NNT: 2023EMSEM002

THÈSE DE DOCTORAT

de Mines Saint-Étienne - Une école de l'IMT

École Doctorale N° 488
(Sciences, Ingénierie, Santé)

Spécialité de doctorat: Génie Industriel

Soutenue publiquement le 20 janvier 2023 par:

Chen HE

Analyzing, optimizing, and explaining hospital miscoding for coding practice improvement

(Analyser, optimiser et expliquer le mauvais codage hospitalier pour l'amélioration des pratiques de codage)

Devant le jury composé de:

VERDIER, Christine	Professeur	Université Grenoble Alpes	Présidente du jury
DUCLOS, Antoine	Professeur	Université Claude Bernard Lyon 1	Examinateur
JOURDAN, Laëtitia	Professeur	Université de Lille	Rapporteur
LENCA, Philippe	Professeur	IMT Atlantique	Rapporteur
BOUSQUET, Cédric	Docteur	CHU de Saint-Étienne	Co-encadrant
DALMAS, Benjamin	Maître assistant	Mines de Saint-Étienne	Encadrant
TROMBERT-PAVIOT, Béatrice	Professeur	Université Jean Monnet	Co-directeur de thèse
XIE, Xiaolan	Professeur	Mines de Saint-Étienne	Directeur de thèse

Affidavit

Je soussigné, Chen HE, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Prof. Xiaolan XIE et Prof. Béatrice TROMBERT-PAVIOT, dans le respect des principes d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisées dans le respect de la charte nationale de déontologie des métiers de la recherche.

Ce travail n'a pas été précédemment soumis dans sa globalité en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Saint-Étienne, le 01 janvier 2023

Abstract

The primary objective of this thesis is to develop a series of methodologies based on optimization, machine learning, and data mining techniques, which are then adopted in the context of PMSI (i.e., Information Systems Medicalization Program) for the hospital miscoding problem. A series of data-driven optimization approaches are developed to identify hospital miscoding behaviors, analyze them, model them, and correct them in order to improve the operational efficiency of the University Hospital of Saint Etienne (CHU-SE). This thesis puts a particular emphasis on model transparency and interpretability since they are critical elements to guarantee the fairness of developed applications. An evaluation of the proposed approaches and an analysis of the benefits of the optimization approaches are performed. Experimental results show that the proposed approaches are able to decrease the hospital miscoding rate without increasing the workload of medical coding staff in the hospital. The results are promising and reveal that potential benefits for all private clinics and public hospitals in France are achievable.

Résumé

L'objectif principal de cette thèse est de développer une série de méthodologies basées sur des techniques d'optimisation, d'apprentissage automatique et de fouille de données. Ces méthodes sont ensuite appliquées dans le contexte du PMSI (i.e., Programme de Médicalisation des Systèmes d'Information) pour répondre au problème de mauvais codage hospitalier. Une série d'approches d'optimisation guidées par les données est développée pour identifier les comportements de mauvais codage hospitalier, les analyser, les modéliser et les corriger afin d'améliorer l'efficacité opérationnelle du CHU de Saint Etienne (CHU-SE). Cette thèse met un accent particulier sur la transparence et l'interprétabilité des modèles car ce sont des éléments critiques pour garantir l'équité des applications développées. Une évaluation des approches proposées et une analyse des avantages des approches d'optimisation sont réalisées. Les résultats expérimentaux montrent que les approches proposées sont capables de diminuer le taux de mauvais codage dans les hôpitaux sans nécessairement augmenter la charge de travail du personnel de codage médical dans l'hôpital. Les résultats sont prometteurs et révèlent que des bénéfices potentiels pour toutes les cliniques privées et les hôpitaux publics en France sont réalisables.

Acknowledgements

At the beginning of the thesis, I would like to take this opportunity to thank a few people who have influenced me a lot during the preparation of my thesis.

First of all, I would like to thank Prof. Xiaolan XIE and Ass Prof. Benjamin DALMAS, my thesis director and supervisor, respectively, for their patient guidance and invaluable feedback, on moments when I could not see clearly the way forward. Without their guidance, I would not have the courage to complete projects for which the results are not guaranteed. Uncertainty is exactly what I have been taught to avoid in scientific research. Prof. XIE suggested looking beyond the code (or the programming task) and refining an idea before writing a program, which has had a long-term positive impact on my thesis writing. In addition, this work would not have been possible without their financial support.

During my studies, I have been thinking about the differences between my graduate studies (graduate, and post-graduate level) and my non-graduate studies (elementary, secondary, and undergraduate level). Here are the initial answer I found. At the non-graduate level, teachers will give students a specific problem and then ask the students to master some efficient solution approaches. In other words, the problems themselves are provided in advance.

At the graduate level, the faculty members will teach you how to identify and formulate a problem. The choice of solution approaches is just a secondary issue. This is mainly due to (i) once the problem has been defined, students can refer to and make use of some known solution approaches, and (ii) students can design their own heuristic algorithms even though no known solution is available for the problem at hand. At this stage, the most important thing is the problem itself rather than the solution approaches. Therefore, we must avoid the phenomenon revealed by the law of the hammer - "If the only tool you have is a hammer, you tend to treat every problem as a nail.". I am very grateful to my thesis advisors for making me aware of this problem.

Apart from that, I would also like to thank my thesis co-directory Prof. Béatrice TROMBERT-PAVIOT. She trusted me for my observation internship in the Medical Information Department (SSPIM) of the University Hospital of Saint Etienne (CHU-SE) at the beginning of the thesis preparation. Without this apprenticeship, I would not have had an in-depth understanding of the actual coding process in a French healthcare institution and, therefore, would not have had the chance to achieve this work.

I also would like to thank my thesis co-supervisor, Dr. Cédric BOUSQUET. He gave me a lot of help during my observation internship and thesis preparation. He taught me a lot about healthcare management, the organizational structure of French medical institutions, and hospital operations.

Without the support of working staff at the SSPIM of the CHU-SE, this thesis might have ended with abstract and theoretical approaches to a practical problem without considering the particularities of real-life hospital data. Theophile and Chrystel, thank you for letting me aware of this problem.

Thanks to my colleagues from the Center for Biomedical and Healthcare Engineering (CIS), École des Mines de Saint-Étienne (EMSE): Vincent Augusto, Thierry GARAIX, Laurent NAVARRO, Ramsey PHAN, Julia FLECK, Marianne SARAZIN, Colin RIVIERE, Solmaz FARZANEH, Ahmed BAKALI EL KASSIMI, Asma GASMI, Mohamed El Habib MESSABIS, Long NGUYEN-PHUOC, Daniëlle Hooijenga, Jules LE-LAY, Omar RIFKI, Zhihao PENG, Nilson Herazo-Padilla, Marius Huguet, Cyriac Azefack, and Oussama Batata, especially for the mini-discussions, the coffee breaks, and the nice evenings that we shared in Saint-Étienne. I am lucky to have those beautiful memories and times. In particular, thanks to Camille BREEN, with whom I had the chance to share my office during these three years.

I also want to express my gratitude to all the friends I met during my study in France. Thanks for their help in my daily life, for their kindness, and their continuous encouragement.

除此之外，我需要特别感谢，在法国六年里的求学岁月中，给与我支持和帮助的朋友：路遥，Kevin Delmares，刘念，罗素，曹健，马洪锋，徐思翔，张国栋。六年海外，三年疫情，感谢你们的陪伴和在生活上给与我的帮助，我才得以顺利完成硕士博士学业。你们之中有的已经离开法国，有的仍留在法国战斗。未来大家注定会分散在世界的各个角落，我借此机会，祝愿你们生活美满、前程似锦。我相信未来一定会有更优秀的朋友，陪伴你们度过更加美好的时光。

最后，我想感谢我的父母和家人，有你们在背后的无条件支持，我才能心无旁骛地投入到学习以及工作中。我不知道如何向你们表达我的感谢，因为我欠你们太多。

Thanks!

Chen HE
January 2023
at Saint-Étienne

Seasons in Sainté

- Adapted from "Seasons in the Sun - Westlife"

To I4S France staff and all my friends:

goodbye to you, my dear colleagues
we've known each other in I4S France team
together we've enjoyed coffees and breaks
shared our ideas and memories
learned of Python and C

goodbye to you, my old friends
we've known each other behind Centre Deux
together we've traveled and land and sea
wish you joy and happiness
when you call me, I'll be there

we had such a beautiful scene, we had seasons in Sainté
but the hills that we climbed were just seasons out of time

goodbye Xiaolan, our little one
you gave us love and helped us find the way
and every time that we were lost
you will always come around
get our feet back on the ground

goodbye Sainté, it's bittersweet
when all the birds are singing near Cité-du-design
now that's the winter in the air
snowflakes are everywhere
when you see them, I was there

we had such a wonderful time, we had seasons in Sainté
but the pic(nics) that we shared were just seasons out of time

goodbye to you, my dear colleagues
when autumn leaves are falling, please don't go
but it's good to see you find a job
wish you best for your new career
we will always miss you here

goodbye to you, my old friends
when Christ(mas) fireworks are blooming, please don't go
but it's great to see you're chasing your dream
wish you best for your new life
keep in touch as a family

we had such a beautiful scene, we had seasons in Sainté
but the hills that we climbed were just seasons out of time
we had such a wonderful time, we had seasons in Sainté
but the pic(nics) that we shared were just seasons out of time
we had joy, we had fun, we had seasons in Sainté
but the wine and the song like the seasons never gone

Chen HE
January 2023
Saint-Étienne

Contents

List of Figures

List of Tables

Glossary

ARS

A regional health agency (ARS) is a public administrative establishment of the French government, which is responsible for implementing health policy in its region. (Agence régionale de santé, in French). 8

ATIH

Technical agency for information on hospitalization (ATIH) is a French public institution, created in 2000 as a centre of expertise on the four fields of hospital activity: MCO (Medicine, surgery, obstetrics and dentistry), HAD (Hospitalization at home), SSR (Follow-up and Rehabilitation Care) and PHY (Psychiatry), operating under the supervision of the ministers of health and social security. It plays an important role in the field of hospital medical statistics, for the protection of personal health data, and for the analysis and evaluation of the performance of the activities of healthcare institutions. The Agency is located in Lyon, and has an office in Paris., (Agence technique de l'information sur l'hospitalisation, in French). 8

CCAM

The (CCAM) is a French Social Security nomenclature that includes the coding of medical procedures performed by physicians, dental surgeons and midwives. (Classification commune des actes médicaux in French). 1, 3, 8, 10, 11, 13, 26, 35, 90

CHU-SE

The University Hospital of Saint-Étienne (or CHU-SE) is a university hospital in France. The CHU operates in the Loire department but also in the north of the Ardèche in the Annonay basin and the northeast of the Haute-Loire. It employs 7,588 people, making it the leading recruiter in the Loire health sector. It has a total capacity of 1,802 beds and places in 2020, including 42 intensive care beds. (CHU de Saint-Étienne, in French). 2, 4, 20, 44, 90, 100, 138, 156, 167, 168, II–V

CIS

Center for Biomedical and Healthcare Engineering (CIS) is a research center of the École des Mines de Saint-Étienne (EMSE). (Centre Ingénierie et Santé, in French). V

CMA

CMA is the abbreviation of complications and morbidities, consisting of (i) complications = related to the pathology or treatment, e.g.: surgical wound deunion; and (ii) associated morbidity = intercurrent pathology, e.g.: known sickle cell anemia in a patient coming for appendicitis. It is also the significant associated diagnose (DAS) coded in ICD-10-FR, which have been shown to significantly increase the length of stay and therefore the cost of the stay. Consequently, they are taken into account by T2A and generally allow an increase in the cost of the stay. (Complication et morbidité associée in French). 15, 17, 21, 24, 25

CPAM

A primary health insurance fund (CPAM) is an organization, related to health and exercising a public service mission in France. (Caisse primaire d'assurance maladie in French). 16, 22

DAS

The significant associated diagnosis(es) (DAS) concerns any associated morbidity (diagnosis or therapy), which has led to additional management. It can be a disease progression, a new symptom (on the basis of the DP), a new condition (acute, chronic, punctual), an alteration of an organ, etc. (Diagnostic associé significatif , in French). 10, 11, 19, 22, 131

DIM

The Medical Information Department (DIM) of a public or private hospital manages patients' health information. In particular, it is responsible for coding medical activity for the purpose of reimbursement of hospital services by the Health Insurance. (Le Département de l'Information Médicale, in French). 1, 8, 11, 20, 100, 138

DP

The principal diagnosis (DP) is the medical diagnosis that can optimally match medical resources and manage patients, either in terms of medical effort or in terms of the initial management in the medical unit. It is usually assigned after collecting all the medical data for a patient (Diagnostic principal, in French). 10, 11, 18, 19, 22, 131

DR

The related diagnosis (DR) is used to specify the pathological context when the DP is coded with a Z-code (Chapter XXI) in ICD-10-FR. (Diagnostic relié, in French). 10, 11, 19, 22, 131

EHRs

EHRs is an abbreviation of Electronic Health Records. 1, 9, 20, 33, 90, 100

EMSE

The École Nationale Supérieure des Mines de Saint-Étienne (EMSE) is an engineering school in the French higher education system. The school was founded in 1816 and is part of the French Ministry of Economy and Industry. (École des Mines de Saint-Étienne, in French). 30, 133, 138, V

GHM

A homogeneous group of patients (GHM) groups together medical treatments of the same medical and economic nature and constitutes the basic classification category in MCO. Each stay ends up in a GHM according to an algorithm based on the medico-administrative information contained in the standardized discharge summary (RSS) of each patient. GHM corresponds to the concept of Diagnosis-related group (DRG) in North America, (Groupe Homogène de Malades, in French). 9, 14, 16, 24, 137

GHS

In the context of activity-based pricing (T2A) in medicine, surgery, obstetrics and dentistry (MCO), the homogeneous group of stays (GHS) corresponds to the tariff for the homogeneous group of patients (GHM). The vast majority of GHMs have only one GHS, i.e. a single tariff, but in some cases a GHM may have two or more tariffs (depending, for example, on different levels of equipment for the same treatment). (Groupe Homogène de Séjours, in French). 9, 16, 137

HAD

HAD is an abbreviation of Hospitalisation à domicile, in French, which means Hospitalization at home. 8

HAS

French health authority. (Haute autorité de santé, in French). 22, 132

ICD

International Classification of Diseases and Related Health Problems. It published by the World Health Organization (WHO), is used to encode diagnoses in healthcare institutions. 1, 11, 22

ICD-10-FR

ICD-10-FR is the French modification of the International Classification of Diseases, tenth version (ICD-10). It published by the World Health Organization (WHO), is used to encode diagnoses for different PMSI domains. (CIM-10: La Classification internationale des maladies, dixième édition, in French). 3, 8, 10–13, 26, 35, 90

LPP

LPP (La liste des produits et prestations, in French) is a coding system used for encoding reimbursable medical products (i.e., implantable medical devices, wheelchairs, orthoses, external prostheses, etc.). It contains 1800 codes grouped under 60 categories. 3, 9, 16, 26, 35

MCO

MCO is an abbreviation of Médecine Chirurgie Obstétrique, in French, which means Medicine, surgery, obstetrics, and dentistry departments. 2, 8, 9, 11, 14, 20

MIM

Medical Information Physician, (Médecin de l'Information Médicale, in French). 137

PMSI

The information systems medicalization program (PMSI) provides a synthetic and standardized description of the medical activity of health care institutions. It is based on the standardized medico-administrative data collected in a standard manner. It includes 4 "fields": "medicine, surgery, obstetrics and odontology" (MCO), "follow-up or rehabilitation care" (SSR), "psychiatry" in the form of the RIM-Psy (collection of medical information in psychiatry), and "hospitalization at home" (HAD), (Programme de médicalisation des systèmes d'information, in French). 1, 3, 6, 8, 20, 21, 25, 28, 29, 91

PSY

PSY is an abbreviation of Psychiatry department (Psychiatrie, in French). 2, 8

RSA

The production of the anonymous discharge summary (RSA) is based on the standardized discharge summary (RSS) and is the result of an automatic process carried out by a software module provided by the national services. (Résumé de sortie anonyme in French). 11

RSS

The standardized discharge summary (RSS) is a medical document that gathers various types of demographic, diagnostic and therapeutic information about each patient's stay in hospital. The standardised discharge summary (RSS) is made up of all the RUM related to the same patient stay in the MCO sector.(Résumé de Sortie Standardisé, in French). 8, 9, 11, 25, 131

RUM

A Summary of Medical Unit (RUM) is produced at the end of each patient's stay in a medical unit providing medical, surgical, obstetrical and odontological care. It contains administrative and medical information, coded according to standardized nomenclatures and classifications, so that it can be processed automatically. (Résumé d'Unité Médicale, in French). 8, 10, 11

SSPIM

The Unit of public health and medical information (SSPIM) is the Medical Information Department (DIM) of the University Hospital of Saint-Étienne (Le Service de Santé Publique et d'Informatique Medical, CHU de Saint-Étienne, in French). 2, 20, 30, 90, 107, 131, 133, 138, IV, V

SSR

SSR is an abbreviation of Soins de Suite et de réadaptation, in French, which means Follow-up and Rehabilitation Care. 2, 8

T2A

Activity-based pricing (T2A) is the only method of financing public and private healthcare institutions in France. It was launched in 2004 as part of the "Hospital 2007" plan. (La tarification à l'activité, in French). 6–9, 11, 13, 14, 16, 20, 24, 25, 29, 90

TIM

Medical Information Technician, (Technicien de l'Information Médicale, TIM, in French). 2, 3, 20, 133, 136, 137

TSH

Senior Hospital Technician, (Technicien Supérieur Hospitalier, in French). 20, 137

UCD

UCD (Unité commune de dispensation, in French) is a coding system for encoding all drugs delivered by city pharmacies or hospital pharmacies. Drugs are coded in standard dispensing units. 3, 9, 16, 26, 35

UM

A medical unit (MU) corresponds to a homogeneous set of medical activities and is designated as an individualized set of medical resources providing care to hospitalized patients, identified by a specific code in a nomenclature maintained by the institution. (Unité médicale, in French). 8, 10

General introduction

In French healthcare institutions, coded data aim at summarizing the patient stays in terms of healthcare resource consumption. In general, the coded data is generated according to patient histories registered in patients' EHRs (Electronic Health Records) and several conventional coding systems such as the International Classification of Disease (ICD) [1] for diagnoses coding , and the Classification Commune des Actes Médicaux (CCAM) [2] for coding medical acts and procedures.

