbole est lu par une tête de lecture. La tête de lecture peut lire le contenu des cellules, et le mettre à jour. La machine dispose également d'un registre listant son état ainsi que tous les états possibles. Selon l'état actuel et le symbole lu par la tête de lecture, une table d'instruction permet de choisir s'il faut mettre à jour le contenu de la cellule du ruban, et s'il faut déplacer d'une case la tête de lecture, à gauche ou à droite. Plus formellement, la machine de Turing peut être décrite comme un ensemble composé d'un set de symboles Γ, du symbole actuel S_i, d'un set d'états admissibles Q, de l'état actuel q_i, d'une fonction f de décision décidant de l'action à faire (décaler la tête de lecture à Gauche ou à Droite), du nouvel état $q\prime$ et de l'éventuel symbole à écrire $S\prime : f : S_i, q_i \rightarrow \{G, D\} \times q\prime \in Q \times S\prime \in \Gamma$ [26]. Ce modèle de fonctionnement d'une machine a servi de base de raisonnement pour prouver des propriétés inhérentes à toutes les machines, comme le principe selon lequel il est impossible de déterminer si un programme pourra s'arrêter ou pas.

Les premiers ordinateurs sont arrivés pendant la seconde guerre mondiale, notamment pour être au service de la cryptographie, qui demandait de pouvoir tester un nombre important de possibilités de clés de chiffrement. Dans ce contexte, la machine « Bombe » a été la première machine utilisant des algorithmes heuristiques pour déchiffrer des messages. En particulier, elle utilisait le fait que le contenu de certains messages pouvait être deviné en avance (e.g. prévision météorologique) pour comparer le contenu du message chiffré et du message déchiffré et diminuer le nombre de combinaisons à tester. Cette utilisation montre que la machine peut utiliser des stratagèmes pour la rendre « intelligente » et atteindre des objectifs prédéterminés. La figure 1.6 montre une

photographie d'une « Bombe », où on peut voir ses rotors.

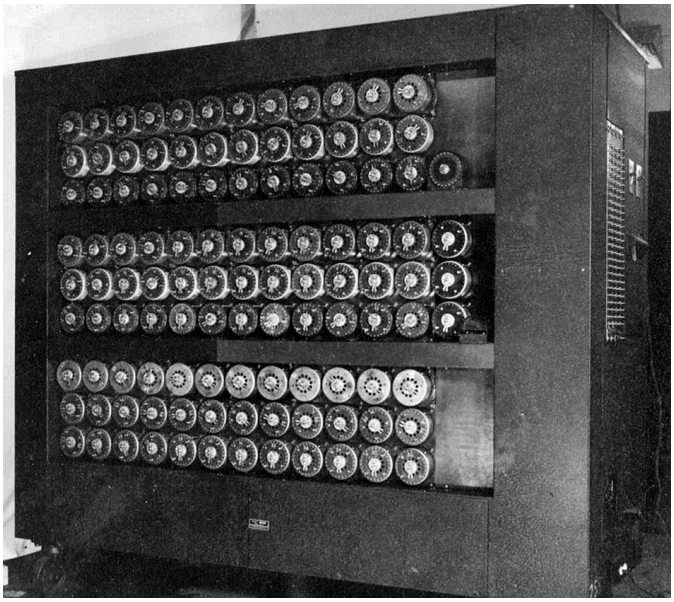


Figure 1.6 – « Bombe », l'un des premiers ordinateurs

Le premier test permettant de définir si une machine peut avoir un comportement intelligent, le test de Turing est créé en 1950 [27]. Dans l'interprétation standard de ce test, un interrogateur parle avec deux participants, et doit déterminer lequel des deux est un ordinateur, mesurant ainsi la capacité d'un ordinateur à imiter un comportement humain, sans toutefois aborder les notions de conscience et de compréhension [28], car une machine peut imiter le langage humain sans pour autant le comprendre ou avoir de conscience propre. Le terme « Intelligence Artificielle » est proposé par John McCarthy lors de l'atelier de Dartmouth de 1956 [29].

Le perceptron [30], le premier neurone artificiel, est inventé en 1958, il permet d'apprendre automatiquement les valeurs de paramètres w et b d'une fonction de seuil. L'équation (1.1) décrit formellement l'usage du perceptron comme fonction de seuil.

$$f(\mathbf{x}) = \sigma(w \cdot x + b) \tag{1.1}$$

où w est un vecteur de poids, x est un vecteur d'entrée, b le biais et σ est une

fonction dite d'activation non-linéaire. Pour obtenir une fonction de seuil ressemblant à un neurone artificiel, la fonction échelon d'Heaviside est utilisée comme fonction d'activation (voir figure 1.7), ce qui permet d'obtenir (1.2) :

$$f(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{w} \cdot \mathbf{x} + b > 0, \\ 0 & \text{sinon} \end{cases} \tag{1.2}$$

L'apprentissage des paramètres se fait de manière itérative en se basant sur la différence entre la sortie du perceptron et la sortie désirée. Les auteurs du perpectron étaient très optimistes sur ses possibilités, même s'il a été critiqué [31] notamment sur le fait qu'un perceptron seul ne puisse pas implémenter une fonction OU exclusif (*X*OR).

L'algorithme de rétropropagation des erreurs utilisant la règle de dérivation des fonctions composées permettant d'utiliser des réseaux de neurones multicouches (aussi appelé apprentissage profond, *deep learning*) a été inventé en 1974 [32] et utilisé expérimentalement en 1986 [33]. Il permet de calculer la valeur de la dérivée partielle des poids de chaque perceptron d'un réseau de neurones multicouche par rapport à une fonction de perte. Une étape clé du développement technique des ordinateurs et des systèmes experts est atteinte en 1997, avec la victoire d'une IA aux échecs face au meilleur joueur de l'époque [34].

L'usage de l'apprentissage profond connait un essor important à partir des années 2010s, notamment grâce à une augmentation progressive de la puissance de calcul [35] et la capacité de stockage des données, mais aussi d'améliorations techniques permettant d'entrainer des réseaux de neurones composés de nombreuses couches. En particulier, la fonction d'activation Unité Linéaire Rectifiée a permis d'améliorer les performances des réseaux de neurones [36] et de remplacer la fonction sigmoid qui a été considérée comme non-optimale car (i) sa moyenne est différente de zéro [37], ce qui n'est pas recommandé dans les réseaux de neurones [38], et (ii) sa sortie sature vers $-\infty$ et $+\infty$, ce qui a pour conséquence des gradients très faibles, réduisant la vitesse de l'entrainement du réseau de neurones. De plus, un changement dans la façon d'initialiser les poids du réseau de neurones a permis d'augmenter la vitesse de convergence

de la descente de gradient [37]. La figure 1.7 montre la fonction ReLU, la tangente hyperbolique (une fonction d'activation populaire), la fonction échelon d'Heaviside et sigmoid.

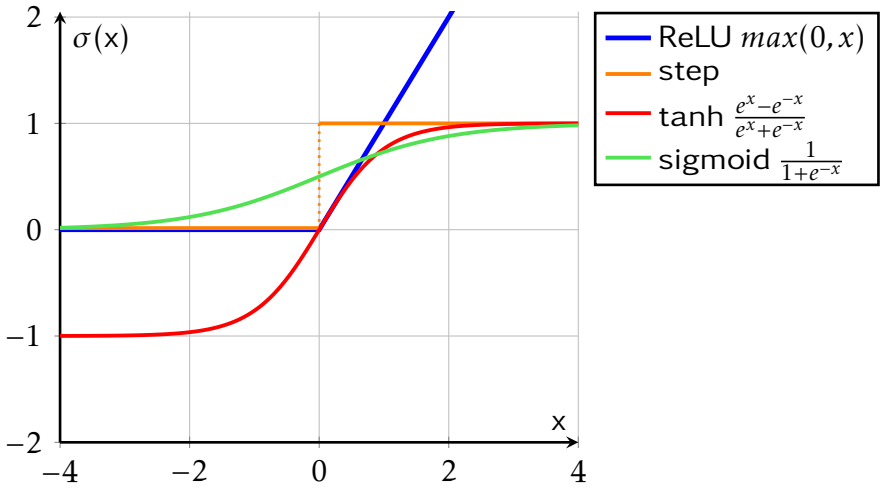


FIGURE 1.7 – Fonctions d'activation Unité Linéaire Rectifiée (*Rectified Linear Unit*, ReLU), échellon d'Heaviside (step), tangente hyperbolique (tanh) et sigmoid sur l'intervalle $[-4, 4]$

Les années 2010s marquent aussi le lancement du jeu de données Imagenet, un jeu de données créé en 2009 [39], qui contient plus de 14 millions d'images appartenant à plus de 20 000 catégories et de la compétition annuelle qui l'accompagne, *the ImageNet Large Scale Visual Recognition Challenge*, utilisant un sous-ensemble de 1000 catégories. La compétition a été marquée en 2012 par l'utilisation d'un réseau de neurones convolutif (*Convolutional Neural Network*, CNN, cf. section 1.3.2) qui a engendré une amélioration des performances de classification des images de 12 % (une exactitude de 0,84) [40], et par un réseau de neurones convolutif avec couches résiduelles en 2015 qui a obtenu une exactitude de 0.96 [41]. A partir de 2017, la majorité des modèles dans la compétition a une exactitude supérieure à 0,95.

L'apprentissage par renforcement, une branche de l'IA dont l'objectif est d'apprendre à un agent à choisir les meilleures actions dans un environnement de manière à maximiser un score, a également bénéficié des améliorations en matière d'apprentissage profond et de réseaux de neurones. L'apprentissage pro-

fond a été utilisé pour estimer la valeur de chaque action en considérant l'état de l'agent, ce qui a permis en 2015 à un agent d'apprendre à gagner à des jeux vidéos sans autres instructions que les règles du jeu [42]. Plus formellement, le réseau de neurones va maximiser la somme cumulée du score r avec une importance dégressive γ, en considérant les actions possibles A et les états possibles S. L'état du jeu est représenté par une capture de l'écran de taille 84x84x4.

$$\max_{\forall s \in S; \forall a \in A} \mathbb{E}\left(\sum_i \gamma_i r_i \mid a; s \right) \tag{1.3}$$

Ce succès a été suivi d'une victoire au jeu de Go contre des joueurs professionnels, d'abord avec AlphaGo, un modèle reposant sur connaissances préalables en 2016 puis avec AlphaGoZero, un modèle connaissant uniquement les règles du jeu en 2017 [43, 44]. Ce travail a été amélioré en laissant le modèle comprendre les règles du jeu et en rendant l'apprentissage possible sur des actions non déterministes [45].

Les travaux présentés dans cette thèse étant centrés sur l'apprentissage profond, il convient d'en détailler le fonctionnement.

1.3.2 Apprentissage Profond (*deep learning*)

L'apprentissage profond (*deep learning*) est une sous-branche de l'apprentissage automatique (*machine learning*) qui elle-même est une sous branche de l'IA. L'apprentissage profond regroupe toutes les utilisations des réseaux de neurones (i.e. perceptrons) avec de nombreuses couches.

Réseaux de neurones vers l'avant

La figure 1.8 présente le fonctionnement d'un réseau de neurones vers l'avant. Un réseau de neurones est constitué de perceptrons organisés sous forme de couches successives. Dans cet exemple, chaque perceptron est connecté à tous les perceptrons de la couche précédente et de la couche suivante. Le nombre de perceptrons dans la dernière couche correspond au nombre de catégories pour une classification. En général, la couche de sortie est constituée d'un unique perceptron dans le cadre d'une régression (prédiction d'un nombre continu).

Les données circulent donc de la couche d'entrée jusqu'à la couche de sortie, d'où le terme de réseau de neurones vers l'avant.

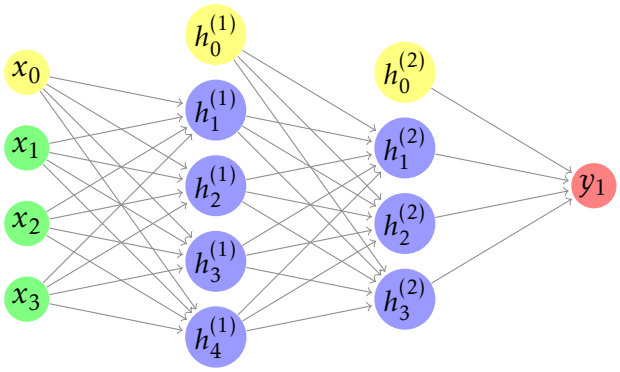


FIGURE 1.8 – Exemple de réseaux de neurones vers l'avant. Les nœuds verts représentent les données d'entrée, les nœuds jaunes représentent les biais des neurones de chaque couche, les nœuds bleus représentent les couches cachées et le nœud rouge représente la couche de sortie.

Descente de gradient et rétropropagation

Pour ajuster les paramètres des perceptrons composants les couches des réseaux de neurones, l'algorithme de descente de gradient est couramment utilisé. Cet algorithme itératif est défini comme (1.4) :

$$w_t = w_{t-1} - \lambda \frac{\partial E}{\partial w_{t-1}} \tag{1.4}$$

où w_t est un paramètre w à l'étape t de la descente de gradient, E est une mesure d'erreur entre la sortie du réseau de neurones y et la vraie valeur et λ est la vitesse d'apprentissage. $\frac{\partial E}{\partial w_{t-1}}$ dépend de E et de $\frac{\partial y}{\partial w_{t-1}}$. Le calcul de la dérivée partielle de la sortie y par rapport à un paramètre w peut être vue sous forme de matrice jacobienne. Pour un réseau de neurones à deux couches cachées $h^{(1)}$ et $h^{(2)}$, le poids $w^{(1)}$ de $h^{(1)}$ de taille n et avec m sorties y est :

$$\frac{\partial y}{\partial w^{(1)}} = \begin{bmatrix} \frac{\partial y_1}{\partial w_1^{(1)}} & \cdots & \frac{\partial y_1}{\partial w_n^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial w_1^{(1)}} & \cdots & \frac{\partial y_m}{\partial w_n^{(1)}} \end{bmatrix} \tag{1.5}$$

$$= \begin{bmatrix} \sum_{i=0}^{m} \frac{\partial y_i}{\partial w_1^{(1)}} & \cdots & \sum_{i=0}^{m} \frac{\partial y_i}{\partial w_n^{(1)}} \end{bmatrix} \tag{1.6}$$

Comme $\frac{\partial y}{\partial w^{(1)}}$ dépend de $\frac{\partial y}{\partial w^{(2)}}$, la mise à jour des paramètres se fait en partant de la sortie jusqu'à l'entrée. C'est pourquoi la propagation de l'erreur se fait en sens inverse de la progression dans le réseau de neurones, d'où le terme rétropropagation.

Réseaux de neurones récurrents et mécanisme d'attention

Un des défauts du réseau de neurones vers l'avant est qu'il ne conserve pas d'état. Il est donc difficile de prendre en compte le contexte historique d'une donnée, et cela oblige à résumer l'information passée plutôt que l'utiliser telle quelle. Les réseaux de neurones récurrents (*Recurrent Neural Networks*, RNN) n'ont pas ce problème et peuvent conserver des informations entre chaque entrée *via* un état caché. Par conséquent, ils peuvent facilement traiter les données organisées sous forme de séquences, comme le langage ou les séries temporelles. La figure 1.9 et les équations (1.7) montrent le fonctionnement d'un réseau de neurones récurrent. Les états cachés sont calculés en fonction de l'état caché précédent, et les sorties dépendent de l'état caché courant. Les 3 vecteurs de poids w_{xh}, w_{hh} et w_y sont utilisés pour l'importance des données d'entrée actuelle à conserver pour l'état caché, l'importance de l'état caché précédent et l'importance des données actuelles pour la sortie, respectivement. Au global, ces poids permettent de déterminer l'importance relative des données d'entrées composant la séquence. La fonction σ est une fonction d'activation non-linéaire.

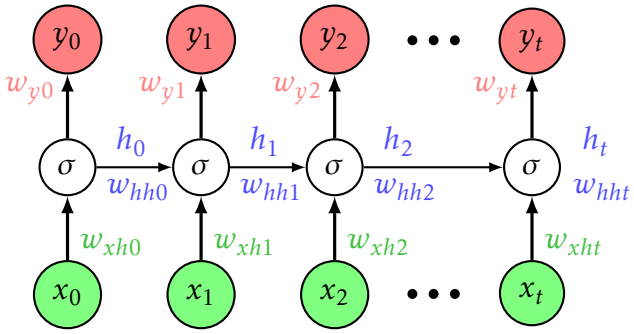


FIGURE 1.9 – Fonctionnement d'un réseau de neurones récurrent pour une séquence à t éléments. Les nœuds verts représentent les entrées successives (e.g. les éléments atomiques d'une séquence), les nœuds rouges représentent les sorties et le texte en bleu représente les états cachés transmettant l'état du réseau. En général, seule la dernière sortie est conservée.

$$y_t = \sigma_y\left(w_y h_t + b_y\right) \tag{1.7a}$$

$$h_t = \sigma_h\left(w_{xh} x_t + w_{hh} h_{t-1} + b_h\right) \tag{1.7b}$$

De par leur nature séquentielle, il est difficile de paralléliser les réseaux de neurones récurrents, ce qui rend leur entrainement chronophage. Une autre méthode, *Attention*, propose de pallier ce problème en répliquant le concept d'attention cognitive, qui détermine comment prioriser certaines entrées par rapport à d'autres dans le cadre du traitement d'un problème [46]. Pour obtenir un vecteur d'attention $A = \alpha_1, \ldots, \alpha_n \in \mathbb{R}$ d'une séquence $S = s_1, \ldots, s_n$ de taille n, le mécanisme d'attention utilise 3 vecteurs de poids : (i) Q représente comment chaque élément dans S est important par rapport aux autres éléments ; (ii) K représente l'importance dans le sens inverse ; (iii) V représente l'importance globale de chaque élément vis-à vis de la séquence S. Avoir 2 matrices Q et K permet une relation non-symétrique entre les éléments d'une séquence : l'importance de x_1 par rapport à x_2 peut être différente de l'importance de x_2 par rapport à x_1. Le calcul de A est obtenu dans (1.8). Le produit scalaire $Q \cdot K$ est divisé par $\sqrt{d}$. Le dénominateur est proportionnel à la taille d des vecteurs d'*embeddings* pour éviter qu'il soit trop grand. L'activation softmax (1.9) permet

de déterminer les valeurs les plus importantes, en plus de normaliser le vecteur, car $\sum_i^n \sigma(s) = 1 \ \forall n \in \mathbb{R}_{>0}$.

$$Attention(Q, K, V) = \sigma\left(\frac{Q \cdot K}{\sqrt{d}}\right)V \tag{1.8}$$

$$\sigma(t_i) = \frac{e^{t_i}}{\sum_{j=1}^{N} e^{t_j}} \tag{1.9}$$

Réseaux de neurones convolutifs

Il est difficile pour un réseau de neurones classique de traiter des images, car les données d'entrées peuvent prendre n'importe quelles valeurs, en fonction du positionnement des éléments composant l'image. Les poids d'un réseau de neurones classique ne peuvent être agnostiques au positionnement des éléments des images. Un autre type de réseau de neurones est donc utilisé pour traiter les images, les réseaux de neurones convolutifs (*Convolutional Neural Network*, CNN) qui s'inspirent du traitement des images par le cerveau. Ce type de réseau repose sur la convolution, qui est un produit scalaire entre un filtre et une image. Cette opération est répétée de manière à ce que le filtre couvre un maximum de la surface de l'image, ce qui permet de détecter des caractéristiques d'une image indépendamment de leurs positions possibles.

La figure 1.10 illustre le principe de fonctionnement d'une convolution, où un filtre passe le long d'une image en la découpant en régions, de manière à détecter des caractéristiques. Le découpage en régions peut optionnellement avoir des chevauchements ou des écarts, et toutes les régions sont de même taille. Pour que des caractéristiques toujours plus complexes soient détectées, les convolutions sont empilées les unes sur les autres, et séparées par une phase de *pooling* qui permet d'accroitre le contraste tout en simplifiant l'image de manière successive.

La figure 1.11 montre un exemple de *max pooling* et d'*average pooling*, où la valeur maximale / moyenne de chaque région d'une image est retenue, augmentant ainsi le contraste entre deux régions de l'image tout en les simplifiant, ce qui accélère l'apprentissage. De la même manière que la convolution, le *pooling* est réalisé plusieurs fois par blocs de manière à couvrir l'intégralité de l'image.

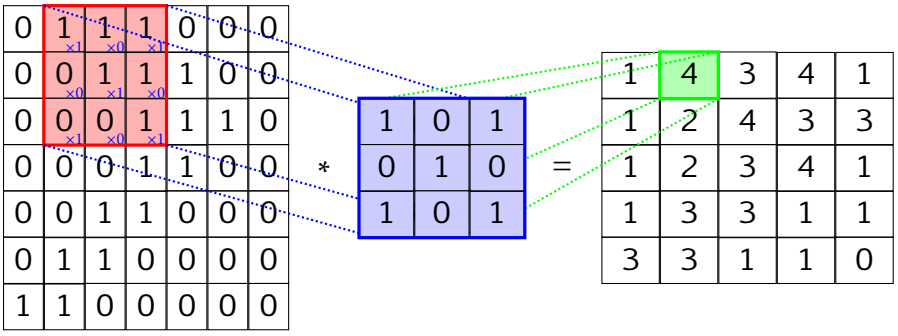


FIGURE 1.10 – Exemple d'une étape d'une convolution. Une image en noir et blanc (à gauche) passe par une convolution de taille 3x3 en bleu. Le filtre est déplacé le long de l'image à chaque étape, ce qui produit la matrice de droite. Le passage par la zone en rouge produit le résultat en vert

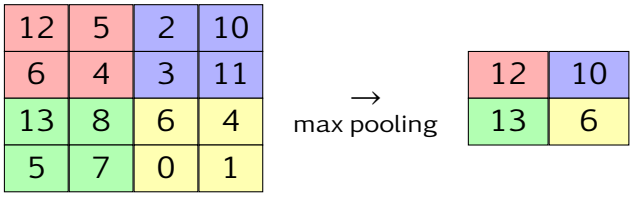
Afin d'aboutir à une prédiction, les étapes de convolution et de *max pooling* sont suivies par un réseau de neurones vers l'avant sur l'image réduite à une dimension.

1.3.3 Représentation des données catégorielles

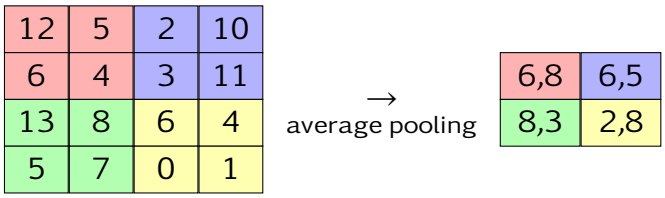
Comme les réseaux de neurones fonctionnent avec des données numériques, les valeurs d'entrée discrètes (catégorielles) sont souvent converties en catégories numériques. Pour cela, la fonction caractéristique (1.10) (aussi appelé *one-hot encoding*) est utilisée pour une variable discrète x dont les valeurs possibles sont dans un set A.

$$\text{car}(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{x} = a, \\ 0 & \text{sinon} \end{cases} \quad \forall a \in A \tag{1.10}$$

Cependant, cette représentation est volumineuse et creuse car elle engendre un vecteur de la taille de A, comporte beaucoup d'éléments vides, et ne permet pas d'obtenir une représentation sémantique des catégories. C'est pourquoi une autre représentation, s'appuyant sur les *embeddings*, est utilisée. Dans cette représentation, les catégories sont transformées en vecteurs numériques. Les valeurs des catégories correspondent à des vecteurs d'*embeddings* différents. A une variable $A = a_1, \ldots, a_n$ correspond donc des vecteurs $v_1, \ldots, v_n$ d'une taille



(a) Exemple de *max pooling*



(b) Exemple d'*average pooling*

FIGURE 1.11 – Une image en niveaux de gris 16 bits de dimension 4x4 passe par un *max pooling* ou un *average pooling* 2x2, divisant globalement le nombre de pixels par 4.

arbitraire et identique quelle que soit la valeur de A. Les valeurs des vecteurs sont initialisées de manière aléatoire, puis modifiées pendant l'entrainement du réseau de neurones au même titre que les poids de ce dernier. Contrairement à la fonction caractéristique, la taille de la représentation numérique ne dépend pas du nombre de valeurs possibles dans A. Elle permet donc une représentation compressée d'une variable, en plus d'en fournir une représentation riche et dense car sur plusieurs dimensions (i.e. la taille du vecteur) et avec aucun élément vide, et de rapprocher numériquement les concepts similaires. Par exemple, des *embeddings* encodant des mots ont appris à représenter des concepts et déterminer des relations implicites entre eux, comme la relation entre un pays et sa capitale [47].

1.3.4 L'apprentissage profond en santé

L'intelligence artificielle et l'apprentissage profond fournissant des méthodes algorithmiques génériques, agnostiques à la nature du problème, ils ont de nombreuses applications dans le domaine de la santé [48]. De plus, l'utilisation de l'IA en santé connaît un intérêt croissant depuis plusieurs décennies [49].

L'utilisation de réseaux de neurones convolutifs a permis d'améliorer les performances pour la détection de la tuberculose avec des radiographies [50], en détection du cancer de la peau [51] et en identification de la rétinopathie diabétique [52] ; pathologies pour lesquelles des données d'imagerie peuvent suffire à l'établissement d'un diagnostic. La création des scores de risques a également été influencée par l'arrivée de méthodes complexes en intelligence artificielle. Par exemple, un autoencodeur a été utilisé pour estimer l'âge biologique d'une personne [53], et un CNN pour le risque lié à la maladie d'Alzheimer [54]. Ces scores ont respectivement été évalués en regardant la corrélation avec le risque d'hypertension et de diabète, et en s'assurant que le taux d'erreur d'identification des risques reste bas tout en identifiant correctement tous les risques avérés. Les réseaux de neurones récurrents et le système d'attention ont été très utilisés. Par exemple, ils ont été utilisés pour apprendre à représenter des données sur les patients sous forme de vecteurs numériques d'*embeddings* [55]. Les *embeddings* peuvent ensuite être utilisés en tant que variables d'entrées pour d'autres tâches.

L'IA permet également de faciliter la recherche dans le domaine biomédical. L'utilisation d'algorithmes identifiant des concepts dans des articles scientifiques, sous forme de graphes [56] ou de vecteurs d'*embeddings* [57, 58] permet de construire des moteurs de recherche et des outils identifiant des rapprochements thématiques entre articles.

Les réseaux de neurones récurrents (RNN) ont été utilisés pour prédire la durée de séjour, où chaque étape du réseau de neurones représente un événement dans une hospitalisation comme par exemple la prise d'un médicament, la réalisation d'un acte médical ainsi que les mutations et transferts [59]. Les RNNs ont également été utilisés pour prédire la durée de séjour à partir de séries temporelles d'indicateurs biologiques et physiologiques, comme les électrocardiogrammes ou les signes vitaux [60, 61].

Il est difficile d'utiliser un RNN pour représenter uniquement les mutations (transfert intra-hospitalier) (i.e. les RUM), car une majorité de patients ont des séjours hospitaliers mono-RUM, ce qui rend le concept d'état des RNNs moins intéressant pour ces séjours. Dans le contexte de la prédiction des durées de séjours toutes hospitalisations confondues, il est donc nécessaire d'avoir une granularité plus fine que les RUM pour bénéficier de la représentation des RNNs. Sans cette granularité, et comme les CNNs ne sont pas adaptés à des données tabulaires, un réseau de neurones vers l'avant semble donc être le plus adapté dans ce contexte.

L'apprentissage profond est donc utilisé en santé pour des tâches diverses, notamment pour prédire la durée de séjour à partir de données administratives, biologiques ou physiologiques.

Chapitre 2

Objectifs

La durée de séjour est un indicateur clé pour la gestion des hôpitaux. La prédire précisément et correctement est un enjeu important, pour lequel des solutions techniques basées sur les sciences de la donnée sont requises. L'intelligence artificielle est un domaine scientifique large et récent dont l'évolution est rapide, en particulier depuis la dernière décennie. Les techniques d'apprentissage profond sont utilisées dans de nombreux domaines, dont la santé. Il convient d'explorer comment l'intelligence artificielle peut contribuer à améliorer la qualité des prédictions de durée de séjours, permettant ainsi de potentiellement améliorer la qualité des soins. L'objectif de la thèse est de donc développer une méthode de prédiction des durées de séjours pour tout type de patients et d'hospitalisation, fonctionnant à chaque étape d'un séjour avec des données médico-administrative standardisées. Afin d'organiser le travail, cet objectif principal a été scindé en trois sous-objectifs réalisés dans un ordre chronologique.

2.1 Faire un état de l'art de la littérature

Comprendre les spécificités des méthodes utilisées, tant pour la préparation des données que la prédiction ou l'évaluation des performances, était indispensable. Il était donc cardinal de commencer par une revue systématique de la littérature sur le sujet. En particulier, nous avons cherché à comprendre les ten-

dances en matière de méthodes de prédiction et de schémas d'étude pour identifier des bonnes pratiques et méthodes populaires. Même si comparer des études entre elles est toujours une tâche complexe comportant de nombreux écueils, identifier les techniques de prédiction et d'évaluation les plus prometteuses sur lesquelles baser notre réflexion était important. Proposer une méthode de prédiction des durées de séjour sans connaître et identifier les précédentes études sur le sujet aurait pu rendre cet exercice beaucoup plus difficile.

2.2 Proposer une méthode innovante pour prédire la durée de séjour

Après avoir effectué une revue de la littérature et compris les principaux écueils méthodologiques, ainsi qu'avoir obtenu l'autorisation d'accéder et de traiter les données, il convenait de développer notre propre méthode de prédiction des durées de séjour, incluant à la fois la préparation des données, l'utilisation d'une méthode d'apprentissage et une évaluation rudimentaire des performances. Plus spécifiquement, nous avons voulu comprendre (i) comment préparer un jeu de données de manière à le rendre utilisable par un réseau de neurones et (ii) comment les techniques de représentation des données par vectorisation des variables catégorielles des bases de données hospitalières administratives peuvent améliorer les performances des réseaux de neurones vers l'avant.

2.3 Évaluer les performances de la méthode de prédiction et la comparer avec l'état de l'art

Afin d'avoir confiance en l'outil de prédiction et ainsi en faciliter l'adoption, rigoureusement évaluer ses performances et les comparer avec d'autres algorithmes mentionnés dans notre revue de la littérature était indispensable. De plus, des pistes d'améliorations ont été identifiées lors du développement de la méthode de prédiction, et il convenait de déterminer si ces pistes d'améliorations apportaient de réels bénéfices.

Chapitre 3

État de l'art de la littérature

Nous avons souhaité commencer ce travail de thèse par un tour d'horizon des méthodes déjà utilisées pour comprendre les tendances méthodologiques en matière de prédiction des durées de séjours, afin de pouvoir ensuite proposer notre propre méthode. Plus précisément, nous nous sommes intéressés à comment les jeux de données sont construits, quelles variables sont retenues et comment elles sont prétraitées de manière à pouvoir être utilisées dans le cadre d'une prédiction, quelles méthodes de prédictions sont employées, quels sont les schémas d'études et mesures de performances.

Nous avons donc effectué une revue systématique de la littérature selon la méthodologie PRISMA [62], qui est une méthode pour conduire des revues systématiques de la littérature. Une recherche de mots clés autour de la prédiction de la durée de séjour a été réalisée sur les bases de données PubMed, ScienceDirect et arXiv, permettant une recherche à partir de sources hétérogènes. Suite à cette recherche, la sélection d'articles a été réalisée par deux personnes selon des critères définis préalablement. Les désaccords sur les inclusions d'articles ont été résolus par l'intermédiaire d'une troisième personne. Nous avons choisi les variables d'intérêt que nous voulions extraire pour tous les articles retenus, puis procédé à une analyse de données servant à répondre à la problématique. L'analyse de données comportait une description des données extraites, des visualisations et des tests statistiques. Cela nous a permis de dégager du sens des données extraites, d'en tirer des conclusions et d'apporter des perspectives sur

les choix méthodologiques présentés dans les articles sélectionnés.

Article 1 : Vincent Lequertier et al. « Hospital Length of Stay Prediction Methods : A Systematic Review ». *In* : *Medical Care* 59.10 (oct. 2021), p. 929-938. ISSN : 0025-7079. DOI : 10.1097/MLR.0000000000001596

Présentation orale lors d'une réunion scientifique : Vincent Lequertier. « Méthode globale de prédiction des durées de séjours avec intégration des données incrémentales et évolutives ». Réunion scientifique RESHAPE. Lyon, 9 oct. 2020

Hospital Length of Stay Prediction Methods
A Systematic Review

Vincent Lequertier, MEng,†‡ Tao Wang, PhD,§ Julien Fondrevelle, PhD,‡ Vincent Augusto, PhD,‖ and Antoine Duclos, MD, PhD*†*

Objective: This systematic review sought to establish a picture of length of stay (LOS) prediction methods based on available hospital data and study protocols designed to measure their performance.

Materials and Methods: An English literature search was done relative to hospital LOS prediction from 1972 to September 2019 according to the PRISMA guidelines. Articles were retrieved from PubMed, ScienceDirect, and arXiv databases. Information were extracted from the included papers according to a standardized assessment of population setting and study sample, data sources and input variables, LOS prediction methods, validation study design, and performance evaluation metrics.