With the widespread adoption of the PMSI (i.e., Information Systems Medicalization Program) [3] health information system, coded data has become more accessible in patients' EHRs in French hospitals. Under the PMSI program, *French public hospitals and private clinics have to report their medical activities (in the form of coded data) to the French Health Authority every year to get the fiscal budget and revenue.* This process is often performed by coding staff in the medical information department (DIM) of hospitals and **highly depends on coded data** such as ICD codes and CCAM codes. The advantage of using these coded data is twofold: (i) from the perspective of public health, they help monitor health trends, medical activities, and healthcare practices efficiency; (ii) from an institutional perspective, since the coded data are supposed to represent the medical services provided by the hospitals, they are considered a basis to ration hospital budget.

However, the production of coded data involves not only well-trained medical coders but also less performing physicians. The coding task consumes a lot of human resources in healthcare institutions. *Due to varying coding habits or missing information, a large number of coding errors appear in patients' EHRs.* **Hospital miscoding has multiple negative impacts on**:

1. The public health system, e.g.,:
    a) An inaccuracy in reporting epidemiological statistics;
    b) A disruption in the scheduling of health service resources;
    c) A deviation in health budget allocation;
2. Healthcare institutions, e.g.,:
    a) Leading to reduced efficiency in health service operations;
    b) A deviation of healthcare services reimbursement;
    c) An increase in potentially preventable health services costs;
3. And hospital inpatients, e.g.,:

a) Generating incorrect medical histories of patients and indirectly impacts patient health;

b) Patients receive inaccurate medical reimbursement from insurance companies.

This research project is in collaboration with the medical information department (abbreviated as SSPIM) of the University Hospital of Saint Etienne (abbreviated as CHU-SE). The CHU-SE is the largest regional hospital in Loire, France. *To ensure the quality of coded data, a quality control (QC) process is periodically performed in CHU-SE, and a subset of EHRs is selected for review.* The primary objective of such a QC process is to increase hospital fiscal revenue by correctly reporting hospital medical activities and also prevent potential financial penalties caused by hospital miscoding. The secondary objective of the QC process is to decrease the miscoding rate in hospitals.

Implementation of the QC process raises challenges. Given the heterogeneity of the population (i.e., patients are admitted to the hospital for different reasons) and the coded data (i.e., medical information is often coded by multiple coders in a different manner), the current practice of SSPIM relies mainly on simple heuristics (i.e., a set of rules of thumb based on the experience of the coding staff) to select a sub-population for review. Such a practice is often too coarse, and many unnecessary medical reviews are assigned to correctly coded hospital stays. On the other hand, miscoded cases with significant negative impacts might not be taken into account. In simple terms, the QC process is not precise enough and is not able to efficiently screen miscoded cases with significant negative impact.

The data contained in EHRs come in multiple forms: structured tabular data, semi-structured statistical chart data, unstructured textual data, etc. According to the initial monitoring phase performed in SSPIM at the beginning of this project, it generally takes about 3 minutes to search and extract relevant information from a patient's EHR. The data contained in EHRs are heterogeneous. Although these medical reviews are not expensive, reviewing EHRs of a large number of hospital stays leads to a significant budget in the context of quality control. For example, if we set the review time of a hospital stay as 3 minutes, then for a medical coder (TIM) who works 7 hours a day, he (or she) can only review $(7 \times 60)/3 = 140$ hospital stays per day without interruption. If the workday is 229 days per year [1], a medical coder can review $229 \times 140 = 32,060$ hospital stays per year. The number of full and part-time hospital stays in the CHU-SE was 93,304 in 2021 [2]. Considering the most conservative case in which medical coders have to review all these hospital stays, the CHU-SE would need at least 3 medical

[1] If we consider that an institution is closed 2 days a week on average, and if we take into account the paid vacations, we find that an employee works approximately 299 days a year (i.e., 365 days - 104 Saturdays and Sundays - 25 holidays - 7 holidays not falling on a Saturday or Sunday).

[2] According to the Medical activity report of the CHU-SE (2021), the number of hospital stays at the CHU-SE in 2021 = 58,164 stays in medical department (MCO) + 20,295 stays in surgical department (MCO) + 5,001 stays in obstetrics (MCO) + 3,469 stays in psychiatry (PSY) + 4,375 stays in Follow-up and rehabilitation care (SSR) = 91,304 hospital stays

coders. Actually, as illustrated in Figure 1.7, there are only 2 medical coders (TIMs) in SSPIM who are responsible for the quality control. **Therefore, there is a need to completely or partially automate the QC process.** The proposed techniques are required to save reviewing time and reduce the waste of human resources.

The design of appropriate profiling rules allows a personalized portfolio of medical reviews that best fits personalized characteristics of hospital stays, including some administrative information as well as relevant medical and clinical information. **Intuitively, we believe that with appropriate profiling of miscoded instances, it would be possible to significantly increase the hospital's fiscal revenue without increasing the review efforts.**

To improve the cost-efficiency of the QC process, we have investigated the design and implementation of several data-driven optimization techniques. We address the fundamental problem that whether a medical review is needed for a specific hospital stay for the increase in hospital fiscal revenue. More specifically, in the context of hospital miscoding audits, instead of simply reviewing all features (i.e., descriptive variables related to a stay code) of a selected subpopulation, the proposed approaches screen a set of miscoded cases resulting in maximum negative impacts on the healthcare institution and then select the set of most relevant features from them to review. Medical review rationing is a meaningful work of the QC process since only a small part of hospital stays need these reviews. Such a rationing process is, of course, subject to the chance constraint, i.e., only a limited number of hospital stays are selected and reviewed. In addition to the economic benefit (i.e., the increase in hospital fiscal revenue), the saved human resources and time can also be devoted to other important issues.

This manuscript concerns a series of data-driven approaches based on optimization, machine learning, or data mining techniques to optimize the quality control process (or miscoding audit process). **In this research project, we aim at (i) improving the current medical coding practices** *(i.e., enhancing the quality of coded data and decreasing the quantity of coding errors.),* **while (ii) reducing the workload of the medical coding staff in hospitals** *(i.e., increasing the efficiency of the miscoding audit task and avoiding repetitive miscoding behaviors of medical coders.).* The proposed techniques reveal some novel insights of which medical coders are unaware. These novel insights are provided to stakeholders and lead to the improvement of the current coding practices in the CHU-SE and, subsequently, could benefit other health institutions.

Thesis outline

This manuscript is organized as follows. In Chapter 1, we present an overview of the PMSI program as well as the challenges that arise from this program. We first introduce the PMSI program as well as various coding conventions (i.e., ICD-10-FR, CCAM, UCD, LPP) adopted under this program. Next, we present the hospital coding task performed in various healthcare institutions and the hospital miscoding problem arising from the coding task. We highlight direct and indirect negative impacts caused

by hospital miscoding and the need to provide medical coders with automation tools for hospital miscoding audits. At the end of the first chapter, the general research objectives are set, and an overview of the proposed methodologies is also provided.

Then, the manuscript is divided into two parts. In part I, which contains chapters 2 and 3, we are interested in the profiling of hospital miscoding and providing innovative techniques for the modeling of miscoding behaviors. In part II, which contains chapters 4 and 5, we propose two optimization approaches to the hospital miscoding problem.

In Chapter 6, we conduct an on-site evaluation of the integrated optimization approach proposed in Chapter 5. The protocol for method validation and some practical considerations are also provided. The experimental results show that the proposed approach is promising and can be implemented in the CHU-SE for future use.

1. Hospital miscoding and relevant research questions

Summary

Abstract of the chapter

This chapter introduces the research background and objectives of this thesis. This research project is under the French Information System Medicalization Program (PMSI). The PMSI program is often adopted to describe medical activities performed and health resources consumed by French healthcare institutions to finance these healthcare institutions. In Section 1.1, we first give a brief introduction to the PMSI program, including the PMSI framework, PMSI data, the coding systems (or coding conventions) adopted in the PMSI program, and the funding system - T2A (i.e., the activity-based pricing) for financing hospital medical activities. On top of this, we introduce various hospital coding tasks under the PMSI program (see Section 1.2), as well as the hospital miscoding problem that arises in the hospital coding process (refer to Section 1.3). The primary research background is given. In Section 1.4, we introduce the motivation and the necessity of miscoding audits, including the state-of-the-art methods for miscoding review, shortcomings of the existing methods, as well as the research direction of this thesis. Finally, in section 1.5, we set the general research objectives of this study, which will be discussed again at the end of this manuscript - the general conclusion part.

Keywords: PMSI program, hospital miscoding, hospital code audits.

Résumé du chapitre

Ce chapitre présente le contexte de recherche et les objectifs de cette thèse. Ce projet de recherche s'inscrit dans le cadre du Programme de Médicalisation des Systèmes d'Information français (PMSI). Le programme PMSI est souvent adopté pour décrire les activités médicales réalisées et les ressources de santé consommées par les établissements de santé français, afin de déterminer l'enveloppe budgétaire pour chacun de ses établissements. Dans la section 1.1, nous présentons d'abord brièvement le programme PMSI, notamment son cadre et les données qui le composent. Les systèmes de codage (ou conventions de codage) adoptés dans le programme PMSI ainsi que le système de financement - T2A (c'est-à-dire la tarification à l'activité) pour le financement des activités médicales hospitalières sont également introduits. En outre, nous présentons diverses tâches de codage hospitalier dans le cadre du programme PMSI (voir section 1.2), ainsi que le problème de mauvais codage hospitalier qui survient dans le processus de codage hospitalier (voir section 1.3). Le contexte de la recherche primaire est présenté. Dans la section 1.4, nous exposons la motivation et la nécessité des audits de mauvais codage, y compris les méthodes de pointe pour l'examen du mauvais codage, les lacunes des méthodes existantes, ainsi que l'orientation de la recherche de cette thèse. Enfin, dans la section 1.5, nous définissons les objectifs généraux de recherche de cette étude, qui seront à nouveau discutés à la fin de ce manuscrit - la partie conclusion générale.

Mots-clés: Programme PMSI, mauvais codage hospitalier, audits de codes hospitaliers.

1.1. Hospital coding systems

1.1.1. Information systems medicalization program (PMSI)

In order to measure the medical activities performed and resources consumed by healthcare institutions, the PMSI program is proposed to produce quantified and standardized health information - the PMSI data. The PMSI program is a national program in France and is part of the reform of the French healthcare system aimed at reducing inequalities in resources between healthcare institutions. The collection and processing of the PMSI data is managed by the Agence technique de l'information sur l'hospitalisation (ATIH).

The implementation of the PMSI program is mandatory for all French healthcare institutions (e.g., public hospitals or private clinics). The medical information department (DIM) is a functional department in French healthcare institutions. Often, the DIM is committed to collecting and processing PMSI data in order to transmit them to the ATIH. PMSI data are then de-identified and released by the ATIH. The two main objectives of the AITH are to (i) describe hospital activities (in the broadest sense) in detail; (ii) serve as a support agency for the financing of healthcare institutions (T2A).

Specifically, the PMSI data production and processing chain consist of the following steps: (i) local processing: the PMSI data are collected, processed, and de-identified within each French healthcare institution. *A part of medical information is encoded as medical codes according to the following medical coding conventions: ICD-10-FR (diagnosis), CCAM (medical procedures)*; (ii) regional processing: the PMSI data are then transmitted to Regional Health Agencies (ARS) and Health insurance agencies for quality control and regional processing; (iii) national processing: the processed PMSI data are then transmitted to the ATIH for national processing and dissemination;

The PMSI data are medico-administrative data and can be divided into four categories according to the four fields of hospital activity: MCO (Medicine, surgery, obstetrics, and dentistry), SSR (Follow-up and rehabilitation care), HAD (Hospitalization at home), PSY (Psychiatry). **In this thesis, we focus on the hospital activities performed in MCO departments**. The description of medical activities in MCO departments is based on the collection of the PMSI data and partially automated processing of these data.

As a simplified example (illustrated in Figure 1.1), imagine an outpatient is admitted to a hospital after a medical consultation. After this hospital admission, the patient becomes an inpatient rather than an outpatient. The patient may be transferred between multiple medical units (UMs). During the hospital stay, physicians will give the patient a series of medical treatments until the patient is discharged from the hospital or dies in the hospital. The medical unit summary (RUM) is established when the patient is discharged from a medical unit (UM). *The RUM contains coded medical information such as ICD-10-FR codes and CCAM codes.* At the end of the patient's hospitalization, a standardized discharge summary (RSS) must be produced. The RSS is composed of one or more medical unit summaries (RUMs) produced during the

hospital stay. All these medical histories, including RUM and RSS, are registered in the patient's EHRs.

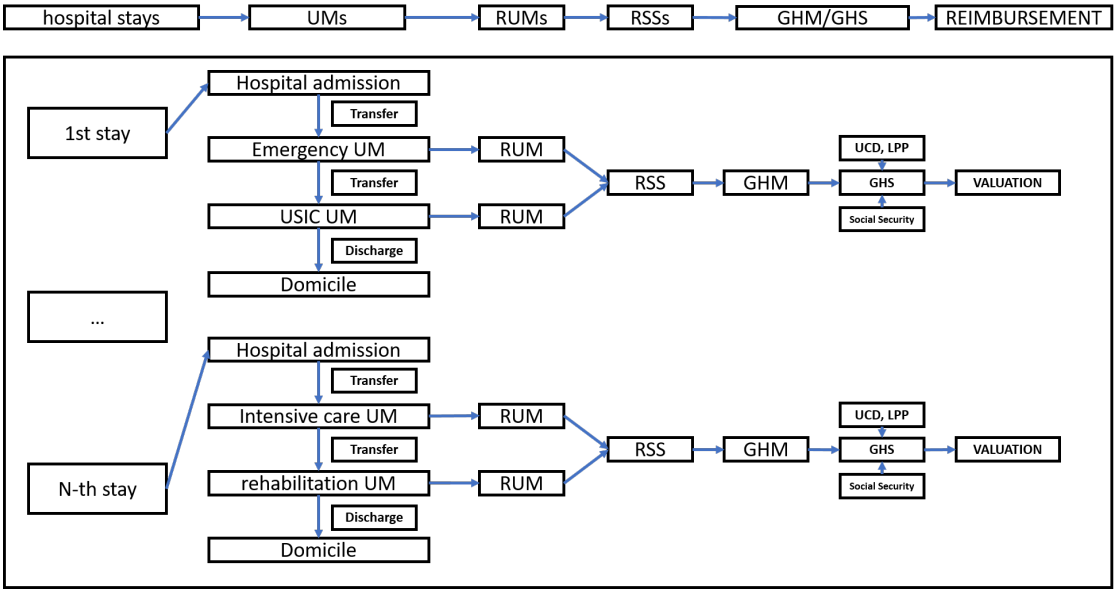


Figure 1.1.: A standardized pricing process for healthcare services.

After that, GHM codes are assigned to patient stays by a grouping algorithm based on ICD codes, CCAM procedures, sex, age, discharge status, and the presence of comorbidities or complications. Hospital stays within each GHM group are expected to be clinically similar and consume the same level of health service resources. Note that GHM groups correspond to Diagnosis-related groups (DRGs) in North America.

The concept of GHS and GHM are almost identical. A GHM gives a hospital stay a medical label, while a GHS allows attaching a tariff to a given hospital stay. Often, 1 GHM corresponds to 1 GHS. In some special cases, 1 GHM corresponds to several GHSs. Next, GHM, UCD (drug information), LPP (medical device information), and Social Security information, are aggregated and used to generate a GHS code. A pricing model (called T2A) is applied to the GHS group to obtain the final price of health services provided for each hospital stay (€ / hospital). The health institution receives fiscal revenue from the French Health Authority by performing medical activities and reporting them (in the form of PMSI data) to the French Health Authority.

1.1.2. Production of the PMSI data

The description of medical activities within the PMSI framework in the (MCO) sector is based on the systematic collection of minimal and standardized medico-administrative data contained in the standardized discharge summary (RSS) and on the methodical

and partially automated processing of these data. Every hospitalization, with or without accommodation, in the MCO sector of a healthcare institution is subject to an RSS, consisting of one or more medical unit summaries (RUM). An RSS must be produced each time a patient has left the MCO hospitalization sector of a hospital legal entity.

Medical Unit Summary (RUM)

The medical unit summary (RUM) is established at the time of the patient's discharge from a medical unit (UM). Each medical unit corresponds to a homogeneous set of medical activities and is designated as an individualized set of resources providing healthcare services to hospitalized patients. Each medical unit is identified by a specific code in a nomenclature determined by the institution. The RUM contains a limited number of columns that must be systematically filled in. The information provided by a RUM is of an administrative and medical nature:

Administrative information:

1. Patient identification information
   a) RSS number: the identifier (or the hospitalization ID) of hospital stays.
   b) Gender: male = 1, female = 2;
   c) Date of birth (day, month, year);
   d) Postcode: postal code of the place of residence.
2. Other administrative information
   a) RUM number: the identifier of medical units (UM).
   b) Type of medical units: (01) Intensive care unit (ICU), (02) Continuous monitoring unit, ..., etc.
   c) Date of entry in the medical unit (day, month, year);
   d) Date of discharge from the medical unit (in the form of - day, month, year);
   e) Entry and exit modes: Internal transfer (6), Normal transfer (7), Domicile (8), Death (9), ..., etc.
   f) etc.

Medical information:

1. *Diagnostics (coded data): relevant ICD-10-FR codes including 1 principal diagnosis code (DP), 0 - 1 related diagnosis (DR), and 0-N significant associated diagnoses (DAS);*
2. *Medical acts and procedures (coded data): CCAM codes.*
3. etc.

In MCO units, medical coders mainly consider three types of diagnoses. All three types of diagnoses are coded using the ICD coding system: (i) *1 principal diagnosis (DP)*: The DP indicates the reason for admission to a medical unit (UM). It is assigned at the end of each hospitalization; (ii) *0-1 related diagnosis (DR)*: DR is filled only when the DP begins with Z. The DR indicates chronic disease or permanent health condition possibly related to the DP or disease explaining palliative care; (iii) *0-N significant associated diagnoses (DAS)*: Any other diagnosis active during the stay;

Standardized discharge summary (RSS)

Every hospital stay in the MCO units of a public or private health institution must be attached with a RSS, made up of one or more medical unit summaries (RUM) under the control of physicians (in DIM).

If the patient has attended only one medical unit during his stay: (i) the stay is said to be single-unit; (ii) the RSS is strictly equivalent to the RUM produced for this hospital stay; (iii) the RSS then contains only one record (called a "monoRUM" RSS).

If the patient has attended several medical units: (i) the stay is said to be multi-unit; (ii) the RSS corresponds to all the RUMs produced by each of the units attended by the patient during the hospitalization, ordered chronologically by the responsible physician; (iii) the RSS is then made up of a set of RUMs (called a "multiRUM" RSS) which are all identified by the same RSS number.

The RSS constituted contains a set of joined RUMs: (i) All RUMs have the same RSS number; (ii) the entry date of the first RUM in the RSS is the date of entry into the MCO sector; (iii) the discharge date of the last RUM is the date of discharge from the MCO sector; (iv) the entry date of an intermediate RUM is the exit date of the previous RUM; (v) Identity information in RUMs is consistent (gender, postal code, etc.); (vi) *The RSS contains all CCAM codes contained in RUMs*; (vii) *The RSS contains all ICD codes contained in RUMs*; (viii) In the case of multi-unit RSS, one of the RUMs' principal diagnoses (DP) becomes the DP of the RSS (a complex algorithm);

After a de-identification procedure, the RSS becomes the anonymous discharge summary (RSA) that can be released and distributed by the ATIH.

1.1.3. Coding systems adopted in PMSI program

In the context of MCO, the ICD diagnostic is the essential element of the T2A pricing completed by the CCAM classification of medical acts. In the following two subsections, we briefly introduce two relevant coding systems, ICD-10-FR and CCAM.

1.1.3.1. ICD coding system

ICD coding system is an internationally unified method of disease classification developed by the World Health Organization (WHO), which classifies diseases according to their characteristics such as etiology, pathology, clinical situation, and anatomical

location, making them into a systematic hierarchical coding system (Refer to Figure 1.2).

- IV Endocrine, nutritional and metabolic diseases
  - E00-E07 Disorders of thyroid gland
  - E10-E14 Diabetes mellitus
  - E15-E16 Other disorders of glucose regulation and pancreatic internal secretion
  - E20-E35 Disorders of other endocrine glands
  - E40-E46 Malnutrition
    - E40 Kwashiorkor
    - E41 Nutritional marasmus
    - E42 Marasmic kwashiorkor
    - E43 Unspecified severe protein-energy malnutrition
    - E44 Protein-energy malnutrition of moderate and mild degree
      - E44.0 Moderate protein-energy malnutrition
      - E44.1 Mild protein-energy malnutrition
    - E45 Retarded development following protein-energy malnutrition
    - E46 Unspecified protein-energy malnutrition
  - E50-E64 Other nutritional deficiencies

Figure 1.2.: An excerpt from chapter IV of the ICD-10 coding system.

In France, the ICD-10-FR is widely adopted by healthcare institutions for various purposes such as statistical purposes, decision support, and reimbursement purposes. ICD-10-FR contains 22 chapters and approximately 16,000 distinct codes for compactly representing diseases, symptoms, abnormal findings, external causes of injury or diseases, social circumstances, epidemiology, etc. The coding system is divided into hierarchical levels, where each level corresponds to a category or a subcategory. As shown in Figure 1.2, ICD-10 allows the coding staff to encode the diagnosis with strong specificity.

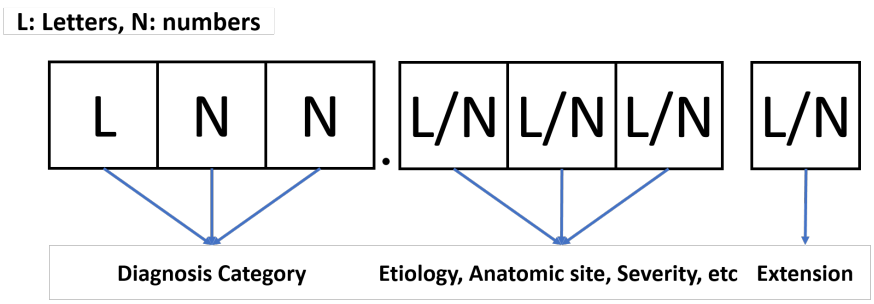


Figure 1.3.: The composition of ICD codes.

In general, ICD codes consists of three to seven characters. The more characters in an ICD code, the more precise the diagnosis. As illustrated in Figure 1.3, the first three characters indicate the diagnosis category; The middle three characters designate the related etiology or other additional diagnostic information. The seventh character extension provides additional information about the visit encounter.

1.1.3.2. CCAM coding system

The Common Classification of Medical Procedures (CCAM) is a French Social Security nomenclature that includes the coding of medical procedures performed by physicians, dental surgeons, and midwives. It is intended to be exhaustive, manageable, and evolving. This taxonomy is used to establish: (i) in public hospitals, the GHS and its T2A pricing of hospital stays transmitted to the health insurance system; (ii) in private clinics, the fees for medical acts and medical procedures performed during consultations;

Table 1.1.: CCAM code: HHFA001 (Appendectomy)

HH.	F	A.	001
action	technical	topography	counter

In its V2 version (i.e., CCAM V2), the CCAM included 7,623 codes. Each code is accompanied by a phrase intended to specify its meaning unequivocally, followed by its tariff in euros. The Table 1.1 shows an example of CCAM code. Each CCAM code is composed of four letters and three numbers: (i) the first letter designates a major anatomical system; (ii) the second letter specifies the organ (or function) in the apparatus corresponding to the first letter; (iii) the third letter designates the action performed; (iv) the fourth letter identifies the approach or technique used; (v) The next three digits are used to differentiate acts with four identical key letters;

1.1.3.3. Advantages and disadvantages of using coded data

The primary purpose of adopting medical coding conventions in MCO departments is to report medical activities performed and get hospital reimbursement [4]. Apart from that, the adoption of medical coding systems has many other merits: (i) *coded data can be processed automatically compared to raw data (which is usually in textual format)*; (ii) the assigned medical codes can be used as a statistical tool for efficient medical resource scheduling; (iii) adopting medical coding conventions is a way to structurally represent patients' medical histories for clinical monitoring, efficient information exchange between physicians in different medical units, and further epidemiological studies;

Structural representation such as ICD-10-FR and CCAM is denser and therefore inherits all the pros (and cons) of a discrete representation. The discrete representations are unified and comparable. For example, the same code can be assigned to all patients

with the same pathology, while there are many ways to describe a pathology for textual representation. In addition, we can compare two diseases by hierarchically measuring the distance between two ICD codes, i.e., 'A05.0' - Staphylococcal food poisoning and 'A05.1' - Botulism (Clostridium botulinum).

However, *coded data loses some specificity*, i.e., disease variants in textual format can be easily distinguished, while the distinction of disease within an ICD code is not always possible since ICD codes only have limited granularity (or length). For example, the ICD-10-FR code 'D43.2' groups several unspecified tumors in the brain and central nervous system. It is worth noting that the ICD coding system is mainly designed for reporting and reimbursement purposes and sometimes can not accurately describe certain diseases and pathologies.

1.1.4. Financing of French healthcare institutions under PMSI program

In France, activity-based pricing (T2A) was launched in 2004 as part of the "Hospital 2007" plan. It aims to allocate health service revenue based on an estimation of the nature and volume of medical activities carried out by healthcare institutions. Since then, T2A has been the sole method of financing medical, surgical, and obstetrical (MCO) activities in public hospitals (i.e., formerly DG, also called, ex-DG[1]) and private clinics (i.e., formerly OQN, also called, ex-OQN[2]).

Homogeneous group of patients (GHM)

GHM is a classification system that classifies hospital stays into one of the clinically defined groups. This system was developed to achieve two sub-objectives: i) grouping hospital stays into diagnosis-related homogeneous groups such that hospital stays within each group are clinically similar; ii) hospital stays in a given GHM group are expected to consume the same level of hospital resources (or medical treatment costs).

GHMs are assigned by a grouping program based on medical information such as ICD diagnostic codes, CCAM medical procedure/act codes, the presence of complications or comorbidities, and administrative data such as age, gender, mode of entry & discharge.

Structure of GHM codes:

A GHM code is composed of six characters that can be broken down as follows: (i) the first two are numeric and indicate the *major category (CM)*; (ii) the third is alphabetical and characterizes the GHM according to the logic of the classification; this character indicates the *type of GHM*; (iii) the fourth and fifth are numeric and are used as a counter to distinguish all GHM codes that have the same first three

[1] ex-DG is the former method for financing public healthcare institutions before the T2A reform.
[2] ex-OQN is the former method for financing private healthcare institutions before the T2A reform.

characters. These characters correspond to the *GHM number in this type*; (iv) the sixth character is alphabetical and indicates the *level of "complexity"* of the GHM, but this can also be numerical when indicating the *level of severity*.

The Table 1.2 shows an example of GHM code **01C031** - craniotomies for trauma, age > 17 years, level 1, where the GHM root (the first three levels or the first five characters) is **01C03_** - craniotomies for trauma, age > 17 years.

Table 1.2.: An example of GHM code

01	C	03	1
CM	Type of GHM	GHM number	complexity or severity
Major category	C: surgical K: non-operative procedure M: medical Z: undifferentiated H: error	counter	1-4: 4 levels of severity A-D: 4 levels of severity Z : not segmented J: ambulatory activity T: very short duration E : with death

Major category (CM), also called major diagnostic category (CMD), is the first level of classification of RSS, which is often determined by the principal diagnosis (DP) in the RSS. Often, each CM corresponds to a functional system, i.e., (i) CMD 01: Nervous system disorders; (ii) CMD 02: Diseases of the eye; (iii) CMD 04: Diseases of the respiratory system. Apart from that, a number of RSS are classified in CM in which the classification is not based on principal diagnosis (DP), as is the case for: (1) CM 27: Organ Transplants; (II) CM 90: Errors and other unclassifiable stays.

GHM grouping:

The GHM root consists of a complex tree structure based entirely on medical knowledge. Figure 1.4 illustrates the tree structure of GHM roots. The assignment of a GHM root is based on the ICD diagnoses codes, CCAM codes, age, mode of discharge, etc. A final level of segmentation (the last character) is determined by LOS, CMAs (i.e., complications or morbidities)[1], age, etc. Most GHM roots are thus broken down into four levels of "severity" numbered 1, 2, 3, and 4 and therefore associated with a tariff. Level 1 corresponds to the "no CMA" level, i.e., without significant severity. Finally, a GHM is labeled with a specific GHS (most often, only one GHS per GHM). In short, a grouping program first directs the stay in one of the 669 GHM roots (2013). Next, a final branching of GHM roots (often in 4 branches) produces 2,588 different GHM.

Therefore, the GHM is a homogeneous group of hospital stays in terms of (i) medical content such as principal diagnosis, medical procedures performed, comorbidities, associated diagnosis, (ii) and administrative information such as gender, age, length of stay (LOS).

[1]Certain CMAs result in a significant increase in cost or length of stay (LOS). Therefore, they should be coded as good as possible.

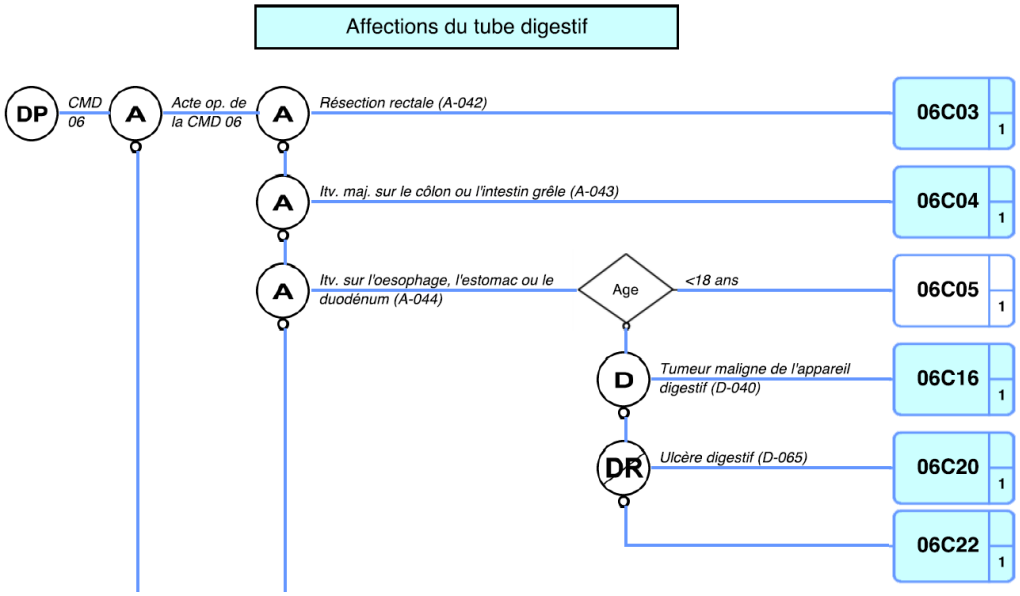


Figure 1.4.: The tree structure for determining GHM root of a hospital stay.

Homogeneous group of stay (GHS)

In the context of T2A, the homogeneous group of stays (GHS) corresponds to the tariff for the homogeneous group of patients (GHM). In France, T2A is used to determine how much the Primary Health Insurance Fund (CPAM) pays to healthcare institutions. Note that T2A does not cover services provided to outpatients, including emergency room visits, outpatient procedures, consultations, etc.

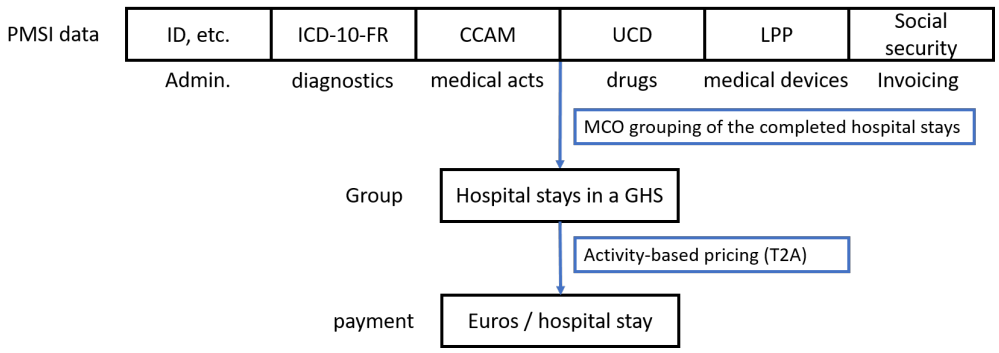


Figure 1.5.: The T2A pricing of a given hospital stay

The vast majority of GHMs correspond to a single GHS, i.e., a single tariff. However, in some cases, a GHM may have two or more tariffs (depending, for example, on different levels of medical resources for the same treatment). GHS groups are assigned based on patients' GHMs, UCDs, LPPs, social security, etc. As illustrated in Figure 1.5, in MCO units, hospital stays in a GHS group share the same health service price

regardless of the length of stay (LOS). Note that the health service price of a hospital stay is evaluated right after the end of a hospitalization.

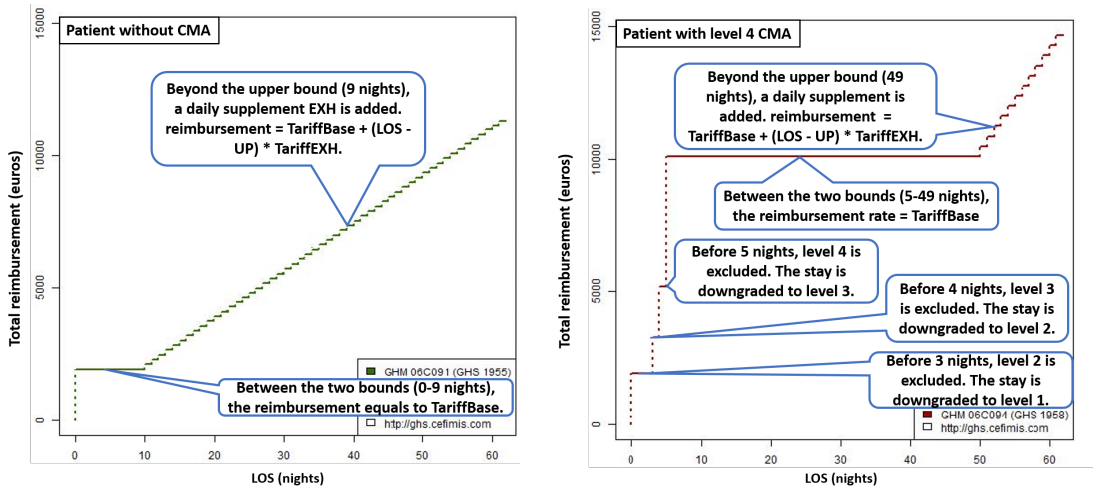
From GHM to GHS

The table 1.3 is an excerpt of GHS codes from a public hospital, in which **TariffEXH** (EXH=extremely high, extrêmement haut in French) is a daily supplement for each day of hospitalization beyond the upper bound (UB). Note that there is no lower bound on these GHMs in this example.

Table 1.3.: An example of GHS codes

GHS	GHM	Description	Upper bound	TariffBase (€)	TariffEXH (€)
1955	06C091	App. non compliquees, niveau 1	9	1930.13	180.78
1956	06C092	App. non compliquees, niveau 1	15	3291.58	157.71
1957	06C093	App. non compliquees, niveau 1	27	5184.11	222.81
1958	06C094	App. non compliquees, niveau 1	49	10086.00	382.54

Based on the information provided by Table 1.3, Figure 1.6 (a) shows the change in reimbursement rate of a hospital stay (with GHM 06C091) with the increase of his or her length of stay (LOS). A more complex hospital stay with GHM 06C094 and a CMA (i.e.., complications and morbidities) with a severity level of 4 is also provided and discussed in Figure 1.6 (b). Apart from these intuitive examples, in the next subsection, we introduce and explain the method for computing the GHS tariff and reimbursement rate.



(a) Tariff for the GHM code 06C091 without CMA.

(b) Tariff for the GHM code 06C094 with a level 4 CMA.

Figure 1.6.: The reimbursement rate for a patient without CMA (a) and a patient with a level 4 CMA (b).

From GHM to GHS and tariff

Table 1.4 is an example extracted from a public hospital on which the computation of reimbursement rate is based. The LB and UB are abbreviations of Lower Bound and Upper Bound, respectively. The general objective of the **ForfaitEXB** (EXB=extremely low, extrêmement bas in French) is to downgrade the reimbursement rate to a lower price. The general objective of **TariffEXB** is to compute the reimbursement rate according to the length of stay (LOS).