Results: Among 74 selected articles, 98.6% (73/74) used patients' data to predict LOS; 27.0% (20/74) used temporal data; and 21.6% (16/74) used the data about hospitals. Overall, regressions were the most popular prediction methods (64.9%, 48/74), followed by machine learning (20.3%, 15/74) and deep learning (17.6%, 13/74). Regarding validation design, 35.1% (26/74) did not use a test set, whereas 47.3% (35/74) used a separate test set, and 17.6% (13/74) used cross-validation. The most used performance metrics were R^2 (47.3%, 35/74), mean squared (or absolute) error (24.4%, 18/74), and the accuracy (14.9%, 11/74). Over the last decade, machine learning and deep learning methods became more popular ($P = 0.016$), and test sets and cross-validation got more and more used ($P = 0.014$).

Conclusions: Methods to predict LOS are more and more elaborate and the assessment of their validity is increasingly rigorous. Reducing heterogeneity in how these methods are used and reported is key to transparency on their performance.

Key Words: data analysis, epidemiology, heath service research, quality of care, decision-making

(*Med Care* 2021;59: 929–938)

Length of stay (LOS) prediction accuracy is critical for hospital management and bed capacity planning, which influences health care delivery access, quality, and efficiency.[1,2] In addition to blocking and wasting inpatient bed days, incorrect prediction can jeopardize medical services and cause the dissatisfaction of patients and health care professionals. Conversely, accurate LOS prediction allows better resource allocation and care organization from patient admission to discharge preparation.[3,4]

Human-made LOS prediction is poorly reliable because of lack of background information on patients or heterogeneousness among health care professional opinions.[5,6] A patient may be assigned different LOS estimates depending on the person making the prediction. Drawing from this conclusion emerged the potential value of automated predictions. A lot of research aimed to model LOS and determine what statistical technique would provide the best predictions.

Previous reviews have covered the topic of LOS prediction in specific contexts of care: intensive care or neonatal units,[7,8] cardiac surgery,[9,10] or thermal burns.[11] Another review focused on a particular aspect of modeling, such as risk adjustment.[12] To our knowledge, no systematic review attempted to consider this topic from a broader spectrum, without restrictions related to the context of care or prediction methods. We sought to establish the overall picture in scientific literature of the LOS prediction methods based on available hospital data and the related study protocols designed to measure their performance.

MATERIALS AND METHODS

Search Strategy

This systematic review was conducted under the Preferred Reporting Items for Systematic Reviews and Meta-Analyses[13] (PRISMA) framework. Search queries for English articles were issued on PubMed, ScienceDirect, and arXiv from their inception until September 10th, 2019. Articles on LOS prediction were sourced from those platforms to consider both

From the *Research on Healthcare Performance (RESHAPE), Université Claude Bernard Lyon 1, INSERM U1290; †Health Data Department, Lyon University Hospital, Lyon; ‡Univ Lyon, INSA Lyon, Université Claude Bernard Lyon 1, Univ Lumière Lyon 2, DISP, EA4570, 69621 Villeurbanne, France; §University of Lyon, INSA Lyon, Université Claude Bernard Lyon 1, Univ Lumière Lyon 2, UJM-Saint-Etienne, Decision and Information Systems for Production systems (DISP), Villeurbanne Cedex; and ‖Mines Saint-Etienne, University of Clermont Auvergne, CNRS, UMR 6158 LIMOS, Centre CIS, Saint-Etienne, France.

Supported by the European Research Ambition Pack 2018 grant, distributed by the French Auvergne-Rhône-Alpes region.

The authors declare no conflict of interest.

Correspondence to: Vincent Lequertier, MEng, 162 Avenue Lacassagne, Batiment A, 6ème étage, Lyon 69003, France. E-mail: vincent.lequertier@chu-lyon.fr.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

ISSN: 0025-7079/21/5910-0929

medical and computer science perspectives. For PubMed, the Medical Subject Heading (MeSH) terms were "Length of Stay," "Hospitals," and "Forecasting," whereas "length of stay," "length of hospitalization," and "length of hospital stay" were searched in the title and "predict," "hospital" in the title and the abstract. For ScienceDirect, the terms "length of stay," "length of hospitalization," and "length of hospital stay" were searched in the title and "predict," "hospital" in the title, abstract, and keywords. For arXiv, the articles had to contain "length of stay prediction" in the title or the abstract. The exact search terms are available in Supplementary Table 1 (Supplemental Digital Content 1, http://links.lww.com/MLR/C279).[1]

After gathering the search results, duplicates were removed by ensuring that the punctuation-free, lowercased titles of all publications were unique. Then, the abstracts were independently reviewed by 2 investigators (T.W. and V.L.). Disagreements on abstract inclusion were straightened out by a third party (A.D.), who reviewed the publications excluded by only 1 of the 2 original reviewers. After abstract screening, the study selection was further refined by considering the full text version of preselected articles. Inclusion criteria were as follows: (1) the articles must have been original; (2) the article goal must have been to provide LOS prediction (eg, studies investigating LOS determinants); (3) the population size must have been > 1000 to make sure prediction methods may be applicable to a large patient cohort; (4) predictions had to be made for patients individually, not for an aggregated population; (5) the articles' scope must have been broad enough and include methods applicable to acute care with length of stay measured in days (eg, excluded articles were those focusing on stays in emergency departments, rehabilitation, or psychiatric care centers); (6) the predictions must have been made by a computer program, not by humans and the data must come from a database, not a survey.

Literature Data Extraction and Analysis

For each included article, the following information was extracted using a standardized protocol with categories determined a priori: journal of the published article with its corresponding Web of Science category, publication date and country, population setting and study sample with number of hospitals and inpatient stays, data sources and input variables used by the prediction models, reimputation strategies for missing data, LOS modeling format with potential transformations, validation study design, employed LOS prediction methods, and performance evaluation metrics.

Two main data sources were considered. Administrative data were those collected for billing and care reimbursement purposes (eg, Diagnosis-related Groups like datasets), whereas clinical data referred to routine care information stored in Electronic Health Records (eg, biological examinations or drug prescriptions). Utilized variables from those data sources were gathered into 12 categories: (1) patient administrative data, (2) patient demographic and anthropometric, (3) patient diagnoses and medical history, (4) patient care with performed procedures, (5) biological examinations and physiological parameters, (6) drugs administered, (7) adverse events occurrence, (8) patient risk scores, (9) timing and frequency of patients admissions, (10) hospitals' characteristics, (11) health care professionals characteristics, and (12) clinical notes.

Prediction methods were classified into 3 categories: (1) regression model depicted statistical analyses trying to find an association between LOS and input variables, with different regression types according to LOS distribution, (2) machine learning encompassed a set of methods learning how to predict LOS with minimal human intervention and allowed complex relationships between input variables, (3) deep learning was a subcategory of machine learning algorithms, which used deep artificial neural networks and leveraged a rich representation of the input variables to handle intricate datasets.

Regarding the validation study assessing the prediction method performance, 3 designs were possible: (1) no split design corresponded to a validation based on the same dataset for training the model and evaluating its predictions. (2) Train-test split design corresponded to a validation based on a randomly selected subset of the initial dataset for evaluating the predictions of the model (test set) while the model has been developed from another part of the initial dataset (train set). (3) Cross-validation[14] split design was close to the train-test approach, but to avoid potential bias in selection of the test set, one could evenly divide the dataset in k non-overlapping parts and select k times a different part as the train or test sets.

The metrics used to estimate the model performance included: (1) the coefficient of determination (R^2) as the sum of squared residuals divided by the total sum of squares, (2) mean squared error as the mean squared sum of residuals, (3) mean absolute error as the mean of the absolute sum of residuals, (4) accuracy as the number of correctly classified samples, (5) sensitivity as the number of correctly classified positives cases among all positives cases in a binary classification, (6) and the area under the curve (AUC) score.[15]

The data extracted from the selected articles was analyzed using the Python programming language,[16] version 3.7, the "Pandas" library[17] version 1.0 for data manipulation, "matplotlib"[18] version 3.2 for producing the figures, and "tableone"[19] version 0.7.6 was used for creating Table 1. The data analysis was done using descriptive statistics and χ^2 tests were used for independence testing.

RESULTS

Population Settings and Study Samples

The search queries yielded 2033 results (1744 from PubMed, 268 from ScienceDirect, and 21 from arXiv), from which 74 articles (62 from PubMed, 4 from ScienceDirect, and 8 from arXiv) were selected in this review (Fig. 1) across the fields of medicine (48.7%, 36/74), computer science (24.3%, 18/74), public health (20.3%, 15/74), or multidisciplinary fields (6.8%, 5/74) (Table 1). Studies between 1972 and 2010 accounted for 35.1% (26/74) of the corpus, while 64.9% (48/74) of the articles had been published between 2010 and 2019.

Overall, 29.7% (22/74) of the articles did not report the number of hospitals and 2.7% (2/74) did not report the number of inpatient stays in the study sample. Accordingly, 53.8% (28/52) were monocentric and the known median

TABLE 1. Extracted Data Summary (n = 74)

Variables	N (%)
Web of Science category	
Medicine	36 (48.65)
Computer	18 (24.32)
Public health	15 (20.27)
Multidisciplinary	5 (6.76)
Time period	
2010–2019	48 (64.86)
1972–2010	26 (35.14)
Continent	
North America	49 (66.22)
Europe	12 (16.22)
Oceania	6 (8.11)
Asia	6 (8.11)
Africa	1 (1.35)
Population setting	
Monocenter	31 (41.89)
Multicenter	24 (32.43)
Unknown	19 (25.68)
No. inpatient stays, median [min, max]* (missing N = 2)	16292 [1065, 3517950]
Inclusion period duration in months, median [min, max] (missing N = 3)	59 [1, 215]
Population selection criteria based on specific diagnoses†	
Yes	38 (51.4)
No	36 (49.6)
Data sources	
Clinical	37 (50.00)
Administrative	25 (33.78)
Clinical and administrative	12 (16.22)
Length of stay format	
Continuous	40 (54.05)
Continuous with log or polynomial transformation	16 (21.62)
Categorical	18 (24.32)
No. input variables, median [min, max]	8 [1, 26]
No. categories of input variables for predictions, median [min, max]	4 [1, 8]
Categories of variables input for prediction, n (%)	
Patient administrative data‡	40 (54.05)
Patient demographics and anthropometrics§	44 (72.97)
Patient diagnoses and medical history‖	54 (60.81)
Patient care with performed procedures¶	28 (37.84)
Patient biological exams and physiological parameters‖,#	29 (39.19)
Patient drugs administered**	10 (13.51)
Patient adverse events occurrences††	10 (13.51)
Patient risk scores‡‡	37 (50.00)
Timing and frequency of patient admissions§§	20 (27.03)
Hospital characteristics‖‖	13 (17.57)
Health care professional characteristics¶¶	7 (9.46)
Clinical notes##	2 (2.70)
Used a reimputation method for missing variable	14 (18.92)
Timing of the prediction	
Before patient admission	3 (4.05)
At patient admission	22 (29.73)
Throughout inpatient stay	28 (37.84)
Unknown	24 (32.43)
No. prediction methods used, median [min, max]	1 [1, 7]
Used a regression method	48 (64.86)
Used a machine learning method	15 (20.27)
Used a deep learning method	13 (17.57)
Validation study design	
No split	26 (35.14)
Train/test	35 (47.30)
Cross-validation	13 (17.57)
Test dataset percentage, median [min, max] (missing N = 15)	10 [0, 50]
No. performance evaluation metrics, median [min, max]	1 [1, 6]
Performance evaluation metric	
R^2	35 (47.30)
Accuracy	11 (14.86)
Mean absolute error	9 (12.16)
Mean squared error	9 (12.16)
Area under the curve	8 (10.81)
Sensitivity	5 (6.76)

*Studies may have reported the number of patients, number of stays, or both. If only the number of patients was reported, the number of stays was assumed to be the number of patients.

†Studies with inclusion criteria based on specific diagnoses[10,20–54] or without inclusion criteria based on specific diagnoses.[7,12,55–90]

‡Administrative data about the patient stay. Example: discharge location.[7,20,22,24–26,28,30–32,35,40,41,43–46,48,49,52,53,55–57,63,67,69–71,74,76,77,80,81,83,84,86,88–90]

§Patient demographics and anthropometrics, for example, the age.[7,20–26,28–32,34,35,38–41,43–49,51–55,57,58,63,65,67–72,74,76,78,79,81–83,85–88,90]

‖Patient diagnoses and medical history, for example, ICD-10 codes.[7,20–26,30,31,39,40,42–44,46–49,51–57,63,64,66–70,72,74,76,79–84,86,89,90]

¶Patient procedures, for example, type of surgery.[7,21–23,26,30,32,34,41–44,46,48,52,56,65–67,69,72,74,78,81,84,86,87,90]

#Patient biological exams and physiological parameters, for example, heart rate.[22,24,26,28,30,38,43,46,51,54,56–60,62,63,65–67,69,70,73,82]

**Patient drugs administered.[20,32,46,48,49,51,66,76,81,84]

††An event occurring during the hospital stay, for example, development of complications.[21,23,24,26,30,42,47,48,55,81]

‡‡Patient risk scores, for example, Injury Severity Score.[7,21,24–26,29,33,34,36–38,41,43,46,49,50,54,58,59,63–65,69,72–75,78–81,83,85,87–89,91]

§§Data about the timing, for example, the day of the admission.[25,32,40,43,45,48,49,55,57,64,66,69,70,79,83,84,86,87,89,92]

‖‖Hospital data, for example, whether a hospital is public or private.[39,45,49,52,54,64,68,70,77,84–86,90]

¶¶Data about the health-care professionals who care for the patient, for example, the surgeon.[39,40,44,45,83,84,86]

##Clinical notes.[27,66]

dataset size from these studies was 16,292 stays with an inclusion period duration of 4.9 years. Regarding population selection, 51.4% (38/74) of the studies had an inclusion criteria based on a principal diagnosis (eg, congestive heart failure,[51] traumas,[38] burns,[47] spinal cord injury,[21] or acute myocardial infarction[22]) and 29.7% (22/74) on the patient's age. Furthermore, 24.3% (18/74) of them excluded any stays in which an input variable was missing; 18.5% (24/74) excluded patient stays based on their pathway (eg, patient transfer from or to another hospital, or readmission); and 13.5% (10/74) excluded stays in which the patient died. To further reduce LOS variance, 8.1% (6/74) of the studies also excluded LOS outliers[23] and 4.1% (3/74) excluded patients admitted in specific hospital wards.[76]

Data Sources and Input Variables

The data sources were only clinical for 50.0% (37/74) of the studies, only administrative for 33.8% (25/74) of the studies, and both administrative and clinical for 16.2% (12/74). Figure 2 displays the inclusion period of every study according to related data sources and multicenter design. While the density of studies seemed to increase over time, there were no temporal trends with regard to the durations of inclusion periods (χ^2, $P = 0.1$) or the data sources used (χ^2, $P = 0.8$). However, there were fewer multicenter studies over the years (χ^2, $P = 0.006$).

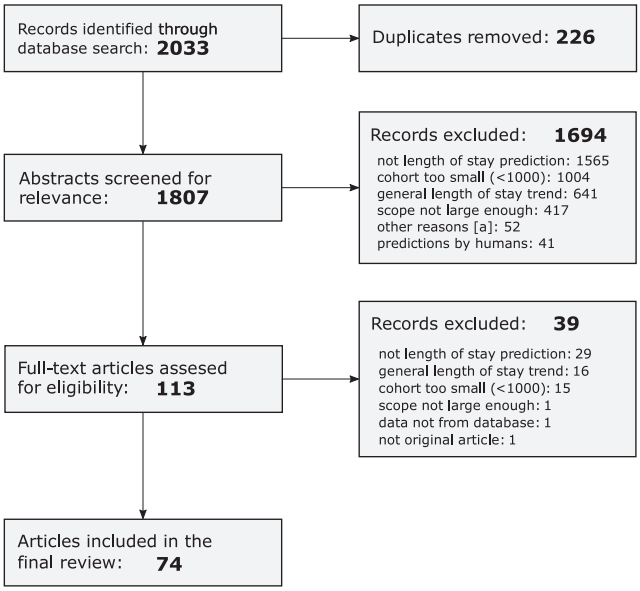


FIGURE 1. Flowchart of the study selection. Exclusion criteria were not mutually exclusive.

LOS modeling format was a continuous variable without transformation, a continuous variable with transformation (logarithmic or polynomial), or a discrete variable in 54.1% (40/74), 21.6% (16/74), and 24.3% (18/74) of the articles, respectively. Among the 24.3% (18/74) of articles considering LOS as discrete, the goal of prediction was to distinguish between short versus long stays.

Distribution of input variables across predefined categories is presented in Table 1. Overall, there was a median of 8 variables used per article: 31.1% (23/74) of articles used between 1 and 5 variables, 35.1% (26/74) between 5 and 10, 18.9% (14/74) between 10 and 15, and 14.9% (11/74) used more than 15 variables. The most used variables were the patient age (68.9%, 51/74 of the articles), sex (39.1%, 29/74), ethnicity (17.6%, 13/74), insurance status (14.9%, 11/74), and diagnosis (13.4%, 10/74). Only 1 study of 74 attempted to predict LOS without using individual patients' data,[77] while another study highlighted the importance of including the time elapsed since the last admission for each patient in the input variables.[57]

Reimputation of missing variables was employed in 18.9% (14/74) of the articles, as follows: carry over the last observation,[57–63] the observed mean as a replacement value,[20,43,45,59,62,63] machine learning technique,[20,25,59] and regression estimate.[22,44]

Prediction Methods, Study Designs, and Performance Metrics

Table 2 summarizes the methodological approaches employed for prediction. Although 74.3% (55/74) of articles reported the use of only one method, 25.7% (19/74) reported 2 methods or more. In detail, 64.9% (48/74) of articles used the regression method, 20.3% (15/74) machine learning method, and 17.6% (13/74) the deep learning method.

Utilization of those prediction methods evolved over time (Fig. 3A), with an increase in machine learning contribution over regression modeling before and after 2010 (χ^2, $P = 0.016$). When reported in the publication, prediction was made before patient admission (6.0% of studies, 3/50), at admission 44.0% (22/50) or continuously during hospital stay (56.0%, 28/50); 3 articles out of 50 (6.0%) made predictions at multiple timings.

Regarding validation study design, 35.1% (26/74) of the articles used the same dataset for models' training and evaluation; 47.3% (35/74) used a separate train and test sets; and 17.8% (13/74) used cross-validation. Those evolved (Fig. 3B) toward more rigorous study designs over time, with an increase in randomly splitting datasets either with a train-test or cross-validation (χ^2, $P = 0.014$). Several studies also used bootstrapping for features selection and internal calibration of their prediction models.[25,30,44]

Employed metrics to evaluate the performance of prediction methods were adapted to the LOS format. Performance was mostly assessed with R^2 (47.3%, 35/74) with a median of 0.31 (range, 0.02 to 0.84), the mean squared error (12.2%, 9/74) with a median of 0.37 (range, 0.035 to 1514.09), or the mean absolute error (12.2%, 9/74) with a median of 4.68 (range, 0.8 to 94) among studies considering LOS as a continuous variable, and with accuracy (14.9%, 11/74) with a median of 0.80 (range, 0.35 to 0.96), the AUC curve (10.8%, 8/74) with a median of 0.80 (range, 0.62 to 0.94), or sensitivity (6.8%, 5/74) with a median of 0.82 (range, 0.66 to 0.98) among studies treating LOS as categorical. In the article corpus, there was no evaluation of prediction models' efficiency, except in one study where the time required to train the model and the hardware specifications were reported.[63]

DISCUSSION

This systematic review summarized available literature on hospital LOS predictions methodology. It highlights that scientific efforts to provide accurate prediction of LOS have been conducted worldwide for half a century. Publications escalation on this topic over the past decade suggests that planning bed capacity and patient discharge remains a matter of concern in health care delivery. However, reviewed studies did not frequently report several pieces of information, making it harder to understand their model's implications and hindering reproducibility. Identifying an optimal prediction methodology is unclear due to the heterogeneity of available evidence with variability regarding data sources and population selection choices, the input variables and models employed for prediction and how their performance is evaluated. Notable trends exist toward the use of more sophisticated methods for predicting LOS and more rigorous study designs for evaluating their validity. In particular, regression methods are superseded by machine learning techniques and using the same dataset for training and testing a model is being deprecated in favor of having a separate test set or using cross-validation.

Other literature reviews covered the topic of LOS prediction, but with a narrower scope in cardiac surgery,[9,10] thermal burns,[11] neonatal,[8] and intensive care units.[7] Conclusions drawn from these past works have similarities with ours regarding the observed disparities around the selection process, the definition of the pre-

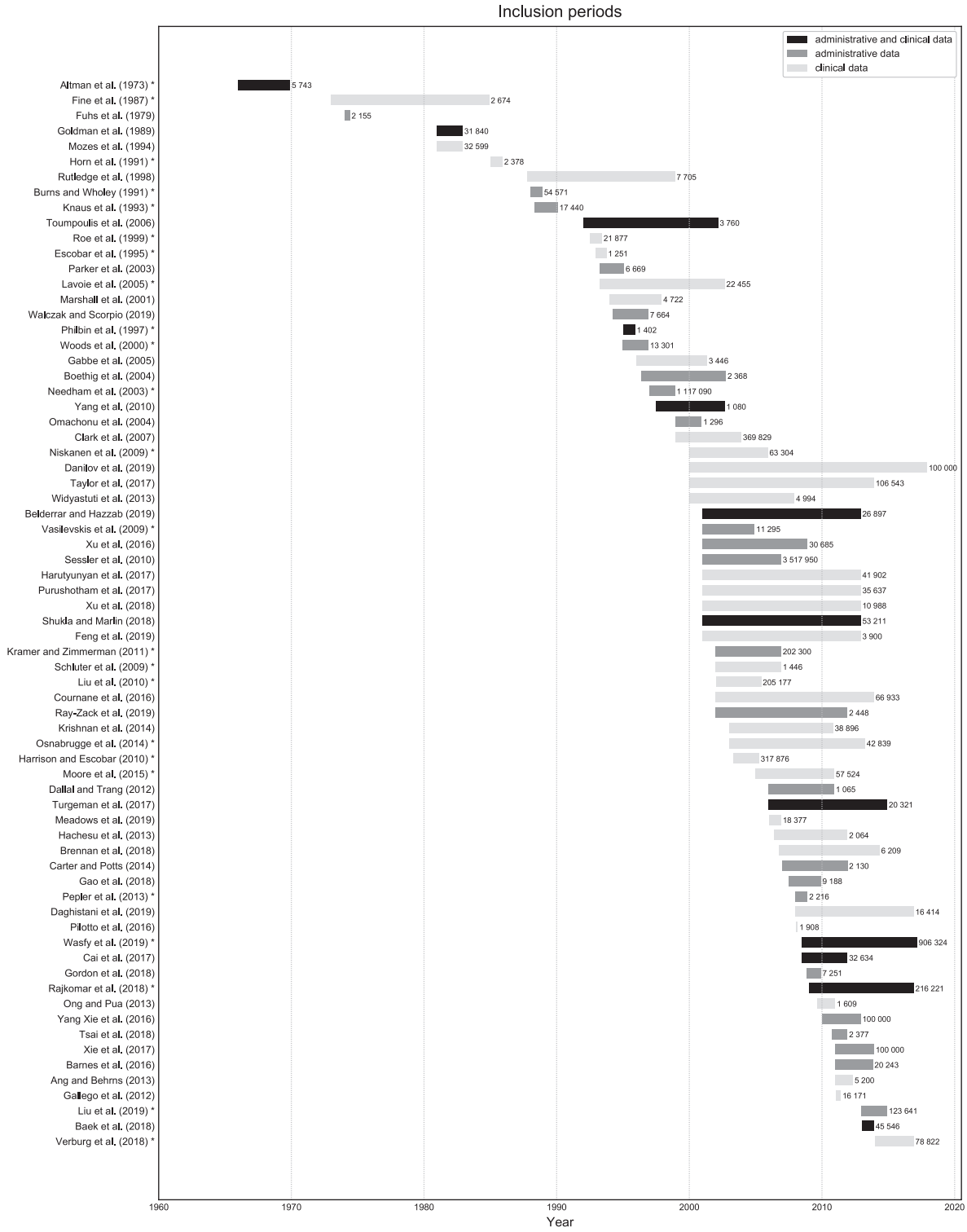


FIGURE 2. Inclusion periods of each study. An asterisk is used to mark multicenter studies. The number at the right of each bar shows the number of inpatient stays for the study.

diction goal and the input variables. Specifically, the exclusion of LOS outliers was also reported as a potential source of bias in performance reporting and articles of the corpus were deemed unsuitable for benchmarking, because they did not always report the variables used, were not free of organizational characteristics, and did not produce accurate predictions without calibration bias.[7]

However, a majority of the studies from these reviews used a binary classifier of LOS and regression-based methods for

TABLE 2. List of the Prediction Methods Regrouped into Categories, Along With Their Descriptions and the Papers That Used Them

Method Category	Description	References
Regression		
Regression models	This category contains linear regression, binomial regression, ridge regression, Lasso regression, logistic regression	21,23–26,32–37,39–52,55,59,61,74–86,88,91
Linear discriminant analysis (LDA)	A model finding the best linear combination of continuous data to approximate a categorical dependent variable	31
Mixed model	Models taking into account both fixed and random effects	7,38,73
Survival model	Models analyzing the impact of input variables on an event over time: Cox regression	30,71,72
Machine learning		
Noncompound	Models not combined with others: decision tree, support vector machine, fuzzy artmap	20,22,23,28,29,55,59,65,92
Tree-based composite	A class of machine learning models which use a combination of multiple decision trees: random forest, bagged regression trees	28,55–57,61,68,89,90
Non–tree-based composite	A class of machine learning models which break down the output space into pieces and/or uses a combination of multiple underlying models: ensemble learning, piece-wise exponential model, restricted cubic splines	20,54,87
Probabilistic	Models relying on probabilities: Bayesian Networks, Markov Models	59,69,70
Deep learning		
Artificial neural network	A model defining a network composed of perceptrons organized in multiple layers	20,28,29,52,53,64,67
Recurrent neural network	Neural network models dealing with temporal data: recurrent neural network keeping a state along a time dimension	27,58,60–63,66
Attention-based neural network	Neural network models dealing with temporal data by leveraging the attention mechanism[93]	60,66
Multimodal deep learning	A combination of neural network architectures	60,63,66

prediction,[9,10] while we reported many studies modeling LOS as a continuous outcome based on a more diverse set of methods centered around machine and deep learning approaches.

The surge in machine learning and deep learning in recent years has been reported by other studies.[94–96] Deep learning is increasingly used with imaging data,[94] and artificial intelligence and machine learning are growing in biomedical research for the last 20 years.[95] This might be catalyzed by technical improvements, advances in hardware, and the ability to handle increasingly large, complex, and unstructured datasets.[97] Deep neural networks could outperform other types of models[98,99] and leverage complex health care datasets,[100] which is key to improve LOS prediction accuracy.[12]

Strengths and Limitations

This study complies with the guidelines of the PRISMA framework, which provides a rigorous methodology to report the information needed to reproduce the results and support

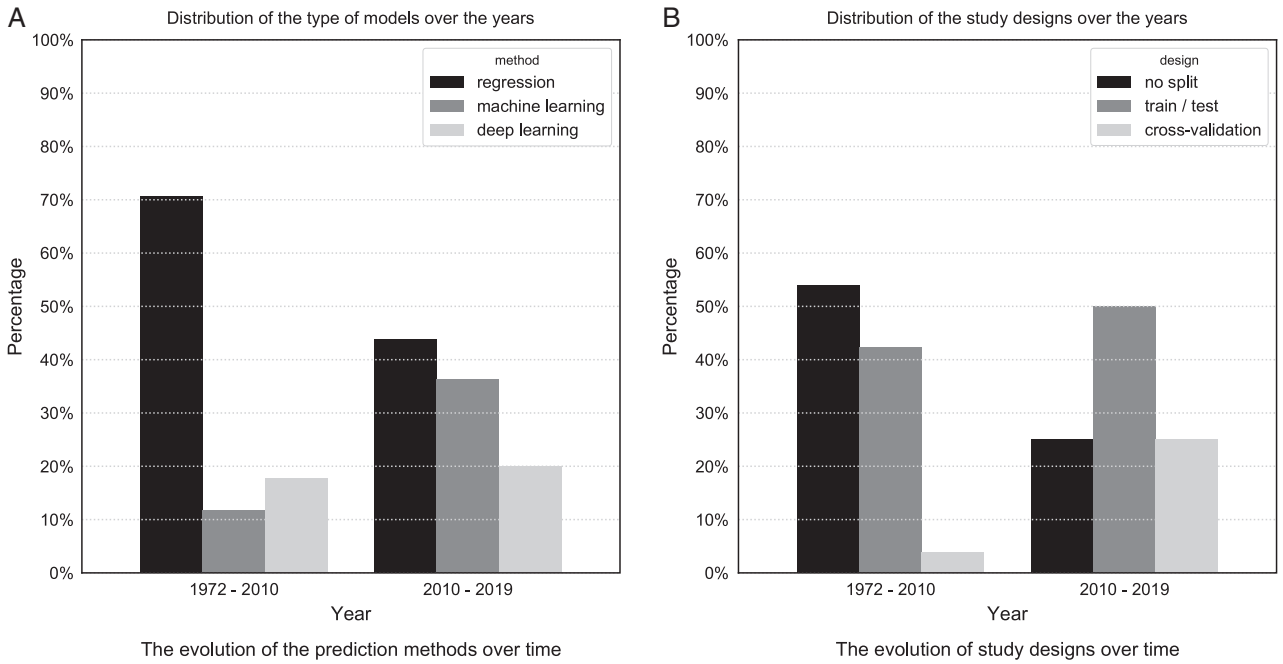


FIGURE 3. A and B, Evolution of prediction methods and testing methodology.

its conclusions. Search strategy was conducted by investigators with complementary backgrounds ranging from health services research to data science and using distinct article sources for covering the topic of LOS prediction from both biomedical and engineering perspectives. Nevertheless, the selection process inherent to literature reviews limits generalization of those results. Because selected studies used different datasets and validation metrics, we were unable to accurately benchmark the prediction methods. Furthermore, almost no articles provided information on the time periods used to define input variables,[25,57,58,61,90] making it impossible to restitute. Likewise, only a few articles mentioned efficiency measures, and only several studies justified their use of prediction methods.[57,63]

Finally, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)[101] checklist was not employed to assess the quality of reported articles in this review, because several topics of the checklist did not apply to LOS prediction algorithms using artificial intelligence based on large datasets or were lacking.[102]

Practical Implications and Directions for Future Research

The lesson learned from this review is that improving LOS prediction through the identification and sharing of the best artificial intelligence methods implies better reporting of research performed internationally to allow an accurate benchmarking of model performance. The need of standardization includes the transparent restitution of population sample selection, data sources, and input variables used by the prediction models, reimputation strategies for missing data, LOS modeling format with potential transformations, the employed LOS prediction methods, the validation study design, and performance evaluation metrics. For all these topics, several suggestions can be expressed to improve the quality of studies investigating LOS prediction.

To ensure models work in multiple contexts, independently of the specificities of a single hospital (eg, care policies), they should be evaluated on multicenter datasets. Choosing population selection criteria that do not arbitrarily exclude outliers is also important to evidence how a prediction model would behave in real-world situations. This point is crucial to avoid overstated conclusions on the validity of a prediction method, considering that prediction will perform well for high-volume and standardized care with short LOS, but more poorly in case of "outlier" patients requiring complex care with extended LOS. Accommodating them should therefore be considered, and reimputation methods were leveraged for missing data. Regarding data used for LOS prediction, an accurate description of their sources is necessary with exhaustive listing of all input variables used for prediction.

Improvements in study designs for evaluating the validity of predictions can also be suggested. The simplest and easiest approach, having the same dataset for training the model and testing its performance, cannot be used to check whether the model is overfitting the data. It should be avoided because one ought to make sure the model reacts properly to novel data for making predictions. Conversely, splitting the data to have a separate test set makes it possible to measure the model's generalization. If the test set is selected randomly, it may only contain the simplest or hardest cases to predict, giving an optimistic or pessimistic performance estimate, respectively. Bias in test set selection may therefore be a major design pitfall for the evaluation process. To overcome this concern, k cross-validation design shall be preferred.[103]

Regarding the metrics to evaluate performance of prediction methods, the accuracy allows one to estimate the quality of a model fit but does not guarantee its usefulness for the purpose of making predictions because it does not account for the distribution of classification labels. To avoid misleading decision-makers, using a metric agnostic to the label distribution like the AUC score shall be preferred, as well as the sensitivity and specificity values. Furthermore, because R^2 depends on the variance in the data used by the model, it may not be the best metric for indicating its usefulness in the context of making predictions.[104] Future work may opt for metrics not depending on the dataset variance, such as the mean squared error or the mean absolute error, and consider using metrics for models comparison such as the Akaike Information Criterion. Furthermore, only one of the articles from this review reported the time required to train the prediction model, and none of them gave an easy way to assess their efficiency, although this information is important in the context of digital resources minimization. Even though having an objective measure of model efficiency is complex,[105] giving a detailed description of the LOS prediction algorithms would simplify the comparison of models through this criterion.

Improving the methods for LOS prediction and evaluating their performance enable improvements in health care resources management. Better bed management and hospital discharge planning allow health care professionals to organize better patient care. Integrating LOS predictions into the hospital information system also permits optimization in the scheduling of care delivery and therefore avoids resource waste.[3] More generally, facilitating the comparison between prediction methods would help to improve them and foster innovations beneficial to patient care. To reliably quantify prediction performance improvements, the use of open and freely available datasets, and of a shared performance evaluation framework shall be recommended. Compared with restricted datasets, open datasets enable a larger group of people to study health care questions, including LOS prediction. They would therefore have to cover administrative and clinical data and include stays from all care units to represent the data that may be used for LOS predictions. Albeit centered on intensive care unit stays, there has been some efforts in these directions, with the development of a freely available database and of a benchmarking task.[29,106] Common data models such as the Observational Medical Outcomes Partnership Common Data Model (OMOP) would also be useful in making data more usable. Moreover, open tools and algorithms in artificial intelligence already exist[107–109] and make it easy to reproduce works and propose improvements on them.

CONCLUSIONS

Research on LOS prediction is headed toward the use of more and more elaborate approaches such as machine learning

methods and assessment of their performance is based on increasingly rigorous study designs. However, the information required to reproduce these algorithms and assess the validity of their predictions is not systematically provided and deserves standardization. Reducing heterogeneity in how these methods are used and reported is key to gaining transparency on their performance.

REFERENCES

1. Weissman J, Rothschild J, Bendavid E, et al. Hospital workload and adverse events. *Med Care*. 2007;45:448–455.
2. Tibby S, Correa-West J, Durward A, et al. Adverse events in a paediatric intensive care unit: relationship to workload, skill mix and staff supervision. *Intensive Care Med*. 2004;30:1160–1166.
3. Schmidt R, Geisler S, Spreckelsen C. Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources. *BMC Med Inform Decis Mak*. 2013;13:3.
4. Hills R, Kitchen S. Satisfaction with outpatient physiotherapy: a survey comparing the views of patients with acute and chronic musculoskeletal conditions. *Physiother Theory Pract*. 2007;23:21–36.
5. Nassar AP, Caruso P. ICU physicians are unable to accurately predict length of stay at admission: a prospective study. *Int J Qual Health Care*. 2015;28:99–103.
6. Durstenfeld MS, Saybolt MD, Praestgaard A, et al. Physician predictions of length of stay of patients admitted with heart failure. *J Hosp Med*. 2016;11:642–645.
7. Verburg IWM, Atashi A, Eslami S, et al. Which models can I use to predict adult ICU length of stay?: a systematic review. *Crit Care Med*. 2017;45:e222–e231.
8. Seaton SE, Barker L, Jenkins D, et al. What factors predict length of stay in a neonatal unit: a systematic review. *BMJ Open*. 2016;6. Available at: https://bmjopen.bmj.com/content/6/10/e010466.
9. Atashi A, Verburg IW, Karim H, et al. Models to predict length of stay in the intensive care unit after coronary artery bypass grafting: a systematic review. *J Cardiovasc Surg*. 2018;59:471–482.
10. Almashrafi A, Elmontsri M, Aylin P. Systematic review of factors influencing length of stay in ICU after adult cardiac surgery. *BMC Health Serv Res*. 2016;16. Available at: https://pubmed.ncbi.nlm.nih.gov/27473872/.
11. Hussain A, Dunn K. Predicting length of stay in thermal burns: a systematic review of prognostic factors. *Burns*. 2013;39:1331–1340.
12. Lu M, Sajobi T, Lucyk K, et al. Systematic review of risk adjustment models of hospital length of stay (LOS). *Med Care*. 2015;53:355–365.
13. Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. *PLoS Med*. 2009;6:e1000097.
14. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Series B Methodolo*. 1974;36:111–147.
15. James G, Witten D, Hastie T, et al. *An Introduction to Statistical Learning (Springer Texts in Statistics; Vol 103)*. New York, NY: Springer New York; 2013.
16. Van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace; 2009.
17. McKinney W. Pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*. 2011; 14. Available at: https://www.semanticscholar.org/paper/pandas%3A-a-Foundational-Python-Library-for-Data-and-McKinney/1a62eb61b2663f8135347171e30cb9dc0a8931b5.
18. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9:90–95.
19. Pollard TJ, Johnson AEW, Raffa JD, et al. Tableone: an open source Python package for producing summary statistics for research papers. *JAMIA Open*. 2018;1:26–31.
20. Hachesu PR, Ahmadi M, Alizadeh S, et al. Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthc Inform Res*. 2013;19:121.
21. Fine P, Stover S, DeVivo M. A methodology for predicting lengths of stay for spinal cord injury patients. *Inquiry*. 1987;24:147–156.
22. Wasfy JH, Kennedy KF, Masoudi FA, et al. Predicting length of stay and the need for postacute care after acute myocardial infarction to improve healthcare efficiency. *Circ Cardiovasc Qual Outcomes*. 2018;11. Available at: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.118.004635.
23. Yang C-S, Wei C-P, Yuan C-C, et al. Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages. *Decis Support Syst*. 2010;50:325–335.
24. Gabbe BJ, Cameron PA, Wolfe R, et al. Predictors of mortality, length of stay and discharge destination in blunt trauma. *ANZ J Surg*. 2005;75:650–656.
25. Moore L, Stelfox HT, Turgeon AF, et al. Derivation and validation of a quality indicator of acute care length of stay to evaluate trauma care. *Ann Surg*. 2014;260:1121–1127.
26. Toumpoulis IK, Anagnostopoulos CE, DeRose JJ, et al. Does EuroSCORE predict length of stay and specific postoperative complications after coronary artery bypass grafting? *Int J Cardiol*. 2005;105:19–25.
27. Danilov G, Kotik K, Shifrin M, et al. Prediction of postoperative hospital stay with deep learning based on 101 654 operative reports in neurosurgery. *Stud Health Technol Inform*. 2019;258:125–129.
28. Daghistani TA, Elshawi R, Sakr S, et al. Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. *Int J Cardiol*. 2019;288:140–147.
29. Rutledge R, Osler T, Emery S, et al. The end of the Injury Severity Score (ISS) and the Trauma and Injury Severity Score (TRISS): ICISS, an International Classification of Diseases, ninth revision-based prediction tool, outperforms both ISS and TRISS as predictors of trauma patient survival, hospital charges, and hospital length of stay. *J Trauma*. 1998;44:41–49.
30. Widyastuti Y, Stenseth R, Wahba A, et al. Length of intensive care unit stay following cardiac surgery: is it impossible to find a universal prediction model? *Interact Cardiovasc Thorac Surg*. 2012;15:825–832.
31. Altman H, Angle HV, Brown ML, et al. Prediction of length of hospital stay. *Compr Psychiatry*. 1972;13:471–480.
32. Levin SR, Harley ET, Fackler JC, et al. Real-time forecasting of pediatric intensive care unit length of stay using computerized provider orders. *Crit Care Med*. 2012;40:3058–3064.
33. Pilotto A, Sancarlo D, Pellegrini F, et al. The Multidimensional Prognostic Index predicts in-hospital length of stay in older patients: a multicentre prospective study. *Age Ageing*. 2016;45:90–96.
34. Meadows K, Gibbens R, Gerrard C, et al. Prediction of patient length of stay on the intensive care unit following cardiac surgery: a logistic regression analysis based on the cardiac operative mortality risk calculator, EuroSCORE. *J Cardiothorac Vasc Anesth*. 2018;32:2676–2682.
35. Omachonu VK, Suthummanon S, Akcin M, et al. Predicting length of stay for Medicare patients at a teaching hospital. *Health Serv Manage Res*. 2004;17:1–12.
36. Boethig D, Jenkins KJ, Hecker H, et al. The RACHS-1 risk categories reflect mortality and length of hospital stay in a large German pediatric cardiac surgery population. *Eur J Cardiothorac Surg*. 2004;26:12–17.
37. Lavoie A, Moore L, LeSage N, et al. The Injury Severity Score or the New Injury Severity Score for predicting intensive care unit admission and hospital length of stay? *Injury*. 2005;36:477–483.
38. Schluter PJ, Cameron CM, Davey TM, et al. Using Trauma Injury Severity Score (TRISS) variables to predict length of hospital stay following trauma in New Zealand. *N Z Med J*. 2009;122:65–78.
39. Burns LR, Wholey DR. The effects of patient, hospital, and physician characteristics on length of stay and mortality. *Med Care*. 1991;29:251–271.
40. Dallal RM, Trang A. Analysis of perioperative outcomes, length of hospital stay, and readmission rate after gastric bypass. *Surg Endosc*. 2012;26:754–758.
41. Pepler PT, Uys DW, Nel DG. Predicting mortality and length-of-stay for neonatal admissions to private hospital neonatal intensive care units: a Southern African retrospective study. *Afr Health Sci*. 2012;12:166–173.
42. Ang DN, Behrns KE. Using a relational database to improve mortality and length of stay for a department of surgery: a comparative review of 5200 patients. *Am Surg*. 2013;79:706–710.
43. Krishnan KR, Bhattacharya R, Pereira A, et al. The HALOS-ND model: a step in the journey of predicting hospital length of stay after liver transplantation. *Clin Transplant*. 2013;27:809–822.
44. Ong P-H, Pua Y-H. A prediction model for length of stay after total and unicompartmental knee replacement. *Bone Joint J*. 2013;95-B:1490–1496.
45. Carter EM, Potts HWW. Predicting length of stay from an electronic patient record system: a primary total knee replacement example. *BMC Med Inform Decis Mak*. 2014;14:26.
46. Osnabrugge RL, Speir AM, Head SJ, et al. Prediction of costs and length of stay in coronary artery bypass grafting. *Ann Thorac Surg*. 2014;98:1286–1293.

47. Taylor SL, Sen S, Greenhalgh DG, et al. Not all patients meet the 1day per percent burn rule: a simple method for predicting hospital length of stay in patients with burn. *Burns*. 2017;43:282–289.
48. Brennan A, Gauvreau K, Connor J, et al. A method to account for variation in congenital heart surgery length of stay. *Pediatr Crit Care Med*. 2017;18:550–560.
49. Ray-Zack MD, Shan Y, Mehta HB, et al. Hospital length of stay following radical cystectomy for muscle-invasive bladder cancer: development and validation of a population-based prediction model. *Urol Oncol*. 2019;37:837–843.
50. Escobar GJ, Fischer A, Li DK, et al. Score for neonatal acute physiology: validation in three Kaiser Permanente neonatal intensive care units. *Pediatrics*. 1995;96:918–922.
51. Philbin EF, Rocco TA, Lynch LJ, et al. Predictors and determinants of hospital length of stay in congestive heart failure in ten community hospitals. *J Heart Lung Transplant*. 1997;16:548–555.
52. Tsai P-FJ, Chen P-C, Chen Y-Y, et al. Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. *J Healthc Eng*. 2016;2016:1–11.
53. Belderrar A, Hazzab A. Hierarchical genetic algorithm and fuzzy radial basis function networks for factors influencing hospital length of stay outliers. *Healthc Inform Res*. 2017;23:226–232.
54. Clark DE, Lucas FL, Ryan LM. Predicting hospital mortality, length of stay, and transfer to long-term care for injured patients. *J Trauma*. 2007;62:592–600.
55. Barnes S, Hamrock E, Toerper M, et al. Real-time prediction of inpatient length of stay for discharge prioritization. *J Am Med Inform Assoc*. 2015;23:e2–e10.
56. Mozes B, Easterling MJ, Sheiner LB, et al. Case-mix adjustment using objective measures of severity: the case for laboratory data. *Health Serv Res*. 1994;28:689.
57. Turgeman L, May JH, Sciulli R. Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. *Expert Syst Appl*. 2017;78:376–385.
58. Harutyunyan H, Khachatrian H, Kale DC, et al. Multitask learning and benchmarking with clinical time series data. *Sci Data*. 2019;6:96–114.
59. Sotoodeh M, Ho J. Improving length of stay prediction using a hidden Markov model. *AMIA Jt Summits Transl Sci Proc*. 2019;2019:425–434.
60. Xu Y, Biswal S, Deshpande SR, et al. RAIM: Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London, UK: ACM; 2018:2565–2573.
61. Nestor B, McDermott MBA, Boag W, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. 2019. Available at: http://arxiv.org/abs/1908.00690. Accessed April 15, 2020.
62. Shukla SN, Marlin BM. Modeling irregularly sampled clinical time series. 2018. Available at: http://arxiv.org/abs/1812.00531. Accessed April 15, 2020.
63. Purushotham S, Meng C, Che Z, et al. Benchmark of deep learning models on large healthcare MIMIC datasets. 2017. Available at: http://arxiv.org/abs/1710.08531. Accessed April 15, 2020.
64. Suresh A, Harish KV, Radhika N. Particle swarm optimization over back propagation neural network for length of stay prediction. *Procedia Comput Sci*. 2015;46:268–275.
65. Feng J, Sondhi A, Perry J, et al. Selective prediction-set models with coverage guarantees. 2019. Available at: http://arxiv.org/abs/1906.05473. Accessed May 11, 2020.
66. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning for electronic health records. *NPJ Digit Med*. 2018;1:18.
67. Walczak S, Scorpio RJ. Predicting pediatric length of stay and acuity of care in the first ten minutes with artificial neural networks. *Pediatr Crit Care Med*. 2000;1:42–47.
68. Gao C, Kho AN, Ivory C, et al. Predicting length of stay for obstetric patients via electronic medical records. *Stud Health Technol Inform*. 2017;245:1019–1023.
69. Marshall AH, McClean SI, Shapcott CM, et al. Developing a Bayesian belief network for the management of geriatric hospital care. *Health Care Manag Sci*. 2001;4:25–30.
70. Cai X, Perez-Concha O, Coiera E, et al. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *J Am Med Inform Assoc*. 2016;23:553–561.
71. Gordon AS, Marshall AH, Zenga M. Predicting elderly patient length of stay in hospital and community care using a series of conditional Coxian phase-type distributions, further conditioned on a survival tree. *Health Care Manag Sci*. 2018;21:269–280.
72. Sessler DI, Sigl JC, Manberg PJ, et al. Broadly applicable risk stratification system for predicting duration of hospitalization and mortality. *Anesthesiology*. 2010;113:1026–1037.
73. Gallego B, Perez-Concha O, Lin F, et al. Exploring the role of pathology test results in the prediction of remaining days of hospitalisation. *Stud Health Technol Inform*. 2012;178:45–50.
74. Vasilevskis EE, Kuzniewicz MW, Cason BA, et al. Mortality probability model III and simplified acute physiology score II: assessing their value in predicting length of stay and comparison to APACHE IV. *Chest*. 2009;136:89–101.
75. Woods AW, MacKirdy FN, Livingston BM, et al. Evaluation of predicted and actual length of stay in 22 Scottish intensive care units using the APACHE III system. Acute physiology and chronic health evaluation. *Anaesthesia*. 2000;55:1058–1065.
76. Parker JP, McCombs JS, Graddy EA. Can pharmacy data improve prediction of hospital outcomes? Comparisons with a diagnosis-based comorbidity measure. *Med Care*. 2003;41:407–419.
77. Needham DM, Anderson G, Pink GH, et al. A province-wide study of the association between hospital resource allocation and length of stay. *Health Serv Manage Res*. 2003;16:155–166.
78. Niskanen M, Reinikainen M, Pettilä V. Case-mix-adjusted length of stay and mortality in 23 Finnish ICUs. *Intensive Care Med*. 2009;35:1060–1067.
79. Liu V, Kipnis P, Gould MK, et al. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Med Care*. 2010;48:739–744.
80. Harrison GW, Escobar GJ. Length of stay and imminent discharge probability distributions from multistage models: variation by diagnosis, severity of illness, and hospital. *Health Care Manage Sci*. 2010;13:268–279.
81. Kramer AA, Zimmerman JE. The relationship between hospital and intensive care unit length of stay. *Crit Care Med*. 2011;39:1015–1022.
82. Goldman ES, Easterling MJ, Sheiner LB. Improving the homogeneity of diagnosis-related groups (DRGs) by using clinical laboratory, demographic, and discharge data. *Am J Public Health*. 1989;79:441–444.
83. Cournane S, Byrne D, O'Riordan D, et al. Factors associated with length of stay following an emergency medical admission. *Eur J Intern Med*. 2015;26:237–242.
84. Baek H, Cho M, Kim S, et al. Analysis of length of hospital stay using electronic health records: a statistical and data mining approach. *PloS One*. 2018;13:e0195901.
85. Liu J, Larson E, Hessels A, et al. Comparison of measures to predict mortality and length of stay in hospitalized patients. *Nurs Res*. 2019;68:200–209.
86. Fuhs PA, Martin JB, Hancock WM. The use of length of stay distributions to predict hospital discharges. *Med Care*. 1979;17:355–368.
87. Knaus WA, Wagner DP, Zimmerman JE, et al. Variations in mortality and length of stay in intensive care units. *Ann Intern Med*. 1993;118:753–761.
88. Roe CJ, Kulinskaya E, Dodich N, et al. Comorbidities and prediction of length of hospital stay. *Aust N Z J Med*. 1998;28:811–815.
89. Yang X, Neubauer S, Schreier G, et al. Impact of hierarchies of clinical codes on predicting future days in hospital. *Annu Int Conf IEEE Eng Med Biol Soc*. 2015;2015:6852–6855.
90. Xie Y, Schreier G, Hoy M, et al. Analyzing health insurance claims on different timescales to predict days in hospital. *J Biomed Inform*. 2016;60:187–196.
91. Horn SD, Sharkey PD, Buckle JM, et al. The relationship between severity of illness and hospital length of stay and mortality. *Med Care*. 1991;29:305–317.
92. Xu H, Wu W, Nemati S, et al. Patient flow prediction via discriminative learning of mutually-correcting processes. 2016. Available at: http://arxiv.org/abs/1602.05112. Accessed June 25, 2020.
93. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 2017. Available at: http://arxiv.org/abs/1706.03762. Accessed April 15, 2020.

94. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6:94–98.
95. Rong G, Mendez A, Bou Assi E, et al. Artificial intelligence in healthcare: review and prediction case studies. *Engineering*. 2020;6:291–301.
96. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet*. 2020;395:1579–1586.
97. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2:719–731.
98. Le QV, Ranzato M, Monga R, et al. Building high-level features using large scale unsupervised learning. 2011. Available at: http://arxiv.org/abs/1112.6209. Accessed August 21, 2020.
99. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60:84–90.
100. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inform Sci Syst*. 2014;2:3.
101. Collins G, Reitsma J, Altman D, et al. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594–g7594.
102. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393:1577–1579.
103. Hastie T, Tibshirani R, Friedman J. Model assessment and selection. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer, New York; 2009:219–259.
104. Alexander DLJ, Tropsha A, Winkler DA. Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J Chem Inform Model*. 2015;55:1316–1322.
105. Schwartz R, Dodge J, Smith NA, et al. Green AI. arXiv:190710597. 2019. Available at: http://arxiv.org/abs/1907.10597. Accessed July 1, 2020.
106. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.
107. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc FD, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems 32*. Vancouver, Canada: Curran Associates Inc.; 2019:8026–8037.
108. Abadi M, Barham P, Chen J, et al. *TensorFlow: A System for Large-Scale Machine Learning*. Savannah, GA: USENIX Association; 2016:265–283.
109. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.

Ce travail suggère que (i) il est difficile de comparer les résultats des études entre elles du fait des différences dans les jeux de données, les prétraitements, les schémas d'études et les mesures de performances, ainsi que la façon dans, (ii) les méthodes d'apprentissage automatique (*machine learning*) et d'apprentissage profond (*deep learning*) sont de plus en plus utilisées, (iii) les schémas d'études sont de plus en plus rigoureux, avec une augmentation du nombre de validations croisées et de l'utilisation d'une séparation jeu d'entrainement / jeu de test par rapport à ne pas utiliser de jeu de test.

Ce travail a permis d'identifier des pistes d'amélioration dans la façon de rapporter les résultats d'une étude sur un système de prédiction. les disparités en termes de critères d'inclusion et d'exclusion du jeu de données, de cibles de prédiction (i.e. classification binaire, multi-classe ou régression) et de mesures des performances rendent difficile une comparaison qui pourrait identifier les meilleurs modèles de prédiction des durées de séjours. Certaines études parmi notre sélection ne donnaient pas assez d'informations pour pouvoir constituer un jeu de données équivalent permettant de reproduire leurs résultats. Tout cela nuit à la reproductibilité des études et à la capacité à baser de futurs travaux sur les modèles les plus performants, rendant donc la recherche dans ce domaine plus incertaine qu'elle le devrait [65]. Un jeu de données accessible à tous permettant de facilement comparer les différentes approches pallierait ce problème, comme il a pu provoquer des avancées techniques dans la classification d'images (cf. section 1.3.1).

De plus, les performances ne doivent pas être les uniques points de comparaison entre des modèles de prédiction. L'efficacité du modèle, que ce soit en temps nécessaire pour l'entrainement, associé aux capacités informatiques requises, ainsi que la complexité en termes de structures ou de nombres d'opérations peuvent également être comparées si spécifiées dans les articles, même s'il est difficile de trouver une mesure juste et applicable à tout type de modèle [66], et qu'il peut exister un compromis entre efficacité et précision [67]. Cela peut rendre l'analyse de la complexité hasardeuse et contraignante. Ces critères sont génériques et peuvent donc avoir des équivalents pour d'autres tâches de prédiction en santé qui ignorent aussi l'efficacité.

L'analyse de données nous a permis de connaître les types d'algorithmes

d'apprentissage automatique et d'apprentissage profond utilisés pour prédire la durée de séjour, ainsi que le contexte dans lequel ils sont utilisés. Cela nous a notamment permis de confirmer que les réseaux de neurones récurrents sont utilisés pour traiter des séries temporelles d'événements médicaux et de mesures biologiques et physiologiques. Cette recrudescence de l'apprentissage automatique et de l'apprentissage profond peut être causé par les récents succès de ces méthodes dans d'autres domaines et champs d'application, pour lesquels l'apprentissage automatique a produit des résultats très encourageants.

Une fois utilisé en routine en production, un modèle fera des prédictions sur des données nouvelles, qu'il n'a pas vu pendant l'entrainement. Il est donc souhaitable de mesurer la capacité d'un modèle à avoir de bonnes performances sur de nouvelles données, (i.e sa capacité de généralisation). Utiliser l'intégralité du jeu de donnée pour l'entrainement ne permet pas de mesurer la capacité du modèle à généraliser, et c'est pourquoi cette approche est de moins en moins utilisée. Ces améliorations suivant une tendance plus générale à appliquer les bonnes pratiques en termes de sciences de la donnée et intelligence artificielle, il est probable que les futurs progrès dans ces domaines aient également une influence.

Lire les articles sélectionnés pour la revue de littérature a également permis d'approfondir nos connaissances en matières de techniques de mesure des performances et autres points méthodologiques intéressants en termes de culture scientifique au sens large. Enfin, conduire une revue de la littérature et chercher des données précises sur des études a permis de se sensibiliser davantage sur la nécessité de fournir toutes les informations aux lecteurs de manière à ce qu'ils puissent correctement juger le travail, en reproduire les résultats et les améliorer.

Tout cela a été utile dans le cadre de cette thèse et pour développer notre propre méthode de prédiction des durées de séjours.

Chapitre 4

Méthode de prédiction de la durée de séjour

L'objectif de cette thèse était de proposer une méthode innovante pour prédire la durée de séjour, à chaque instant du séjour du patient à l'aide de données médico-administratives. Les données médico-administratives étant structurées mais complexes, un prétraitement des données était nécessaire avant de pouvoir les utiliser. C'est pourquoi une procédure détaillée de nettoyage des données devait tout d'abord être développée. Comme le jeu de données contient un nombre de diagnostics principaux important, et que certains sont rarement utilisés, chercher à les regrouper et les simplifier permet de simplifier le jeu de données, et donc de faciliter l'apprentissage d'un modèle d'apprentissage automatique pour prédire la durée de séjour. L'algorithme 1 donne les détails de l'algorithme de simplification des diagnostics. Les diagnostics sont simplifiés jusqu'à ce qu'ils soient présents dans un nombre minimal de séjour, ou d'une longueur minimale. Les diagnostiques associées ont eux aussi été simplifiés, en conservant uniquement ceux utilisés dans le score d'Elixhauser [68].

Algorithme 1 : simplificationDiagnostic

Données : Une liste D des diagnostics principaux pour tous les séjours
Une fréquence minimale N et une taille minimale l pour chaque diagnostic
Résultat : La liste D mise à jour

Function *cardDiag* $d1$ **:**
    $n \leftarrow 0$
    pour $d2 \in D$ **faire**
        si $d2$.`commencePar(`$d1$`)` **alors**
            $n \leftarrow n+1$
    retourner n / `Taille(`D`)`

début
    pour $d \in D$ **faire**
        $diagSimple \leftarrow d$
        tant que `cardDiag(`$diag$`)` $> N$ **and** `Taille(`d`)` $> l$ **faire**
            $diagSimple \leftarrow$TrimR(diagSimple)
    retourner D

Comme les actes médicaux sont également très nombreux, il a fallu également les filtrer. Pour cela, une mesure du gain d'information des actes médicaux par rapport à la durée de séjour, et la fréquence des actes ont été calculés puis ajoutés entre eux, formant un score. Les actes médicaux avec un score supérieur à deux fois la moyenne ont été conservés.

La durée d'un séjour étant liée à son contexte, ce travail a nécessité de prendre en compte les informations passées et de les résumer. Pour cela, pour chaque RUM (étape d'une hospitalisation), les données des autres séjours des 30 derniers jours sont agrégées. Des statistiques sont ensuite calculées sur ces agrégats glissants, comme la moyenne de la durée de séjour ou le nombre d'admissions, à l'échelle de l'hôpital, de l'unité médicale, du diagnostic principal du patient pour laquelle la durée de séjour est prédite, et à l'échelle du patient lui-même

pour prendre en compte ses hospitalisations passées.

Une méthode de prédiction devait ensuite être proposée. Les *embeddings* ayant engendrés des résultats intéressants pour d'autres tâches de prédiction ou des tâches de prédictions similaires à la nôtre, nous avons fait l'hypothèse qu'ils produiraient de bons résultats pour la prédiction des durées de séjour avec des données administratives tabulaires. Pour tester cette hypothèse, un modèle de prédiction utilisant des *embeddings* a été testé sur un échantillon des données mises à disposition et les résultats ont été comparés avec ceux d'une méthode utilisant la fonction *one-hot encoding* (cf. partie 1.3.3) à la place des *embeddings* ainsi qu'à une prédiction aléatoire basée sur la distribution des durées de séjour dans le jeu de données. Comme les *embeddings* produisent une représentation riche et complexe des données et que les valeurs de ces *embeddings* peuvent être optimisés par descente de gradient, un réseau de neurones a été entrainé pour réaliser la classification.

Comme l'outil de prédiction sera utilisé à des fins organisationnelles, la cible de prédiction devait être choisie de manière à obtenir une vision détaillée des futurs parcours hospitaliers au sein d'une unité médicale. C'est pourquoi nous avons fixé dans cet objectif une prédiction jour après jour de la durée de séjour, et ceci jusqu'à 14 jours ou plus, engendrant ainsi une classification à 15 classes différentes : une classe par jour de 0 jour à 13 jours et une classe pour 14 jours ou plus.

Article 2 : Vincent Lequertier et al. « Predicting length of stay with administrative data from acute and emergency care : an embedding approach ». *In : 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. Août 2021, p. 1395-1400. DOI : 10.1109/CASE49439.2021.9551429

Présentation accompagnée d'un poster lors d'un séminaire doctoral : Vincent Lequertier. « Global method for predicting the length of stay in hospital using incremental and evolutionary data ». Doctoral Seminar DISP. Lyon, 3 déc. 2021

2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)
August 23-27, 2021. Lyon, France

Predicting length of stay with administrative data from acute and emergency care: an embedding approach

Vincent Lequertier[1,2,3] Tao Wang[3] Julien Fondrevelle[3] Vincent Augusto[4] Stéphanie Polazzi[1,2]

Antoine Duclos[1,2]

Abstract— **Hospital beds management is critical for the quality of patient care, while length of inpatient stay is often estimated empirically by physicians or chief nurses of medical wards. Providing an efficient method for forecasting the length of stay (LOS) is expected to improve resources and discharges planning. Predictions should be accurate and work for as many patients as possible, despite their heterogeneous profiles.**

In this work, a LOS prediction method based on deep learning and embeddings is developed by using generic hospital administrative data from a French national hospital discharge database, as well as emergency care.

Data concerned 497 626 stays of 304 931 patients from 6 hospitals in Lyon, France, from 2011 to 2019. Results of a 5-fold cross-validation showed an accuracy of 0.73 and a kappa score of 0.67 for the embeddings method. This outperformed the baseline which used the raw input features directly.

I. INTRODUCTION

Hospitals are bound to optimize the use of their resources to meet the growing healthcare needs of the population and budget constraints. Inpatient bed is one of the critical resources in the hospital. Most hospital activities interact with bed management, such as caregivers' staffing, scheduling, ward services and logistics. For each inpatient admission to a medical or surgical unit, a bed should be assigned to the patient for an estimated duration, called length of stay (LOS). The LOS is often estimated empirically by physicians or chief nurses of medical wards. The estimation depends on their experience, since a decision is made by matching current admission demand with one of the patient profiles they already followed or studied. The decision is local and can be different from one hospital to another, or from one healthcare professional to another. Especially for non-frequent patient profiles, estimation error will be large due to the lack of information.

Many factors impact the LOS, such as socio-administrative data (age, sex, residential area, profession, lifestyle choices) and medico-administrative data (past medical history, principal diagnosis, adequacy of care unit, emergency level, admission period, care pathway). As administrative data are well standardized, they are a compelling base for LOS prediction. LOS prediction models have already been made using administrative data with specific patient cohorts. For example, data from 22 Scottish intensive care units were used to predict LOS, with the Acute Physiology and Chronic Health Evaluation III system [1]. In [2], a model for LOS prediction was built for patients with bladder cancer and another one was built for patients undergoing gastric bypass in [3]. A LOS prediction study focused on the United States' national social insurance Medicare, for patients over 65 years old [4].

Various methods have been developed to predict the LOS, most of them focusing on small cohort studies with specific diseases. Few studies were interested in developing general prediction methods. According to our to be published systematic review on LOS prediction on 74 selected articles. 64.9% of these articles applied statistical methods based on regressions to extract mathematical models and rules, 20.1% were interested in machine learning methods to discover knowledges from data mining, and 17.6% developed deep learning models based on artificial neural networks, including articles using methods from multiple categories. Regression methods were being superseded by machine learning and deep learning over time. Since the authors used different data sources and metric measures, it is very hard to give an unbiased performance benchmark. However, the results showed that deep learning is promising for improving prediction performance on big medical data in computer vision [5], time series processing [6] and NLP [7].

The administrative data are heterogeneous and multimodal, because of the large number of patient conditions and medical pathways, thus forming a high dimensional feature space. An efficient method should be identified to deal with this complexity. Embeddings methods substantially improved performance of models in Natural Language Processing (NLP) by creating a feature-rich representation of textual data [8, 9]. The breakthrough inspired the use of embeddings in healthcare, where embeddings were built to represent patients [10] or medical concepts [7]. Embeddings have therefore been used to represent categorical features with numerical vectors which can be fed to neural networks, and can map the high-dimensional space into a low-dimensional one.

The purpose of this work was to make a global LOS prediction model using a neural network with embeddings vectors and compare it to a model without embedding vectors. Key challenges revolved around feature engineering, because of the heterogeneity of the data.

[1]Université Claude Bernard Lyon 1, Research on Healthcare Performance (RESHAPE), INSERM U1290, Lyon, France
[2]Heath data department, Lyon University Hospital, Lyon, France
[3]Univ Lyon, INSA-Lyon, UJM-Saint-Etienne, UCBL, Univ Lumière Lyon 2, DISP EA4570, 69621 Villeurbanne, France
[4]LIMOS CNRS UMR 6158, Ecole Nationale Supérieure des Mines, 158 cours Fauriel, 42023 Saint-Etienne, France

978-0-7381-2502-2/21/$31.00 ©2021 IEEE

The rest of this work is organized as follows. Section II explains the data extraction step, data preprocessing, prediction methods and the evaluation of model performances, Section III presents the experimental results and Section IV concludes with a discussion and practical implications for future work.

II. MATERIAL AND METHODS

A. Data extraction

The data came from the French national standard discharge database (PMSI), akin to Diagnosis-Related Group (DRG) data, and emergency care data encoded with the Emergency visit summary (RPU), a French standard for storing data about emergency care stays. Emergency care data were included because it could add more insights to the context of the inpatient stay. Included data were from inpatient stays in 6 hospitals of Lyon, France, with admission date from January 1st, 2011 to December 31st, 2019. Inpatient stays were identified with a inpatient identifier, a stay identifier and an identifier for the Medical Unit Summary (RUM), corresponding to a step within the inpatient stay. Each transfer to a new medical unit led to a new RUM. Since length of stay is the outcome of interest, data extracted concerned only non ambulatory. It included 1 798 447 RUM. Exclusion criteria were as follows: 1) patients under eighteen years old at the age of admission, 2) medical units with less than 100 stays over the time period, 3) stays with a nominal duration and 4) stays with erroneous information. This led to 1 593 241 RUM, corresponding to 1 233 360 hospital stays of 535 267 patients. Inpatient stays longer than 31 days were discarded, and the dataset was randomly sampled down to 551 684 RUM, 497 626 inpatient stays of 304 931 patients because of resources constraints. LOS was defined at the level of the RUM by computing the difference in days between the discharge date and the admission date.