Table 1.4.: From GHM to GHS and tariff

GHM	Description	LB	UB	TariffBae (€)	ForfaitEXB (€)	TariffEXB (€)	TariffEXH (€)
06C081	Appendicectomies compliqueés, niveau 1	3	13	3166	-	655	176
06C082	The same as above, niveau 2	4	21	4577	1411	-	247
06C083	The same as above, niveau 3	5	30	6802	2225	-	261
06C084	The same as above, niveau 4	-	40	9410	-	-	441

In general, the final reimbursement rate of a hospital stay can be calculated using the following method:

- If LB ≤ LOS ≤ UB: Tariff = TariffBase;
- If LOS > UB: Tariff = TariffBase + (LOS - UB) * TariffEXH;
- If LOS < LB:
  - If death: Tariff = TariffBase;
  - If "Forfait EXB" is specified : Tariff = TariffBase - ForfaitEXB;
  - If "Tariff EXB" is specified: Tariff = TariffBase - (LB - LOS) * TariffEXB;

1.2. The organization of hospital coding tasks

In France, the hospital coding task can be defined as the process of assigning medical codes to an inpatient stay according to some coding conventions and coding guidelines, i.e., the coding guideline for ICD-10-FR [5], the unofficial reference book CoCoA [6], etc. In this process, diagnoses, medical acts, complications, morbidities, and more are abstracted from patient histories and are translated into medical codes. The following example shows a complete ICD coding process: (i) Mr. Dominique comes for initial care of prostate cancer, the responsible physician assigns relevant medical codes to the patient, i.e., DP = C61 - malignant prostate tumor; (ii) Mr. Dominique

returns for radiation treatment of his prostate cancer, the responsible physician assign relevant medical codes to the patient, i.e., DP = Z51.01 - radiation session, DR = C61 - malignant prostate tumor; (iii) Mr. Dominique returns for a post-surgical checkup, the responsible physician assign relevant medical codes to the patient, DP = Z08.0 - post-surgical checkup of malignancy, DR = C61 - malignant prostate tumor; (iv) Mr. Dominique returns for appendicitis, the responsible physician assign relevant medical codes to the patient, DP = K35.8 - acute appendicitis, other and unspecified, DAS = C61 - malignant tumor of the prostate.

1.2.1. Hospital coding organization in practice

With the increase in the type and volume of medical activities, healthcare operation and management is becoming more and more dependent on coded data. In addition to financial and reimbursement purposes, medical codes are also increasingly being used as benchmark data to assess the efficiency of medical units and even individual coding practitioners. This also increases the need to make medical codes more accurate for each hospital stay.

Given a set of EHRs of inpatient hospitalizations, the coding tasks are mainly performed by hand by the physicians or a specialized coding team (making the task *subjective, error-prone, time-consuming*) and require high technical and medical expertise (making the task *complex, expertise-required, labor-intensive*). The coding task is complex since it is difficult to screen a set of relevant codes that can approximately describe the corresponding disease or pathology, i.e., (i) inappropriate medical terms are selected for medical code search, i.e., often, the textual description in EHRs contain plenty of medical terminologies, and each medical coder has his or her own interpretation of them; (ii) sometimes there is no code in the coding system that can precisely describe the disease (i.e., some rare diseases); (iii) sometimes too many codes can describe the same disease (etiological description, physiopathological description, clinical description, different levels of precision, etc.).

Apart from that, the selection of the appropriate code from the candidate set is even more difficult since: (i) it is difficult to sort the set of codes found in the coding system since the selection of DP, DAS, or DR is to some extent subjective; (ii) keep in mind that medical coders can not fully remember all the coding rules. In other words, medical coders are not familiar with the coding rules of uncommon codes.

Therefore, by making the coding staff aware of these coding issues caused by medical terminology, poor coding habits, or even missing data in patient EHRs that could be easily found in health information systems, a part of the problem can be solved. One possible solution is to implement a *double-coding process*, where two medical coders assign medical codes to the same hospital stay, and the coding agreement between two coders is reached. After that, the quality control process is conducted and is considered the last line of defense to control the miscoding rate of hospitals. The discussion about the accuracy of assigned medical codes certainly makes the task of

medical coding more complex, as this also introduces doubt about the quality of the manually assigned medical codes.

The accuracy of manually assigned medical codes is an enduring hot topic and largely depends on hospital departments, and individual medical coders [7, 8]. Pieces of literature on the accuracy of human medical coding in the surgery department show that 51% - 93% of patient stays have at least one code change when assessing the assigned codes [4, 9, 10]. The largest study case [4], with a sample size close to our study, reviewed 30,127 patient stays and found 13% of primary diagnoses and 12% of medical procedures are miscoded. **Such a survey shows that there is certainly much more room for improving the coding practice of coding practitioners**.

1.2.2. The current coding practice in the University Hospital of Saint Etienne

The University Hospital of Saint-Étienne (CHU-SE) is composed of 3 inpatient centers situated in three different locations across Saint-Étienne, including North Hospital, Bellevue Hospital, and Charité Hospital. As the largest hospital in the Loire region, France, the CHU-SE has a total of 1,549 inpatient beds[1] in 68 distinct medical service units. It is also the support (or core) establishment of a regional healthcare cooperation network[2] around the Loire region.

This research project is in collaboration with SSPIM - the medical information department (DIM) of the CHU-SE. The Figure 1.7 illustrates the organizational structure of the SSPIM. Specifically, the SSPIM has about ten professional medical coders (TIMs), 8 of them are responsible for medical coding tasks in different medical units, and 2 of them are responsible for Quality Control (QC) of the assigned medical codes. In addition, two physicians from the CHU-SE (2 PH (hospital practitioners) in MCO units) are responsible for part of the coding task in MCO units. 2 TSH are in charge of the final quality control of assigned code and the hospital reimbursement based on T2A, respectively. In total, **our research influences 14 working staff** in SSPIM, CHU-SE. In SSPIM, the coding staff has access to patient histories (i.e., EHRs, PMSI data, etc.) of inpatients in the CHU-SE.

Patient histories contain health information of various modalities (i.e., structured tabular data, unstructured textual data, medical images, etc.) related to hospital stays and are widely used in various healthcare applications. With the migration to the new health information system "Easily", huge amounts of digitized patient histories are becoming more accessible for the working staff in SSPIM. The "Easily" is designed by the University Hospital of Lyon[3], and gradually replaced the previously used web platform "Cristal-Net" in the CHU-SE since Oct 2016.

[1] Please refer to the Medical activity report of the CHU-SE (2021) for more details

[2] GHT-Loire: Le Groupement Hospitalier de Territoire de la Loire, in French, The aim of constructing such a cooperation network is to pool and share healthcare resources and facilitate the coordination of health resources between different agencies inside the Loire region.

[3] University Hospital of Lyon: Hospices Civils de Lyon - HCL, Rhône, in French

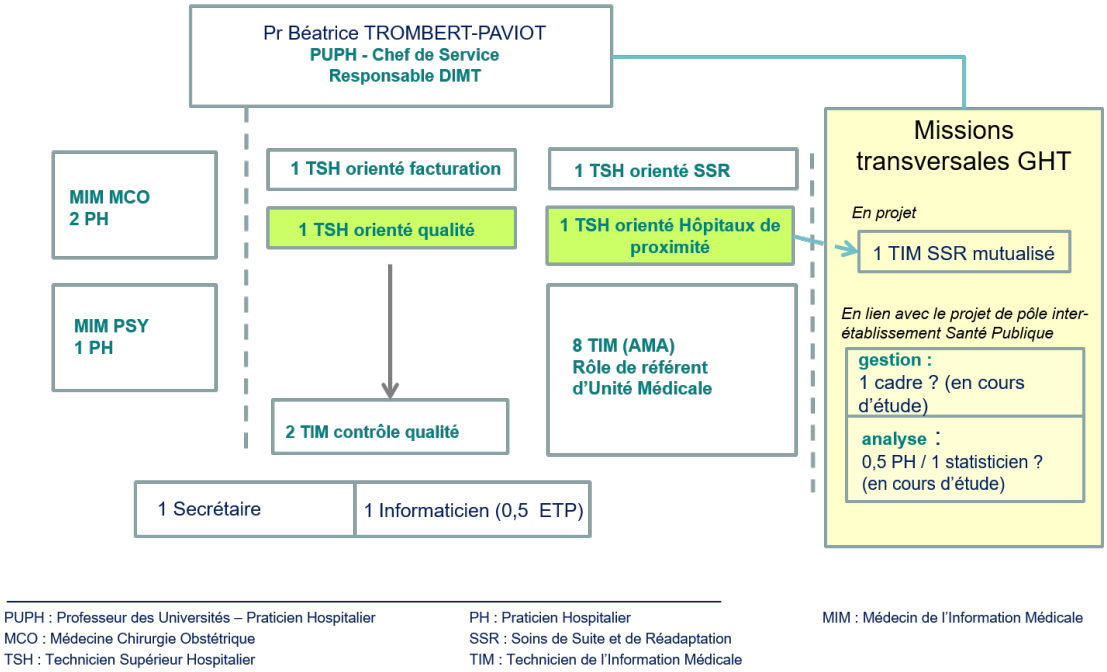


Figure 1.7.: Organizational chart of the SSPIM.

The existing French-based coding assistants include an internal keyword search tool inside the "Easily" platform and a publicly available keyword search tool - Aide au codage[1]. These tools search for related medical codes through keyword matching. These tools slightly reduce the workload of practitioners. Nevertheless, the complexity and intensity of the coding task still result in a non-negligible miscoding rate. Taking the CHU-SE as the research object, **the estimated miscoding rate is between 10% to 15%**, which includes errors in coding rule application, insufficient information in patient histories, etc.

1.3. Hospital miscoding

In the context of the PMSI program, coding errors are mainly composed of two parts, including subjective and objective errors. *Objective errors are mainly caused by various objective reasons*, such as failure to identify the relevant history of patients (i.e., cross-indexing of data in multiple databases) and quality of documented patient information (i.e., inaccurate or insufficient data in patient EHRs), while *subjective errors are mainly caused by poor coding habits, or poor coding experience of medical coders*, which include omission of codes for CMAs (complications or morbidities), the coder's knowledge with the disease (i.e., incorrect interpretation of coding rules, misjudgment of causality), the coders' attention to the coding task (i.e., the excessive workload

[1]Aide au codage: https://www.aideaucodage.fr

will lead to lower coding quality, ambiguous medical codes with low specificity are selected), medical coders do not keep up with the latest version of coding guidelines (i.e., insufficient coding training and information exchange, variation in coding standards between institutions, variation in coding standards over time), handover of coding work between coders and physicians (i.e., disagreement on nomenclature and abbreviation.), etc.

Main error sources along the 'paperwork' include communication among patients and physicians, variance in the hand-written and electronic health records, coding training and coding experience, unintentional coding errors such as unbundling and misspecification, intentional coding errors such as upcoding or over-coding (to obtain more health service reimbursement).

In addition, it is worth noting that *the coding rules are often complex and come from multiple sources*: (i) methodological guide for the field concerned (often used for PMSI or reimbursement purpose, i.e., the CocoA guideline [6]); (ii) coding guides to clinical situations (often used for the clinical purpose, i.e., coding guidelines issued by HAS); (iii) jurisprudence or case law: assessment of ATIH referrals during external audits by the Assurance Maladie (CPAM) (often used for handling special cases). Apart from that, in MCO units, instead of assigning one code to each patient stay, medical coders have to select 1 DP, 0-1 DR, and 0-N DAS for each hospitalization, which makes the coding task more complex.

1.3.1. Types of hospital miscoding

Among all those coding errors, in this subsection, we only introduce several typical errors in hospital coding. We did not exhaustively enumerate every possible type of coding error. However, these examples are sufficient to justify the complexity and error-prone nature of hospital coding tasks.

Examples of miscoding subtypes

Miscoding subtype I

A medical coder tries to assign an ICD code to a patient with "pneumopathie à Streptococcus pneumoniae". The word "pneumopathie" is searched in the coding system. In the table 1.5, we shows the set of relevant codes found. Therefore, the medical coder selects the code J154 from the list since the code J154 contains the term "pneumopathies bactériennes" and "à streptocoque".

However, J154 is not the correct code since it is customary to call "pneumopathies à S. pneumoniae" as "pneumonies" (although it is "pneumopathies"). Therefore, we have to search for the word "pneumonie", and we obtain the following:

Miscoding subtype II

The same patient now has meningitis with the same bacteria. So let's search for the word "Streptococcus pneumonia", and we get the following set of codes:

Table 1.5.: The inappropriate codes under the situation I

J14	Pneumopathie due à Haemophilus influenzae
J150	Pneumopathie due à Klebsiella pneumoniae
J151	Pneumopathie due à Pseudomonas
J152	Pneumopathie due à des staphylocoques
J153	Pneumopathie due à des streptocoques, groupe B
J154	Pneumopathie due à d'autres streptocoques
J155	Pneumopathie due à Escherichia coli
J156	Pneumopathie due à d'autres bactéries aérobies à Gram négatif
J690	Pneumopathie due à des aliments et des vomissements
J851	Abcès du poumon avec pneumopathie

Table 1.6.: The relevant codes for situation 1

J13	Pneumonie due à Streptococcus pneumoniae

Table 1.7.: The inappropriate codes under situation 2

A403	Septicémie à Streptococcus pneumoniae
B953	Streptococcus pneumoniae, cause de maladies classées dans d'autres chapitres
J13	Pneumonie due à Streptococcus pneumoniae

The medical coder may select the code B953. However, B953 is not the correct code since the bacteria " Streptococcus pneumoniae" can also be called "Pneumocoque". We search for the word "Pneumocoque" and we obtain relevant codes in table 1.8. In the end, we retain the code G001 for the patient.

Medical codes are often very similar. Nonprofessional persons would not be able to distinguish one code from another. These examples show that the medical coding task is error-prone and, to a certain extent, subjective.

Summary

Given the complexity of modern medical diagnoses and medical acts, it is not surprising that hospital miscoding is common and hard to avoid and prevent. In addition, regional or demographic variations in miscoding rates have also been observed within a country, which reflects differences in coding habits in different healthcare institutions. For instance, Lorence and Chen [11] found a wide variation in miscoding rate across the USA. Moje et al. [12] revealed that the depth of hospital coding varies across healthcare institutions in Victoria. Besides, Pervez et al. [13] compared outcomes of two databases and reported that using ambiguous medical codes was the primary source of hospital miscoding.

Table 1.8.: The relevant codes for situation 2

G001	Méningite à pneumocoques
M001	Arthrite et polyarthrite à pneumocoques
M0010	(...) - Sièges multiples
M0011	(...) – Articulations acromio-claviculaire, scapulo-humérale, et sterno-claviculaire
M0012	(...) - Articulation du coude
M0013	(...) - Articulation du poignet
M0014	(...) - Articulations de la main
M0015	(...) - Articulations de la hanche et sacro-iliaque
M0016	(...) - Articulation du genou
M0017	(...) - Articulations de la cheville et du pied
M0018	- Autres articulations
M0019	(...) - Siège non précisé

Obviously, regular monitoring and evaluation of coding accuracy are necessary to maintain the operational efficiency of healthcare institutions and the quality of the coded data used for various purposes. In addition to that, the recognition of the evidence of errors in medical code assignments would help medical coders to improve their coding habits and coding practices. However, before these medical codes can be used as a quality measure of a medical specialty (or medical unit), the coding events (i.e., diagnostic coding or coding of medical procedures) must be defined in advance.

1.3.2. Consequence of hospital miscoding

Data-driven health services provision is becoming increasingly dependent on coded health data [14, 9]. The applications based on coded hospital data drive the need for accurate and reliable hospital coding and for the development of automated tools to identify coding errors. Specifically, the accuracy and completeness of assigned medical codes affect hospital operational efficiency, reimbursement rate, and other areas [14]. This increases the need to have accurate and complete medical codes for patient stays.

For example, the appropriate funding of health services via the activity-based funding system (i.e., T2A) is dependent on the reliable grouping of medical diagnosis and procedures into homogeneous groups of patients (i.e., GHM), which relies on accurate medical coding and in turn depends on the quality of physicians' documentation. Health information and patient histories (including coded data), which drive medical and economic research and the development of health management regulations, are extracted and used under the assumption that they are the gold standard and accurate. Monitoring of health information provided by patient histories, for example, the nature and incidence of complications or morbidities (CMAs), is crucial to patient health and safety, clinical risk management, the conduct of health prevention programs,

health service quality assessment, and the evaluation of hospital performance.

Funding of health services

In France, the accuracy of hospital coding is critical to the proper funding of medical activities performed and health services provided by healthcare institutions under the T2A funding mechanism. A study by Reid et al. [15] is designed to verify whether all supplementary diagnoses are recorded and correctly coded in a Sydney hospital. A 'gold standard' medical coder reviews and corrects miscoded cases where the re-coded codes are not consistent with the original codes. In this study, the under-coding of CMAs (complications or comorbidities) is found to be more frequent than over-coding, which results in an under-funding of health services. Marshall et al. [16] found that the overall miscoding rate in a healthcare institution in Florida is about 84.5%, resulting in a significant loss of health services reimbursement.

Although *the primary concern of this thesis is the effect of hospital miscoding on the funding of healthcare institutions*, the following subsections present other potential negative impacts caused by hospital miscoding.

Hospital operational efficiency and league tables

In French, coded patient discharge summaries (RSSs) are pillars of the national strategy (i.e., PMSI program) designed to reduce inequalities in healthcare resources between healthcare institutions and improve the quality of healthcare services [17]. Errors in medical coding are often common and multifactorial. Santos et al. [18] investigate various organizational factors affecting the quality of medical coding. A substantive statistical association is found between structural characteristics of healthcare intuitions and medical coding error rates in this study. A study in Taiwan [19] audits death certificates following the guideline of the ICD-9-CM and shows different levels of coding accuracy for different diseases between original coders and reviewers.

Ballaro et al. [20] conducted a coding audit in a urological clinic in the United States of America and reported a high miscoding rate by some urology physician trainees. The authors also claim that "the medical codes assigned by them" do not reflect the actual clinical practice. On top of that, they declare that league tables produced based on routinely collected clinical data are not totally reliable. In addition, Jameson et al. [21] also emphasize the importance of coding accuracy since the coded data are considered the cornerstone for the creation of league tables in the United Kingdom. In addition, assigned medical codes are often used as benchmark data to assess the efficiency of specialties in hospitals and even individual medical coders [14, 9, 7, 4].

Epidemiology, disease surveillance, and prevention

Patient histories (including coded health data) are a key element for designing epidemiological surveillance programs, which are therefore used in the design of epidemic

prevention programs and the prescription of medical exams. In addition, a number of public health programs are based on standardized health information. For example, Schoenman et al. [22] presents a series of healthcare applications using patient histories in the United States: (i) epidemiological surveillance and disease prevention; (ii) studies on the economic burden of diseases; (iii) reports on public health service provision; and (iv) environmental health, etc. They state that these health data are used by various users such as government agencies, healthcare institutions, individual health service providers, patients, insurance companies, policymakers, researchers, etc.

Moje et al. [12] conducted research on coding accuracy in Victorian health institutions in Australia. Nelson et al. [23] also emphasize the importance of identifying and correcting coding errors for providing coded data to the development of health applications such as disease surveillance, reporting the effectiveness of vaccination, and disease prevention programs.

1.4. Hospital code audits

1.4.1. State-of-the-art solutions

With the wide adoption of EHRs and various coding conventions in healthcare institutions, computer-assisted medical coding tools are in high demand. In this section, we sort the task into two categories: code recommendation and code audit. The code recommendation task is defined as a multi-label classification process that translates available health information into given coding taxonomies (i.e., ICD-10-FR, CCAM, UCD, LPP), while the code audit is a miscoding detection and correction process that tracks coding errors occurring in a dataset and corrects them.

Applications related to medical code recommendation

A comprehensive survey on medical coding automation [24] reveals an emerging trend toward applying AI methodologies, especially natural language processing (NLP) and deep learning (DL) techniques, to the medical code recommendation task since 2009. However, DL techniques adopt neural networks, perform non-causal reasoning, and are likely to extract underlying false patterns [25]. Another comprehensive survey [26] reveals that DL-based hospital coding models are usually not generalizable over different datasets, which suggests practitioners dynamically adjust and update models to react to local variations in data format.

To deal with such a drawback, the systematic review [27] shows multiple enhancement techniques to improve the interpretability of DL-based techniques, including knowledge distillation, dimensionality reduction, attention mechanism, and feature interaction and importance. In addition, the attention mechanism has been proposed to improve the transparency of DL models by locating related snippets from clinical notes for a specific code [28]. Nevertheless, as the DL techniques are not explainable

at the modular level, we still cannot locate and remove underlying incorrect patterns. To date, these DL-based techniques cannot provide sufficiently accurate codes and corresponding evidence and thus cannot be used alone without the intervention of human coders [29].

The intrinsically interpretable techniques are mainly less-performant statistical learning methods (i.e., Decision Tree, Generalized Additive Models (GAM), etc. [30]) coupled with performance enhancement techniques, which are rarely employed in the code recommendation task due to the intrinsic complexity of NLP-based tasks. Code recommendation techniques are on areas of interest but do not fit our objective – coding practice improvement. In the next part, we will discuss relevant techniques for the code audit task.

Applications related to medical code audit

Often, the patient stays' EHRs contain plenty of medical terminologies, and each medical coder has his or her own interpretation of the EHR of a patient stay. These facts show that the medical coding task is error-prone and subjective. The proposed solution is to implement the proposed automated audit tools in this thesis to eliminate coding errors.

Someone would argue that creating an accurate code recommendation system is an ideal way to avoid the miscoding problem. However, in that case, we have to first guarantee that all assigned medical codes are the gold standard and train models on the annotated data for code prediction. However, real-life data often contain incorrect medical codes, and incorrect patterns are possibly learned from the annotated codes. In any case, **the miscoding audit is an inevitable task and should be prioritized.**

1.4.2. Contribution of this study

In most cases, audits of medical codes are mainly performed by hand by human coders [7, 8, 31]. Lusignan et al. [32, 33] introduce a semi-automated miscoding audit tool[1] for the identification and correction of hospital miscoding in diabetes. The intrinsic complexity of coding conventions and the subjectivity of individual medical coders make the audit task labor-intensive and time-consuming. To the best of our knowledge, most existing data-driven methodologies focus on the medical code recommendation task, while **only a few solutions are proposed to automate the code audit process** [32, 33]. The construction of the concept of miscoding behaviors based on profiles of miscoded patient stays to personalize the medical review rationing has not been considered in previous studies.

Therefore, **the techniques proposed in this thesis are implemented to fill this gap in data-driven solutions for the hospital miscoding problem, accounting for the intrinsic complexity of miscoding behaviors.** In addition, in the context of the

[1]This website (https://clininf.eu/cod) provides a miscoding audit tool as a web service without technical details.

hospital miscoding problem, no relevant applications have been proposed to explore the causes of miscoding. In this thesis, some extracted patterns and knowledge are provided to reveal potential causes of miscoding.

1.5. Introduction to the thesis

1.5.1. Scientific objectives

The primary objective of this thesis is to develop methodologies based on optimization, machine learning, and data mining techniques, which will be used in the context of the PMSI program for the hospital miscoding problem. The set of approaches proposed will allow us to identify hospital miscoding behaviors, analyze them, model them, and correct them in order to improve the operational efficiency of the CHU-SE. Since the hospital miscoding problem is in the healthcare domain, this thesis puts special emphasis on model interpretability. Transparency is critical for healthcare applications in order to guarantee fairness and discuss it with medical coders and stakeholders.

More specifically, the scientific objective can be divided into several sub-objectives, each one is formulated as a research question:

- **How to appropriately profile hospital miscoding behaviors for optimizing miscoding review budget?** The data is extracted from PMSI national database, including administrative and medical information in various formats. Taking the heterogeneity of the data into account, two axes are explored in this manuscript: (i) the profiling of miscoding behaviors or the design of profiling rule (i.e., searching for an appropriate representation of hospital miscoding behaviors), (ii) the modeling of uncertainty in hospital miscoding, (i.e., the miscoding risk of a given hospital stay).
- **How to optimize the miscoding review budget from the analysis of complex miscoding behaviors while maintaining fairness and transparency?** In this work, we proposed various algorithms from optimization, machine learning, and data mining fields to the miscoding review rationing problem while taking the specificity of hospital stays into account.

1.5.2. Methodology overview

Three components are critical to the hospital miscoding correction problem: miscoding profiling, miscoding review budget optimization, and miscoding explanation. The remainder of this manuscript is organized as follows:

A general introduction is first given at the beginning of this thesis. In the general introduction part, we briefly introduce the hospital miscoding problem, the necessity to address the problem (i.e., existing research gap or research opportunities), and the general objective of this thesis.

Chapter 1 presents the research context and scientific objectives of this thesis. This chapter first introduces the PMSI program (i.e., French Information Systems Medicalization Program), various hospital coding tasks, as well as the hospital miscoding problem under the PMSI framework. On top of that, the activity-based pricing system (T2A) is introduced, and the financial impact of miscoding is discussed. The primary research context is presented. In the second part, we discuss the motivation and the necessity of this research project. Finally, we introduce the scientific objectives of this work, which will also be discussed in the general conclusion part of the thesis.

Chapter 2 proposes a novel methodology for identifying underlying miscoding subtypes from PMSI data. Our methodology focus on (i) the identification of miscoding subtypes, (ii) the adoption of statistical techniques for validating identified miscoding subtypes, and (iii) the application of a novel optimization model for optimizing the medical review budget. The numerical results on a real-life dataset show that this approach is able to understand and identify significant miscoding subtypes with high precision.

Chapter 3 introduces a novel approach for modeling the uncertainty of hospital miscoding. The contribution of the Chapter 3 is muti-fold. First of all, we make use of a clustering algorithm to split the whole population into homogeneous subgroups, each of which represents a miscoding subtype. Next, we proposed Bayesian inference to estimate the risk of miscoding based on the descriptive characteristics of hospital stays. Each subgroup is associated with a miscoding risk. Finally, we assign medical reviews to hospital stays in high-risk subgroups, while hospital stays in risk-free subgroups are excluded. The application of the proposed approach on a real-life dataset shows a promising result and reveals the success in miscoding risk modeling.

However, the miscoding profiling approaches proposed in Chapter 2 (i.e., topological space projection) and Chapter 3 (i.e., clustering of patient stays) face an obvious limitation - the discovered profiling rules are not precise enough for miscoding review budget optimization. In Chapter 2 and Chapter 3, clustering-based profiling techniques are used to dissect miscoding subtypes present in a dataset. In this case, hospital stays in a subgroup are similar in terms of patient characteristics (i.e., descriptive features of patients). However, this type of profiling technique (i.e., clustering-based profiling algorithms) is not appropriate to describe the miscoding behaviors of medical coders. In other words, the concept of miscoding subtype is not equivalent to the concept of miscoding behavior.

In Chapter 4, we propose a new profiling rule called correction set (or feature combination). Each correction set represents a miscoding behavior. A coding error can be eliminated by removing features in a given correction set from the subject's feature vector and recoding the corresponding subject. In this way, the profiling rule is well-defined. In Chapter 4, a clustering algorithm is first adopted to identify miscoding subtypes. Each miscoding subtype (i.e., miscoding subgroup) is characterized by multiple miscoding behaviors (i.e., correction sets). An optimization model is then proposed to automate the hospital miscoding correction task based on the concepts

of miscoding subtype and miscoding behavior. A two-stage clustering-based optimization approach is proposed for the hospital miscoding correction problem. We also provide additional counterfactual explanations that tell medical coders what and how to change patient data of miscoded cases (i.e., descriptive features of patients) to get the correct medical codes. The numerical results on a real-life dataset show that the proposed approach is efficient and is able to identify and correct 100% of coding errors with only 36% of patient descriptive features reviewed (please refer to experimental results in Table 4.4 for more details).

In addition to the concept of correction set proposed in the previous chapter, an essential concept, called hypercube, is proposed in Chapter 5 to refine the miscoding profiling approach. On top of that, a completely novel optimization program is proposed to adapt to the improved profiling approach and takes some new suggestions from stakeholders in the CHU-SE into account. The concepts of histogram statistics and error distribution are proposed for the miscoding explanation. Finally, an integrated optimization approach is proposed to tackle the miscoding correction problem, which can jointly explore miscoding behaviors and optimize the miscoding correction budget.

Chapter 6 conducts a retrospective evaluation of the approach proposed in the previous chapter. The numerical results show that the proposed approach is promising and can be implemented in the hospital for future use. The proposed approach is efficient and is able to correct 100% of coding errors with only 24% of patient descriptive features reviewed (please refer to experimental results in Table 6.6 for more details). We also evaluate the economic impacts of the integrated optimization approach. The experimental results show that the proposed approach is able to efficiently increase health services reimbursement of hospitals and also prevent potential finical penalties from the French Health Authority. A method evaluation protocol involving the EMSE and SSPIM is also provided for practical implementation.

At the end of this thesis, a general conclusion is presented, and the contribution of the proposed approaches is presented. Some possible future research directions are also discussed.

Part I.

Profiling hospital miscoding behaviors

2. A topological analysis of hospital miscoding

Summary

Abstract

This chapter applies topological data analysis (TDA) techniques to investigate the nature of complex high-dimensional data by extracting global shape information (patterns) and gaining novel insights from them. The objective is to characterize miscoding subtypes, identify patterns for miscoding subtypes, and select specific groups of subjects for which the electronic health records (EHRs) are worth giving an additional review. Our method combines a TDA technique and an optimization-based model to provide a geometric representation of interrelated hospital stays while permitting the audit of miscoded subjects by preferentially selecting subgroups with more coding errors. Through the proposed method[1], we successfully identified and validated multiple miscoding subtypes that traditional methodologies fail to find. Furthermore, with only 20% of the subjects reviewed, the proposed approach reduces coding errors by 64% of the whole population. Experimental results indicate that the proposed method is promising and can reduce coding errors efficiently, thereby eliminating the negative impacts caused by hospital miscoding.

Keywords: Hospital miscoding, topological data analysis, budget optimization.

[1]This chapter is based on our previous work published in the proceedings of IEEE 17th International Conference on Automation Science and Engineering (CASE) as "A topological and optimization based methodology to identify and correct ICD miscoding behaviors" by Chen HE, Benjamin DALMAS, Cedric BOUSQUET, Beatrice TROMBERT-PAVIOT, and Xiaolan XIE [34].

Résumé du chapitre

Ce chapitre applique des techniques d'analyse topologique des données (ATD) pour étudier la nature complexe des jeux de données en haute dimension via l'extraction d'informations de forme globale (modèles) et l'obtention de nouvelles informations à partir de celles-ci. L'objectif est de caractériser les sous-types de mauvais codage, d'identifier les motifs caractérisant ces sous-types de mauvais codage et de sélectionner des groupes spécifiques de sujets pour lesquels les dossiers médicaux électroniques (DME) méritent un examen supplémentaire. Notre méthode combine une technique d'ATD et un modèle d'optimisation pour fournir une représentation géométrique des séjours hospitaliers interdépendants tout en permettant l'audit des sujets mal codés en sélectionnant en priorité les sous-groupes présentant le plus d'erreurs de codage. Grâce à la méthode proposée[1], nous avons réussi à identifier et à valider de multiples sous-types distincts de mauvais codage hospitalier que les méthodologies traditionnelles ne parviennent pas à trouver. De plus, avec seulement 20% des sujets examinés, l'approche proposée réduit les erreurs de codage de 64% de l'ensemble de la population. Les résultats expérimentaux indiquent que la méthode proposée est prometteuse et peut réduire efficacement les erreurs de codage, limitant ainsi leurs impacts négatifs.

Mots-clés: Mauvais codage hospitalier, analyse des données topologiques, optimisation du budget.

[1] Ce chapitre est basé sur notre travail précédent publié dans les actes de la 17e conférence internationale de l'IEEE sur la science et l'ingénierie de l'automatisation (CASE) sous le titre "A topological and optimization based methodology to identify and correct ICD miscoding behaviors" par Chen HE, Benjamin DALMAS, Cedric BOUSQUET, Beatrice TROMBERT-PAVIOT et Xiaolan XIE [34].

2.1. Introduction

Hospital coding is traditionally defined as the process of assigning standard codes to each inpatient stay according to conventional coding taxonomies. The coding process heavily relies on healthcare professionals and well-trained coding staff to encode relevant information from structured tabular data as well as unstructured clinical notes to sophisticated coding taxonomies, i.e., ICD-10-FR, CCAM, UCD, LPP. Consequently, coding practices are often subjective, time-consuming, error-prone, and even inconsistent between well-trained human coders. Increased attention to coding errors has occurred as a result of the widespread adoption of coding systems for computing statistics in epidemiology (i.e., morbidity, mortality, etc.), designing medical resource allocation programs, making medical reimbursement decisions, and other purposes [35]. Even though the consequences of miscoding rarely fall directly on individuals, it does have many negative effects on healthcare institutions, including (i) an inaccuracy in reporting epidemiological statistics, (ii) a disruption in medical resource scheduling, (iii) a deviation of healthcare reimbursement, (iv) an increase in potentially preventable costs.

To solve such a problem, most methods currently being adopted aim to automate the hospital coding task by mining large healthcare databases. Although numerous approaches have been developed in this direction [26], their objective is mainly to provide a list of potential codes for a specific hospital stay regardless of the interpretability of models. However, in order for the coding staff to improve coding practices and fully benefit from the insights provided by such models, we emphasize the need to identify the causes of miscoding first. The problem thus drifts from a miscoding prediction task to a comprehensive analysis of miscoding subtypes. Our work aims at providing an independent perspective to the hospital coding task, where the goal is to identify miscoding subtypes and adjust the current coding practices.

Most methods currently adopted for comprehensive analysis of potential high-dimensional datasets run as a mechanism of hypothesis verification and therefore depend upon researchers to formulate appropriate hypotheses based on relevant domain knowledge. However, the hypothesis space for many complex datasets is quite large, which makes the task of proposing appropriate hypotheses quite tricky and challenging. In this chapter, we deal with the question of recognizing the 'shape' of high-dimensional health data in order to identify subspaces prone to hospital miscoding without formulating a hypothesis beforehand.

The fundamental mathematical challenges ultimately come from the difficulty in understanding the inherent shape of high-dimensional health data. First, the high dimensionality nature of health data leads to mathematical difficulties in recognizing its geometric shapes. Also, health statuses are inherently highly variable and dynamic from one to another, and the notion of health similarity is less rigid. Therefore, data-analytic methods for tracking hospital miscoding are required to identify shape characteristics in health data while maintaining its robustness to noise and changes by distance scaling. These constraints lead us to explore methods from the branch

area of mathematics called topology, which studies geometric structures that are not rigid. Originated from the study of Leonard Euler [36], topology has gradually begun to be introduced for studying shape characteristics of large-scale and high-dimensional data and dealing with qualitative geometric structures [37]. This study field is often referred to as Topological Data Analysis (TDA).

In this study, we leveraged the Mapper algorithm [38], a TDA technique that has emerged in recent years as a new field whose aim is to uncover and understand the topological and geometric structure of potentially high-dimensional data [39]. The Mapper can identify both coarse-grained and fine-grained geometric information that often cannot be identified by other methods, such as cluster analysis and more distance metric sensitive dimension reduction techniques like multi-dimensional scaling (MDS) [40] and Principal component analysis (PCA) [41]. Cluster analysis generates distinct subgroups, while PCA and MDS generate scatterplots without connectivity information. All these methods cannot capture the shape characteristics of high-dimensional data that Mapper does. Specifically, Mapper has the following features: (i) it constructs a network topology in a coordinate-free manner, which suggests that Mapper can be applied to any situation in which a distance function is defined, not only in Euclidean space; (ii) Mapper is robust to noise and small deformations, which means that topologically, a circle, a triangle, and a hexagon are all identical. By stretching one of these shapes, one can definitely get any of the others; (iii) The generated network topology is a compressed representation of data shapes and has a multiresolution form. The multiresolution nature enables Mapper to distinguish between actual patterns and artifacts.

Recognizing shape characteristics or patterns is crucial to uncovering novel insights in the data and discovering meaningful subspaces. The objective of TDA techniques is to extract the "signature" of the data, which will then be interpreted to identify hidden patterns. In the Mapper, the signature takes the form of a graph, a representation that is easy to visualize and more suitable for interpretation than the original high-dimensional data. Consequently, the Mapper is an excellent tool for investigating the structure of a data set [42]. Typical structures in topological graphs such as "loops" (circles) and "flares" (curve segments) are often used as patterns for identifying interesting subtypes [43]. Besides, Lum et al. [43] provide a formal definition of flares and verify that flares cannot be generated from random data. Naturally, we might group samples in the nodes concentrated solely on the tail of a flare. These samples are then studied with standard statistical methods, including descriptive statistics and statistical inference.

Specifically, we propose a topology-based approach. The first step is applying the Mapper algorithm [38] to a coded population to highlight the mismatch between coding guidelines and coding practices and subsequently identify miscoding subtypes and the features responsible for it. Given the constructed topological graph, in an additional step, we introduce statistical hypothesis-testing approaches to validate the TDA-identified miscoding subtypes and search for features responsible for them. In the last step, we propose an optimization-based method to recommend groups of

subjects for which the health records are worth being given an additional review to adjust the incorrect code. As a result, this hybrid approach is able to extract patterns or shape characteristics within health data and identify the risk distribution of miscoding in the whole population. We identified multiple subtypes of coding errors and several distinct subgroups with a higher possibility of being miscoded and whose health signature is distinct from well-coded individuals.