Table I lists the data extracted for each hospital stay, and Table II lists the data extracted for each emergency care stay.

TABLE I
DESCRIPTION OF THE DATA EXTRACTED FOR EACH HOSPITAL INPATIENT STAY

Data	Description
Patient ID	Patient identifier
Stay ID	Stay identifier
RUM ID	The RUM identifier, corresponding to a step in the patient pathway
Hospital	The identifier of the hospital where the patient was admitted
Medical Unit ID	The identifier of the medical unit
Medical Unit type	Type of the medical unit, e.g. surgical, intensive care, palliative care
Entry date and time	The date and time at which the patient was admitted to the medical unit
Discharge date and time	The date and time at which the patient was discharged from the medical unit
Age	Age of the patient
Gender	Gender of the patient
Principal Diagnosis	The principal diagnosis encoded with the International Classification for Diseases, 10th revision (IDC-10)
Associated Diagnoses	The associated diagnoses encoded with the ICD-10
Medical Procedures	The medical procedures encoded with the French Common Classification of medical acts (CCAM)
Mode of entry	How the patient was admitted: from home, mutation or transfer

B. Data preprocessing

The associated diagnoses were simplified by keeping only the ones used in the Elixhauser comorbidity score [11], which is a score identifying 31 comorbidities associated with LOS. The score was validated on French healthcare data [12]. "AIDS/HIV", "Drug abuse", "Psychose", "Peptic ulcer disease excluding bleeding" and "Lymphoma" were removed from the Elixhauser comorbidities because they were seldom represented in the dataset. This preprocessing simplified associated diagnosis data.

Because the medical procedures were encoded with the French CCAM (Common classification of medical acts) coding, to ensure the study was reproducible, preprocessing did not take advantage of the CCAM structure. Instead, the information gain with regard to LOS and the procedures' frequencies were taken into account, with the following score: $score(act) = InfoGain_{act} * (freq_{act} + 1)$. Medical procedures with a score superior to twice the mean were considered worthy and therefore selected. Moreover, only the medical procedures that had an impact on the grouping of similar stays for reimbursement purposes were kept. To simulate a real simulation and avoid leaking information, procedures done during a hospitalization step were not known until the next one.

The date of each stay was cut down into the time of day (i.e., morning, afternoon, night), day of week, month number and year number and were considered as categorical. Number of hospital and emergency admissions, total LOS and mean LOS were computed at the main diagnosis level, the medical unit level, the patient level and as a whole for the last 30 days, to account for environmental context. The time spent at the hospital since an admission started was also computed. The patient ID and Stay ID were removed afterwards. This preprocessing allowed the deep learning model to get information about the state of inpatient stays without having to resort to stateful models like Recurrent Neural Networks, which would perform badly for inpatient stays with a single step.

For each hospital admission following emergency care, the main diagnosis identified at emergency care and the one identified at the inpatient admission were deemed similar if 1) They had the same ICD-10 chapter (the first three characters) and 2) The other characters composing the code

TABLE II
DESCRIPTION OF THE DATA EXTRACTED FOR EACH EMERGENCY CARE STAY

Data	Description
Patient ID	Patient identifier
Stay ID	Stay identifier
Stay duration	Emergency care stay duration, in minutes
Principal Diagnosis	The principal diagnosis encoded with the ICD-10
Associated Diagnoses	The associated diagnoses encoded with the ICD-10
Medical Procedures	The medical procedures encoded with the French Common Classification of medical acts (CCAM)

had a Levenshtein edit distance inferior to 2. This reduced the heterogeneity of emergency care diagnosis. Emergency care duration was discretized into less than 8 hours, less than 12 hours and more than 12 hours.

Categorical features were encoded with an ordinal encoder, multi label features (i.e., associated diagnoses and medical procedures) were encoded with one-hot encoding. Continuous features were standardized between -1 and 1 and normalized with a mean of 0 and a standard deviation of 1.

C. Methods

To evaluate the capacity of models to perform well on unseen data, k-cross validation was employed, with $k = 5$, so that the models were trained and tested 5 times on different datasets, with each row of the dataset being part of the test set exactly once. The dataset was thus split 5 times into a training set and test set, representing 80% and 20% of the dataset, respectively. The training set and test set were not in a chronological order. Cross-validation was deemed more systematic and less prone to bias than having a single testing dataset.

To create numerical vectors out of categorical features such diagnosis codes, embeddings were used. Embeddings consisted of a (N, M) table for each categorical feature, where N is the number of distinct categories and M is the size of the embedding vectors. Embeddings can be concatenated together before being fed to neural networks. Therefore, if a dataset is entirely composed of D categorical features, the size of the input of the neural network would be $\sum_{i=0}^{D-1} M_i$, where M_i represents the embedding vector size of the ith categorical feature.

Embeddings helped with dimensional feature spaces, because it transformed categories to fixed-length vectors. This made it possible to encode categorical variables without having to resort to one-hot encoding, where every possible value of a category is represented for each sample, with a "one" denoting the category value. For example, if there are 1000 different principal diagnoses in a dataset of 1000 samples, a one-hot encoded vector would be composed of 999 zeros and 1 one, whereas an embedding vector could represent each category value with 200 numerical values. Preprocessing the dataset for training a deep learning model would therefore lead to a $(1000, 1000)$ matrix for one-hot encoding and a $(1000, 200)$ matrix for embeddings thereby compressing the data while providing a rich representation of each diagnosis.

Compared to one-hot encoding, another benefit of embeddings is that it preserves similarities between concepts, because instead of flattening hierarchical ontologies it tends to represent them in a way that keeps relationships intact [13].

In this work, embedding vectors were created for all categorical features, concatenated together and concatenated with the continuous features. The vectors' size was determined according to the cardinality of each categorical variable's domain. Table III shows the size of the embedding vectors. Categories marked with an asterisk (*) were multi-label and therefore one-hot encoded before being turned into embedding vectors.

TABLE III
STRUCTURE OF THE EMBEDDINGS FOR CATEGORICAL FEATURES.

Category	# distinct values	Embedding vector size
RUM ID	39	43
Hospital	6	15
Medical unit type	6	15
Medical unit	173	99
Entry month	12	22
Entry year	10	20
Entry day of week	7	16
Entry time of day	3	10
Gender	2	8
Principal Diagnosis	708	218
Associated diagnoses*	2	8
Medical procedures*	2	8
Principal Diagnosis for emergency care	279	129
Mode of entry	5	14
Emergency care stay duration	3	10

Embedding vectors were initialized randomly and considered as parameters of the model optimized during the training. The deep learning models were based on a Feed-Forward Neural Network (FFNN), using either the embedding vectors as inputs or the raw features directly. After preprocessing, the dataset had 110 features, the neural network was composed of linear layers of width 512, 256, 192, 128, 64, 32, 15. Scaled Exponential Linear Unit (SELU) activations [14] were used between the linear layers instead of the more commonly used Rectified Linear Unit (ReLU) to help preserving the normalization of the linear layers and to avoid the "dying

ReLU" issue, where neurons of the neural network stop learning. A dropout layer [15] randomly deactivating neurons was used between the embedding and the linear layers to help with generalization. Dropout rate was set to 0.05.

The neural networks' logits were transformed using LogSoftmax so that they could be interpreted as probabilities:

$$\text{LogSoftmax}(x_i) = \log\left(\frac{e^{x_i}}{\sum_{j=0}^{i} e^{x_j}}\right) \quad (1)$$

where x was a logit from the neural network and i represented the classes of the classification.

Figure 1 summarizes the architecture of the FFNN with embeddings.

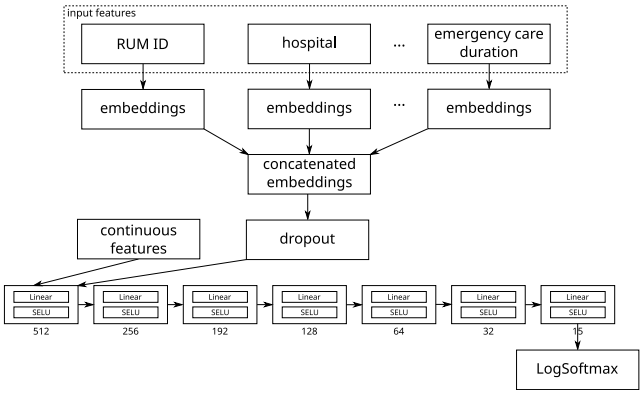


Fig. 1. Architecture of the FFNN with embeddings. Each value of the categorical features were mapped to an embedding vector. For example, the hospital "A" is mapped to a numerical vector of length 15. The hospital "B" would get mapped to another numerical vector of the same length. The values of the embedding vectors were optimized during the training of the FFNN through gradient descent.

The parameters of the models were optimized by stochastic gradient descent with the AdamW optimizer [16]. The learning rate and weight decay, a regularization term aiming at preventing overfitting, were both set to $1e{-}3$ after basic hyperparameter tuning. The loss function was the Negative Log-Likelihood. Both models were trained for 60 epochs.

The LOS prediction was modeled as a classification, with a category per day in the medical unit (RUM) from 0 day to 13 days, and a category for 14 days or more. As a multiclass classification problem, the performances of the models were evaluated using the accuracy and Cohen's kappa [17], because the accuracy may be misleading due to the non-uniformity of the LOS distributions. A dummy classifier giving random predictions according to the LOS distribution is also used for comparison.

The whole experiment was done with Python 3.7. Plots were made using Matplotlib [18] version 3.3.3, the prediction models used PyTorch [19] version 1.8.0 and performance evaluation was done with Scikit-learn [20] version 0.24.0. All steps were done on a GNU/Linux server with 32 gigabytes of memory and an Intel(R) Xeon(R) Gold 6132 CPU with 16 processors. Training took approximately one hour and a half.

III. RESULTS

The LOS distribution was skewed to the right, with a mean of 4 days and a standard deviation of 7. This skew was reported in other studies [1, 3, 6] as well.

The test accuracy and the training loss of the two FFNN (i.e. with and without embeddings) for one fold of the cross-validation are shown in Figure 2. Subfigure A, subfigure B showed the test accuracy and training loss for the FFNN with embeddings, and subfigure C, subfigure D showed the test accuracy and training loss for the FFNN without embeddings. Loss decreased steadily during the first training steps and converged. The FFNN model with embeddings had a better training loss and test accuracy for all epochs during the training process, and the learning was smoother.

A confusion matrix of a fold of the cross validation is shown in Figure 3, normalized over the ground truth labels. The accuracy decreased for longer stays and had an increasing tendency to be exactly 7 days off, for stays shorter than a week. The decrease in accuracy could be linked to the class imbalance, as short, frequent stays may be easier to model and therefore focusing on them may have been an efficient way to minimize the loss function.

The results of the 5-fold cross-validation are shown in Table IV for the FFNN with embeddings, the FFNN without embeddings and a dummy classifier. This showed that neural networks with embeddings had good overall performance, with kappa value considered substantial [21].

TABLE IV
MEAN PREDICTION PERFORMANCES WITH STANDARD DEVIATION ON 5-FOLD CROSS-VALIDATION.

Model type	Accuracy	Kappa
Dummy classifier	$0.13{\pm}1e{-}3$	0
FFNN without embeddings	$0.47{\pm}1e{-}2$	$0.38{\pm}2e{-}2$
FFNN with embeddings	**0.73**$\pm4e{-}3$	**0.67**$\pm6e{-}3$

IV. CONCLUSIONS

This work analyzed the benefits of using embeddings (i.e. creating a rich representation of input features while reducing its size) for LOS prediction using neural networks. It highlights that the promising results of embeddings found in other domains also applies to epidemiological questions such as length of stay (LOS) prediction, and that neural networks on administrative data can give satisfactory results.

A. Strength and limitations

This work focused on generic, well-standardized administrative data from hospital and emergency care, available for all patients. Preprocessing was done with conventional methods. This allows this work to be useful for a majority of inpatient stays and makes it possible to reproduce. By using embeddings, the model could handle both categorical and continuous input variables within the same models. The

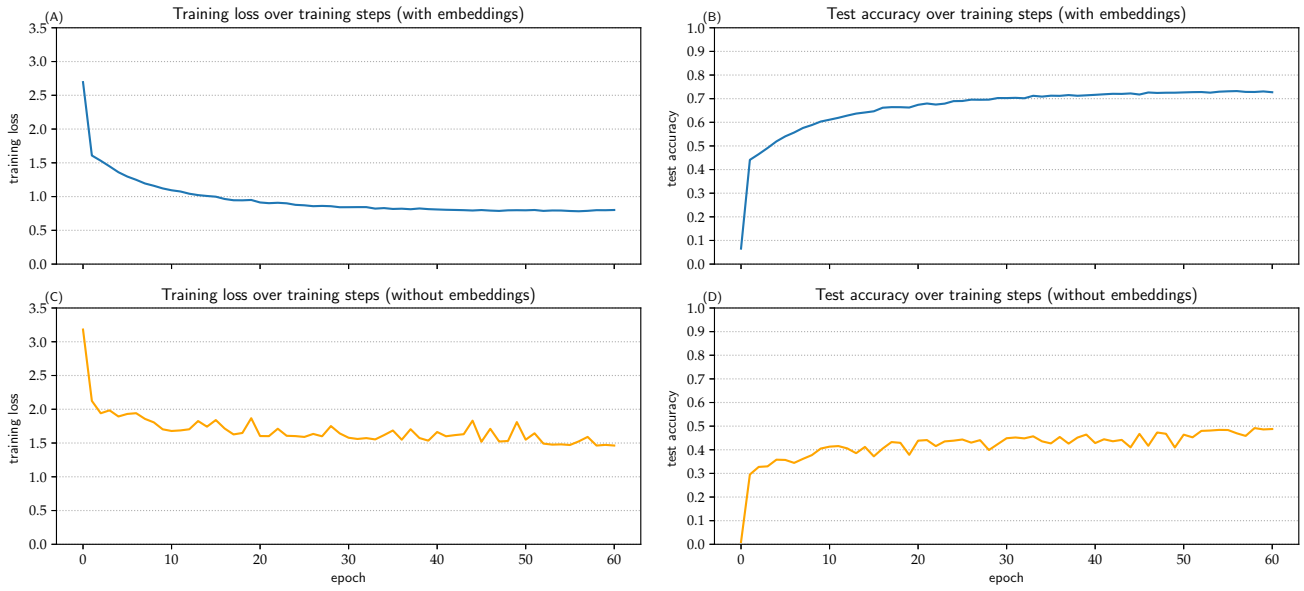


Fig. 2. Training loss and test accuracy of the FFNN model with embeddings (**A** and **B**) and without embeddings (**C** and **D**).

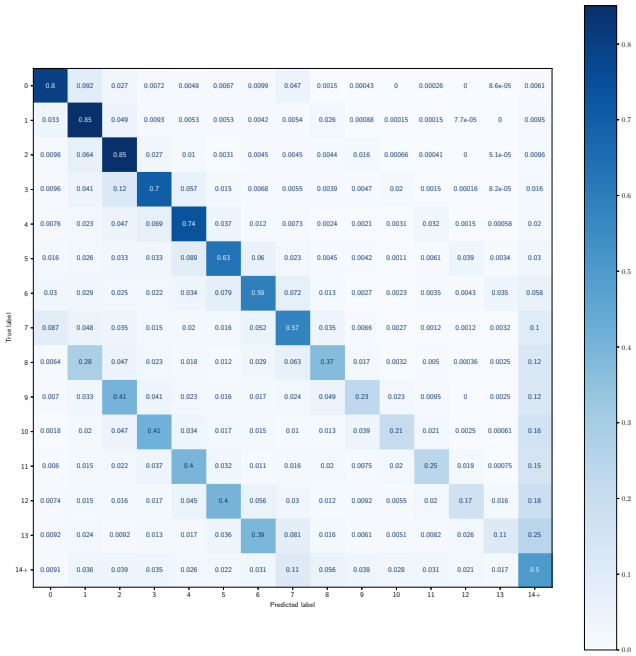


Fig. 3. Confusion matrix for a fold of the cross-validation for the FFNN model with embeddings. Test accuracy decreased for long stays.

precise prediction target gave informative estimation, thereby facilitating bed capacity planning.

On the other hand, using only administrative data restricted the amount of information available for each stay, preventing the model to access information specific to particular healthcare domains or diseases. Moreover, estimating LOS was made difficult by the heterogeneity of the patient profiles. The model's accuracy decreased sharply for longer stays.

B. Implications for future work

Several methodological improvements can be suggested to increase the performance of the model. Preprocessing of the data could be improved by better taking into account the environmental context surrounding an inpatient stay. The number of days used to compute the moving average of LOS and admission numbers at the hospital, medical unit and disease level could be made dynamic depending on some characteristics of the inpatient stays. Diagnosis code and medical procedures data preprocessing could be improved by better exploiting the structure of the medical ontologies. Because it can impact the hospital length of stay, data about access to rehabilitation care could be used to take into account factors influencing the discharge date [22].

While the parameters of the model are optimized by gradient descent, hyperparameter controlling the convergence of optimization process are not automatically optimized. Several strategies exist to find the best combination of hyperparameters leading to the best optimization process in term of prediction performances. The most common one, grid search, tests all possible combination of a set of hyperparameters of a random sample. More advanced techniques such as Nevergrad [23] and Asynchronous Hyperband a scheduler stopping early unpromising combinations [24] could be leveraged to obtain an optimal set of hyperparameters, which is more efficient than a brute-force approach. Using methods that take advantage of the ordinal nature of LOS could also improve the accuracy of the predictions.

Methods to address class imbalance can also be suggested. If a model does not take class imbalance into account, it can have a tendency to focus exclusively on increasing the accuracy of overrepresented classes and ignore the underrepresented ones. Therefore, one could rely on sampling strategies to either downsample the overepresented classes

to a number of samples similar to the one in the underrepresented classes or oversample the underrepresented classes by using some samples multiple times or generating new ones. Another technique would be to deal with class imbalance directly during the training of the model by rewarding the model more for correct classifications of an underrepresented class than correct classifications of an overrepresented one.

Finally, improvements could be made on the interpretability of the deep learning models, which is critical for adoption and trust by healthcare professionals as part of their daily work [25].

ACKNOWLEDGMENT

This work is partially supported by the European Research Ambition Pack 2018 grant, distributed by the French Auvergne-Rhône-Alpes region.

REFERENCES

[1] A. W. Woods, F. N. MacKirdy, B. M. Livingston, J. Norrie, and J. C. Howie, "Evaluation of predicted and actual length of stay in 22 scottish intensive care units using the APACHE III system. acute physiology and chronic health evaluation," *Anaesthesia*, vol. 55, no. 11, pp. 1058–1065, Nov. 2000, ISSN: 0003-2409. DOI: 10.1046/j.1365-2044.2000.01552.x.

[2] M. D. Ray-Zack, Y. Shan, H. B. Mehta, X. Yu, A. M. Kamat, and S. B. Williams, "Hospital length of stay following radical cystectomy for muscle-invasive bladder cancer: Development and validation of a population-based prediction model," *Urologic Oncology: Seminars and Original Investigations*, vol. 37, no. 11, pp. 837–843, Nov. 2019, ISSN: 10781439. DOI: 10.1016/j.urolonc.2018.10.024.

[3] R. M. Dallal and A. Trang, "Analysis of perioperative outcomes, length of hospital stay, and readmission rate after gastric bypass," *Surgical Endoscopy*, vol. 26, no. 3, pp. 754–758, Mar. 2012, ISSN: 0930-2794, 1432-2218. DOI: 10.1007/s00464-011-1947-z.

[4] V. K. Omachonu, S. Suthummanon, M. Akcin, and S. Asfour, "Predicting length of stay for medicare patients at a teaching hospital," *Health Services Management Research*, vol. 17, no. 1, pp. 1–12, Feb. 2004, ISSN: 0951-4848. DOI: 10.1258/095148404322772688.

[5] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, Aug. 2017, ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.2017162326. (visited on 03/12/2021).

[6] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun, "RAIM: Recurrent attentive and intensive model of multimodal patient monitoring data," *arXiv:1807.08820 [cs, stat]*, Jul. 2018. [Online]. Available: http://arxiv.org/abs/1807.08820.

[7] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–10, May 2018, ISSN: 2398-6352. DOI: 10.1038/s41746-018-0029-1.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv:1301.3781 [cs]*, Jan. 2013. [Online]. Available: http://arxiv.org/abs/1301.3781 (visited on 03/10/2021).

[9] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv:1310.4546 [cs, stat]*, Oct. 2013. [Online]. Available: http://arxiv.org/abs/1310.4546.

[10] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record," *IEEE Access*, vol. 6, pp. 65 333–65 346, 2018, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2875677.

[11] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Medical Care*, vol. 36, no. 1, pp. 8–27, Jan. 1998, ISSN: 0025-7079. DOI: 10.1097/00005650-199801000-00004.

[12] S. Haviari, F. Chollet, S. Polazzi, C. Payet, A. Beauveil, C. Colin, and A. Duclos, "Effect of data validation audit on hospital mortality ranking and pay for performance," *BMJ Quality & Safety*, vol. 28, no. 6, pp. 459–467, Jun. 2019, ISSN: 2044-5415, 2044-5423. DOI: 10.1136/bmjqs-2018-008039. [Online]. Available: https://qualitysafety.bmj.com/lookup/doi/10.1136/bmjqs-2018-008039.

[13] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, pp. 41–50, Jul. 20, 2016, ISSN: 2153-4063.

[14] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *arXiv:1706.02515 [cs, stat]*, Sep. 2017. [Online]. Available: http://arxiv.org/abs/1706.02515.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html.

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv:1711.05101 [cs, math]*, Jan. 2019. [Online]. Available: http://arxiv.org/abs/1711.05101.

[17] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, ISSN: 0013-1644, 1552-3888. DOI: 10.1177/001316446002000104.

[18] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8026–8037.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, Nov. 2011. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/1953048.2078195.

[21] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977, Publisher: [Wiley, International Biometric Society], ISSN: 0006-341X. DOI: 10.2307/2529310. [Online]. Available: https://www.jstor.org/stable/2529310.

[22] A. Duarte, C. Bojke, W. Cayton, A. Salawu, B. Case, L. Bojke, and G. Richardson, "Impact of specialist rehabilitation services on hospital length of stay and associated costs," *The European Journal of Health Economics*, vol. 19, no. 7, pp. 1027–1034, Sep. 1, 2018, ISSN: 1618-7601. DOI: 10.1007/s10198-017-0952-0. (visited on 03/14/2021).

[23] J. Rapin and O. Teytaud, *Nevergrad - a gradient-free optimization platform*, 2018. [Online]. Available: https://GitHub.com/FacebookResearch/Nevergrad.

[24] L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, M. Hardt, B. Recht, and A. Talwalkar, "A system for massively parallel hyperparameter tuning," *arXiv:1810.05934 [cs, stat]*, Oct. 2018.

[25] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Computing and Applications*, vol. 32, no. 24, pp. 18 069–18 083, Dec. 2020, ISSN: 0941-0643, 1433-3058. DOI: 10.1007/s00521-019-04051-w.

GLOBAL METHOD FOR PREDICTING THE LENGTH OF STAY IN HOSPITAL USING INCREMENTAL AND EVOLUTIONARY DATA

Laboratoire Décision et Information Pour les Systèmes de Production

Vincent LEQUERTIER
vincent.lequertier@chu-lyon.fr
3rd year of PhD

Supervisors :
A. Duclos, T. Wang, J. Fondrevelle, V. Augusto (collab.)
Project PREDIM

RESHAPE
HCL HOSPICES CIVILS DE LYON
European Commission

Objectives: Patient length of stay (LOS) is the number of days of an inpatient stay. Length of stay prediction accuracy is critical for hospital management and bed capacity planning, which influences healthcare delivery, quality and efficiency. The objective of this work was to predict LOS using deep learning on administrative data from acute and emergency care, and compare this method to machine learning and regression methods.

Methods: A deep learning model leveraging the embeddings mechanism was trained on administrative healthcare data to predict LOS at each step of the patient pathway, and compared random forest and logistic regression.

Results: The deep learning model achieved an accuracy of 0.949 and a linear kappa of 0.948. For the same metrics, random forest yielded 0.597 and 0.643, respectively, and 0.312 and 0.421 for the logistic regression.

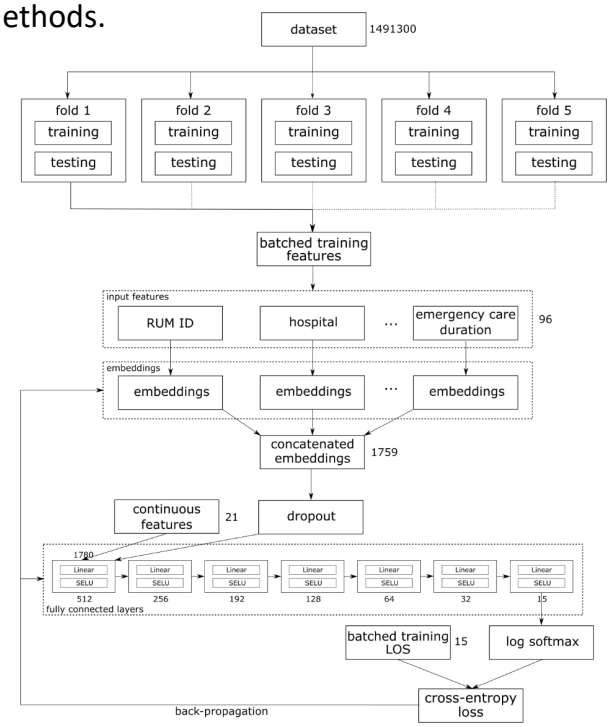
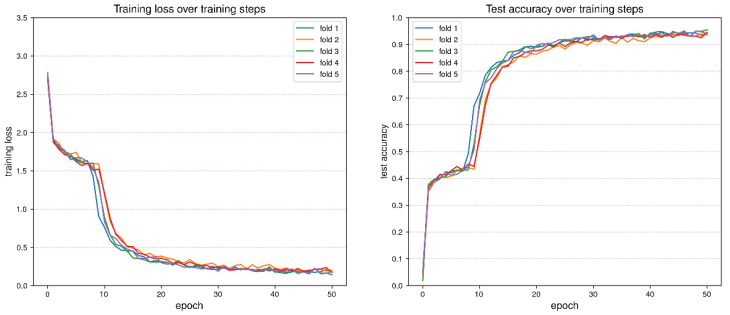


Diagram showing how the deep learning model woks



Training loss and test accuracy of the deep learning model for 50 epochs

	accuracy	Linear Kappa	training time
Deep learning	0.949±7e-3	0.948±6e-3	3 hours 48 minutes
Random Forest	0.597±1e-2	0.643±1e-2	3 hours 52 minutes
Regression	0.312±2e-2	0.421±2e-2	4 hours 15 minutes

Perspectives: We will evaluate the performances more thoroughly and improve the model's usability

V. Lequertier, T. Wang, J. Fondrevelle, V. Augusto, and A. Duclos, "Hospital Length of Stay Prediction Methods: A Systematic Review," *Medical Care*, vol. 59, no. 10, pp. 929–938, Oct. 2021.

V. Lequertier, T. Wang, J. Fondrevelle, V. Augusto, S. Polazzi, and A. Duclos, "Predicting length of stay with administrative data from acute and emergency care: an embedding approach," in *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, 2021, pp. 1395–1400.

2021

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON
Université Claude Bernard Lyon 1
Université Lumière Lyon 2
UNIVERSITÉ DE LYON

Cette étude a suggéré que la méthode des *embeddings* ainsi que les réseaux de neurones vers l'avant ont le potentiel de prédire la durée de séjour avec des données administratives. Dans ce contexte de prédiction des durées de séjours, les *embeddings* engendrent une augmentation importante des performances des réseaux de neurones vers l'avant. Plutôt que d'utiliser les données brutes comme entrées du réseau de neurones, et laisser les poids des couches représenter comment les données d'entrées interagissent avec la tâche de prédiction, représenter les données d'entrées comme des vecteurs numériques permet une représentation multidimensionnelle de chaque variable. On peut supposer que c'est cela qui explique la différence importante d'exactitude entre les deux modèles comparés dans cette étude.

Les performances ont été mesurées par l'exactitude (i.e. le taux de prédictions correctes) et le kappa de Cohen [71]. Les visualisations ont permis de suivre l'évolution de la perte du réseau de neurones vers l'avant ainsi que l'exactitude sur le jeu de test au cours du temps, et de déterminer l'exactitude de la classification par classe, au moyen d'une matrice de confusion, rendant ainsi compte de la capacité du modèle à généraliser et à fonctionner pour toutes les classes.

Ce travail d'analyse a révélé le problème de l'opacité des réseaux de neurones, dont le comportement et la logique est difficile à comprendre. Ce problème s'est manifesté lors de l'interprétation de la matrice de confusion, qui permettait de déterminer l'exactitude de chaque classe de prédiction ou, en d'autres termes, analyser la distribution de l'erreur. En effet, la matrice de confusion a montré que le modèle de prédiction avait tendance à se tromper de 7 jours. Cela illustre bien d'ailleurs pourquoi la mesure de l'exactitude globale peut-être trompeuse, car elle peut cacher des disparités importantes entre les performances pour des classes de prédiction différentes. Étant opaque, comprendre les raisons qui poussaient le modèle à se tromper de 7 jours était difficile, et il a fallu « deviner » les raisons et faire des hypothèses plutôt que d'une manière analytique.

Plusieurs limitations peuvent être constatées. L'évaluation des performances effectuée dans cet article ne permet pas de s'assurer de la qualité des prédictions, notamment en comparaison avec d'autres modèles reposant sur d'autres paradigmes. De plus, il convient de présenter des résultats plus étoffés prenant en compte les considérations techniques évoquées comme (i) améliorer les prédic-

tions en corrigeant les décalages de 7 jours ou (ii) optimiser la recherche d'hyperparamètres et (iii) en améliorant le prétraitement des variables. De plus, utiliser l'intégralité du jeu de données à notre disposition est crucial pour s'assurer de la capacité de notre modèle à généraliser correctement, c'est à dire à fonctionner pour tout type de séjour, incluant des séjours dont il n'a pas eu connaissance pendant la phase d'entrainement.

C'est pourquoi une autre étude a été réalisée, prenant en compte ces éléments.

Chapitre 5

Évaluation des performances de la méthode de prédiction et la comparer avec l'état de l'art

Suite aux résultats encourageants mais insuffisants de l'étude précédente, il était important de corriger les erreurs présentes dans le modèle de prédiction des durées de séjour, d'utiliser l'intégralité du jeu de données et de correctement évaluer ses performances. Cela a fait l'objet du troisième et plus important travail de cette thèse. Il a tout d'abord cherché à identifier les raisons qui poussaient le réseau de neurones à se tromper de 7 jours en prédisant les durées de séjours. Comme ce décalage de 7 jours n'avait pas été relevé à la lecture des articles pour la revue de la littérature (chapitre 3) et n'avait pas d'interprétation médicale, cela ne pouvait être que causé par une erreur facile à corriger. Nous avons donc cherché à identifier la cause de cette erreur. Pour cela, des informations relatives aux dates ont été enlevées itérativement du jeu de données, jusqu'à la disparition de l'erreur. Cela a permis d'identifier qu'utiliser le jour de la semaine comme variable d'entrée cause ce problème, et engendre des décalages de 7 jours entre la durée de séjour réelle et la durée de séjour prédite.

Pour évaluer rigoureusement les résultats du réseau de neurones et obtenir des points de comparaison, une forêt aléatoire et une régression logistique ont également été utilisées. La forêt aléatoire était un modèle d'apprentissage

automatique (*machine learning*) dans lequel plusieurs arbres de décisions sont entrainés. Chaque arbre n'a accès qu'à une sous-partie du jeu de données, à la fois en termes de variables d'entrée et de RUMs (lignes du jeu de données représentant une étape de séjour hospitalier pour laquelle on souhaite prédire la durée). La prédiction finale est celle prédite par une majorité des arbres qui composent la forêt. La régression logistique était un algorithme ajustant les paramètres d'une courbe sigmoid ainsi qu'un seuil de manière à modéliser la distribution de probabilités d'un événement binaire. La fonction sigmoid est donnée par l'équation (5.1) :

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}} \tag{5.1}$$

où p est la probabilité d'une durée de séjour pour une entrée x, β_0 est l'intercepte et β_n est le poids associée à la variable x_n. Un seuil, au-delà duquel on considère que la classe prédite est la classe positive, est également choisi. Comme, dans le cadre de cette thèse prédire la durée de séjour est considéré comme une classification multi-classe, l'approche « OnevsAll » est utilisée, dans laquelle chaque régression logistique binaire prédit si la durée de séjour sera un jour donné ou bien un des autres. La forêt aléatoire et la régression logistique avaient la même cible de prédiction que le réseau de neurones (i.e. prédire si la durée de séjour sera de 0 jour, 1 jour, 2 jours, jusqu'à 14 jours ou plus).

Article 3 : Vincent Lequertier et al. « Length of stay prediction with standardized hospital data from acute and emergency care using a deep neural network ». *In* : *(soumis pour publication)* (2022)

Présentation lors d'un congrès scientifique ADELF - EMOIS : Vincent Lequertier. « Prédiction des durées de séjours avec des données médico-administratives à l'aide d'un réseau de neurones. » Congrés ADELF EMOIS. Dijon, 31 mars 2022. url : https://doi.org/10.1016/j.respe.2022.01.071

Présentation lors d'une réunion scientifique : Vincent Lequertier. « Length of stay prediction with standardized hospital data from acute and emergency care using a deep neural network ». Réunion scientifique RESHAPE. Lyon, 8 avr. 2022

Length of stay prediction with standardized hospital data from acute and emergency care using a deep neural network

Vincent Lequertier[1,2], Tao Wang[3], Julien Fondrevelle[2], Vincent Augusto[4], Stéphanie Polazzi[1,5], and Antoine Duclos[1,5]

[1] Université Claude Bernard Lyon 1, Research on Healthcare Performance (RESHAPE), INSERM U1290, Lyon, France;

[2] Univ. Lyon, INSA Lyon, Univ Lyon 2, Université Claude Bernard Lyon 1, DISP, UR4570, 69621 Villeurbanne cedex, France;

[3] Univ. Lyon, UJM-Saint-Etienne, DISP, UR4570, 69621 Villeurbanne cedex, France;

[4] Mines Saint-Etienne, Univ. Clermont Auvergne, CNRS, UMR 6158 LIMOS, Centre CIS, F - 42023 Saint-Etienne France;

[5] Health data department, Lyon University Hospital, Lyon, France;

Corresponding author: Vincent Lequertier

Address: 162 avenue Lacassagne, Batiment A, 6ème étage, 69003 Lyon, France

Phone number: +33642693530

email: vincent.lequertier@laposte.net

Key words: length of stay, forecasting, artificial intelligence, patient care, clinical decision support, electronic health record

Background: Length of stay (LOS) is an important metric for the organization and scheduling of care activities. This study sought to propose a LOS prediction method based on deep learning using widely-available administrative data from acute and emergency care, and compare it with other methods.

Methods: All admissions between January 1, 2011, and December 31, 2019, at 6 university hospitals of the Lyon metropolis were included, leading to a cohort of 1,140,100 stays of 515,199 patients. Data included demographics, primary and associated diagnoses, medical procedures, the medical unit, the admission type, socio-economic factors, and temporal information. A model based on embeddings and a Feed-Forward Neural Network (FFNN) was developed to provide fine-grained LOS predictions. Performances were compared to random forest and logistic regression, with the accuracy, Cohen's kappa and a Bland-Altman plot, through a 5-folds cross-validation.

Results: The FFNN model achieved an accuracy of 0.944 [CI: 0.937, 0.950] and a kappa of 0.943 [CI: 0.935, 0.950]. For the same metrics, random forest yielded 0.574 [CI: 0.573, 0.575] and 0.602 [CI: 0.601, 0.603], respectively, and 0.321 [CI: 0.319, 0.322] and 0.431 [CI: 0.428, 0.434] for the logistic regression. The FFNN model had a limit of agreement ranging from -2.73 to 2.67, which was better than random forest (-6.72 to 6.83) or logistic regression (-7.53 to 8.20).

Conclusions: The FFNN model was better at predicting LOS than random forest or logistic regression. Implementing the FFNN model for routine acute care could be useful for improving the quality of patients' care.

BACKGROUND

Inpatient Length Of Stay (LOS) represents the number of hospital days between a patient admission and discharge. Because anticipating LOS of patients is helpful for managing bed occupancy and patients' discharge, predicting it accurately contributes to better resource allocation and improves the quality of patient care delivery (1,2). Conversely, incorrect predictions may disturb the organization of medical units and potentially lead to an increased patient waiting time and exposure to hospitalization-induced complications. Moreover, because patients have diverse profiles and LOS is associated with a large number of factors, predictions made by healthcare professionals may lack accuracy (3) and depend on the experience of the healthcare professionals.

Automated LOS predictions have therefore been of high interest. Because artificial-intelligence-based prediction methods such as deep learning have shown great results in other areas such as computer vision (4,5) or natural language processing (6), using them for predicting LOS has been compelling. Due to the sparsity and complexity of medical concepts, and the quality of results it yielded in other data science areas, embeddings, a method for transforming categorical data into numerical vectors, have seen a lot of use for tackling data science problems in healthcare. Examples include medical events processing (7), where embeddings were trained using an attention mechanism (8), diagnosis codes assignment (9), or early detection of heart failure (10). More specifically on LOS prediction, machine learning and deep learning are more and more used (11). For example, embeddings with either a Recurrent Neural Network or a Feed-Forward Neural Network coupled with an attention mechanism were used to predict LOS (12,13). Handling multimodal data and making the models interpretable were important goals in those studies. However, their prediction targets were binary discrimination between short and long duration inpatient stays, thus limiting their usefulness for the purpose of improving healthcare resource management, which would need a more fine-grained prediction of LOS. Regression models were

used to predict LOS and compared together (14). However, predictions were only made for the Intensive Care Units. Machine learning methods like random forests, support vector machine, Bayesian network, and artificial neural network were also used and combined (15). Another ensemble model based on decision trees was used to make explainable predictions, but the corresponding predictions were only made at the time of admission (16), thus ignoring possible new information about the patient's diagnosis, complications and adverse events occurring along the patient's pathway. This study tried to address these points by predicting LOS multiple times during a hospitalization with data available for all patients and by having a precise prediction target, potentially allowing for improvements in hospital's resources management.

The objective of this work was to predict LOS for all hospitalized patients using deep learning on administrative data from acute and emergency care, which are well-standardized and thus widely available in hospitals for day-to-day care. In this work, a deep learning model leveraging the embeddings mechanism was trained on administrative healthcare data to precisely predict LOS at the start of each step of the patient pathway (i.e. mutation or transfer to a new medical unit), and was compared to other approaches from the regression and machine learning domains.

METHODS

Data extraction

The data were extracted from the medico-administrative data warehouse of the Hospices Civils de Lyon, a multicenter institution including 6 acute care university hospitals located across the Lyon metropolis, France. Any inpatient stays whose entry date was between January 1, 2011, and December 31, 2019, were included, leading to a dataset of 667,090 patients, accounting for 1,403,112 stays and 1,798,447 steps in a patient pathway (hereafter RUM). Exclusion criteria were as follows: 1) Exclusion of patients less than 18 years old at entry date, 2) Exclusion of medical units with fewer than 100 stays performed in the time period, or stays with a nominal duration, in order to focus on patient stays whose LOS is unknown beforehand, 3) Exclusion of stays with data coding errors, 4) Exclusion of RUMs longer than 31 days because they were infrequent. After applying the exclusion criteria, there were 1,140,100 stays of 515,199 patients, for a total of 1,491 300 RUMs (see study flow chart in Figure 1). This study was based on anonymous data. In accordance with the French ethical directives, it was declared to the French National Data Protection Commission (CNIL) with reference DR-2020-196.

The data were using the Information Systems Medicalization Program (Programme de Médicalisation des Systèmes d'Information in French) standard, akin to Diagnosis-Related-Groups (DRGs) which are medico-administrative data used for reimbursement purposes. The data were linked to the emergency care summary database (Résumé de Passage aux Urgences in French) at the stay level. The emergency care data were left blank for patients who were not coming from emergency care. Data extracted for hospitalization were gender, age, hospital center, medical unit, time of entry, time of discharge, type of hospitalization, mode of entry, primary and associated diagnoses in the International Classification of Disease, 10th edition (ICD-10), medical procedures in the French Classification of Medical Procedures (CCAM), universal healthcare coverage, and the

Glasgow Coma Scale score. Data extracted for emergency stays were primary and associated diagnoses in the ICD-10, medical procedures in the CCAM, stay duration, and urgency category. There were no missing data allowed in this national database according to the regulations.

Data preprocessing

To simplify the processing of associated diagnoses, only the diagnoses included in the Elixhauser risk score calculation were retained, as it has been suggested that the Elixhauser comorbidities are associated with LOS (17) and have been validated using French data (18). “AIDS/HIV”, “Drug abuse”, “Psychoses”, “Peptic ulcer disease excluding bleeding”, and “Lymphoma” were removed from the Elixhauser comorbidities because they had a very low prevalence, leading to 26 comorbidities finally retained.

To select worthy medical procedures, the mutual information gain (19) with regard to LOS and the frequency of each medical procedure were added together. Procedures with a score lower than twice the mean were removed, thus retaining 53 medical procedures out of 6,866.

To account for temporality (i.e. time and date of admission), seasonality, and concept drift, the mean LOS overall, by diagnosis, medical unit, and the number of admissions overall by diagnosis, medical unit, and patient were computed for the last 30 days, as well as the time already spent in the hospital within the same hospitalization. Because of the heterogeneous nature of the dataset, there were a lot of different primary diagnoses, with a right-skewed distribution. Moreover, it was known that learning from rare events was difficult (20), and removing these rare events reduced the model’s usefulness as it made it unusable for those rare events (i.e. rare diagnoses). As a workaround, rare diagnosis were therefore regrouped together, following this logic: trim the diagnosis to the right of the ICD-10 code until the diagnosis was present in at least 1% of the dataset or the diagnosis is 3 letters long or less. Because of the hierarchical nature of ICD-10 codes, removing letters was equivalent to reducing the granularity. To accommodate for the fact that

regrouped diagnoses may end up being different and lead to a wide range of LOS, the quartiles of LOS by diagnosis and by diagnosis for the medical unit in the last 30 days were added to the model. Diagnoses in emergency care and acute care were considered identical if they met the following two conditions: 1) the Levensthein distance (21) between them was lower than two after the three first characters (forming the ICD-10 chapter) and 2) if they had the same first three characters.

Table 1 provides descriptive statistics after feature selection. Across the dataset, the arithmetic mean LOS was 4.1 and the median was 3. Overall, there were a total of 702 different diagnoses in the dataset. 23.4% of the RUMs were linked to emergency care data. Appendix A shows the details of the data preprocessing process.

To make the predictions useful for the purpose of potentially improving the hospital's resources management, length of stay was categorized as unique classes from 0 to "14 days and more", yielding a multi-class classification task with 15 categories. To make LOS predictions useful for the purpose of improving bed management, they were made at the level of the RUM rather than the whole hospital stay. This also allowed predictions to be refined as more information about the patient's condition were made available.

Prediction methods

The prediction model used a Feed-Forward Neural Network (FFNN), as a Recurrent Neural Network would not be useful for cases where the hospital admission had a single step. Because neural networks were designed to work on numerical inputs, categorical features needed to be converted to numerical ones. Moreover, there was a need for a rich representation of medical concepts which preserved their relationships. A tool called "embeddings" was used to these purposes, which acted as a lookup table between each value of the categorical input and the corresponding numerical vector. The numerical vectors values were initialized with random numbers following the normal distribution and incrementally refined during the model's training

through stochastic gradient descent. Because all categorical features did not have the same complexity (e.g. days of the week were simpler concepts than main diagnoses), the sizes of the numerical vectors were proportional to the number of unique values of the categorical inputs, thus encouraging the model to learn a rich, multi-dimensional and numeric representation for complex concepts and a simpler one for more basic concepts. The embedded representations of the categorical inputs were concatenated together before being fed into the neural network. Continuous inputs were concatenated to the numerical vectors produced by the embeddings and fed into 7 blocks of fully connected layers and Scaled Exponential Linear Unit (SELU) activation functions (22), whose last size was the number of categories (i.e. 15). The SELU activation function was used to ensure the distribution remained normal throughout the neural network layers. The last layer of the neural network was a log Softmax function, which allowed interpreting the output of the neural network as the probability of each class (i.e. the LOS), thus allowing making predictions based on the neural network's output. Stochastic gradient descent was used to optimize the weights and biases of the neural network's neurons (23). More specifically, the Adamw (24) variant was used to take advantage of its fast and stable convergence properties. The loss function used was the cross-entropy loss:

$$H(p,q) = -\sum_{i=1}^{N} w_i . p_i \log(q_i)$$

where p_i is the ground truth probability of the i^{th} category, q_i is the predicted probability of the i^{th} category, and w_i is the weight of the i^{th} category. The weights were used to reward more correct prediction of the underrepresented classes (i.e. long LOS), thus mitigating class imbalance. They were computed as follows:

$$w_i = \frac{n_{samples}}{n_{classes} \times count(i)}$$

Figure 2 summarizes the training of the deep learning model.

To compare the deep learning model with other popular classification methods, a random forest model and a logistic regression were also trained. Random Forest was an ensemble learning model where multiple decision trees were trained with bagging (25) as a mean to reduce the individual trees' variance. The final prediction was the class predicted by most trees composing the random forest. Logistic regression was a prediction method which tried to approximate the relationship between a target variable and its potential predictors by adjusting the coefficient of each input variable in the dataset. Because the target of the prediction was categorical, multinomial logistic regression (26) was used, with L2 regularization. As the logistic regression was not efficient at handling many different one-hot encoded diagnoses, the primary diagnoses were only represented as mean and quartiles of LOS per diagnosis for this model.

Hyperparameters

Because the three methods had various configurations, hyperparameter tuning was done on a random subset of 10% of the training test. Table 2 shows the hyperparameters for the FFNN model and the random forest models, respectively, along with a description and candidate and chosen values. Specifically, a system including early stopping and parallel processing of the configurations (27) was used for the FFNN, as well as the OnePlusOne optimizer of Nevergrad (28) for finding the optimal configuration. The reduction factor was set to 4, the grace period was set to 2, and max budget to 50. A parallel grid search was conducted for identifying the best hyperparameters for the random forest and logistic regression models.

Study design and evaluation methods

Once the hyperparameters were selected, training of the models and performance evaluation could be conducted. To have a rigorous measure of the model's performance, 5-folds cross-validation was employed, where the models were trained on 5 different random training sets so that each RUM

belonged to a held-out test set exactly once across the 5 folds. Ensuring that each data point was part of the test sets exatly once was important to guarantee an unbiased evaluation. The evaluation of performance was only done on the 5 held-out test sets.

Because the prediction task was a multi-class classification, the ratio of correct classifications (i.e. accuracy) and linear Cohen's kappa were used (29) to determine the FFNN model's performance compared to random forest and logistic regression. A Bland-Altman chart (30) was also plotted for the FFNN, random forest, and logistic regression models for the 5 folds of the cross-validation. Each point's coordinates were therefore the mean between the predicted class and the ground truth, and their difference. As the prediction task was categorical, the mean and difference had discrete values. The points were thus colored according to their number of occurrences. The limits of agreement were shown with a 95% confidence level.

A confusion matrix was plotted for the FFNN model, where each prediction was added to a count table where the rows represented the predicted class and the columns represented the actual, ground truth one. The table was then normalized across the columns. This provided a way to analyze the prediction results in terms of error distribution.

All steps were performed on a GNU/Linux server with 128 gigabytes of memory and an Intel(R) Xeon(R) Gold 6132 CPU with 16 processors. The FFNN model was developed using PyTorch 1.9 (31) and the machine learning and logistic regression models used Scikit-Learn 0.24 (32).

RESULTS

Table 2 shows the selected hyperparameters. For the FFNN model, a wider model with large embedding vectors with a factor of 7, a rather large 1,024 batch size with a dropout of 1e-1, a learning rate and regularization of 1e-3 yielded the best results. Batch normalization did increase the FFNN accuracy. Increasing the number and depth of the trees composing the random forest to 600 and 50, respectively, improved the accuracy. An L2 regularization strength of 1e-2 was chosen for the logistic regression.

Table 3 shows the evaluation metrics for the three models considered. The accuracy and linear kappa were 0.944±5e-3 and 0.943±5e-3 for the FFNN model, 0.574±1e-2 and 0.602±1e-2 for the random forest, and 0.312±2e-2 and 0.431±2e-2 for the logistic regression. Training times were comparable. Figure 3 shows the training loss (panel A) and test accuracy (panel B) of the FFNN model for each fold of the cross-validation. The training was done for 50 epochs. Training loss decreased, while the test accuracy increased. Figure 4 shows a confusion matrix for the FFNN model which allows analyzing the accuracy in more detail. The ratios of correct predictions were above 0.80 for all classes, and above 0.90 for 12 out of 15.

Table 3: Results of the 5 folds cross-validation. ACC: Accuracy. Training times were shown for one fold of the 5-folds cross-validation.

	ACC	**Linear Kappa**	**training time**	**Bland-Altman Limits of Agreement**
FFNN model	**0.944±5e-3 [CI: 0.937, 0.950]**	**0.943±5e-3 [CI: 0.935, 0.950]**	**3 hours 48 minutes**	**[-2.73, 2.67]**
Random forest	0.574±9e-3 [CI: 0.573, 0.575]	0.602±1e-2 [CI: 0.601, 0.603]	3 hours 52 minutes	[-6.72, 6.83]
Logistic	0.321±2e-1 [CI:	0.431±2e-2 [CI:	4 hours 15	[-7.53, 8.20]

	ACC	Linear Kappa	training time	Bland-Altman Limits of Agreement
regression	0.319, 0.322]	0.428, 0.434]	minutes	

Figure 5 shows a Bland-Altman (30) plot for the FFNN, random forest, and logistic regression models, in panels A, B, and C, respectively. The FFNN model had a limit of agreement ranging from -2.73 to 2.67, compared to random forest (-6.72 to 6.83), or logistic regression (-7.53 to 8.20).

DISCUSSION

This work presented a prediction model for length of stay (LOS) using standardized medico-administrative data and compared it to other state-of-the-art methods from the machine learning and statistics domains. The confusion matrix showed that predictions were accurate for frequent inpatient length of stay, for up to 14 days or more, with a slight accuracy drop for longer RUMs. The prediction model could therefore be used throughout the whole patient pathway. The FFNN model outperformed the random forest and logistic regression models across all metrics considered. The training speed of the FFNN model was as fast as other models and there was a negative correlation between the training loss and the test accuracy. Furthermore, the test accuracy kept increasing, therefore suggesting a good generalization capacity.

The approach of using embeddings for encoding categorical features to create a representation suitable for neural networks has been deemed superior to other approaches (33). Automated LOS prediction has been a research topic for 5 decades (11). Studies using deep learning reported an AUC of 0.86 with data up 24 hours after the patient's admission (13) or an accuracy of 0.86 (12). Studies leveraging machine learning methods reported a MAE of 0.80 (16) using the ensemble, trees-based method Cubist, or a MAE of 8.99 with Support Vector Machine (34). Logistic regression was also used in several studies, which reported an AUC of 0.71 (35) or a sensitivity of 0.72 and specificity of 0.55 (36). However, these results were hard to interpret and compare with this study because of the discrepancies in input data size, time period, inclusion criteria, nature of the prediction target, study design evaluation metrics, and way of reporting.

Data were extracted from standardized, widely available medico-administrative data from 6 hospitals in Lyon, France, making the study easily reproducible in other hospitals. This LOS prediction methodology can thus be reproduced worldwide based on inpatient abstracts using a common set of data that are routinely collected in many countries (37). Moreover, because the data

included all medical units and there were no exclusion criteria based on medical characteristics, the study cohort was large and diverse, and predictions could be made for all patients. Preprocessing was required in order to simplify the data and to take seasonality and temporality into account. Because embeddings mapped categories to dense vectors of arbitrary sizes, they suffered neither from the sparseness of one-hot encoding nor from the semantic issues of ordinal encoding which stemmed from the fact that they could lead the neural network to interpret categorical values as raw numbers. Considering LOS as a categorical outcome allowed to compute the accuracy and kappa, which may have more meaning and easier interpretation than performance measures for continuous outcomes. Making sure that stakeholders outside of the data science community can be confident about the performance of the prediction model is key to ensure the model may be deployed and used on a daily basis. The prediction model was based on a Feed-Forward Neural Network (FFNN), and a 5-folds cross-validation design was set up in order to evaluate it with several metrics computed over the 5 mutually-exclusive test sets. However, this work suffered from several shortcomings. Data were extracted from 6 hospitals in a single city, thus limiting our ability to assess the generalization of the model to other hospitals and policies thereof. In France, during the time period of this study, the primary diagnosis was coded after the patient's discharge. Using the LOS prediction model in routine care would therefore require coding a projected diagnosis at the patient's admission that could be corrected at the discharge if needed. This projected diagnosis may use pre-admission information from scheduled care, the diagnosis from the previous steps in the hospitalization, or data from emergency care. Given that the projected diagnosis may not be correct, this could decrease the predictions performances. Because the FFNN model was complex, completely explaining its predictions and displaying them in an interpretable way to end-users (i.e. healthcare professionals) was non-trivial, and could hinder trust and adoption. While the cross-validation study design allowed to prevent bias in the dataset selection and detect overfitting, it

could induce an optimistic evaluation of the performances. Given the temporal nature the data used and the sampling of the cross-validation, the 5 test sets may contain temporal bias (38). The performance of the model was slightly reduced for longer LOS presumably because they were infrequent and more complex. The prediction model was not yet used in routine care. Using the model on a daily basis in routine care would allow to validate the practical usefulness of the LOS prediction model.

Directions for future research

Using more detailed data, having better preprocessing and improving the prediction method could have increased the performance of the model. Specifically, the integration of patients' physiological data or data specific to patients' diagnoses could provide more detailed pictures of their condition, thus paving the way for better predictions. Additionally, because patients going to rehabilitation care could not be discharged from the hospital if a bed for rehabilitation care was not available, thus inducing delays (39), using rehabilitation care data to determine rehabilitation care availability could improve the prediction's accuracy. Moreover, it is known that LOS is correlated to healthcare professionals' experience (40). Adding more data about the team and professionals carrying out the medical procedures could potentially improve the quality of LOS predictions.

Moreover, the number of past days used for computing aggregated temporal statistics (i.e. 30 days) could be made dynamic, based on the patient's and hospital's characteristics and optimized through gradient descent. This would allow for a better representation of seasonality and seasonality in the model. Having a more thorough exploration of hyperparameter combinations by expanding the hyperparameter space being considered and increasing the budget of each trial could improve the training of the neural network. Pairing the FFNN model with one designed to model sequential, temporal data such as a recurrent neural network (RNN) or an attention mechanism (8) could also lead to improvements in the performances of the model, especially for longer stays which may

contain a lot of information, thereby making them harder to summarize with aggregated temporal statistics. Finding a better loss function weighting scheme or leveraging other methods aiming at mitigating class imbalance such as weighted sampling or oversampling (41,42) has the potential of improving the model performances for long LOS. A method leveraging the ordinal nature of LOS was tried, by training binary classifiers that told whether a stay was shorter or longer than a given value until reaching a final prediction, but did not yield satisfactory results. More work would be needed to fully explore this approach.

Using an attribution algorithm such as the integrated gradients (43) would help making the predictions more interpretable, thus fostering trust in them. This would allow for the identification of factors which are the most associated with LOS in different contexts, thereby allowing for better healthcare policies (44) and decision-making. A similar analysis has already been done for hospital mortality (45). Analyzing the various learned embeddings spaces would allow to check if the relationships between concepts have properly been encoded. For example, checking that similar medical units or similar diseases are also similar in the embedded space may give more confidence in the training of the neural network.

Finally, a continuous prediction target or one that is more precise for stays longer than 14 days could produce a more detailed picture of long-term bed availability for medical units where stays tend to last longer. The prediction model was not deployed yet. Putting the model in production and integrating it into the hospital's information system used by the healthcare professionals would allow for a qualitative assessment of the model's usefulness and an adjustment of the prediction target according to the needs of healthcare practitioners and bed managers. Using a temporal-based study design would help fostering trust in the performances of the model once in production.

Practical implications

Accurate LOS predictions for steps in the patient's journey were helpful from an organizational perspective. As the patient moved across medical units in the hospital, predicting LOS within a medical unit allowed the bed manager to have an understanding of future bed occupancy. Therefore, he or she could better schedule bed use, as well as other resources. Resource planning could also be done automatically. In this case, the LOS prediction of each patient's stay could be used as an input for constraint programming (1).

Moreover, being able to estimate patients' LOS made it possible to prepare the discharge, allowing smoother transfers to other medical units and transition to the next steps in the healthcare pathway, or hospital discharge. More generally, LOS prediction could play a role in patient flow improvements. Furthermore, because LOS predictions helped to better organize medical units, this could reduce resources waste and limit overcrowding which is associated with bad scheduling (46) and leads to long waiting time for patients as well as dissatisfaction and anxiety for healthcare professionals (47). Quality of care could consequently be better with correct LOS predictions.

Because the data used were standardized and readily available, and the model was tested on a large and diverse cohort, the prediction model could be implemented in other hospitals with ease. Given the relative complexity of the preprocessing process, the number of input variables involved and the number of parameters of the FFNN model, it would be inconvenient to use a hard-coded results table. The prediction model would therefore have to be used in production, and its accuracy should be monitored to detect concept drift. The model could be retrained regularly to account for changes in healthcare policies and practices. Moreover, simplifying the model could make it easier to interpret and faster to train. A simpler version of it, with less layers and of narrower sizes, could also be deployed in production.

CONCLUSION

Length of stay can be effectively predicted with administrative data from acute and emergency care using a feed-forward neural network. The neural network outperformed random forest and logistic regression in all metrics considered. Performances can be further improved by integrating more data and through methodological improvements. Integrating the prediction model into the hospital's information system and day-to-day use by healthcare practitioners would complete the assessment of the model's practical usefulness.

FIGURES

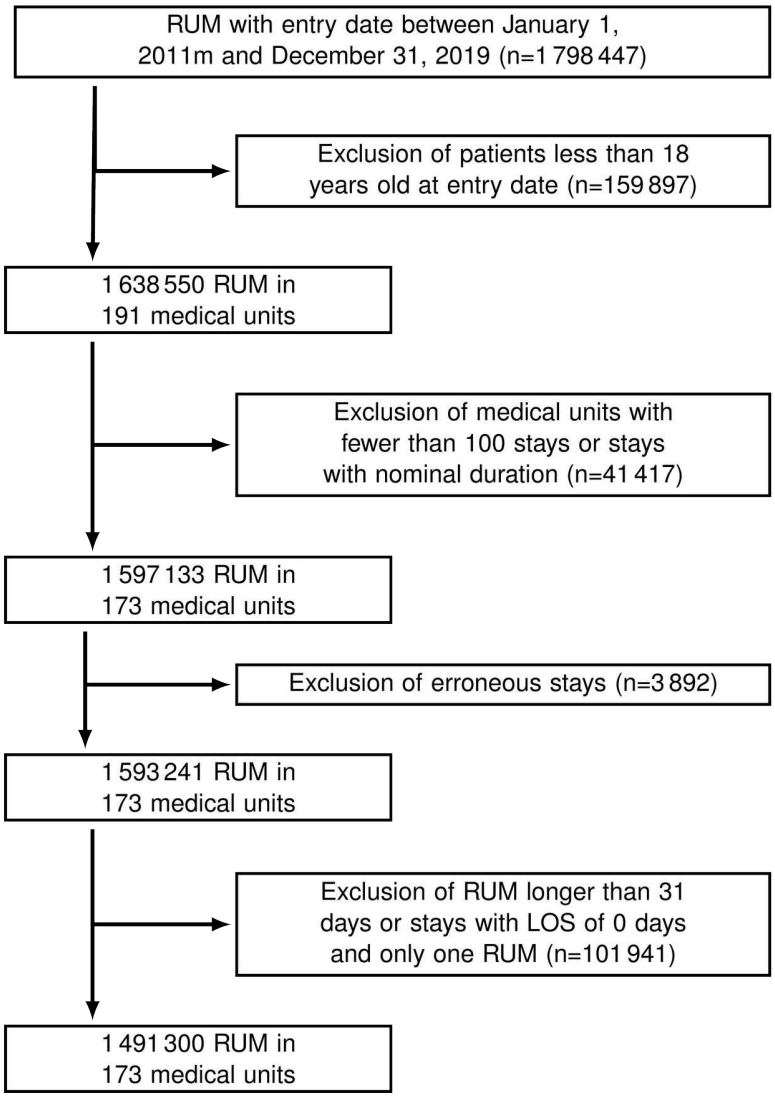


Figure 1: Flowchart of the exclusion criteria

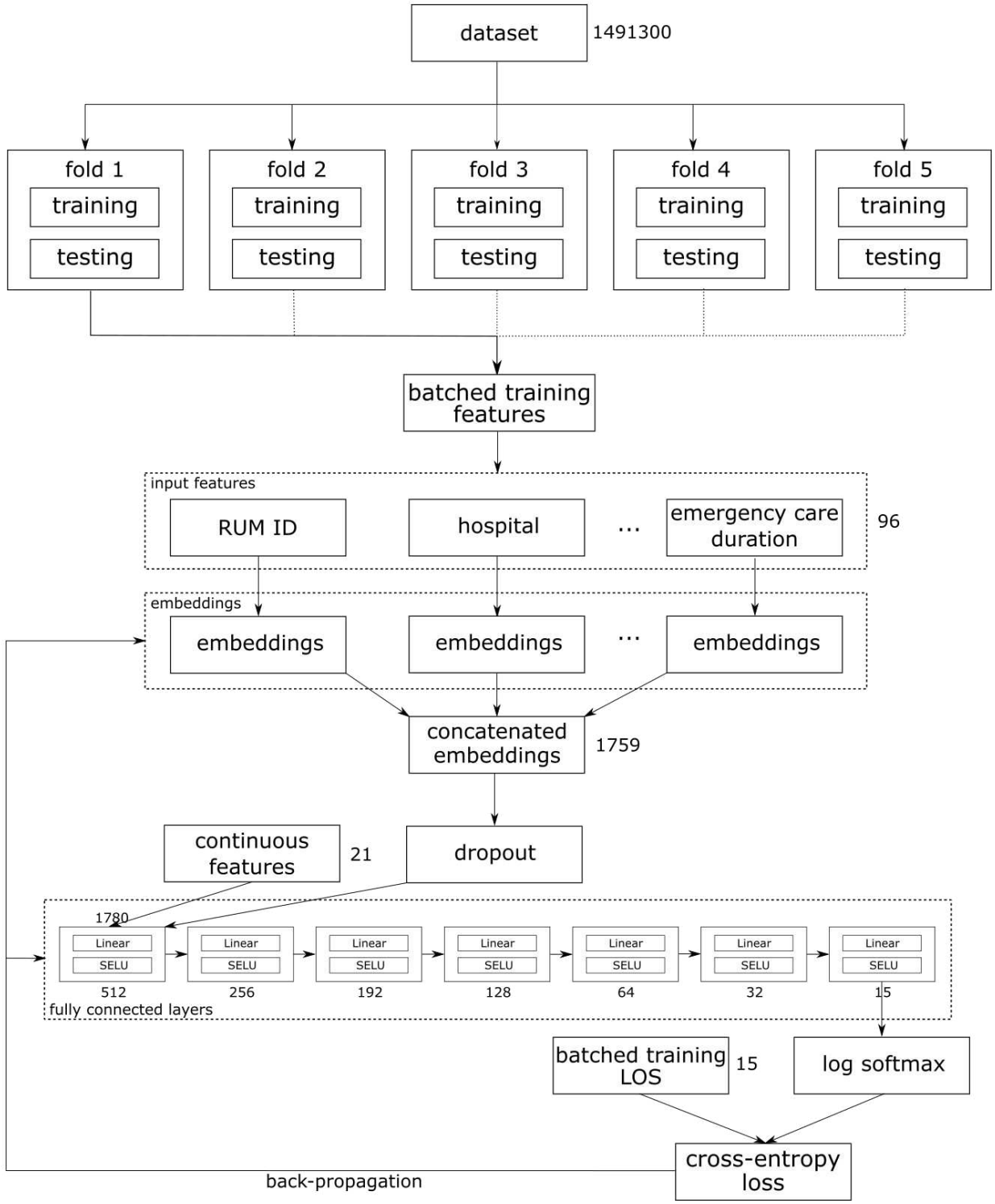


Figure 2: Training of the deep learning model. The dataset was randomly split into 5 folds. Then, for each fold, a deep learning model composed of embeddings followed by a feed forward neural network was trained. A loss function compared the probabilities coming from the neural network to the ground truth probabilities. This loss was in turn used to iteratively update the neural network parameters and the embeddings vectors.

Figure 3: Training loss (A) and test accuracy (B) of the FFNN model for the 5 folds of the cross-validation, over the 50 epochs. The training loss was minimized during training, therefore, a low loss meant a good model training. The accuracy was the number of correct predictions. The training loss and test accuracy were negatively correlated throughout the training for the 5 folds.