A novel topology-based method is proposed, and our contribution is multifold: (i) A holistic and transparent framework is proposed to visualize and identify hospital coding errors; (ii) In combination with statistical hypothesis testing, we propose an interpretable approach that generates a representative portrait and a list of potential risk factors for each identified miscoding subtype; (iii) We also provide a simple but efficient optimization model that treats subgroups with more coding errors preferentially. In the case study in section 2.5, with the audit of 20% cases, the proposed method yields up to a 64% reduction in miscoding rate. Results show that the proposed approach is able to reduce the miscoding rate efficiently, which subsequently eliminates the negative impact of hospital miscoding.

The remaining part of this chapter is organized as follows. Section 2.2 reviews existing TDA techniques. The proposed methodology is introduced in Section 2.4 while Section 2.5 presents experimental results on a coded dataset. Finally, Section 2.6 points out limitations and possible future refinements of the presented work.

2.2. Literature review on topological data analysis (TDA)

The Mapper algorithm [38] has been widely used for data analysis in several domains, e.g., in sports science (e.g., redefining the positions of basketball players in the well-known work of [44], in bio-medical science (e.g., revealing phenotype-biomarker associations in traumatic brain injury [45], diagnosing pulmonary embolism [46]), in soil science (e.g., uncovering pedogenetic rules of topsoil system [47]), or even in space science (e.g., solar magnetic eruption prediction [48]). Some systematic surveys [49, 50, 51] on the TDA technique are available for its development history and more techniques details.

It has also proved to be a valuable tool in biology and health domains. In [52], the authors analyzed transcriptionally genomic data about breast cancer and identified a unique subset of breast cancers with a very high survival rate. In [53], the authors developed a precision medicine approach to characterize the complexity of type 2 diabetes (T2D) patient populations based on electronic medical records and genotype data. Their analysis helped identify three distinct subgroups of T2D from topology-based patient-patient networks. Longitudinal and cross-sectional data about malaria are used in [54] to map the routes infected individuals take through their responses to pathogens. The authors showed that less resilient individuals have

a longer infection time with more severe symptoms. Nielson et al. [55] construct syndromic networks based on preclinical spinal cord jury (SCI) and traumatic brain injury (TBI) datasets through topological data analysis. The constructed networks reveal interactions between TBI and co-occurring SCI. In the follow-up work [56], Nileson et al. collected data about imaging, genetics, and clinical outcomes for Traumatic Brain Injury (TBI). With the mapper, they identified a unique diagnostic subgroup of patients with unfavorable outcomes after mild TBI.

These different studies validate the effectiveness of the mapper algorithm in providing new insight by identifying specific structures in complex datasets.

2.3. Problem definition

This study aims to propose a precise approach to better understand and quantify the complexity of miscoding subtypes through topological analysis of subject-subject similarity across available features. Each subject s is characterized by a z dimensional vector of qualitative and quantitative features and is labeled with a binary ground-truth $G(s)$ that equals one if the subject is miscoded and zero otherwise.

Formally speaking, the problem consists of two consecutive steps, which first constructs a topological graph $G = (V, E)$ to understand coding errors across the topological space, and then determines whether to review subjects of each node $v_i \subset V$ in order to maximize the number of miscoded subjects reviewed with limited human resources.

2.4. A TDA-based approach for subject profiling

2.4.1. Topological space projection

The profiling of hospital miscoding can be defined as a space partitioning problem. Intuitively, supervised classification technique such decision tree [57] and unsupervised clustering techniques such as k-mean clustering [58] (i.e., a centroid-based clustering algorithm), and DBSCAN clustering [59] (i.e., a density-based clustering algorithm) can be used to identify subgroups (or subspaces) prone to hospital miscoding. However, these well-known methods are too simplistic and do not work well on high-dimensional complex datasets. One may ask if there is a more informative and intuitive stratification of hospital coding based on subjects' profiles. To answer such a question, we constructed miscoding profiles for each of the subjects using the mapper algorithm [38].

The algorithm aims at generating a graph signature where the nodes are groups of subjects, and the links between nodes represent feature relations between nodes (Figure 2.1). The first step consists in **defining a *filter function* f - or *lens*** - (Fig 2.1.a). With the objective of spreading the data points over the topological space, a filter

function assigns each data point a real number, which will be used to construct the graph.

Then, the next step is about **covering**, where the f-projected space is divided into overlapping fragments - or cubes - (Fig 2.1.b), within which the data points are grouped together based on a clustering algorithm and a similarity metric (Fig 2.1.c). Each group will be a node in the graph (Fig 2.1.d). Since the overlaps can result in a single subject being included in two (or more) clusters, edges are created between nodes if they share common subjects. Formally, the generated graph can be expressed as a set of ordered pair $G = (V, E)$ consisting of a set of nodes $V = \{v_1, v_2 \ldots, v_m\}$ and a set of unordered pairs $E \subseteq \{\{a, b\} \mid a, b \subset V \text{ and } a \neq b\}$, called edges.

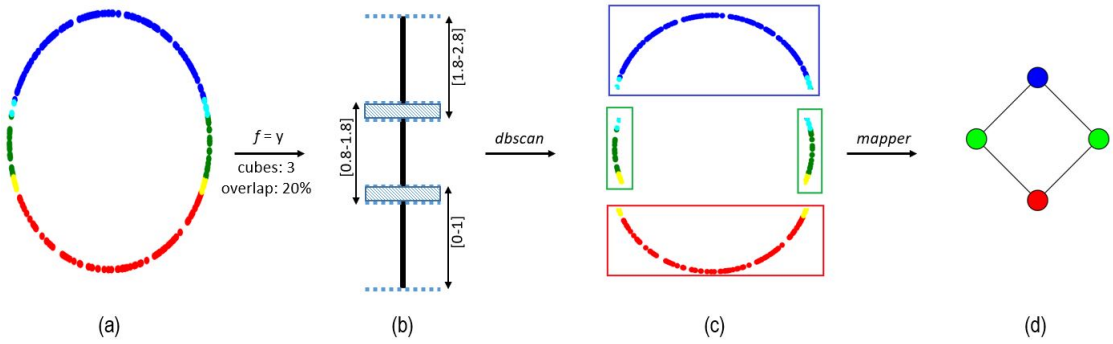


Figure 2.1.: Example of the Mapper filter function.

Formally, given a d dimensional point cloud $\mathbb{X} = \{X_1, X_2, \ldots, X_i, \ldots, X_n\}, X_i \in \mathbb{R}^d$ and a pairwise dissimilarity measurement, the analyst has to choose a filter function f and computes on each point X_i. The application of the Mapper algorithm on the point cloud X with the filter function f is presented below:

1. Covering the embeddingd $\mathbb{Y}_n = f(\mathbb{X}_n)$ by a series of consecutive and overlapping intervals $\{I_1, I_2, \ldots, I_M\}$.

2. Selecting and applying a clustering algorithm on the pre-image of each interval, i.e., $f^{-1}(I_m), m \in \{1, 2, \ldots, M\}$. Such a process produces a *pullback cover* $C = \{C_{1,1}, \ldots, C_{1,k_1}, \ldots, C_{M,1}, \ldots, C_{M,K_M}\}$ of the original point could $\mathbb{X}$. The underlying idea is to group together, in the original space, data points that are close to each other in the projected space.

3. The Mapper algorithm is the *nerve* of the cover C. A vertex $v_{m,k}$ of the Mapper algorithm is equivalent to a component $C_{m,k}$. A connection is created between two vertices $v_{m,k}$ and $v_{m',k'}$ only if the intersection of $C_{m,k}$ and $C_{m',k'}$ is not empty, i.e., $C_{m,k} \cap C_{m',k'} \neq \emptyset$.

The construction of the Mapper projection depends on the personalized selection of filter functions, intervals that cover the image of the filter function f, and the clustering algorithm used to group the pre-images $f^{-1}(I_m), m \in \{1, 2, \ldots, M\}$.

With regard to the choice of filter functions, it strongly depends on the features we aim to highlight [51]. It is a common rule to use topological metrics to position the data points respectively to the others. In the literature, the following metrics are often used [51]:

- Density estimates: these approaches can help understand the structure and connectivity of high-density areas (clusters).
- Coordinates from linear or nonlinear dimensionality reduction technique, eigenfunctions of graph laplacians.
- The centrality function and the eccentricity function are also graph-oriented methods that may help reveal and understand the geometry of the data.

We mention that these methods are suitable in the first approach since they do not require any specific knowledge about the data. A deeper investigation would require the use of application-oriented filter functions. The functions are driven by the domain knowledge that explicitly derives the concept of similarity between data points in a specific context (e.g., ph level for aqueous solutions, disease severity for patients, etc.). In several studies, the combination of both topology-oriented and application-oriented filter functions brought the best results since, first, it gathers together application-wise *similar* data points and then spreads them out in the projected space based on their topological features.

Moreover, the filter functions traditionally operate in the full data space based on a single representation of the data points. This aspect raises two limits. First, from a structural perspective, the analysis of complex high-dimensional datasets is subject to the *curse of dimensionality*, where the data points become sparser as the dimensionality increases. In other words, different data points are "lost in space" and make the task of grouping them into meaningful structures difficult. Secondly, from a conceptual perspective, it faces the "*single-representation*" problem. Traditional techniques performing space partitioning or data clustering are based on the assumption of an appropriate, and sufficient single representation [60]. However, in many application domains, it is possible to derive multiple descriptions for an object. This results in separate descriptor spaces within which the objects behave differently. Therefore, we believe it is more suitable to use several lenses on feature subspaces, but it raises the question of building relevant subspaces.

Finally, once the topological graph is generated, it can be colored by a customized color function, and thereby, numerical values are transformed into colors. The idea is to visually emphasize the presence of feature ranges in specific regions of the topological space. In this study, the color value of a node v_i is defined in Equation 2.1 where the numerator and the denominator indicate the number of miscoded subjects in the node v_i and in the whole population, respectively.

$$y_i = [\sum_{j=1}^{|v_i|} 1\left(s_j \in v_i\right) * G(s)] \,/\, N^+ \tag{2.1}$$

A critical step of the Mapper is to apply clustering algorithms to subsets of the original dataset. In this step, we employ DBSCAN [59], a density-based clustering algorithm, to search for local clusters in each subset. Compared to centroid-based clustering algorithms requiring one to specify the number of clusters for each subset in advance, DBSCAN requires only two global parameters for all subsets: (i) ϵ, the maximum distance between two data points for one to be in the neighborhood of the other, (ii) minPts, the minimum number of data points in a neighborhood for a data point to be a core point. Also, without domain knowledge, the number of clusters is often hard to know in advance. Schubert et al. [61] provide some heuristics for choosing these two parameters for DBSCAN. Another advantage of DBSCAN is its ability to discover clusters with arbitrary shapes because clusters can be in various shapes in real-life datasets. Besides, DBSCAN treats unreachable data points as noise, which improves its robustness to outliers.

Within the clustering algorithm, the distance metric used for the similarity measurement is correlation distance. Correlation-based distances are widely used for biological data analysis and gene expression data analysis, which measures the degree of dependence between two attribute vectors instead of the spatial distance defined by Euclidean distance. The correlation distance between two data points X and Y with z dimensional attribute vectors is defined by:

$$\text{corr}(X, Y) = 1 - r(X, Y) \tag{2.2}$$

where $r(X, Y)$ is the element-wise, mean-centered correlation between X and Y. The term $\bar{X}$ is the mean of the elements of X, and $X \cdot Y$ is the scalar product of X and Y.

$$r(X, Y) = \frac{(X - \bar{X}) \cdot (Y - \bar{Y})}{\|X - \bar{X}\|_2 * \|Y - \bar{Y}\|_2} \tag{2.3}$$

On top of the constructed topological graph $G = (V, E)$, we face two challenges in applying hypothesis testing and optimization models to samples in the topological space. Firstly, the Mapper algorithm discards unreachable samples when applying DBSCAN for partial clustering. These unreachable samples are grouped into a heterogeneous node of outliers, which is then appended to the graph G as an isolated node. Secondly, within the constructed topological graph, an edge is created only if two nodes have at least one shared sample. This mechanism produces overlapped nodes that contain shared samples. When selecting nodes, i.e., groups of patients to review, the optimization algorithm will face the problem of duplicate samples and might be impacted by converging toward a sub-optimal solution. To solve this problem, we iterate over edges instead. We detect shared samples between two endpoints or nodes of a given edge. The smaller node still retains shared samples (i.e., a node with fewer subjects), and shared samples from the larger node are removed. For two endpoints of equal size, we randomly select one of the endpoints and remove shared samples from the selected endpoint. In this way, we generate a revised topological graph $G' = \left(V', E\right)$, which can avoid the duplication of miscoding reviews.

To identify and validate subtypes of coding errors appearing in the topological space

$G' = (V', E)$, an approach based on hypothesis testing is proposed in the next section.

2.4.2. Miscoding sutype identification and statistical hypothesis testing

To identify miscoding subtypes, we leverage a simple thresholding technique. The thresholds $T = \{t_1, t_2, \ldots\}$ are selected manually according to the distribution of the color values defined by Equation 2.1. Given the revised topological graph G', we define a subgraph as a coding error subtype if the subgraph satisfies the following conditions: i) each node must be a high-risk node whose color value is greater than or equal to a predefined threshold $t, t \in T$, ii) each node has to be connected to at least one other high-risk node. In other words, a miscoding subtype is a set of interconnected nodes whose color value is higher than a predefined threshold $t, t \in T$.

Once the miscoding subtypes are defined, the next step is to evaluate and validate these discovered miscoding subtypes. To validate the discovered subtypes, the χ^2 test of independence is performed to compare the statistical difference of coding errors between each subtype and the whole population, i.e.,

- H_0 : {*The frequency distribution of the ground truth* $G(s)$ *of a subtype is independent of the frequency distribution of the ground truth of the whole population*}.
- H_1 : {*the mismatch indicator of a subtype is not independent of the mismatch indicator of the whole population*}.

A subtype is validated only if the null hypothesis H_0 is supported.

To sketch a validated subtype, we applied the χ^2 test and the two-sided Kolmogorov-Smirnov (KS) test combined with the p-value to compare categorical or numerical features. Specifically, We compare each feature's statistical difference between the miscoded population and the ordinary population for each identified subtype. The non-parametric KS test investigates the probabilistic distribution of samples across each feature, while the χ^2 test is performed to compare categorical features. We select a categorical feature as a marker of a subtype if the null hypothesis of the χ^2 test is accepted, i.e.,

- H_0 : {*The frequency distribution of a feature of the miscoded population is independent of the frequency distribution of the same feature of the ordinary population*}.

Likewise, a numerical feature is selected as a maker when the null hypothesis of two samples KS test is rejected, i.e.,

- H_0 : {*The distribution of a feature of the miscoded population and the same feature of the ordinary population are from the same distribution*}.

In addition to these markers, we use a mean vector and a covariance matrix to portray a typical portrait for each subtype, which can provide more interpretability to the coding staff.

With the revised topological graph G', in the next section, we build a mathematical programming model to determine nodes $V^+ \subset V'$ that need to be sent to coding staff for correcting coding errors in it. Due to limited human resources, a trade-off should be made: nodes with more coding errors are preferentially targeted, while the remaining nodes are treated with lower priority.

2.4.3. Optimal censoring budget rationing

For a given set of nodes $V\prime$, each node is characterized by a value ρ_i indicating the number of miscoded subjects in that node, and a weight ω_i representing the total number of subjects in that node. We seek to select a subset of nodes with the maximum number of miscoded subjects such that the total number of selected subjects is less than or equal to the miscoding censoring budget λ. Therefore, the problem is modeled as a 0-1 knapsack problem, i.e.,

Goal:

$$\max \sum_{i \in V\prime} \rho_i \, x_i \tag{2.4}$$

Subject to:

$$\sum_{i \in V} \omega_i \, x_i \leq \lambda \tag{2.5}$$

$$x_i \in \{0,1\}, \forall \, i \in V\prime \tag{2.6}$$

where x_i is a binary indicator equal to one if the subgroup v_i is selected, and zero otherwise. Subjects in selected subgroups are sent to the coding staff for review.

2.4.4. Performance evaluation

Given the results from the mathematical programming model, the False Negative Rate (FNR), False Omission Rate (FOR), Reduction of Coding Errors (RCE), and Efficiency Score (EFF) are used for the performance evaluation:

$$FNR = FN/(TP+FN), \; FOR = FN/(TN+FN) \tag{2.7}$$

$$RCE = [\sum_{j=1}^{|v_i|} 1\left(s_j \in v_i\right) * G(s) * x_i] \, / \, N^+ \tag{2.8}$$

In Equation 2.7, TP, FP, TN, and FN stand for true positive, true negative, false positive, true negative, and false negative, respectively. The numerator and the denominator of the Equation 2.8 denote the number of selected miscoded subjects and the number of miscoded subjects in the whole population, respectively. The efficiency of the mathematical programming model is assessed by EFF = RCE / λ, where RCE $\in$ [0, 1],

$\lambda \in (0, 1]$ and EFF $\in [0, \infty)$. A higher EFF score reflects a better efficiency of the model. These metrics are then used to assess the performance of the proposed method.

2.5. Case study: subject profiling for miscoding screening

2.5.1. Data description

In this study, we evaluate our methodology on a real-world dataset drawn from the University Hospital of Saint-Etienne, France (CHU-SE). After filtering subjects under the age of eighteen, a dataset with 33143 de-identified health records is formed. These records represent one year of hospital stays. This study focuses on hospital stays with undernutrition; each stay is assigned to one of the three categories by human coders: without undernutrition, moderate undernutrition (E44.0 or E44.1), and severe undernutrition (E43). The category of a stay assigned by human coders is called a man-made label. Moreover, according to coding rules defined in the guideline in Chapter 6, Section 6.4, we create a rule-based label representing the true category of a stay and a binary ground truth $G(s)$ equal to one if the man-made label and the rule-based label are mismatched and zero otherwise. Based on the ground truth, around 11% of the stays are miscoded.

Features known for each stay include demographics and behavioral information such as gender, age, length of stay (LOS), and the number of visits to various medical departments. Biological information is also considered, including body mass index (BMI), weight evolution, pre-albumin, albumin, and C-reactive protein (CRP). In total, 24 features are taken into account. The dataset is highly imbalanced and includes only 8.9% subjects with undernutrition. Among all these stays, 2.9% of subjects died during their hospitalization.

2.5.2. Data pre-processing

Since it is not necessary for a human coder to take into account all features to diagnose each subject with undernutrition, around 29.2% of cells in the dataset are not measured and filled. Unfortunately, as part of the Mapper algorithm, most clustering algorithms are not robust to missing data and often require a complete matrix as input, which makes missing data imputation an inevitable step. Table 2.1 shows percentage of missing values about the missing data.