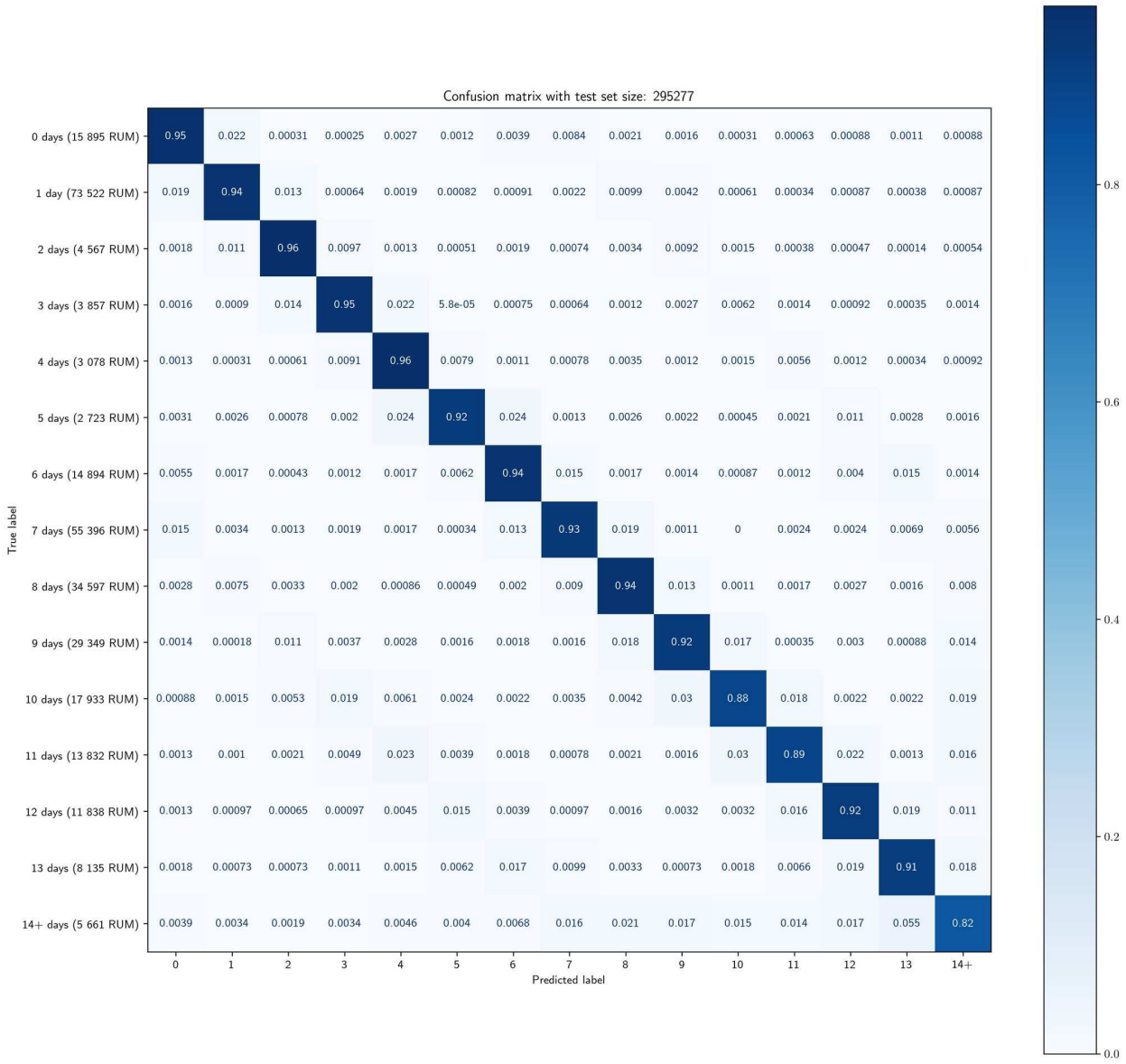


Figure 4: Confusion matrix of the FFNN, normalized by columns. Performances were fairly stable for all classes, with a small decrease for longer stays

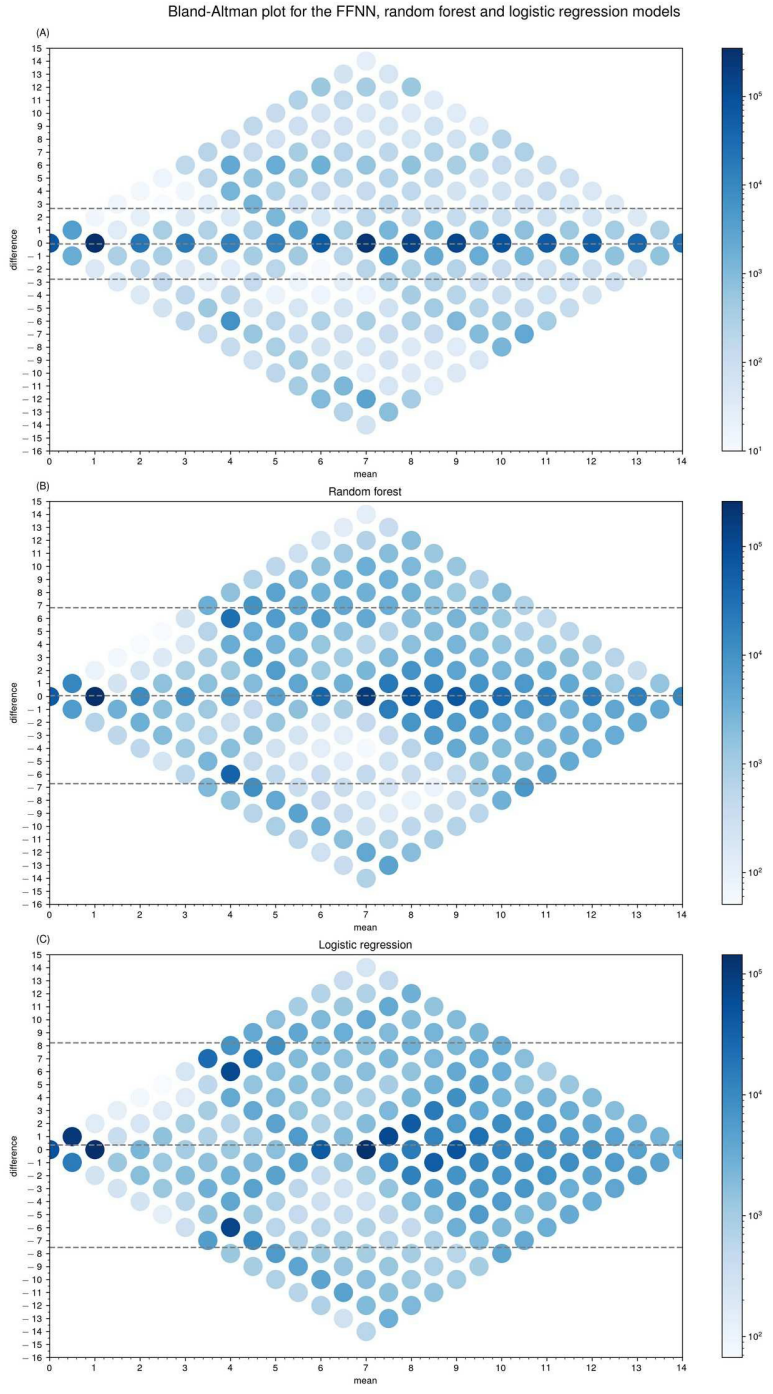


Figure 5: Bland-Altman plot for the FFNN (A), random forest (B) and logistic regression (C) models, with the difference and mean of the predicted and ground truth classes. For example, for the FFNN model, there were about 250,000 points where the mean between predicted class and the ground truth was 7. Each point was colored according to its number of occurrences (i.e. the number of points with the same difference and mean) in a logarithmic scale. The numbers of occurrences were not homogeneous across the 15 classes, hence the color discrepancies. The dashed lines represented the difference's mean and the limit of agreement at a 95% confidence level, therefore showing an estimation of the differences mean. A difference of 0 meant a perfect prediction.

TABLES

Table 1: Summary of the data selected after preprocessing. The data also contained 26 one-hot encoded associated diagnoses and 53 medical procedures.

		Overall
n		1,491,300
RUM number (step in the patient pathway), mean [min,max]		1.4 [1.0,45.0]
Hospital center, n (%)	A	79645 (5.3)
	B	331,696 (22.2)
	C	136,404 (9.1)
	D	376,022 (25.2)
	E	398,963 (26.8)
	F	168,570 (11.3)
Hospitalization type, n (%)	Medicine	610,834 (41.0)
	Surgery	440,425 (29.5)
	Obstetrics	175,896 (11.8)
	Reanimation	171,521 (11.5)
	Emergency	886,79 (5.9)
	Palliative	3,945 (0.3)
Gender, n (%)	Female	7,74114 (51.9)
	Male	7,17186 (48.1)
Age, mean [min,max]		58.3 [18.0,118.0]
Mode of entry, n (%)	Came from home	702,056 (47.1)
	Hospital transfer	372,613 (25.0)
	Emergency	336,510 (22.6)
	Ward transfer	79,665 (5.3)
	Other kind of transfer	456 (0.0)
Day of entry, median [min,max]		16 [1,31]
Month of entry, n (%)	January	135,715 (9.1)
	February	123,103 (8.3)
	March	134,047 (9.0)
	April	125,814 (8.4)
	May	125,213 (8.4)
	October	133,643 (9.0)

		Overall
	June	129,260 (8.7)
	July	117,360 (7.9)
	August	94,148 (6.3)
	September	126,443 (8.5)
	November	127,498 (8.5)
	December	119,056 (8.0)
Year of entry, median [min,max]		2016 [2011.0,2020.0]
Whether Glasgow Coma Scale was present, n (%)	No	1,350,464 (90.6)
	Yes	140,836 (9.4)
Healthcare universal coverage, n (%)	No	1,420,475 (95.3)
	Yes	70,825 (4.7)
Time of day of entry [a], n (%)	Afternoon	746,804 (50.1)
	Morning	486,120 (32.6)
	Night	258,376 (17.3)
Went to emergency care, n (%)	No	1,141,641 (76.6)
	Yes	349,659 (23.4)
Time spent in emergency care, n (%)	less than 8hours	1,408,504 (94.4)
	less than 12 hours	50,101 (3.4)
	more than 12 hours	32,695 (2.2)
Urgency category [b] , n (%)	None	1,141,641 (76.6)
	1	12,103 (0.8)
	2	141,147 (9.5)
	3	163,786 (11.0)
	4	19,887 (1.3)
	5	8,985 (0.6)
	P	3,622 (0.2)
	D	129 (0.0)
Length of stay, n (%)	0	81,125 (5.4)
	1	370,577 (24.8)
	2	280,550 (18.8)
	3	175,049 (11.7)
	4	148,777 (10.0)
	5	90,635 (6.1)

		Overall
	6	69,528 (4.7)
	7	59,276 (4.0)
	8	41,270 (2.8)
	9	27,963 (1.9)
	10	22,986 (1.5)
	11	19,164 (1.3)
	12	15,247 (1.0)
	13	13,901 (0.9)
	14+	13,501 (0.9)

[a] Morning was considered to start at 6 a.m. and night was considered to start at 8 p.m.

[b] Category of the urgency. An urgency of 1 meant a stable clinical condition, while an urgency of 5 meant a life-threatening condition. “P” denoted a mental health problem and patients marked with “D” were deceased upon arrival.

1 Table 2: Hyperparameters for the prediction models along with the values tested and selected.

Hyperparameter name	**Description**	**Possible values**	**Selected value**
FFNN model			
Learning rate	Learning rate of the gradient descent algorithm (or initial learning rate for adaptive gradient descent algorithms)	From 1e-6 to 1	1e-3
Weight decay	Weight of the regularization term in the loss function	From 1e-6 to 1	1e-3
Layers	Number and width of layers	Between 3 and 6 layers, with sizes from 1,024 to 32	[512, 256, 192, 128, 64, 32]
Use batch normalization	Whether to use batch normalization (48)	True or False	False
Batch size	Size of the batch in the training loop	Power of 2 between 128 and 2,048	1,024
Gradient descent algorithm	Gradient descent algorithm	Momemtum (49), Adam (50), AdamW (24)	AdamW
Embeddings size factor	Used to compute the ratio between embeddings size and number of unique values of the categorical features	From 1 to 10 with a step size of 0.5	7
Dropout	The rate of dropout (51)	0, 0.05 or 0.1	1e-1
Random forest			
Number of trees	Number of decision trees in the random forest	From 100 to 600 with a step size of 100	600
Criterion	Criteria for measuring the quality of each candidate split	Gini impurity or entropy	Gini impurity
Max features	Number of features by tree	Base 2 logarithm or square root of the total number of features	square root

Hyperparameter name	Description	Possible values	Selected value
Max depth	Depth of the trees	From 10 to 50 with a step size of 10	50
min samples split	Minimum number of samples in a tree's node to be considered as a potential split	Powers of 10 from 0.001 to 0.1	1e-3
Bootstrap	Whether to use training set sampling	True or False	True
Logistic regression			
C	Strength of the L2 regularization	From 1e-4 to 1e4 with a step size of 100	1e-2

1

DECLARATIONS

Ethics approval and consent to participate

Ethical approval was granted and consent to participate was reviewed by the French national commission governing the application of data privacy laws (number DR_2020-196).

Due to the retrospective nature of the study and the large number of cases analysed, the informed consent was waived for patients.

Consent for publication

Not applicable.

Availability of data and materials

The data underlying this article cannot be shared publicly due to privacy policies. Anonymised participant data extracted from the nationwide hospital data warehouse are available from the ATIH Institutional Data Access Platform for researchers who meet the legal and ethical criteria set by the French national commission governing the application of data privacy laws. To obtain this dataset for an international researcher, send an email to demande_base@atih.sante.fr.

Competing interests

All the authors do not have any conflict of interest.

Funding

This work is partially supported by the European Research Ambition Pack 2018 grant, distributed by the French Auvergne Rhône-Alpes region.

Authors' contributions

Stéphanie Polazzi did the data preparation, Vincent Lequertier did the data analysis, wrote the article and made the figures. Antoine Duclos, Julien Fondrevelle, Tao Wang, Stéphanie Polazzi and Vincent Augusto gave research guidance and contributed to the article content.

Acknowledgements

Richard L. Schmeidler participated in the proofreading of this paper.

REFERENCES

1. McRae S, Brunner JO. Assessing the impact of uncertainty and the level of aggregation in case mix planning. Omega [Internet]. 2020 Dec;97:102086. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0305048318307059

2. Schmidt R, Geisler S, Spreckelsen C. Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources. BMC Med Inform Decis Mak. 2013 Dec;13(1):3.

3. Praestgaard A, Saybolt MD, Durstenfeld MS, Kimmel SE, Kimmel SE. Physician predictions of length of stay of patients admitted with heart failure. Journal of Hospital Medicine. 2016 Sep 1;11(9).

4. Esteva A, Kuprel B, Novoa R, Ko J, Swetter S, Blau H, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature [Internet]. 2017 Feb 2; Available from: https://pubmed.ncbi.nlm.nih.gov/28117445/

5. Lakhani P, Sundaram B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology [Internet]. 2017;284(2):574–82. Available from: https://doi.org/10.1148/radiol.2017162326

6. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: A methodical review. Journal of the American Medical Informatics Association. 2020 Mar 1;27(3):457–70.

7. Zhang J, Kowsari K, Harrison JH, Lobo JM, Barnes LE. Patient2Vec: A personalized interpretable deep representation of the longitudinal electronic health record. IEEE Access. 2018;6:65333–46.

8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv:170603762 [cs] [Internet]. 2017 Jun 12; Available from: https://arxiv.org/abs/1706.03762

9. Schafer H, Friedrich CM. UMLS mapping and word embeddings for ICD code assignment using the MIMIC-III intensive care database. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). Berlin, Germany: IEEE; 2019. p. 6089–92.

10. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association. 2017 Mar 1;24(2):361–70.

11. Lequertier V, Wang T, Fondrevelle J, Augusto V, Duclos A. Hospital length of stay prediction methods: A systematic review. Medical Care. 2021 Oct;59(10):929–38.

12. Xu Y, Biswal S, Deshpande SR, Maher KO, Sun J. RAIM: Recurrent attentive and intensive model of multimodal patient monitoring data. arXiv:180708820 [cs, stat] [Internet]. 2018 Jul 23; Available from: http://arxiv.org/abs/1807.08820

13. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Liu PJ, et al. Scalable and accurate deep learning for electronic health records. npj Digital Med [Internet]. 2018 Dec;1(1):18. Available from: https://arxiv.org/abs/1801.07860

14. Moran JL, Solomon PJ, the ANZICS Centre for Outcome and Resource Evaluation (CORE) of the Australian and New Zealand Intensive Care Society (ANZICS). A review of statistical estimators for risk-adjusted length of stay: Analysis of the australian and new zealand intensive care adult patient data-base, 2008–2009. BMC Medical Research Methodology. 2012 May 16;12(1):68.

15. Hachesu PR, Ahmadi M, Alizadeh S, Sadoughi F. Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthcare Informatics Research. 2013;19(2):121.

16. Turgeman L, May JH, Sciulli R. Insights from a machine learning model for predicting the hospital length of stay (LOS) at the time of admission. Expert Systems with Applications. 2017;78:376–85.

17. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Medical Care. 1998 Jan;36(1):8–27.

18. Haviari S, Chollet F, Polazzi S, Payet C, Beauveil A, Colin C, et al. Effect of data validation audit on hospital mortality ranking and pay for performance. BMJ Qual Saf. 2019 Jun;28(6):459–67.

19. Shannon CE. The mathematical theory of communication. Urbana: University of Illinois Press; 1949. v (i.e. vii), 117.

20. Zhao Y, Wong ZSY, Tsui KL. A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection. J Healthc Eng. 2018 May 22;2018.

21. Navarro G. A guided tour to approximate string matching. ACM Comput Surv. 2001 Mar;33(1):31–88.

22. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. arXiv:170602515 [cs, stat] [Internet]. 2017 Jun 8; Available from: https://arxiv.org/abs/1706.02515

23. Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics. 1952 Sep;23(3):462–6.

24. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv:171105101 [cs, math] [Internet]. 2019 Jan 4; Available from: http://arxiv.org/abs/1711.05101

25. Breiman L. Random forests. Machine Learning [Internet]. 2001 Oct 1;45(1):5–32. Available from: https://doi.org/10.1023/A:1010933404324

26. Engel J. Polytomous logistic regression. Statistica Neerland. 1988 Dec;42(4):233–52.

27. Li L, Jamieson K, Rostamizadeh A, Gonina E, Hardt M, Recht B, et al. A system for massively parallel hyperparameter tuning. arXiv:181005934 [cs, stat]. 2018 Oct;

28. Rapin J, Teytaud O. Nevergrad - a gradient-free optimization platform [Internet]. Facebook; 2018. Available from: https://GitHub.com/FacebookResearch/Nevergrad

29. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960 Apr;20(1):37–46.

30. Martin Bland J, Altman DouglasG. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. The Lancet. 1986 Feb;327(8476):307–10.

31. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc Fd{\textbackslash}textbackslashtextbackslashtextquotesingle, Fox E, Garnett R, editors. Advances in neural information processing systems 32. Curran Associates, Inc.; 2019. p. 8026–37.

32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. The Journal of Machine Learning Research. 2011 Nov 1;

33. Lequertier V, Wang T, Fondrevelle J, Augusto V, Polazzi S, Duclos A. Predicting length of stay with administrative data from acute and emergency care: An embedding approach. In: 2021 IEEE 17th international conference on automation science and engineering (CASE). 2021. p. 1395–400.

34. Yang CS, Wei CP, Yuan CC, Schoung JY. Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages. Decision Support Systems [Internet]. 2010 Dec;50(1):325–35. Available from: https://doi.org/10.1016/j.dss.2010.09.001

35. Toumpoulis IK, Anagnostopoulos CE, DeRose JJ, Swistel DG. Does EuroSCORE predict length of stay and specific postoperative complications after coronary artery bypass grafting? International Journal of Cardiology [Internet]. 2005 Oct;105(1):19–25. Available from: https://doi.org/10.1016/j.ijcard.2004.10.067

36. Barnes S, Hamrock E, Toerper M, Siddiqui S, Levin S. Real-time prediction of inpatient length of stay for discharge prioritization. Journal of the American Medical Informatics Association. 2015 Aug;23:e2–10.

37. Health at a glance 2021 : OECD indicators. OECD publishing, Paris; 2021.

38. Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. arXiv:190800690 [cs, stat]. 2019 Aug;

39. New PW, Andrianopoulos N, Cameron PA, Olver JH, Stoelwinder JU. Reducing the length of stay for acute hospital patients needing admission into inpatient rehabilitation: A multicentre study of process barriers: Rehabilitation admission barriers. Intern Med J. 2013 Sep;43(9):1005–11.

40. Rajpal S, Shah M, Vivek N, Burneikiene S. Analyzing the correlation between surgeon experience and patient length of hospital stay. Cureus. 2020 Aug 28;

41. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Int Res. 2002 Jun;16(1):321–57.

42. Gosain A, Sardana S. Handling class imbalance problem using oversampling techniques: A review. In: 2017 international conference on advances in computing, communications and informatics (ICACCI). 2017. p. 79–85.

43. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. arXiv:170301365 [cs] [Internet]. 2017 Mar 3; Available from: http://arxiv.org/abs/1703.01365

44. Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S. Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. Abe T, editor. PLoS ONE. 2018 Apr 13;13(4):e0195901.

45. Stenwig E, Salvi G, Rossi PS, Skjærvold NK. Comparative analysis of explainable machine learning prediction models for hospital mortality. BMC Medical Research Methodology. 2022 Feb 27;22(1):53.

46. Bahadori M, Teymourzadeh E, Ravangard R, Raadabadi M. Factors affecting the overcrowding in outpatient healthcare. J Educ Health Promot. 2017;6:21.

47. Virtanen M, Pentti J, Vahtera J, Ferrie JE, Stansfeld SA, Helenius H, et al. Overcrowding in hospital wards as a predictor of antidepressant treatment among hospital staff. Am J Psychiatry. 2008 Nov;165(11):1482–6.

48. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:150203167 [cs]. 2015 Mar;

49. Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. In: Dasgupta S, McAllester D, editors. Proceedings of the 30th international conference on machine learning [Internet]. PMLR; 2013. p. 1139–47. (Proceedings of machine learning research; vol. 28). Available from: http://proceedings.mlr.press/v28/sutskever13.html

50. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y, editors. 3rd international conference on learning representations, ICLR 2015, san diego, CA, USA, may 7-9, 2015, conference track proceedings [Internet]. 2015. Available from: http://arxiv.org/abs/1412.6980

51. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research [Internet]. 2014;15(56):1929–58. Available from: http://jmlr.org/papers/v15/srivastava14a.html

Ce travail a montré qu'il est possible de prédire effectivement la durée de séjour avec un réseau de neurones avec des données médico-administratives standardisées et donc disponibles dans un grand nombre d'hôpitaux, et que cette approche fonctionne mieux que d'autres méthodes populaires comme la forêt aléatoire ou la régression logistique. Ces meilleurs résultats peuvent être interprétés comme venant de l'utilisation des *embeddings*. Leur utilisation permet une augmentation importante des performances, peut-être dû à la représentation riche du jeu de données qu'ils confèrent (cf. chapitre 4). L'approche des *embeddings* peut difficilement être utilisée par des forêts aléatoires et arbres de décisions, qui nécessitent l'utilisation de valeurs catégorielles. Il en est de même pour l'utilisation des *embeddings* pour une régression logistique, qui nécessiterait de modifier la méthode pour la rendre compatible. L'utilisation de plusieurs mesures de performances a permis de s'assurer des bonnes performances du réseau de neurones de manière détaillée et avec plusieurs perspectives.

Le modèle de réseau de neurones peut être entrainé régulièrement avec de nouvelles données, afin de pouvoir prendre en compte les changements en matière de techniques de soin et de politique hospitalière. Pour cela, une fois le modèle utilisé en routine comme partie intégrante du système d'information, son exactitude pourra être suivie à intervalles réguliers. Une baisse pourra déclencher l'intégration de nouvelles données dans celles utilisées pour l'entrainement du modèle de prédiction.

Plusieurs améliorations techniques sont d'ores et déjà identifiées en matières des données utilisées, du prétraitement qui en est fait, de la méthode de prédiction ou de l'évaluation des performances. Elles permettraient probablement de rendre le modèle plus performant et plus utile pour les professionnels de santé. Ces améliorations sont détaillées dans les perspectives scientifiques de la thèse (chapitre 6.2).

Chapitre 6

Discussion

6.1 Synthèse

L'objectif de cette thèse était de produire une méthode globale de prédiction des durées de séjours hospitalières avec des données incrémentales et évolutives. Le travail pour atteindre cet objectif a été organisé en commençant par une revue de la littérature (cf. chapitre 3) afin de mieux comprendre l'objectif de la thèse et faire un état de l'art. Cela nous a permis de réaliser l'importance de fournir des informations exhaustives sur la conduite d'une étude sur les prédictions de durées de séjours lors de sa restitution afin d'être le plus transparent possible, de montrer que les prédictions de durées de séjours sont de plus en plus effectuées au moyen de méthodes d'apprentissage automatique (*machine learning*) ou d'apprentissage profond (*deep learning*) et que les schémas d'études sont de plus en plus rigoureux, notamment en utilisant la validation croisée. Les deuxième et troisième parties de la thèse se sont focalisées sur la conception d'une méthode de prédiction et son évaluation (cf. chapitres 4 et 5). Plus spécifiquement, la méthode de prédiction a été basée sur un réseau de neurones vers l'avant utilisant des *embeddings* pour obtenir une représentation riche des données médicales hospitalières. Dans cette méthode, les données catégorielles telles que les diagnostics et actes médicaux sont converties en vecteurs numériques d'*embeddings*. Ces vecteurs sont ensuite fusionnés et utilisés par un réseau de neurones vers l'avant responsable de la classification de la durée de séjour en 15 classes, de 0

jour à 14 jours ou plus. La valeur des vecteurs ainsi que les poids du réseau de neurones sont ajustés par le biais d'une descente de gradient pendant la phase d'apprentissage du modèle sur des données d'entraînement. Les données ont été séparées en jeux d'entrainement et des jeux de test selon une validation croisée. Les performances ont été mesurées avec l'exactitude, le kappa de Cohen, une matrice de confusion et un graphique de Bland-Altman, permettant ainsi de mesurer la performance dans le détail et selon plusieurs perspectives. Ces mesures ont suggéré des performances satisfaisantes. Des limitations du projet, comme le manque de diversité géographique de la cohorte, concentrée sur la ville de Lyon, ou le fait que le diagnostic principal ne soit codé qu'à la sortie du patient, sont cependant à prendre en compte pour correctement interpréter ces résultats. Plusieurs pistes sont envisagées pour améliorer ces résultats et faciliter l'intégration du modèle de prédiction dans le système d'information de l'hôpital en vue d'une utilisation en routine par les professionnels de santé. Pour cela, le modèle pourrait servir d'aide à la décision pour les professionnels chargés d'organiser la gestion des ressources directement dans les services ou au sein d'une unité dédiée. La gestion des lits et l'organisation des services de santé seraient ainsi potentiellement améliorées. Également, une intégration du système de prédiction peut permettre de mieux anticiper la sortie des patients, rendant leurs sorties ou transferts entre unités médicales plus fluides. La prédiction des durées de séjours pourrait ainsi rendre meilleurs l'accès, l'efficience et la qualité des soins des patients.

6.2 Perspectives

Des améliorations pouvant potentiellement améliorer la qualité du jeu de données et du prétraitement qui en est fait, rendre la méthode plus efficace, plus transparente ou plus simple à intégrer au sein d'un système d'information sont d'ores et déjà identifiées et forment autant d'axes possibles pour la poursuite des travaux effectués pendant cette thèse. Ces axes sont présentés ici ainsi que des exemples concrets d'implémentation.

6.2.1 Améliorer la qualité du jeu de données et du prétraitement

Augmenter la taille du jeu de données ou la richesse des informations qu'il contient peut améliorer les résultats des prédictions. De plus, même l'algorithme le plus perfectionné peut aboutir à des résultats non satisfaisants quand le jeu de données utilisé présente des défauts en termes de taille ou de richesse [75]. Ajouter des données venant d'autres sources pourraient apporter des informations supplémentaires sur le contexte autour de l'hospitalisation. Également, améliorer le prétraitement des données pourrait améliorer les performances prédictives, sans pour autant nécessiter l'emploi de nouvelles données et les complications logistiques et techniques que cela implique.

Données détaillées sur les hospitalisations

Même si les données utilisées dans le cadre de cette thèse peuvent être perçues comme riches dans leurs structures, car les séjours sont constitués de séquences de nombres et de tailles variables, et dans leurs diversités, car la cohorte considérée est composée de séjours et de patients aux caractéristiques hétérogènes, ces données peuvent être vues comme simples par comparaison avec d'autres données en santé, comme les données biologiques, physiologiques ou textuelles qui ont déjà été utilisées pour prédire la durée de séjour [76-78], mais qui sont absentes du jeu de données utilisé pour cette thèse. Pour des diagnostics fréquents pour lesquels beaucoup de données sont disponibles, intégrer des critères spécifiques à ces diagnostics pourraient améliorer l'exactitude des

prédictions en augmentant le niveau de détails disponibles pour faire des prédictions. Investiguer dans des standards de représentation de la donnée allant plus loin que les données administratives tels que le *Fast Healthcare Interoperability Resources* (FHIR) [79] ou le *Observational Medical Outcomes Partnership* (OMOP) [80, 81] qui peuvent contenir des analyses biologiques, les prises de médicaments ou des notes cliniques, constituerait une première étape dans l'adaptation de la méthode de prédiction développée dans le cadre de cette thèse à des standards susceptibles de contenir plus de détails sur les hospitalisations, tout en conservant l'avantage d'utiliser un standard de données « universel ». Plus globalement, une standardisation de la structure des données médicales biologiques, physiologique ou textuelles faciliterait et donc encouragerait le développement d'outils prédictifs dans le domaine de la santé.

L'utilisation de données supplémentaires par le modèle de prédiction peut également s'inscrire dans la démarche du Dossier Patient Informatisé (DPI), un dossier médial permettant de stocker des informations concernant un patient. Le DPI peut être partagé entre les différents acteurs du système de santé français, et peut donc contenir des informations hétérogènes issues des hospitalisations, mais aussi de la médecine générale ou des laboratoires d'analyse et d'imagerie. Un équivalent du DPI a été développé à partir de 2004 [82] avec le Dossier Médical Partagé (DMP), qui a connu plusieurs versions utilisées en médecine de ville. Depuis le début d'année 2022, le DMP est remplacé par l'Espace Numérique de Santé qui est une version plus sophistiquée du DMP [83]. Dans ce cadre, la prédiction des durées de séjours pourraient utiliser des données détaillées sur les antécédents médicaux des patients. Cependant, comme l'utilisation du DPI ou du DMP à des fins épidémiologiques est facultative et demande le consentement des patients, les données qu'ils contiennent peuvent être manquantes ou incomplètes. De plus, les données étant potentiellement non standardisées (e.g. champs de textes libres), leur utilisation par un algorithme d'apprentissage automatique ou d'apprentissage profond peut être difficile. L'usage de l'intégralité du DPI pour prédire les durées de séjours hospitaliers augmenterait donc la complexité du système de prédiction, ce qui peut être problématique pour son déploiement en routine à l'hôpital, et ses apports restent incertains. Dans ce contexte, il peut donc être judicieux de trouver un compromis entre exactitude

et facilité d'utilisation de l'outil de prédiction.

Données de soins de suite et de réadaptation

Certains séjours hospitaliers nécessitent de la réadaptation effectuée lors des Soins de Suite et Réadaptation (SSR). Cependant, si un lit en soin SSR n'est pas disponible, cela peut prolonger le séjour hospitalier du patient, car celui-ci doit attendre la libération d'une place en SSR. La disponibilité et l'occupation des lits SSR a donc un impact sur la durée de séjour [84]. Ajouter des informations relatives à la demande de soins en SSR faits au début de l'hospitalisation, ainsi que des données contextuelles telles que le taux d'occupation des centres SSR ou le temps d'attente moyen avant le transfert en SSR selon les caractéristiques du patient pourrait permettre d'améliorer les performances des modèles de prédiction des durées de séjours. Cependant, le bénéfice serait seulement applicable aux séjours nécessitant un séjour en SSR.

Données sur les professionnels de santé

Les intelligences artificielles ne contrôlant pas (encore) totalement la gestion des soins, la dimension humaine reste présente lors des prises en charge hospitalières et a donc une influence sur la durée de séjour. Cette dimension concerne à la fois le plan individuel et collectif. Sur le plan individuel, les médecins peuvent influencer positivement ou négativement la durée de séjour. Par exemple, le chirurgien faisant un remplacement du genou est associé à une variation de la durée de séjour [85]. Plus généralement, les professionnels de santé adaptent la durée de séjour à leurs entourage et à leur contexte [7, 86], et le ratio entre le nombre de patients et de professionnels de santé a une influence sur d'autres indicateurs, comme la mortalité [87]. Avoir des informations sur les professionnels de santé, brutes ou agrégées, permettrait de prendre en compte l'impact humain des prises en charge, ce qui peut augmenter la précision des modèles de prédiction, et permettrait potentiellement de déterminer les facteurs organisationnels qui influencent la durée de séjour.

La dimension collective des activités des professionnels de santé est également à prendre en compte. La disponibilité du personnel varie en fonction du

temps et des politiques de gestion hospitalières. Comme le manque de personnel (sous effectif) peux induire des fermetures de lits, cela peut réduire les capacités d'accueil et de prises en charges des hôpitaux, ainsi que délayer la sortie des patients, ce qui peut augmenter la durée de séjour [88-90]. En France, entre 1990 et 2019, le nombre de médecins pour 10 000 habitants est passé de 31,9 à 24,8, soit une baisse de presque 1 % par an [91]. Intégrer des variables informant sur le nombre de professionnels de santé disponibles (regroupés en catégories) dans l'unité médicale, ainsi qu'une projection à court terme permettrait de prendre en compte le manque potentiel de personnel dans la prédiction des durées de séjours.

Mesurer la façon dont le prétraitement interagit avec l'exactitude des prédictions

Une fois les données obtenues, il convient de les modifier de manière à en faciliter la représentation par un modèle. Ce prétraitement a été détaillé dans le chapitre 4. Il peut être amélioré en (i) obtenant une sélection des diagnostics et procédures médicales qui maximise mieux les possibilités d'obtenir des prédictions précises et (ii) en rendant la représentation des données autour du séjour d'un patient dans une unité médicale dynamique en fonction des caractéristiques du séjour. En particulier, le prétraitement sélectionnant les procédures médicales dépend de l'importance relative des fréquences des procédures et du gain d'information par rapport à la durée de séjour, ainsi que du seuil au-delà duquel les procédures sont incluses dans le jeu de données. Cela peut être optimisé de manière à diminuer la valeur d'une fonction de perte ou de manière à maximiser la précision du modèle de prédiction. Également, le contexte autour du séjour peut être représenté de manière dynamique. Pour cela, le nombre de jours (i.e. 30) utilisé pour construire des agrégats statistiques autour du séjour pourrait changer en fonction du diagnostic principal du patient, ou de l'unité médicale dans laquelle il est admis. Même cela augmenterait la complexité du prétraitement, cette valeur pourrait être optimisée pendant l'apprentissage du réseau de neurones en utilisant la descente de gradient. Les améliorations en termes de prétraitement doivent se faire collaboration étroite avec les spécialistes en

données de santé, qui pourront identifier les changements les plus prometteurs et ceux qui pourraient avoir un impact néfaste.

6.2.2 Améliorer la méthode de prédiction et l'exploiter dans d'autres contextes

Déterminer l'utilité des *embeddings* pour d'autres tâches

En traitement automatique du langage naturel, il a été déterminé que des *embeddings* représentant des mots et ajustés dans un objectif précis, par exemple dans une optique de prédiction de la suite d'une phrase (entrainement non supervisé), peuvent être utiles à d'autres fins, comme déterminer la similarité entre deux phrases ou la reconnaissance d'entités au sein d'une phrase [92]. Dès lors, on peut faire l'hypothèse que cette capacité des *embeddings* à être utilisés pour plusieurs objectifs peut se retrouver pour les *embeddings* utilisés pour prédire la durée de séjour. Un moyen de vérifier cette hypothèse serait de déterminer si un parallèle peut être fait entre les *embeddings* représentant des mots et ceux représentant des concepts médicaux en mesurant la corrélation entre la similarité des *embeddings* et celle des concepts médicaux associés. Il a en effet été déterminé que les vecteurs d'*embeddings* de concepts similaires sont proches [47], de même pour les mesures de distance des concepts médicaux [93]. Intégrer ces *embeddings* pour des prédictions de durées de séjours dans d'autres contextes, plus précis et spécifiques que celui de cette thèse, ou pour servir d'autres objectifs épidémiologiques (e.g. prédire les complications chirurgicales ou les réadmissions) permettrait de tester à quel point ils sont transverses ou spécifiques aux prédictions de durées de séjours. Cela engendrerait également un gain de temps et une économie d'énergie et ressources humaines, car réutiliser les *embeddings* évite de devoir les reconstituer plusieurs fois.

Utiliser une combinaison de modèles

La méthode de prédiction des durées de séjours proposée dans le cadre de cette thèse n'utilise qu'un seul modèle, le réseau de neurones vers l'avant, quel que soit le séjour du patient. Même si cela a l'avantage pratique de faciliter l'éva-

luation de ses performances et son déploiement hypothétique dans les hôpitaux, cela oblige le modèle de prédiction à fonctionner pour des types de prises en charge et des contextes totalement différents. Par exemple, les séjours courts et les séjours longs présentent des différences importantes dans la structure des données et dans les caractéristiques médicales. Avoir des modèles spécialisés pourrait augmenter les performances du système dans sa globalité. Les méthodes ensemblistes [94, 95] semblent donc judicieuses et dès lors, plusieurs stratégies sont possibles.

Pour les séjours de longues durées, conserver un état rendant compte des précédentes étapes de l'hospitalisation permettrait de fournir des informations utiles qui affecteraient la prise de décision. C'est pourquoi utiliser un réseau de neurones récurrent pourrait être pertinent ici. Le RNN pourrait être utilisé conjointement avec le réseau de neurone vers l'avant pour les séjours de longues durées. L'interaction précise entre le réseau de neurones vers l'avant et le RNN resterait à déterminer, mais plusieurs possibilités peuvent être envisagées. Par exemple, le RNN pourrait représenter la dimension temporelle de l'hospitalisation, et cette représentation (i.e le dernier état caché du RNN) pourrait être ajoutée comme variable d'entrée au réseau de neurones vers l'avant. Également, uniquement les *embeddings* du réseau de neurones vers l'avant pourraient être transmis au RNN. Un autre réseau de neurones pourrait également être entrainé pour choisir lequel du RNN ou du réseau de neurones vers l'avant serait le plus approprié en fonction des caractéristiques du séjour.

La durée de séjour étant une donnée ordinale, elle pourrait être représentée sous forme hiérarchique. Les systèmes hiérarchiques de prédiction ont donné des résultats encourageants dans d'autres domaines [96, 97]. Un système hiérarchique de prédiction binaire pourrait donc également être envisagé. La figure 6.1 illustre son principe de fonctionnement. Des modèles de classification binaire organisés sous forme d'arbre prédisent si la durée de séjour sera plus ou moins une valeur donnée, en commençant par la médiane (e.g. 2 jours). Les prédictions successives de la durée de séjour guident jusqu'à une des feuilles de l'arbre qui donne la prédiction finale. Comme chaque modèle fait une classification binaire, la complexité locale est réduite par rapport à une prédiction multi-classe avec 15 classes allant de 0 jour à 14 jours ou plus, et chaque modèle peut être optimisé

pour traiter une prédiction binaire spécifique indépendante des autres. Les modèles donnant les meilleures performances à chaque prédiction binaire pourront donc être utilisés. Par exemple, des réseaux de neurones récurrents pourraient faire des prédictions pour les séjours longs, tandis que d'autres types de modèles comme les forêts aléatoires, les machines à vecteurs de support (*Support-Vector Machine*) ou les régressions, pourraient faire les prédictions pour les nœuds de l'arbre associés à des durées de séjours plus courtes.

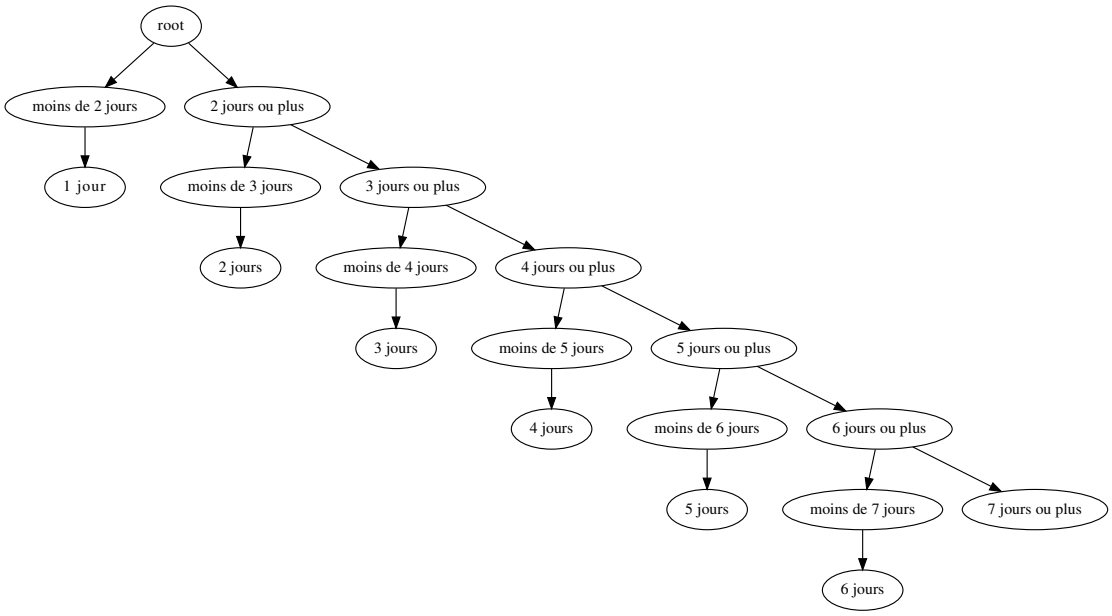


Figure 6.1 – Arbre de décision binaire faisant des prédictions jusqu'à atteindre une feuille donnant la durée de séjour

La topologie de l'arbre dépend du jeu de données, et il conviendra de la déterminer de manière rationnelle ou empirique. L'arbre de décision pourrait être strictement binaire et donc parfaitement symétrique, ou asymétrique. Prédire la durée de séjour avec un arbre comme celui présenté dans la figure 6.1 se ferait au moyen de l'algorithme 2, qui prédit récursivement la durée de séjour jusqu'à arriver à une feuille de l'arbre (i.e. un nœud sans successeurs). Un avantage de cette méthode est qu'il est possible d'ajuster dynamiquement la granularité de la prédiction en fonction de l'incertitude. Comme la prédiction est itérative, celle-ci peut-être stoppée si une prédiction est trop incertaine, avec un seuil prédéterminé. Cela permettrait de donner un intervalle de durée de séjour prédit

avec certitude. Des modèles probabilistes seraient utilisés dans ce cas.

Cette façon de prédire la durée de séjour fait l'hypothèse que la somme des erreurs des modèles composant l'arbre est plus petite que l'erreur globale d'un modèle unique donnant une probabilité pour chaque classe de durée de séjours.

Algorithme 2 : prédire-durée-séjour

Données : Un graphe G, des données D, un nœud n
Résultat : Une estimation de la durée de séjour

```
début
    si Taille(Successeurs(n)) = 0 alors
        retourner n
    model ← G.ObtenirModel(n)
    prediction ← model.prédire(D)
    si prediction = 0 alors
        retourner prédire-durée-séjour(Successeurs(n)[0])
    sinon
        retourner prédire-durée-séjour(Successeurs(n)[1])
```

6.2.3 Faciliter l'utilisation du réseau de neurones en routine dans les unités médicales

Certaines modifications techniques peuvent faciliter l'utilisation du système de prédiction au sein de l'infrastructure technique des hôpitaux. Par exemple, l'utilisation de standards pour échanger des informations entre le système de prédiction et une application l'utilisant faciliterait le déploiement du système de prédiction en routine à l'hôpital. Plus spécifiquement, les échanges entre une application et le système de prédiction pourraient se faire *via* une *Application Programming Interface* (API), où l'application ferait des requêtes au système de prédiction selon un protocole prédéterminé dans lequel les requêtes contiendraient toutes les informations nécessaires pour faire une prédiction et la réponse contiendrait l'estimation de la durée de séjour.

Une architecture de la sorte permettrait de décorréler les prédictions des programmes les utilisant, permettant ainsi de dédupliquer les ressources, car

un seul serveur de prédiction serait utilisé, et de faciliter le déploiement d'améliorations potentielles, car il faudrait ne mettre à jour qu'un seul serveur, indépendant des logiciels applicatifs. Des logiciels utilisés par des publics différents (e.g. infirmières, patients) pourraient ainsi facilement utiliser le même système de prédiction des durées de séjours et adapter leurs interfaces aux utilisateurs.

Plusieurs applications pourront bénéficier d'un système de prédiction des durées de séjours. Par exemple, une application de gestion des calendriers pourrait automatiquement afficher l'estimation de la durée de séjour de tous les patients, permettant ainsi aux gestionnaires des lits de visualiser la disponibilité des ressources et de faciliter la gestion organisationnelle. Un outil de prédiction facilitera donc le travail des gestionnaires des lits et des infirmières coordinatrices en leur faisant gagner du temps par une estimation plus rapide et plus fiable des durées de séjours.

La figure 6.2 montre un logiciel de gestion prévisionnelle des lits. Les prédictions pourraient être accompagnées d'informations complémentaires permettant de mieux comprendre les données utilisées pour les prédictions et le contexte autour de l'hospitalisation. Ces informations pourraient être personnalisées en fonction du profil de l'utilisateur du logiciel ou choisies par l'utilisateur lui-même. La figure 6.3 montre l'utilisation des prédictions de durées de séjours dans un logiciel de gestion des lits hospitaliers des Hospices Civils de Lyon utilisé pour le suivi quotidien des patients au sein d'une unité médicale. La durée de séjour prévisionnelle est indiquée ainsi qu'une échelle probabiliste. La comparaison de la durée de séjour prévisionnelle avec celle d'autres patients présentant des caractéristiques similaires n'entre pas dans le cadre de cette thèse. De même, un logiciel prédisant la durée de séjour pourrait automatiquement notifier les professionnels de santé de la nécessité de préparer la sortie d'un patient en fonction des estimations, de manière à fluidifier le parcours de soin.

Un des défauts de l'apprentissage profond (*deep learning*) est qu'il est difficile d'interpréter les prédictions, de visualiser l'algorithme d'apprentissage et de comprendre les contributions des variables d'entrée ainsi que les relations entre elles [99]. Faire confiance aux résultats d'un modèle de prédiction peut permettre de faciliter son adoption [100, 101]. Il est donc important de donner des explications rattachées à chaque prédiction. Pour cela, et en exploitant la struc-

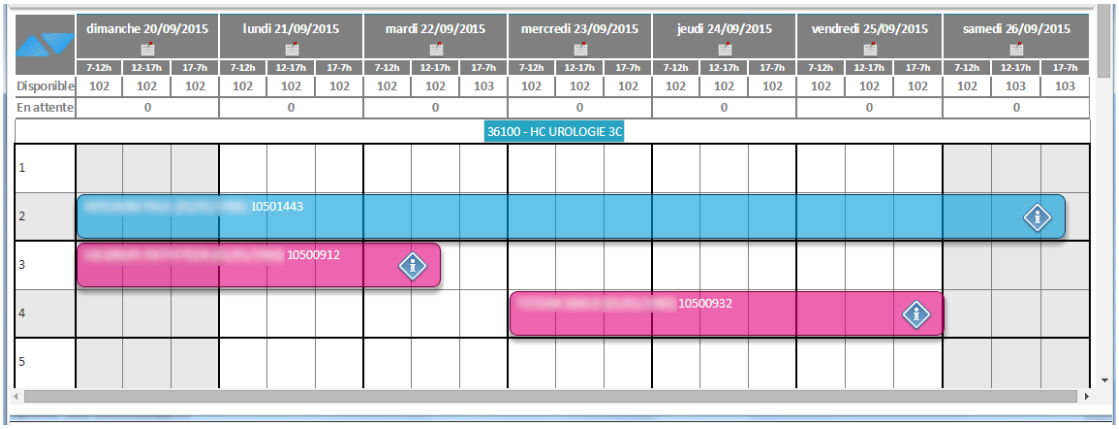


FIGURE 6.2 – Capture d'écran d'une intégration d'un système de prédiction des durées de séjours dans un logiciel de gestion des lits [98] utilisé par les gestionnaires de lits [1].

ture des réseaux de neurones, l'intégration de gradients peut être utilisée [102]. L'équation (6.1) illustre le principe de fonctionnement de l'intégration de gradients, où x est la donnée d'entrée, x' est une donnée pour laquelle la sortie du réseau de neurones est zéro, F est un réseau de neurones, i est une dimension de x et α est un coefficient d'interpolation entre x et x'. Dans l'intégration de gradients, la variation de gradients est mesurée entre une donnée pour laquelle la sortie du réseau de neurones est égale à 0, et une donnée d'entrée à partir de laquelle on veut faire des prédictions et les accompagner d'explications, ce qui permet de comprendre le rôle de chaque entrée sur la sortie. Intuitivement, la variation de la sortie du réseau de neurones selon les changements de l'entrée permet de comprendre l'impact de chaque valeur d'entrée sur la sortie, ce qui est important pour rendre les prédictions faciles à interpréter. Une implémentation de cette méthode avec PyTorch [103] est disponible dans l'annexe A.1. Utiliser cette méthode ou une méthode similaire pourrait faciliter l'adoption de l'outil de prédiction en vue d'une utilisation en routine.

$$IntegratedGrads_i(x) ::= \left(x_i - x_i'\right) \times \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha \times (x - x')\right)}{\partial x_i} \mathrm{d}\alpha \qquad (6.1)$$

Même utilisé en routine à l'hôpital, le modèle de prédiction peut tout de

1. Les auteurs de cette image en autorisent la reproduction.

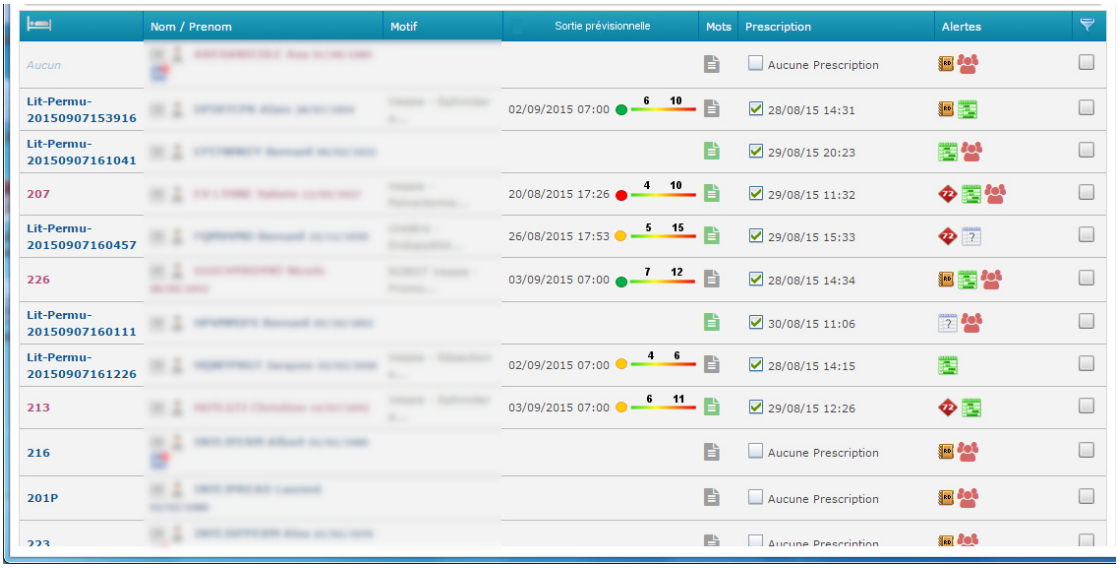


FIGURE 6.3 – Capture d'écran d'une intégration d'un système de prédiction des durées de séjours dans un logiciel de gestion des lits [98] dans la colonne « Sortie prévisionnelle » [2].

même subir des modifications. Le système de prédiction apprenant à modéliser les durées de séjours à partir des précédentes hospitalisations, il est capable à la fois de réapprendre entièrement à prédire les durées de séjours en s'appuyant sur une méthode plus perfectionnée, donnant de meilleurs résultats, ou de modifier ses prédictions existantes en s'appuyant sur des hospitalisations plus récentes. Pour modifier les prédictions effectuées et donc à nouveau entrainer le réseau de neurones, utiliser les poids actuels du réseau de neurones en début d'apprentissage, plutôt que les initialiser aléatoirement, permettra d'augmenter la vitesse de convergence. En cas de changements fondamentaux de la méthode, il est possible que réinitialiser aléatoirement les poids du réseau de neurones puisse conduire à de meilleurs résultats. Prendre en compte de nouvelles données venant des hospitalisations plus récentes nécessite de détecter le moment où il est nécessaire de le faire. Pour cela, l'exactitude des prédictions devra être mesurée régulièrement. Une baisse significative de l'exactitude pourrait déclencher automatiquement le réentrainement du modèle avec des données issues des dernières hospitalisations, sans intervention humaine. De plus, comme il

2. Les auteurs de cette image en autorisent la reproduction.

serait possible que l'exactitude diminue seulement pour certaines catégories de patients ou de prises en charge sans que cela n'entraine une baisse significative globale, il conviendra de surveiller l'exactitude de manière assez précise pour détecter des changements concernant des sous-ensembles du jeu de données, qu'il faudra définir avant la mise en place de l'outil au sein du système d'organisation des activités de soin de l'hôpital. Une approche plus simple consisterait à entrainer le modèle avec de nouvelles données à intervalles réguliers et fixes, de manière trimestrielle, mensuelle, hebdomadaire ou quotidienne, selon la rapidité de l'évolution des prises en charge et selon la sensibilité du modèle à celle-ci.

Intégrer l'outil de prédiction au sein du système d'information de l'hôpital en vue de son utilisation en routine par les professionnels de santé constitue un changement organisationnel majeur. Même si ce changement peut avoir des conséquences positives sur les activités de soins, il peut être néfaste si mal reçu par les professionnels de santé [104]. Par exemple, il peut être difficile d'avoir confiance dans un système de prédiction reposant sur des principes complexes et utilisant des données massives, qui pourra de surcroit faire des prédictions avec lesquelles les professionnels de santé pourraient ne pas d'accord a priori. Dès lors, cette transition vers un système de prédiction de séjour servant d'aide à la décision doit être considérée non pas comme un événement ponctuel, mais comme un processus surveillé qui s'inscrit dans la durée et dans lequel les professionnels de santé sont accompagnés et supportés. Répondre à leurs demandes et interrogations permettra non seulement d'améliorer l'outil de prédiction et son interface utilisateur, mais aussi de les rassurer et faciliter l'adhésion au changement. Comme l'incertitude peut induire du stress [105], mettre l'accent sur la planification et la projection à l'avenir que permet la prédiction des durées de séjours, et expliquer le fonctionnement de l'outil peut être une stratégie de communication convaincante auprès des professionnels de santé. De plus, mettre l'accent sur la dimension « humaine » des activités de soins tout en utilisant des systèmes prédictifs serait un objectif important dans le cadre de leur mise en place.

6.2.4 Évaluation pratique du système de prédiction

Même si une partie importante de cette thèse a été centrée sur l'évaluation des performances des prédictions des durées de séjours, cette évaluation se base sur des mesures numériques évaluant à quel point les prédictions sont proches de la réalité. Ces mesures sont accompagnées de visualisations. Cependant, cela peut ne pas suffire pour déterminer l'efficacité et l'utilité du modèle une fois utilisé en routine dans l'hôpital. Par exemple, sans simulation réaliste ou essais en situation réelle, il est difficile de connaître le taux d'erreur au-delà duquel le modèle perd de son utilité. C'est pourquoi, même si les résultats présentés dans le troisième article (cf. chapitre 5) peuvent être considérés comme satisfaisants, tester le modèle une fois déployé en routine en mesurant à la fois des éléments quantitatifs sur l'efficacité de l'hôpital et sur la qualité des soins et des éléments qualitatifs sur l'expérience des professionnels de santé et des patients permettrait de confirmer les apports énoncés en introduction (chapitre 1), ce qui pourrait constituer une contribution scientifique potentiellement intéressante, par ailleurs peu présente dans la littérature existante à la date de rédaction de cette thèse.

Plus précisément, une approche mixte pourra être envisagée, dans laquelle deux études sont menées (i) une approche quantitative permettra de mesurer l'impact de l'outil de prédiction sur des indicateurs cliniques par le biais d'analyses statistiques [106], et (ii) une approche qualitative au moyen d'entretiens individuels ou collectifs et des questionnaires permettra de comprendre l'expérience des professionnels de santé et éventuellement des patients sur l'utilisation de l'outil et les impacts positifs et négatifs sur leurs activités ou séjours [107]. Ces deux facettes de l'étude permettront d'avoir une perspective pratique concrète sur l'outil de prédiction, et donc de déterminer s'il est utile pour améliorer la qualité des activités de soin ou s'il doit être modifié. Plusieurs schémas d'étude sont possibles. Un schéma d'étude « avant / après » par grappes (*clusters*) randomisées pourra être utilisé, dans lequel les études quantitative et qualitative sont effectuées avant et après la mise en place de l'outil au sein de plusieurs hôpitaux. Les résultats des études quantitative et qualitative seraient ensuite comparées entre elles. Une approche non déterministe (i.e. à l'aveugle) de type série alternée

(« *on* / *off* ») pourrait également être envisagée, dans laquelle les résultats des prédictions sont aléatoirement substitués par des résultats auxquels sont ajoutés un bruit gaussien aléatoire ou des résultats venant du système actuel plus rudimentaire de prédiction des durées de séjours [98]. Cela permettrait de mesurer le seul mérite des prédictions et d'ignorer l'impact du changement lui-même. Ici, les résultats des prédictions authentiques et des prédictions factices seraient comparés. Comme l'influence de l'outil sur des mesures quantitatives liées à l'accès, l'efficience ou la qualité des soins pourra varier entre les différentes unités médicales de l'hôpital, et de même pour le ressenti des professionnels de santé et des patients, il conviendra de considérer un ensemble d'unités médicales suffisamment représentatif de l'ensemble de l'hôpital dans lequel l'outil sera graduellement mis en place.

L'ordre dans lequel les deux volets de l'étude mixte sont effectués sera également à déterminer. Comme il est possible que la partie qualitative de l'étude renseigne sur les informations susceptibles d'être le plus impactées par la mise en place du système de prédiction et donc ait une incidence sur la conduite de l'étude quantitative, l'étude qualitative pourra être menée avant l'étude quantitative. L'impact de la partie quantitative sur la partie qualitative peut aussi être considérée. En effet, il est possible que les résultats d'analyses statistiques semblent contre intuitifs ou soulèvent des interrogations. Dans ce cas, le volet qualitatif pourrait apporter des éclaircissements sur les résultats quantitatifs. Effectuer une approche par croisements, dans laquelle plusieurs phases quantitatives et qualitatives sont effectuées successivement, permettrait de prendre en compte l'apport de l'une sur l'autre et vice versa. Les parties qualitative et quantitative pourraient aussi être effectuées simultanément.

Également, comme les parties quantitative et qualitative de l'étude ne mesurent pas les mêmes éléments, leur convergence n'est pas garantie. Si les conclusions divergent (e.g. l'étude quantitative montre une augmentation significative de la qualité des soins, mais cette hausse n'est pas remarquée par les professionnels de santé), il sera difficile de conclure sur l'utilité du modèle de prédiction. Un choix arbitraire privilégiant la partie quantitative, ou l'expérience des professionnels de santé et des patients pourra être envisagé.

Une fois l'outil déployé en routine dans les unités médicales, ce processus

d'évaluation du système de prédiction des durées de séjours pourra être effectué à nouveau en cas de changements majeurs de la méthode de prédiction ou des données utilisées, afin de comprendre leurs influences et de valider leurs utilités concrètes.

Utiliser le modèle de prédiction des durées de séjours en routine à l'hôpital a donc des composantes techniques, scientifiques et organisationnels. Cette mise en place implique de nombreux éléments délicats demandant la coordination et coopération de spécialistes aux compétences diverses : (i) des designers pour créer l'interface utilisateur affichant les prédictions et leur utilisation dans les outils informatiques, (ii) des ingénieurs informatiques pour l'utilisation du modèle de prédiction au moyen de son API, (iii) des analystes de données pour intégrer d'autres données au modèle et le rendre plus transparent et (iv) des épidémiologiques et chercheurs en santé publique pour son évaluation réaliste et pour l'accompagnement au changement. Cela complique donc la mise en place du modèle de prédiction en routine à l'hôpital.

Chapitre 7

Conclusion

Les progrès techniques et scientifiques liés aux méthodes d'intelligence artificielle, l'augmentation des capacités de calcul et la constitution de jeux de données de taille importante, ont permis une utilisation massive des données et des algorithmes d'apprentissage automatique pour l'aide à la décision. Cela engendre un changement de paradigme important, dans lequel plus de confiance et d'importance sont accordées aux données. Les prises de décisions peuvent ainsi considérer l'ensemble des expériences similaires passées, ce qui permet d'éviter la subjectivité de l'expérience d'un individu et de prendre en compte davantage de paramètres, ce qui pourrait donc constituer une avancée significative. Ces avancées profitent au domaine de la santé de nombreuses manières.

Utiliser l'intelligence artificielle et les données hospitalières pour effectuer des prédictions sur la durée de séjour des patients est un objectif scientifique aux applications pratiques diverses, parmi lesquelles figurent l'amélioration de la gestion des ressources hospitalières et la planification des prises en charge médicales. Cette application pratique constitue l'objectif sous-jacent dans le cadre de cette thèse, qui permet potentiellement de rendre les soins plus accessibles, efficaces et efficients. Comme ce sous-objectif implique de faire une prédiction pour tous les patients de l'hôpital, des données médico-administratives devaient être utilisées pour prédire la durée de séjour à toutes les étapes des hospitalisations, car ce type de données est disponible pour tous les patients. Cette thèse a donc permis de tester l'hypothèse selon laquelle la durée de séjour hospitalière peut

être prédite à partir de données médico-administratives disponibles pour tous les patients et mises à jour à chaque étape de son séjour. Cette hypothèse a pu partiellement être vérifiée en validant les performances d'une méthode utilisant un réseau de neurones et un système de représentation basé sur les *embeddings*. Intégrer la méthode de prédiction dans le système d'information de l'hôpital consisterait une prochaine étape dans la validation de cette hypothèse. Plusieurs pistes permettraient potentiellement d'améliorer les performances du modèle de prédiction, et sont autant de poursuites possibles de ce travail de thèse.

Chapitre 8

Valorisation de la thèse

8.1 Publications scientifiques

— Vincent Lequertier et al. « Hospital Length of Stay Prediction Methods : A Systematic Review ». *In* : *Medical Care* 59.10 (oct. 2021), p. 929-938. ISSN : 0025-7079. DOI : 10.1097/MLR.0000000000001596

— Vincent Lequertier et al. « Predicting length of stay with administrative data from acute and emergency care : an embedding approach ». *In* : *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. Août 2021, p. 1395-1400. DOI : 10.1109/CASE49439.2021.9551429

— Vincent Lequertier et al. « Length of stay prediction with standardized hospital data from acute and emergency care using a deep neural network ». *In* : *(soumis pour publication)* (2022)

8.2 Présentations orales

— Vincent Lequertier. « Méthode globale de prédiction des durées de séjours avec intégration des données incrémentales et évolutives ». Réunion scientifique RESHAPE. Lyon, 9 oct. 2020

— Vincent Lequertier et al. « Predicting length of stay with administrative data from acute and emergency care : an embedding approach ». *In* : *2021*

IEEE 17th International Conference on Automation Science and Engineering (CASE). Août 2021, p. 1395-1400. DOI : 10.1109/CASE49439.2021.9551429

— Vincent Lequertier. « Prédiction des durées de séjours avec des données médico-administratives à l'aide d'un réseau de neurones. » Congrés ADELF EMOIS. Dijon, 31 mars 2022. URL : https://doi.org/10.1016/j.respe.2022.01.071

— Vincent Lequertier. « Length of stay prediction with standardized hospital data from acute and emergency care using a deep neural network ». Réunion scientifique RESHAPE. Lyon, 8 avr. 2022

— Vincent Lequertier et al. « Prédiction des durées de séjours avec des données hospitalières standardisées ». *In* : Congrès National de la Recherche des IUT 2022. Roanne, France, 10 juin 2022, p. 104-106. URL : https://web.archive.org/web/20220613134450/https://cnriut2022.sciencesconf.org/data/Recueil_des_Publications.pdf

8.3 Enseignements

— *L'intelligence Artificielle pour les entreprises (18 heures)*, 2021 - 2022
Institut des hautes études économiques et commerciales - Lyon
Cours pour les étudiants en 3ème année de la spécialisation « Digitalisation, Intelligence Artificielle et Big Data ». Présentation de l'intelligence artificielle centrée artificielle le *machine learning* et le *deep learning*. Présentation d'applications concrètes de l'IA, de l'éthique et projet de groupes.

— *Initiation à la recherche (10 heures)*, 2020 - 2021
Institut National des Sciences Appliquées - Lyon
Projet de groupe sur l'application des méthodologies de recherche pour les étudiants en 3ème année du cycle ingénieur Génie Industriel.

— *Initiation à la recherche documentaire (2 heures)*, 2020 - 2021
Institut National des Sciences Appliquées - Lyon
Projet de groupe sur la méthode pour faire une revue de la littérature pour les étudiants en 1ère année du cycle ingénieur Génie Industriel.

— *Probabilités et Statistiques (12 heures chaque année académique)*, 2019 - 2021
Institut National des Sciences Appliquées - Lyon
Travaux pratique de probabilités et statistiques pour les étudiants en 1ère année du cycle ingénieur Génie Industriel.

8.4 Valorisations diverses

— Représentant des doctorant·e·s du laboratoire RESHAPE à partir du 2 janvier 2022

— Collaboration avec une équipe australienne sur une revue systématique de la littérature et meta-analyse sur la prédiction des durées de séjours prolongées : Swapna Gokhale et al. « Hospital Length of Stay Prediction Tools for General Surgery Populations and Total Knee Arthroplasty Admissions : Systematic Review and Meta-Analysis ». *In* : *(soumis pour publication)* (2022)

Bibliographie

[1] R Schmidt, S Geisler et C Spreckelsen. « Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources ». *In* : *BMC medical informatics and decision making* (7 jan. 2013). doi : 10.1186/1472-6947-13-3. PMID : 23289448 (cf. p. 1).

[2] Joel S. Weissman et al. « Hospital workload and adverse events ». *In* : *Medical Care* 45.5 (mai 2007), p. 448-455. issn : 0025-7079. doi : 10.1097/01.mlr.0000257231.86368.09. PMID : 17446831 (cf. p. 1).

[3] Rosemary Hills et Sheila Kitchen. « Satisfaction with outpatient physiotherapy : a survey comparing the views of patients with acute and chronic musculoskeletal conditions ». *In* : *Physiotherapy Theory and Practice* 23.1 (fév. 2007), p. 21-36. issn : 0959-3985. doi : 10.1080/09593980601147876. PMID : 17454796 (cf. p. 1).

[4] Sebastian McRae et Jens O. Brunner. « Assessing the impact of uncertainty and the level of aggregation in case mix planning ». *In* : *Omega* 97 (2020), p. 102086. issn : 0305-0483. doi : 10.1016/j.omega.2019.07.002 (cf. p. 2).