Table 2.1.: Statistics on missing data

Features	Missing (%)	Features	Missing (%)
BMI	64.6	bodyweight	64.6
weight change in one month	84.5	weight change in six months	99.3
albumin	85.4	prealbumin	93.7
CRP	33.6		

To fill in the missing data, multiple strategies, including constant imputation(i.e., mean imputation, zero imputation, etc), KDE (Kernel Density Estimation) imputation, KNN imputation [62], and multivariate imputation [63] are taken into account. We propose a novel strategy called KDE imputation, which first estimates the Probability Density Function (PDF) of a feature with a Gaussian kernel and then generates random samples from the PDF to fill in missing values of the feature. The bandwidth of the Gaussian kernel is determined by grid search with 10-fold cross-validation.

Table 2.2.: Correlation change after the data imputation

	Mean imputation	KDE imputation	KNN imputation	Multivariate imputation
IMC	0.1049	0.1466	0.04975	0.1117
bodyweight	0.1042	0.1601	0.04606	0.06684
weight change in one month	0.1680	0.1704	0.002747	-0.02691
weight change in six months	0.3792	0.3336	0.09283	0.08671
albumin	0.1052	-0.0576	0.01245	0.08501
prealbumin	0.08862	0.006161	-0.1283	-0.1028
CRP	0.05960	-0.1575	-0.09603	-0.08298
Mean	0.1272	0.08597	-0.002934	0.019644
STD	0.1311	0.1653	0.08063	0.08867

Except for the CRP, which has no apparent correlation with the rule-based label, all other features with missing values are negatively correlated with the rule-based label. We assume that a good data imputation strategy would not change the monotonic correlation between the rule-based label and features with missing values after the data imputation. In other words, the less the correlation changes, the better the performance. The Spearman's rank correlation coefficient (ρ) measures the monotonic correlation (whether linear or not) of two variables X and Y and is defined as the Pearson correlation coefficient between two rank variables, i.e.,

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}\left(\mathrm{rg}_X, \mathrm{rg}_Y\right)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}} \tag{2.9}$$

where ρ denotes the Pearson correlation coefficient, rg_X and rg_Y are the rank variables of X and Y. Subsequently, the difference of the Spearman's ρ between the raw data and the filled data is used to evaluate these strategies' performance. Table 2.2 shows the changes in the Spearman's ρ for each feature after the data imputation.

As shown in Table 2.3, the difference of the Spearman's ρ between the raw data and the filled data is used to evaluate these strategies' performance. Among these strategies, the average change in Spearman's ρ for kNN imputation is the smallest. Both KNN and multivariate imputations far surpass the currently adopted baseline (mean imputation). Finally, the data filled by the KNN imputation is then used in the downstream task.

Table 2.3.: Correlation change after the data imputation

	Mean imputation	KDE imputation	KNN imputation	Multivariate imputation
Mean	0.1272	0.08597	-0.002934	0.01964
STD	0.1311	0.1653	0.08063	0.08867

2.5.3. Application of Mapper to undernutrition data

In addition to data samples, the Mapper takes two inputs, including one or more filter functions and two resolution metrics (i.e., the number of intervals and the overlapping rate), to construct a graph. The filter functions determine the space in which we construct a graph. Different choices of filter functions may produce graphs with different shapes, therefore allowing us to uncover the shape characteristics of the data from various perspectives. For simplicity's sake, you can think of it as a camera with lens adjustments and other settings. In this study, we define four different filter functions based on the categories of features used in the guideline [5] to encode undernutrition.

The first two filter functions are the Singular Value Decomposition (SVD) principle component of the weight-related features (i.e., BMI, body weight, the evolution of weight over the last 1 and 6 months) and bio-related features (i.e., albumin and pre-albumin), respectively. The first filter function is adopted to characterize short-term and long-term weight change in subjects, while the second filter function is used to watch for body exhaustion and undernutrition. The first two filters generate a factorization of the data matrix separately into linearly uncorrelated components, with the first SVD component representing the component of the highest variance. Note that the filer functions are not necessarily linear transformations, although they can be.

The third and fourth filter functions are the normalization of the inflammation-related feature (i.e., CRP) and age. The third filter implies inflammatory syndrome in subjects. As the population of different ages demonstrates different characteristics

of undernutrition, normalized age is selected as the fourth filter to reflect the age distribution of subjects.

The graph is set at a resolution of 40 intervals with an overlap rate of 0.2 for the weight-related filter and the bio-related filter and 20 intervals with an overlapping rate of 0.2 for the inflammation-related filter and the age filter. Such a process is similar to scaling up or down on a microscope. Increasing the number of intervals increases the number of nodes in the graph to uncover a more refined structure of the data manifold, which preserves only strong connections between nodes; weak connections tend to break apart, and smaller subgroups are then generated. Increasing the overlap rate increases the number of edges between nodes to reveal finer relevance between nodes while reducing the overlap rate results in smaller and more isolated nodes.

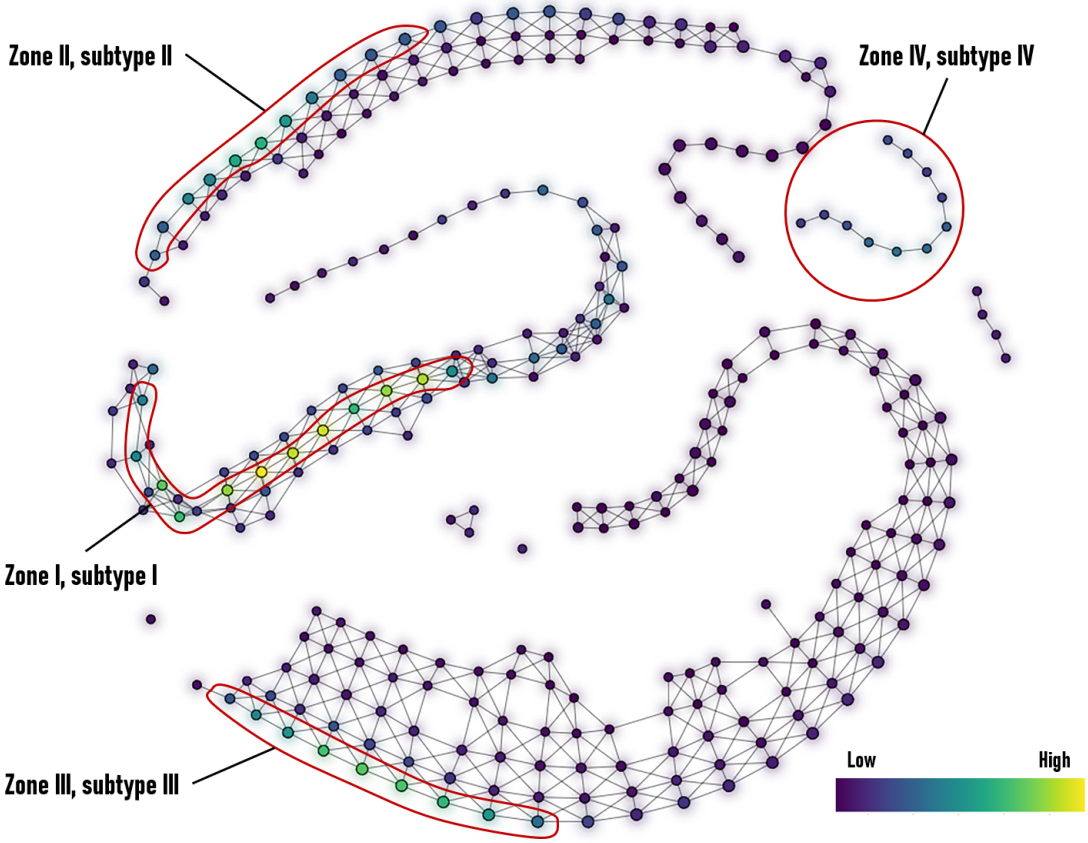


Figure 2.2.: Topological graph derived from the undernutrition data

2.5.4. Coding subtype stratification

Identifying subtypes of coding errors in an intelligible manner is a challenge since subgroups may be small, and their relationships are often complex. In this section, we applied the steps defined in section 2.4.2 to identify subtypes and show that topological graphs can more neatly stratify subjects than the standard decision trees

method [57]. The identified subtypes can be used to design prevention programs and reduce the hospital's coding error rate. The resulting graph in Figure 2.2 represents 292 distinct nodes and 29738 unique samples since 3450 unreachable samples are grouped into a heterogeneous node, which is not shown in the graph. The entire graph has four main networks along with four disconnected components. Each node represents a set of subjects and is colored according to the color function defined in Equation 2.1, with yellow denoting a large value and purple encoding a low value. In this study, the color function is defined as the coding error rate of each node. Therefore, yellow nodes contain a great number of coding errors, whereas nodes that are purple contain subjects whose behavior or expression is close to ordinary subjects. Nodes with large color values correspond to high-risk subgroups with high coding error rates. It is interesting to discover high-risk zones consisting of interconnected high-risk subgroups to characterize sub-types of coding errors.

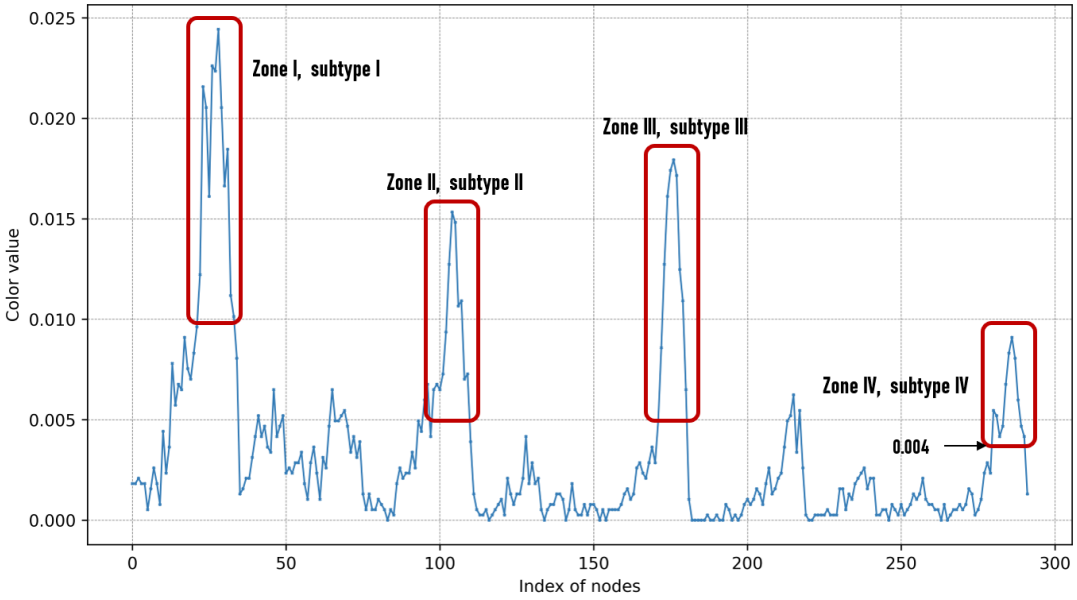


Figure 2.3.: Identification of subtypes of coding errors

We identified and validated four subtypes of coding errors in the graph. As shown in Figure 2.3, coding errors are not uniformly distributed throughout the whole population but frequently occur in certain nodes. To understand the most frequently occurring coding errors (i.e., high-frequency components) while ignoring ambient noise that is approximately evenly distributed throughout the whole population (i.e., low-frequency components), we select three thresholds of the color value (i.e., 0.01, 0.005, and 0.004) and subsequently identified four significant subtypes. The occurrence of ambient noise may be due to missing data in the dataset, e.g., human coders cannot accurately encode a subject if some parts of the data are missing. At a significance level of 99.99% ($\alpha = 0.0001$), we accept the null hypothesis of the χ^2 test

of independence defined in Section 2.4.2 that the frequency distributions of coding errors in each subtype and the whole population are independent.

All identified subtypes are in the shape of a curve segment, implying that most coding errors occur along a particular direction. Each of the four curve segments has a fixed weight-related filter value and a fixed bio-related filter value. However, the values of the inflammation-related filter and age filter vary along the length of each curve segment. Therefore, among the mean vectors provided, more attention should be paid to the weight-related features and the bio-related features since they determine this particular direction or perspective.

Table 2.4.: Basic statistics of subtypes

	No. of nodes	No. of uniq subjects	No. of miscoding	No. of undercoding	No. of overcoding
The whole population	293	33143	3677(11.1%)	1454(39.5%)	2223(60.5%)
Outliers	1	3405	870(25.6%)	355(40.8%)	515(59.2%)
Subtype I	12	1448	692(47.8%)	518(74.9%)	174(25.1%)
Subtype II	12	5036	360(7.1%)	22(6.1%)	338(93.9%)
Subtype III	9	2707	381(14.1%)	2(0.5%)	379(99.5%)
Subtype IV	11	299	206(68.9%)	185(89.8%)	21(10.2%)

From the basic statistics presented in Table 2.4, we can observe that coding errors occurring in the four subtypes account for more than 58% of the total, while the number of subjects in the four subtypes accounts for less than 32% of the total. Besides, the frequency distribution of coding errors for the four subtypes differs significantly from the distribution for the whole population. The hypothesis testing methods mentioned in Section 2.4.2 are applied to select a list of markers that best differentiate miscoded subjects from the ordinary population. These markers can be considered possible sources of coding errors for each identified subtype. It is worth noting that the significance level of a marker can be simply calculated by $1-p$. Table 2.5 shows interesting markers related to each identified subtype.

In zone I, Compared to the whole population, features with a significant difference in standard deviation (STD) and mean are age, CRP, and LOS. Subjects' ages range from 67.2 to 91.3 years old. In contrast, the age range is 18-year-old up to 99.7 for the whole population. Their average LOS is 14.3 days, which is approximately twice longer than the whole population's average LOS (i.e., 7.14 days)

Zone IV is an entire sub-network comprising eleven interconnected nodes. Subjects' ages range from 69.4 to 93.5. Their average LOS is 14.4 days. They lost 3.4% and 6.9% of their weight on average within one month and six months, respectively.

Zone III represents 372 subjects, of which more than 68.8% of subjects are miscoded. Moreover, more than 89% of the miscoded subjects in the zone I are under-coded.

We do not find any significant categorical makers for subtype III, while the discovered numerical markers are the weight evolution within six months(p=0), the pre-albumin(p=0), the weight evolution within one month(p=6.51e-10), the BMI(p=1.98e-9), and the weight(p=2.85e-6). Zone IV consists of eight interconnected nodes, in which subjects' age range from 73.7 to 91.3 years old.

Zone II comprises 8337 subjects, of which about 6.1% of subjects are miscoded. Besides, more than 94% of the miscoded subjects in zone I are over-coded. More than 51.1% of subjects of zone IV have been to the surgery department at least once, which is higher than that of the whole population (i.e., 31.3%). The discovered categorical markers for Type IV coding error are the number of visits to physicians(p=2.98e-55), the number of visits to surgery department(p=5.21e-55), and the binary indicator of death (5.8e-21), while the discovered numerical markers are the LOS(3.70e-79), the age(p=2.16e-36), weight(p=2.41e-33), the BMI(p=8.80e-21), the weigh evolution within one month(p=7.21e-19), and the CRP(p=2.44e-19).

Table 2.5.: Possible sources of coding errors

	Categorical markers	Numerical markers
Subtype I	-	weight(p=6.66e-16), BMI(p=6.66e-16), pct_evol_wgt_1month(p=6.66e-16), pct_evol_wgt_6months(p=6.66e-16), albumin(p=6.66e-16), pre-albumin(6.74e-14)
Subtype II	no_visits_physicians(p=8.99e-45), no_visits_surgery_dept(p=2.42e-43), death_or_not(p=1.19e-16)	LOS(p=1.69e-53), age(p=5.22e-25), weight(p=8.96e-21), BMI(p=2.67e-13), CRP(p=8.58e-13), pct_evo_wgt_1month(p=1.89e-11)
Subtype III	no_visits_physicians(p=1.36e-20), gender(p=4.35e-10), no_visits_ICU(p=6.33e-5)	age(1.55e-15), LOS(p=1.55e-15)
Subtype IV	-	pct_evol_6_months(p=2.56e-14), pre-albumin(p=3.16e-12), pct_evol_wgt_1month(p=3.0e-7), BMI(p=1.13e-6)

We identified and verified four miscoding subtypes by using the hypothesis testing defined in subsection 2.4.2. We have also set up the problem as a binary classification problem and generated decision trees for differentiating miscoded subjects from the ordinary population. However, as shown in Figure 2.4, the CART algorithm [57] generates over-complex decision trees (i.e., a decision tree with more than 2600 decision paths or leaves, the length of the longest path is 44) with an acceptable classification performance (i.e., *Micro f1=0.62, Macro f1=0.55*). The data set contains lots of ambient noise, which leads the algorithm to excessively partition the feature space, generating over-complex decision trees and overfitting. Thus, the generated patterns (decision paths) cannot be trusted and used to understand miscoding subtypes within the dataset. In contrast, the Mapper is robust to noise and small deformations and is an appropriate method to uncover novel insights about miscoding subtypes in the dataset.

In addition, the proposed approach differs from the CART algorithm in several respects: (i) The Mapper, as a clustering-based unsupervised algorithm, measures the similarity between subjects and forms homogeneous subgroups. However, as a supervised classification algorithm, CART forms branches having leaf nodes with similar target variables, which can only guarantee the purity of leaf nodes but not the homogeneity of leaf nodes (i.e., samples in a leaf node are similar in terms of target label, but are not similar in terms of feature vector). (ii) CART is often unstable. A slight deformation in the dataset can lead to an entirely different model. Whereas for the Mapper algorithm, a small variation in the dataset might result in a change in the affiliation of a subject but can hardly affect the global structure of the topological graph. (iii) Outcomes of the CART algorithm are not continuous nor smooth. However, this is not the case for the MIP model. For a given budget between 0 and 1, the MIP model computes a maximum reduction of coding errors, i.e., in the Pareto graph, a decision tree model is represented by a point. However, a MIP model represents a Pareto front. This gives the decision-maker more flexibility. (iv) Heuristics such as the greedy algorithm are often applied to the CART algorithm, from which locally optimal splits are performed at each internal node. In contrast, the Mapper generates a global optimal topological graph with a given set of parameters.