[5] Matthew S. Durstenfeld et al. « Physician predictions of length of stay of patients admitted with heart failure ». *In* : *Journal of Hospital Medicine* 11.9 (sept. 2016), p. 642-645. issn : 15535592. doi : 10.1002/jhm.2605 (cf. p. 2).

[6] Antonio Paulo Nassar et Pedro Caruso. « ICU physicians are unable to accurately predict length of stay at admission : a prospective study ». *In* : *International Journal for Quality in Health Care : Journal of the International Society for Quality in Health Care* 28.1 (fév. 2016), p. 99-103. issn : 1464-3677. doi : 10.1093/intqhc/mzv112. PMID : 26668104 (cf. p. 2).

[7] Judith D. Jong et al. « Variation in Hospital Length of Stay : Do Physicians Adapt Their Length of Stay Decisions to What Is Usual in the Hospital Where They Work ? » *In* : *Health Services Research* 41.2 (avr. 2006), p. 374-

394. ISSN : 0017-9124, 1475-6773. DOI : 10.1111/j.1475-6773.2005.00486.x (cf. p. 2, 97).

[8] *MCO caractéristiques séjours/séances par région*. URL : https://www.scansante.fr/applications/caracteristiques-des-sejours-par-region (cf. p. 2).

[9] *Panorama de la Dress Santé : Les établissements de santé*. Direction de la Recherche, des Études, de l'Évaluation et des Statistiques, 2021, p. 27. URL : https://drees.solidarites-sante.gouv.fr/sites/default/files/2021-07/ES2021.pdf (cf. p. 3).

[10] Roberto IPPOLITI et al. « Neural networks and hospital length of stay : an application to support healthcare management with national benchmarks and thresholds ». *In* : *Cost Effectiveness and Resource Allocation* 19.1 (9 oct. 2021), p. 67. ISSN : 1478-7547. DOI : 10.1186/s12962-021-00322-3 (cf. p. 3).

[11] Lauren DOCTOROFF et Shoshana J. HERZIG. « Predicting Patients at Risk for Prolonged Hospital Stays ». *In* : *Medical care* 58.9 (sept. 2020), p. 778-784. ISSN : 0025-7079. DOI : 10.1097/MLR.0000000000001345. PMID : 32826743 (cf. p. 3).

[12] Mao-Te CHUANG, Ya-han HU et Chia-Lun LO. « Predicting the prolonged length of stay of general surgery patients : a supervised learning approach ». *In* : *International Transactions in Operational Research* 25.1 (2018), p. 75-90. ISSN : 1475-3995. DOI : 10.1111/itor.12298 (cf. p. 3).

[13] Sarah E SEATON et al. « What factors predict length of stay in a neonatal unit : a systematic review ». *In* : *BMJ Open* 6.10 (18 oct. 2016). ISSN : 2044-6055. DOI : 10.1136/bmjopen-2015-010466. PMID : 27797978 (cf. p. 3).

[14] T RICHARDS et al. « The independent patient factors that affect length of stay following hip fractures ». *In* : *Annals of The Royal College of Surgeons of England* 100.7 (sept. 2018), p. 556-562. ISSN : 0035-8843. DOI : 10.1308/rcsann.2018.0068. PMID : 29692191 (cf. p. 4).

[15] Tahani A. DAGHISTANI et al. « Predictors of in-hospital length of stay among cardiac patients : A machine learning approach ». *In* : *International Journal of Cardiology* 288 (2019), p. 140-147. ISSN : 1874-1754. DOI : 10.1016/j.ijcard.2019.01.046. PMID : 30685103 (cf. p. 4).

[16] Wen LIU et al. « Understanding variations and influencing factors on length of stay for T2DM patients based on a multilevel model ». *In* : *PLOS ONE* 16.3 (12 mars 2021). Sous la dir. de Gregor STIGLIC, e0248157. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0248157 (cf. p. 4).

[17] Dale M. Needham et al. « A province-wide study of the association between hospital resource allocation and length of stay ». *In* : *Health Services Management Research* 16.3 (août 2003), p. 155-166. ISSN : 0951-4848. DOI : 10.1258/095148403322167915. PMID : 12908990 (cf. p. 4).

[18] Pria M. D. Nippak et al. « Is there a relation between emergency department and inpatient lengths of stay ? » *In* : *Canadian Journal of Rural Medicine : The Official Journal of the Society of Rural Physicians of Canada = Journal Canadien De La Medecine Rurale : Le Journal Officiel De La Societe De Medecine Rurale Du Canada* 19.1 (2014), p. 12-20. ISSN : 1488-237X. PMID : 24398353 (cf. p. 5).

[19] *Les HCL en chiffres Médiathèque HCL*. 23 juill. 2021. URL : https://www.chu-lyon.fr/les-hcl-en-chiffres (cf. p. 6).

[20] World Health Organization. *Classification statistique internationale des maladies et des problèmes de santé connexes*. Organisation mondiale de la Santé, 2009. ISBN : 978-92-4-254766-5. URL : https://apps.who.int/iris/handle/10665/44082 (cf. p. 7).

[21] *Article L710*. 31 juill. 1991. URL : https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006694595/1993-01-29/ (cf. p. 7).

[22] *Financement de la sécurité sociale pour 2004*. 18 déc. 2003. URL : https://www.circulaires.gouv.fr/jorf/id/JORFTEXT000000249276/ (cf. p. 7).

[23] *CCAM en ligne*. 2022. URL : https://www.ameli.fr/accueil-de-la-ccam/index.php (cf. p. 10).

[24] Christopher D. Green. « Charles Babbage, the Analytical Engine, and the possibility of a 19th-century cognitive science. » *In* : *The transformation of psychology : Influences of 19th-century philosophy, technology, and natural science*. Sous la dir. de Christopher D. Green, Marlene Shore et Thomas Teo. Washington : American Psychological Association, 2001, p. 133-152. ISBN : 978-1-55798-776-1. DOI : 10.1037/10416-007 (cf. p. 14).

[25] Alan Mathison Turing. « On computable numbers, with an application to the Entscheidungsproblem ». *In* : *J. of Math* 58.345 (1936), p. 5 (cf. p. 14).

[26] Harry R. Lewis et Christos H. Papadimitriou. « Elements of the Theory of Computation ». *In* : *ACM SIGACT News* 29.3 (sept. 1998), p. 62-78. ISSN : 0163-5700. DOI : 10.1145/300307.1040360 (cf. p. 14).

[27] Alan Mathison Turing. « COMPUTING MACHINERY AND INTELLIGENCE ». *In* : *Mind* LIX.236 (1er oct. 1950), p. 433-460. ISSN : 0026-4423. DOI : 10.1093/mind/LIX.236.433 (cf. p. 15).

[28] John R. Searle. « Minds, brains, and programs ». *In* : *Behavioral and Brain Sciences* 3.3 (sept. 1980), p. 417-424. issn : 1469-1825, 0140-525X. doi : 10.1017/S0140525X00005756 (cf. p. 15).

[29] John McCarthy et al. « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 ». *In* : *AI Magazine* 27.4 (15 déc. 2006), p. 12-12. issn : 2371-9621. doi : 10.1609/aimag.v27i4.1904 (cf. p. 15).

[30] F. Rosenblatt. *The perceptron - A perceiving and recognizing automaton*. 85-460-1. Ithaca, New York : Cornell Aeronautical Laboratory, jan. 1957. url : https://blogs.umass.edu/brain-wars/files/2016/03/rosenblatt-1957.pdf (cf. p. 15).

[31] Marvin Minsky et Seymour A. Papert. *Perceptrons : an introduction to computational geometry*. 2. print. with corr. Cambridge/Mass. : The MIT Press, 1972. 258 p. isbn : 978-0-262-34393-0 (cf. p. 16).

[32] P. Werbos. « Beyond Regression : New Tools for Prediction and Analysis in the Behavior Science ». *In* : *Ph. D. dissertation, Harvard University* (1974). url : https://ci.nii.ac.jp/naid/10004070196/ (cf. p. 16).

[33] David E. Rumelhart, Geoffrey E. Hinton et Ronald J. Williams. « Learning representations by back-propagating errors ». *In* : *Nature* 323.6088 (oct. 1986), p. 533-536. issn : 1476-4687. doi : 10.1038/323533a0 (cf. p. 16).

[34] Feng-hsiung Hsu. *Behind deep blue : building the computer that defeated the world chess champion*. Princeton : Princeton University Press, 2002. 298 p. isbn : 978-0-691-09065-8 (cf. p. 16).

[35] Gordon E. Moore. « Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff. » *In* : *IEEE Solid-State Circuits Society Newsletter* 11.3 (sept. 2006), p. 33-35. issn : 1098-4232. doi : 10.1109/N-SSC.2006.4785860 (cf. p. 16).

[36] Xavier Glorot, Antoine Bordes et Yoshua Bengio. « Deep Sparse Rectifier Neural Networks ». *In* : *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop et Conference Proceedings, 14 juin 2011, p. 315-323. url : https://proceedings.mlr.press/v15/glorot11a.html (cf. p. 16).

[37] Xavier Glorot et Yoshua Bengio. « Understanding the difficulty of training deep feedforward neural networks ». *In* : *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Sous la dir. d'Yee Whye Teh et Mike Titterington. T. 9. Proceedings of Machine Learning Research. PMLR, 13 mai 2010, p. 249-256. url : https://proceedings.mlr.press/v9/glorot10a.html (cf. p. 16, 17).

[38] Yann LeCun et al. « Efficient BackProp ». *In* : *Neural Networks : Tricks of the Trade*. Sous la dir. de Genevieve B. Orr et Klaus-Robert Müller. Lecture Notes in Computer Science. Berlin, Heidelberg : Springer Berlin Heidelberg, 1998, p. 9-50. isbn : 978-3-540-49430-0. doi : 10.1007/3-540-49430-8_2 (cf. p. 16).

[39] Jia Deng et al. « ImageNet : A large-scale hierarchical image database ». *In* : *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). IEEE, juin 2009, p. 248-255. isbn : 978-1-4244-3992-8. doi : 10.1109/CVPR.2009.5206848 (cf. p. 17).

[40] Alex Krizhevsky, Ilya Sutskever et Geoffrey E. Hinton. « ImageNet classification with deep convolutional neural networks ». *In* : *Communications of the ACM* 60.6 (24 mai 2017), p. 84-90. issn : 0001-0782, 1557-7317. doi : 10.1145/3065386 (cf. p. 17).

[41] Kaiming He et al. « Deep Residual Learning for Image Recognition ». *In* : *arXiv:1512.03385 [cs]* (10 déc. 2015). doi : 10.1109/CVPR.2016.90. arXiv : 1512.03385 (cf. p. 17).

[42] Volodymyr Mnih et al. « Human-level control through deep reinforcement learning ». *In* : *Nature* 518.7540 (fév. 2015), p. 529-533. issn : 1476-4687. doi : 10.1038/nature14236 (cf. p. 18).

[43] David Silver et al. « Mastering the game of Go with deep neural networks and tree search ». *In* : *Nature* 529.7587 (28 jan. 2016), p. 484-489. issn : 0028-0836, 1476-4687. doi : 10.1038/nature16961 (cf. p. 18).

[44] David Silver et al. « Mastering the game of Go without human knowledge ». *In* : *Nature* 550.7676 (oct. 2017), p. 354-359. issn : 0028-0836, 1476-4687. doi : 10.1038/nature24270 (cf. p. 18).

[45] Julian Schrittwieser et al. « Mastering Atari, Go, chess and shogi by planning with a learned model ». *In* : *Nature* 588.7839 (déc. 2020), p. 604-609. issn : 1476-4687. doi : 10.1038/s41586-020-03051-4 (cf. p. 18).

[46] Ashish Vaswani et al. « Attention is All You Need ». *In* : *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Curran Associates Inc., 2017, p. 6000-6010. isbn : 978-1-5108-6096-4 (cf. p. 21).

[47] Tomas Mikolov et al. « Distributed representations of words and phrases and their compositionality ». *In* : *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Red Hook, NY, USA : Curran Associates Inc., 5 déc. 2013, p. 3111-3119 (cf. p. 24, 99).

[48] Kun-Hsing Yu, Andrew L. Beam et Isaac S. Kohane. « Artificial intelligence in healthcare ». *In* : *Nature Biomedical Engineering* 2.10 (oct. 2018), p. 719-731. issn : 2157-846X. doi : 10.1038/s41551-018-0305-z (cf. p. 25).

[49] Guoguang Rong et al. « Artificial Intelligence in Healthcare : Review and Prediction Case Studies ». *In* : *Engineering* 6.3 (1er mars 2020), p. 291-301. issn : 2095-8099. doi : 10.1016/j.eng.2019.08.015 (cf. p. 25).

[50] Paras Lakhani et Baskaran Sundaram. « Deep Learning at Chest Radiography : Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks ». *In* : *Radiology* 284.2 (2017), p. 574-582. doi : 10.1148/radiol.2017162326. PMID : 28436741 (cf. p. 25).

[51] A Esteva et al. « Dermatologist-level classification of skin cancer with deep neural networks ». *In* : *Nature* (2 fév. 2017). doi : 10.1038/nature21056. PMID : 28117445 (cf. p. 25).

[52] V Gulshan et al. « Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs ». *In* : *JAMA* (13 déc. 2016). doi : 10.1001/jama.2016.17216. PMID : 27898976 (cf. p. 25).

[53] Suhyeon Kim et al. « Risk score-embedded deep learning for biological age estimation : Development and validation ». *In* : *Information Sciences* 586 (1er mars 2022), p. 628-643. issn : 0020-0255. doi : 10.1016/j.ins.2021.12.015 (cf. p. 25).

[54] Sanjay Nagaraj et Tim Q Duong. « Deep Learning and Risk Score Classification of Mild Cognitive Impairment and Alzheimer's Disease ». *In* : *medRxiv* (2020). doi : 10.1101/2020.11.09.20226746 (cf. p. 25).

[55] J. Zhang et al. « Patient2Vec : A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record ». *In* : *IEEE Access* 6 (2018), p. 65333-65346. issn : 2169-3536. doi : 10.1109/ACCESS.2018.2875677 (cf. p. 25).

[56] Laura Plaza, Alberto Díaz et Pablo Gervás. « A semantic graph-based approach to biomedical summarisation ». *In* : *Artificial Intelligence in Medicine* 53.1 (1er sept. 2011), p. 1-14. ISSN : 0933-3657. DOI : 10.1016/j.artmed.2011.06.005 (cf. p. 25).

[57] Qingyu Chen, Yifan Peng et Zhiyong Lu. « BioSentVec : creating sentence embeddings for biomedical texts ». *In* : (22 oct. 2018). DOI : 10.1109/ICHI.2019.8904728 (cf. p. 25).

[58] Emeric Dynomant et al. « Doc2Vec on the PubMed corpus : study of a new approach to generate related articles ». *In* : *arXiv:1911.11698 [cs]* (26 nov. 2019). arXiv : 1911.11698 (cf. p. 25).

[59] Alvin Rajkomar et al. « Scalable and accurate deep learning for electronic health records ». *In* : *npj Digital Medicine* 1.1 (déc. 2018), p. 18. ISSN : 2398-6352. DOI : 10.1038/s41746-018-0029-1. arXiv : 1801.07860 (cf. p. 25).

[60] Yanbo Xu et al. « RAIM : Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data ». *In* : *arXiv:1807.08820 [cs, stat]* (23 juill. 2018). DOI : 10.1145/3219819.3220051. arXiv : 1807.08820 (cf. p. 25).

[61] Bret Nestor et al. « Feature Robustness in Non-stationary Health Records : Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks ». *In* : *arXiv:1908.00690 [cs, stat]* (août 2019) (cf. p. 25).

[62] David Moher et al. « Preferred Reporting Items for Systematic Reviews and Meta-Analyses : The PRISMA Statement ». *In* : *PLoS Medicine* 6.7 (juill. 2009), e1000097. DOI : 10.1371/journal.pmed.1000097 (cf. p. 29).

[63] Vincent Lequertier et al. « Hospital Length of Stay Prediction Methods : A Systematic Review ». *In* : *Medical Care* 59.10 (oct. 2021), p. 929-938. ISSN : 0025-7079. DOI : 10.1097/MLR.0000000000001596 (cf. p. 30, 113).

[64] Vincent Lequertier. « Méthode globale de prédiction des durées de séjours avec intégration des données incrémentales et évolutives ». Réunion scientifique RESHAPE. Lyon, 9 oct. 2020 (cf. p. 30, 113).

[65] Hrayr Harutyunyan et al. « Multitask learning and benchmarking with clinical time series data ». *In* : *Scientific Data* 6.1 (déc. 2019), p. 96. ISSN : 2052-4463. DOI : 10.1038/s41597-019-0103-9 (cf. p. 41).

[66] Roy Schwartz et al. « Green AI ». *In* : *Communications of the ACM* 63.12 (17 nov. 2020), p. 54-63. ISSN : 0001-0782, 1557-7317. DOI : 10.1145/3381831 (cf. p. 41).

[67] Jonathan Huang et al. « Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors ». *In* : *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017). doi : 10.1109/CVPR.2017.351 (cf. p. 41).

[68] A. Elixhauser et al. « Comorbidity measures for use with administrative data ». *In* : *Medical Care* 36.1 (jan. 1998), p. 8-27. issn : 0025-7079. doi : 10.1097/00005650-199801000-00004. PMID : 9431328 (cf. p. 43).

[69] Vincent Lequertier et al. « Predicting length of stay with administrative data from acute and emergency care : an embedding approach ». *In* : *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. Août 2021, p. 1395-1400. doi : 10.1109/CASE49439.2021.9551429 (cf. p. 45, 113).

[70] Vincent Lequertier. « Global method for predicting the length of stay in hospital using incremental and evolutionary data ». Doctoral Seminar DISP. Lyon, 3 déc. 2021 (cf. p. 45).

[71] Jacob Cohen. « A Coefficient of Agreement for Nominal Scales ». *In* : *Educational and Psychological Measurement* 20.1 (avr. 1960), p. 37-46. issn : 0013-1644, 1552-3888. doi : 10.1177/001316446002000104 (cf. p. 53).

[72] Vincent Lequertier et al. « Length of stay prediction with standardized hospital data from acute and emergency care using a deep neural network ». *In* : *(soumis pour publication)* (2022) (cf. p. 56, 113).

[73] Vincent Lequertier. « Prédiction des durées de séjours avec des données médico-administratives à l'aide d'un réseau de neurones. » Congrés ADELF EMOIS. Dijon, 31 mars 2022. url : https://doi.org/10.1016/j.respe.2022.01.071 (cf. p. 56, 114).

[74] Vincent Lequertier. « Length of stay prediction with standardized hospital data from acute and emergency care using a deep neural network ». Réunion scientifique RESHAPE. Lyon, 8 avr. 2022 (cf. p. 56, 114).

[75] Lorenzo Brigato et Luca Iocchi. « A Close Look at Deep Learning with Small Data ». *In* : *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, p. 2490-2497. doi : 10.1109/ICPR48806.2021.9412492 (cf. p. 95).

[76] Blanca Gallego et al. « Exploring the role of pathology test results in the prediction of remaining days of hospitalisation ». *In* : *Studies in Health Technology and Informatics* 178 (2012), p. 45-50. issn : 0926-9630. doi : 10.3233/978-1-61499-078-9-45. PMID : 22797018 (cf. p. 95).

[77] Andrew Brennan et al. « A Method to Account for Variation in Congenital Heart Surgery Length of Stay ». *In* : *Pediatric Critical Care Medicine : A Journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies* 18.6 (juin 2017), p. 550-560. issn : 1529-7535. doi : 10.1097/PCC.0000000000001168. PMID : 28437365 (cf. p. 95).

[78] Gleb Danilov et al. « Prediction of Postoperative Hospital Stay with Deep Learning Based on 101 654 Operative Reports in Neurosurgery ». *In* : *Studies in Health Technology and Informatics* 258 (2019), p. 125-129. issn : 1879-8365. PMID : 30942728 (cf. p. 95).

[79] Muhammad Ayaz et al. « The Fast Health Interoperability Resources (FHIR) Standard : Systematic Literature Review of Implementations, Applications, Challenges and Opportunities ». *In* : *JMIR Medical Informatics* 9.7 (30 juill. 2021), e21929. issn : 2291-9694. doi : 10.2196/21929 (cf. p. 96).

[80] George Hripcsak et al. « Observational Health Data Sciences and Informatics (OHDSI) : Opportunities for Observational Researchers ». *In* : *Studies in health technology and informatics* 216 (2015), p. 574-578. issn : 0926-9630. doi : 10.3233/978-1-61499-564-7-574. PMID : 26262116 (cf. p. 96).

[81] J Marc Overhage et al. « Validation of a common data model for active safety surveillance research ». *In* : *Journal of the American Medical Informatics Association : JAMIA* 19.1 (2012), p. 54-60. issn : 1067-5027. doi : 10.1136/amiajnl-2011-000376. PMID : 22037893 (cf. p. 96).

[82] *Loi n° 2004-810 relative à l'assurance maladie*. 13 août 2004. url : https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000625158 (cf. p. 96).

[83] Direction de l'information légale et administrative. *Mon espace santé disponible depuis janvier 2022*. 14 avr. 2022. url : https://www.service-public.fr/particuliers/actualites/A15264 (cf. p. 96).

[84] P. W. New et al. « Reducing the length of stay for acute hospital patients needing admission into inpatient rehabilitation : a multicentre study of process barriers : Rehabilitation admission barriers ». *In* : *Internal Medicine Journal* 43.9 (sept. 2013), p. 1005-1011. issn : 14440903. doi : 10.1111/imj.12227 (cf. p. 97).

[85] I. D. M. Smith et al. « Pre-operative predictors of the length of hospital stay in total knee replacement ». *In* : *The Journal of Bone and Joint Surgery. British volume* 90-B.11 (nov. 2008), p. 1435-1440. issn : 2044-5377. doi : 10.1302/0301-620X.90B11.20687 (cf. p. 97).

[86] Judith D de Jong, Peter P Groenewegen et Gert P Westert. « Mutual influences of general practitioners in partnerships ». *In* : *Social Science & Medicine* 57.8 (oct. 2003), p. 1515-1524. issn : 02779536. doi : 10.1016/S0277-9536(02)00548-8 (cf. p. 97).

[87] Antoine Neuraz et al. « Patient Mortality Is Associated With Staff Resources and Workload in the ICU : A Multicenter Observational Study* ». *In* : *Critical Care Medicine* 43.8 (août 2015), p. 1587-1594. issn : 0090-3493. doi : 10.1097/CCM.0000000000001015 (cf. p. 97).

[88] Linda H Aiken et al. « Hospital nurse staffing and patient outcomes in Chile : a multilevel cross-sectional study ». *In* : *The Lancet Global Health* 9.8 (août 2021), e1145-e1153. issn : 2214109X. doi : 10.1016/S2214-109X(21)00209-6 (cf. p. 98).

[89] Karen B. Lasater et al. « Patient outcomes and cost savings associated with hospital safe nurse staffing legislation : an observational study ». *In* : *BMJ Open* 11.12 (1er déc. 2021). Publisher : British Medical Journal Publishing Group Section : Nursing, e052899. issn : 2044-6055, 2044-6055. doi : 10.1136/bmjopen-2021-052899. PMID : 34880022 (cf. p. 98).

[90] Jack Needleman et al. « Nurse-Staffing Levels and the Quality of Care in Hospitals ». *In* : *New England Journal of Medicine* 346.22 (30 mai 2002), p. 1715-1722. issn : 0028-4793. doi : 10.1056/NEJMsa012247. PMID : 12037152 (cf. p. 98).

[91] GBD 2019 Human Resources for Health Collaborators. « Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from 1990 to 2019 : a systematic analysis for the Global Burden of Disease Study 2019 ». *In* : *Lancet (London, England)* 399.10341 (4 juin 2022), p. 2129-2154. issn : 1474-547X. doi : 10.1016/S0140-6736(22)00532-3. PMID : 35617980 (cf. p. 98).

[92] Joaquim Santos, Bernardo Consoli et Renata Vieira. « Word Embedding Evaluation in Downstream Tasks and Semantic Analogies ». *In* : *Proceedings of the 12th Language Resources and Evaluation Conference*. LREC 2020. Marseille, France : European Language Resources Association, mai 2020, p. 4828-4834. isbn : 979-10-95546-34-4. url : https://aclanthology.org/2020.lrec-1.594 (cf. p. 99).

[93] Zheng Jia et al. « Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity ». *In* : *BMC Medical Informatics and Decision Making* 19.1 (25 avr. 2019), p. 91. issn : 1472-6947. doi : 10.1186/s12911-019-0807-y (cf. p. 99).

[94] Ye Ren, Le Zhang et P.N. Suganthan. « Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article] ». *In* : *IEEE Computational Intelligence Magazine* 11.1 (fév. 2016), p. 41-53. ISSN : 1556-603X. DOI : 10.1109/MCI.2015.2471235 (cf. p. 100).

[95] M. A. Ganaie et al. *Ensemble deep learning : A review*. Number : arXiv:2104.02395. 7 mars 2022. DOI : 10.48550/arXiv.2104.02395. arXiv : 2104.02395[cs] (cf. p. 100).

[96] Carlos N. Silla et Alex A. Freitas. « A survey of hierarchical classification across different application domains ». *In* : *Data Mining and Knowledge Discovery* 22.1 (1er jan. 2011), p. 31-72. ISSN : 1573-756X. DOI : 10.1007/s10618-010-0175-9 (cf. p. 100).

[97] Peter N. Robinson et al. « A Hierarchical Ensemble Method for DAG-Structured Taxonomies ». *In* : *Multiple Classifier Systems*. Sous la dir. de Friedhelm Schwenker, Fabio Roli et Josef Kittler. T. 9132. Series Title : Lecture Notes in Computer Science. Cham : Springer International Publishing, 2015, p. 15-26. ISBN : 978-3-319-20247-1. DOI : 10.1007/978-3-319-20248-8_2 (cf. p. 100).

[98] Antoine Duclos et al. « Développement d'un Outil Informatisé de Prédiction des Durées de Séjour ». Séminaire Gestion des Lits et Anticipation des Sorties. Paris, France, 19 mai 2016 (cf. p. 104, 105, 108).

[99] Zachary C. Lipton. « The Mythos of Model Interpretability : In machine learning, the concept of interpretability is both important and slippery. » *In* : *Queue* 16.3 (1er juin 2018), p. 31-57. ISSN : 1542-7730. DOI : 10.1145/3236386.3241340 (cf. p. 103).

[100] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin. « "Why Should I Trust You?" : Explaining the Predictions of Any Classifier ». *In* : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA : Association for Computing Machinery, 13 août 2016, p. 1135-1144. ISBN : 978-1-4503-4232-2. DOI : 10.1145/2939672.2939778 (cf. p. 103).

[101] Neil Savage. « Breaking into the black box of artificial intelligence ». *In* : *Nature* (29 mars 2022), p. d41586–022–00858-1. ISSN : 1476-4687. DOI : 10.1038/d41586-022-00858-1 (cf. p. 103).

[102] Mukund Sundararajan, Ankur Taly et Qiqi Yan. « Axiomatic attribution for deep networks ». *In* : *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia : JMLR.org, 6 août 2017, p. 3319-3328 (cf. p. 104).

[103] Adam Paszke et al. « PyTorch : An Imperative Style, High-Performance Deep Learning Library ». *In* : *Advances in Neural Information Processing Systems 32*. Sous la dir. de H. Wallach et al. Red Hook, NY, USA : Curran Associates, Inc., 2019, p. 8026-8037. URL : http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (cf. p. 104).

[104] Khai Wah Khaw et al. « Reactions towards organizational change : a systematic literature review ». *In* : *Current Psychology* (13 avr. 2022). ISSN : 1936-4733. DOI : 10.1007/s12144-022-03070-6 (cf. p. 106).

[105] Tonja Blom. « Organisational wellness : Human reaction to change ». *In* : *South African Journal of Business Management* 49.1 (28 juin 2018). ISSN : 2078-5976, 2078-5585. DOI : 10.4102/sajbm.v49i1.2 (cf. p. 106).

[106] Philippe Durieux et Pierre Ravaud. *Méthodes quantitatives pour évaluer les interventions visant à améliorer les pratiques*. Juin 2007. URL : https://www.has-sante.fr/jcms/c_597750/fr/methodes-quantitatives-pour-evaluer-les-interventions-visant-a-ameliorer-les-pratiques (cf. p. 107).

[107] Loraine Busetto, Wolfgang Wick et Christoph Gumbinger. « How to use and assess qualitative research methods ». *In* : *Neurological Research and Practice* 2.1 (27 mai 2020), p. 14. ISSN : 2524-3489. DOI : 10.1186/s42466-020-00059-z (cf. p. 107).

[108] Vincent Lequertier et al. « Prédiction des durées de séjours avec des données hospitalières standardisées ». *In* : Congrès National de la Recherche des IUT 2022. Roanne, France, 10 juin 2022, p. 104-106. URL : https://web.archive.org/web/20220613134450/https://cnriut2022.sciencesconf.org/data/Recueil_des_Publications.pdf (cf. p. 114).

[109] Swapna Gokhale et al. « Hospital Length of Stay Prediction Tools for General Surgery Populations and Total Knee Arthroplasty Admissions : Systematic Review and Meta-Analysis ». *In* : *(soumis pour publication)* (2022) (cf. p. 115).

Annexe A

Implémentation de l'intégration de gradients

```python
import torch

# Example de modèle d'apprentissage profond
class Model(torch.nn.Module):
    def __init__(self):
        super(Model, self).__init__()
        self.lin1 = torch.nn.Linear(20, 10)
        self.relu = torch.nn.ReLU()
        self.lin2 = torch.nn.Linear(10, 3)

    def forward(self, input):
        return torch.nn.functional.log_softmax(
            self.lin2(self.relu(self.lin1(input))), dim=1
        )

model = Model()
# Génère un jeu de donnée avec 50 lignes et 20 colonnes
inputs = torch.rand(50, 20, requires_grad=True)
baseline = torch.zeros_like(inputs, requires_grad=True)
n_steps = 20

# Tableau contenant les gradients à chaque pas de l'interpolation
grads = []
for step in range(1, n_steps + 1):
    model.zero_grad()
    # Interpolation de la baseline à la donnée d'entrée
    baseline_input = baseline + ((step / n_steps) * (inputs - baseline))
    out = model(baseline_input)
    # Obtient les classes prédites et sélectionne les sorties corespondantes
    idx = out.argmax(dim=1).unsqueeze(1)
```

```
31      out = out.gather(dim=1, index=idx)
32      # Rétropropagation pour obtenir les gradients
33      out.backward(torch.ones_like(idx))
34      grads.append(inputs.grad.detach())
35
36  # Calcul la moyenne des interpolations
37  grads = torch.stack(grads, 0).mean(dim=0)
38  # Calcul l'attribution
39  attr = (inputs - baseline).detach() * grads
```

Listing A.1 – Implémentation de l'intégration de gradient

Index

Méthode globale de prédiction des durées de séjours hospitalières avec intégration des données incrémentales et évolutives

Résumé

Prédire la durée de séjour des patients est un enjeu important pour l'organisation des activités de soin dans les hôpitaux, notamment en termes de gestion des lits et de préparation de la sortie des patients. Faciliter l'organisation des activités de l'hôpital influence l'accès, la qualité et l'efficience des soins. Dans cette thèse, nous avons cherché à prédire la durée de séjour pour tous les patients de l'hôpital, à toutes les étapes qui composent leurs parcours de soins, à l'aide de données médico-administratives standardisées de Médecine, Chirurgie, Obstétrique qui sont collectées pour le remboursement des soins. Nous avons commencé par faire une revue systématique de la littérature sur les méthodes de prédiction des durées de séjours, afin de mieux comprendre la préparation des données, les différentes approches de prédiction et la façon de rapporter les résultats. Nous avons ensuite travaillé sur une méthode de prétraitement des données et déterminé si les *embeddings* peuvent représenter les concepts médicaux dans le cadre des prédictions de durées de séjours *via* un réseau de neurones. La capacité du réseau de neurones à correctement prédire la durée de séjour a été évaluée et comparée avec celle d'une forêt aléatoire et d'une régression logistique. Nos travaux montrent que la durée de séjour hospitalière peut être prédite au moyen d'un réseau de neurones avec des données médico-administratives standardisées disponibles pour tous les patients.

Mots clés : durée de séjour, prédiction, intelligence artificielle, aide à la décision, dossier médical informatisé

Global method for predicting the hospital length of stay using incremental and evolutionary data

Abstract

Predicting patient length of stay is an important issue for the organization of care activities in hospitals, especially for beds management the and preparation for patients discharge. Facilitating the organization of hospital activities influences access, quality and efficiency of care. In this thesis, we sought to predict length of stay for all patients in the hospital, at all stages that make up their care pathways, using standardized Medical, Surgical, Obstetric medico-administrative data collected for reimbursement of care.
We began by conducting a systematic review of the literature on methods for predicting lengths of stay, in order to better understand data preparation, the different prediction approaches, and how to report the results. We then worked on a data preprocessing method and investigated the ability of embeddings to represent medical concepts in the context of length of stay predictions *via* a neural network. The ability of the neural network to correctly predict length of stay was rigorously evaluated and compared with a random forest and a logistic regression. This work shows that hospital length of stay can be predicted by a neural network using standardized medical-administrative data available for all patients.

Keywords: length of stay, forecasting, artificial intelligence, clinical decision support, electronic health record

---