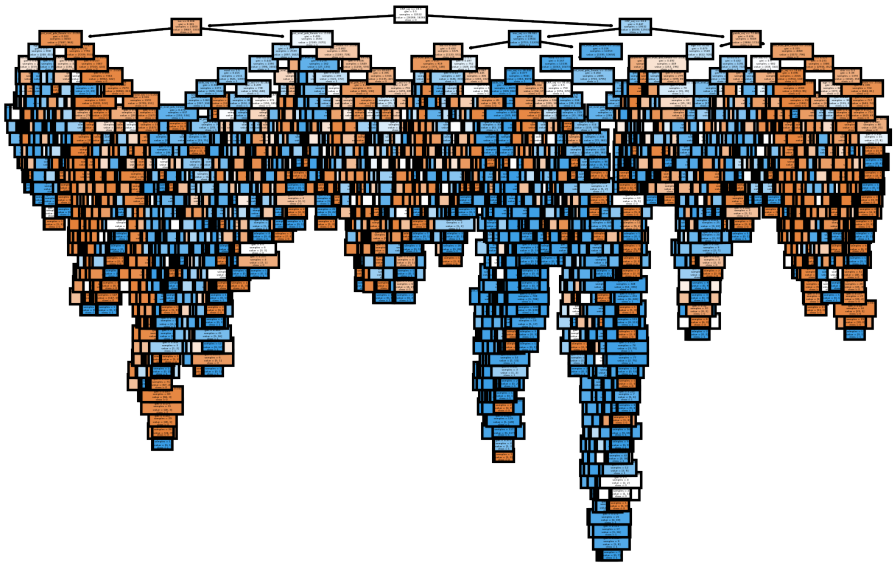


Figure 2.4.: The decision tree model for the binary classification of hospital miscoding

The topological graph is constructed, and the risk score for each node is defined by the equation 2.1. In the next section, we will apply the proposed MIP model to the revised topological graph.

2.5.5. Review budget rationing

The chance-constrained rationing problem is solved by the mathematical programming model defined in section 2.4.3. Health records of a subgroup v_i are reviewed if the decision variable x_i is equal to one. The objective function maximizes the health records reopened for miscoded subjects, while the constraints limit the censoring budget. Table 2.6 shows the results of the programming model.

Table 2.6.: Experimental results of the MIP model

λ	TP	FP	TN	FN	FNR	FOR	RCE	EFF
0.01	6.90e-3	0.105	0.884	3.07e-3	0.308	3.46e-3	0.0615	6.152
0.05	0.0278	0.0843	0.865	0.0221	0.443	0.0249	0.248	4.965
0.1	0.0460	0.0660	0.833	0.0538	0.539	0.0606	0.410	4.108
0.2	0.0725	0.0395	0.760	0.127	0.637	0.143	0.646	3.234
0.3	0.0898	0.0223	0.677	0.210	0.700	0.236	0.800	2.669
0.4	0.0991	0.0130	0.587	0.300	0.752	0.338	0.883	2.209
0.5	0.103	8.19e-3	0.491	0.395	0.792	0.445	0.926	1.853
0.6	0.107	5.09e-3	0.395	0.492	0.821	0.554	0.954	1.590
0.7	0.109	2.68e-3	0.298	0.589	0.843	0.664	0.976	1.394
0.8	0.111	9.03e-4	0.200	0.687	0.860	0.774	0.991	1.239
0.9	0.111	1.80e-4	0.100	0.787	0.875	0.886	0.998	1.109
1	0.112	0	0.0511	0.836	0.881	0.942	1	1

There is a tradeoff between the number of health records reviewed and the true positive rate. The true positive represents miscoded subjects whose health records are reviewed. Intuitively, the more health records are reviewed, the higher the true positive since reviewing the whole population will ensure to cover all miscoded subjects but at the cost of an increased amount of unnecessary reviews (Decrease in EFF score). On the other hand, the less the number of health records reviewed, the less unnecessary expense. Considering a λ value of 0.2, we can reduce more than 64% coding errors at the cost of 20% subjects reviewed.

2.6. Conclusion and perspectives

This chapter focuses on the problem of coding error identification and correction. Through direct visualization of interrelated subjects, TDA detects novel patterns across the topological coding space. The procedure involves not only identifying coding errors but also providing clues that possibly reflect causes resulting in coding errors and underlying miscoding subtypes. The source of coding errors relates to shared biological features that may reflect the health status of subjects as well as shared demographics and behavior information reflecting coding behaviors of coders from different medical units of the hospital. There also exits Besides, reducing the number

of subjects reviewed involves a cost-benefit balance, which should be determined by the decision-maker to ensure the desired performance.

Our research provides a novel insight into identifying and correcting coding errors. Nevertheless, it also faces some limitations. First of all, a naive strategy is proposed to handle unreachable samples. As a future improvement, it would be interesting to append unreachable samples to nearby nodes by comparing the distances between each sample and the mean vector of each node. Secondly, the color function defined in Equation 2.1 does not consider the size of each node. Therefore, a possible future research direction is to evaluate the miscoding probability for each discovered node by considering the size of a node and also the proportion of miscoded subjects in the node. To do so, we can improve the robustness of the proposed approach.

So far, we leverage the TDA approach to discover miscoding subtypes and also a MIP model for review budget optimization. In the next chapter, we propose a clustering and Bayesian-based approach to consider and model the uncertainty of hospital miscoding, which results in a more robust approach for miscoding profiling and identification.

3. Risk analysis of hospital miscoding

Summary

Abstract

Healthcare service systems are progressively becoming more complex and prone to uncertainty. Therefore, a sensible decision requires the decision maker to take uncertainty into account and implement risk control measures. Often, the hospital miscoding problem is considered a binary classification problem, where the objective is to predict if a subject is miscoded or not. Nevertheless, it ignores the uncertainty of hospital miscoding and thus results in poor prediction performance. In this chapter, the problem is defined as a chance-constrained health intervention rationing problem: only subjects with high risk are given supplemental interventions for miscoding correction, while subjects at low risk are excluded. The objective is to profile hospital miscoding, identify miscoding risk, and select subjects at high risk of miscoding to whom supplemental health interventions are recommended while addressing the unknown miscoding risk and the unknown high-risk subgroups. The problem is characterized by two contradictory objectives: (i) maximizing the confidence level of estimated risks; (ii) maximizing the number of significant subgroups which have much higher or lower risks. We propose a novel white-box approach, entitled Alternating Clustering and Bayesian Inference (ACBI)[1], combining clustering and Bayesian inference. We investigate its performance on a real-life miscoding dataset. Results are promising and lead up to a 46.71% reduction in miscoding rate. In combination with Bayesian inference, our approach yields a dynamic system. The experimental results can be further improved if more data becomes available. Another advantage of this approach is its generality. The same approach can be applied directly to other risk control related problems, such as hospital readmission prevention. Experiments on a real-life readmission dataset are also conducted. Experimental results are promising and show a 34.42% reduction in readmission rate.

Keywords: Risk modeling, risk management, hospital miscoding, hospital readmission, machine learning, Bayesian inference, clustering algorithm, budget optimization.

[1]This chapter is based on our previous work published in the proceedings of 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), entitled "ACBI: An Alternating Clustering and Bayesian Inference approach for optimizing medical intervention budget under chance constraints" by He Chen, Dalmas Benjamin, and Xie Xiaolan [64].

Résumé du chapitre

Les systèmes de services de santé deviennent progressivement de plus en plus complexes et sujets à l'incertitude. Par conséquent, pour prendre une décision sensée, le décideur doit tenir compte de l'incertitude et mettre en œuvre des mesures de contrôle des risques. Souvent, le problème du mauvais codage hospitalier est considéré comme un problème de classification binaire, où l'objectif est de prédire si un sujet est mal codé ou non. Néanmoins, cette approche ne tient pas compte de l'incertitude liée à l'erreur de codage de l'hôpital, ce qui entraîne une mauvaise performance de prédiction. Dans ce chapitre, le problème est défini comme un problème de rationnement de l'examen médical sous contrainte de chance : seuls les sujets à haut risque bénéficient d'un examen supplémentaire pour corriger les erreurs de codage, tandis que les sujets à faible risque sont exclus. L'objectif est de dresser le profil des erreurs de codage de l'hôpital, d'identifier le risque d'erreur de codage et de sélectionner les sujets à haut risque d'erreur de codage pour lesquels des examens médicaux supplémentaires sont recommandés, tout en tenant compte de deux probabilités inconnues : (i) celle liée à l'erreur de codage et (ii) celle liée à l'appartenance à un sous-groupe à haut risque. Le problème est caractérisé par deux objectifs contradictoires : (i) maximiser le niveau de confiance des risques estimés ; (ii) maximiser le nombre de sous-groupes significatifs qui ont des risques beaucoup plus élevés ou plus faibles. Nous proposons une nouvelle approche de type boîte blanche, intitulée Alternating Clustering and Bayesian Inference (ACBI)[1], combinant le clustering et l'inférence bayésienne. Nous étudions ses performances sur un ensemble de données réelles de mauvais codage. Les résultats sont prometteurs et conduisent à une réduction de 46,71% du taux de mauvais codage. En combinaison avec l'inférence bayésienne, notre approche permet d'obtenir un système dynamique. Les résultats expérimentaux peuvent encore être améliorés si davantage de données sont disponibles. Un autre avantage de cette approche est sa généralisation. La même approche peut être appliquée directement à d'autres problèmes liés au contrôle des risques. Afin de démontrer cette capacité de généralisation, nous présentons également dans ce chapitre des expériences menées sur un ensemble de données réelles de réadmission de patients. Les résultats expérimentaux sont prometteurs et montrent une réduction de 34,42% du taux de réadmission.

Mots-clés: Modélisation du risque, la gestion du risque, mauvais codage hospitalier, réadmission à l'hôpital, apprentissage automatique, inférence bayésienne, algorithme de regroupement et optimisation budgétaire.

[1]Ce chapitre est basé sur nos travaux antérieurs publiés dans les actes de la 16e conférence internationale IEEE 2020 sur la science et l'ingénierie de l'automatisation (CASE), intitulée "ACBI : An Alternating Clustering and Bayesian Inference approach for optimizing medical intervention budget under chance constraints" par He Chen, Dalmas Benjamin, et Xie Xiaolan [64].

3.1. Introduction

Healthcare service systems are progressively becoming more complex and prone to uncertainty. Since inappropriate risk assessments and misjudgments may lead to unforeseen consequences, uncertainties need to be continuously inspected and controlled. Therefore, making decisions of great and broad impact requires stakeholders to consider uncertainty and implement relevant risk management measures. In healthcare and crisis management, along with the increasing number of uncertainties, a great number of diverse concepts and heterogeneous approaches for risk control have been developed. In recent years, we have perceived the uncertainty being considered in healthcare-related areas such as healthcare management [65, 66], healthcare prevention [67], healthcare emergency planning [68], and psychology [69].

The methodology proposed in chapter 2 has the limitation of not being able to take into account the uncertainty of hospital miscoding. Differences in descriptive features and sample size can introduce uncertainties in the determination of subpopulations as well as the number of subjects at risk of miscoding. This chapter takes into account the uncertainty of hospital miscoding and proposes a Bayesian-based optimization approach by taking into account descriptive features related and sample size related uncertainties.

Based on this, we first put the problem into a general framework of chance-contained health intervention rationing with unknown probability distributions and a given population. We then propose a four-step Bayesian-based optimization approach. The first step is the population partition by a clustering algorithm. The subgroups being determined, in the second step, we leverage the Bayesian inference to address the unknown probabilities, i.e., the probability of a subject belonging to a subgroup and the risk probability of each subgroup. With the given subgroups and inferred probabilities, the third step determines the unknown optimal number of subgroups with an alternating training process. In the fourth step, we build an integer programming model to determine the subgroups to which additional health interventions are given. Besides, to cope with the explainability requirements, the frequent patterns associated with each subgroup are extracted using association rule mining algorithms and provided to stakeholders for final decision-making.

A novel interpretable and dynamic approach is proposed for the health intervention rationing problem. Our contribution is twofold: (i) We provide a representative profile and interpretable risk factors for each discovered subgroup. (ii) With Bayesian inference, the proposed approach can dynamically update the probability distributions as more data becomes available.

In this chapter, a health intervention is defined as a combination of strategies and activities that aims to assess, maintain, or even improve the quality of health services. Health services can include screening programs (i.e., identifying populations with recognized risk factors and connecting them to health resources.), service surveillance (i.e., a systematic collection and analysis of health-related data aims to plan, implement and assess health practices), and health event investigation (i.e., systematically

gathering and analyzing health-related data, identifying un-recognized risk factors, and designing control measures). The health intervention is shaped as much by the corresponding health service (e.g., medical review for the screening and correction of hospital miscoding, follow-up visits that nurses take for the prevention of hospital readmission, etc.).

The remainder of this chapter is organized as follows. Section 3.3 defines the problem, while our methodology is presented in section 3.4. Experiments on real-life miscoding data are conducted and interpreted in Section 3.5. To show the generality of the proposed approach, the proposed approach is then applied to the problem of hospital readmission prevention. Relevant experimental results are presented in Section 3.6. Finally, Section 3.7 concludes and discusses possible future improvements of the proposed approach.

3.2. Literature review on healthcare risk modeling

Recently, holistic solutions for the optimization of health services have been in high demand, especially for the risk modeling and risk management in diverse health services. A recent comprehensive survey on risk modeling can be found in [70]. In this systematic review, the authors selected and reviewed 77 articles and listed a set of risk estimation models including (i) survival analysis, e.g., cox regression (n=10), survival random forest (n=2), (ii) GLM (generalized linear model), e.g., logistic regression (n=52), generalized estimated equation (n=2), hierarchical generalized linear model (n=2), LASSO regression (n=2) and (iii) machine learning, e.g., random forest or decision tree (n=9), SVM - support vector machine (n=4), neural network (n=4). The authors also observed a trend toward the adoption of machine learning methods in recent years.

Walraven et al. [71] proposed a concise index, called LACE, for assessing risk by rounding the regression coefficients of a logistic model. Since linear models [25, 72, 73, 74, 75, 71, 76] are interpretable and model the problem as a regression problem rather than a classification problem, they are widely used for risk modeling. In practice, a standard General Additive Model (GAM) is defined as $g(E(y)) = \beta_0 + \Sigma f_i(x_i)$ where g is the link function that relates descriptive features x_i to a response variable y and $E(f_i) = 0$ for each f_i. Logistic Regression is a simplified form of GAM where each term f_i is restricted to a linear function, i.e. $f(x) = ax + b$. Risk scores of each descriptive feature x_i can be easily expressed as $f_i(x_i)$. The sum of the risk scores is considered the estimated risk. Apart from this, cox regression [77] is another common risk modeling approach, which also takes the time variable into account.

Hoseeeinzadeh et al. [78] applied Naive Bayes, a discriminative model, and a decision tree model for readmission risk modeling and found a part of readmissions in their data set is inherently hard to predict, which is consistent with our experimental results presented in section 3.6. To be specific, the prediction accuracy of certain subgroups (containing hard-to-predict cases) is significantly lower than others, which

undermines the overall accuracy of a simple ensemble method that applies local classifiers for each discovered subgroup. Several studies [79, 80, 81] predict readmission risk using machine learning techniques such as logistic regression, random forest, linear SVM, gradient boost decision tree (GBDT), and neural networks. On top of the risk prediction, the authors suggest that practitioners pay more attention to high-risk patients and take appropriate actions to prevent hospital readmission.

Apart from the data-driven risk modeling methods mentioned above, another area of research in risk modeling is knowledge-driven risk modeling, which often requires domain knowledge to identify relevant risk factors and model the risk of a specific disease or event, e.g., the Framingham CVD risk model [82] for cardiovascular (CVD) diseases, the CAIDE (Cardiovascular Risk Factors, Aging, and Incidence of Dementia) risk model [83, 84, 85] for dementia risk estimation, and ANU-ADRI (Australian National University Alzheimer's Disease Risk Score) risk model [86] for Alzheimer's disease risk estimation. Tang et al. [87] undertook a systematic review on dementia risk modeling. Two studies [88, 89] systematically reviewed a series of risk models for cardiovascular disease. Clifford et al. [90, 82] applied the Framingham risk model and a discrete simulation model to screen and treat personnel with a high CVD risk in the U.S. Air Force.

Existing data-driven methods mostly treat the risk modeling problem as a linear regression problem, which models a pseudo 'risk score' by the sum of the response values of each descriptive feature. Nevertheless, the pseudo 'risk score' cannot be interpreted as a probability, and it is not recommended to make any further statistical analysis and decision-making for individual subjects based on it, but only use it for outcome study, hospital costs calculation, etc. In addition, knowledge-driven risk modeling is event-specific and is hard to be generalized to other problems or events. Therefore, this chapter aims to overcome these limits and investigates the combination of clustering and Bayesian inference to estimate relevant risks for the health intervention rationing problem.

3.3. Problem definition

This chapter considers the problem of identifying at-risk subjects in a homogeneous population P. Each subject $s_i \in P$ is characterized by a tuple $s_i = (x_i, y_i)$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{iz})$ is a z dimensional attribute vector and $y_i \in \{0,1\}$ is the binary class or label of the subject p_i. A subject s is from the attribute space according to the unknown subject flow probability and is at risk with the unknown at-risk probability $F(s)$. The whole population P is partitioned into at-risk subgroups and risk-free subgroups. We assume that the set of health interventions is only recommended for subjects in at-risk subgroups.

The problem consists of two consecutive steps, which firstly partition the whole population into subgroups $C = \{c_1, c_2, \ldots, c_k\}$, then decide whether to give each subgroup health interventions to minimize the total number of health inventions, i.e.

$$\min E_s \left[\sum_{j=1}^{k} 1\left(s \in c_j\right) u_j \right] \tag{3.1}$$

such that the probability of a subject who is at risk but is not given the recommended set of health interventions is at most λ, i.e.

$$E_s \left[\sum_{j=1}^{k} 1\left(s \in c_j\right) F(s) \left(1 - u_j\right) \right] \leq \lambda \tag{3.2}$$

Equation 3.2 is referred to as chance constraint. The term u_j is a set of binary decision variable indicating whether a subgroup c_j is given health interventions, $1\left(s \in c_j\right)$ is an identity function indicating whether a subject s belongs to the subgroup c_j.

3.4. A Bayesian approach for risk modeling

3.4.1. Population partitioning

To split the overall homogeneous population $P = \{s_1, s_2, \ldots, s_{|P|}\}$ into k disjoint subgroups $C = \{c_1, c_2, \ldots, c_k\}$, several clustering algorithms with low computation costs are considered. Specifically, k-means, k-modes, and k-prototypes are adopted for numerical features only, categorical features only, and mixed data types, respectively. In addition, all these clustering algorithms mentioned above require that the number of clusters k be specified. Due to their low (time and space) complexity, all of them scale well to a large number of samples and have been used in the field of big data.

K-means clustering

In case all descriptive features are numerical variables, we should leverage the k-means clustering algorithm [58]. The k-means clustering aims to partition the whole population P into k disjoint subgroups $C = \{c_1, c_2, \ldots, c_k\}$ of equal variance, in which each sample belongs to the subgroup with the nearest mean. On the other hand, this algorithm splits the complete data space into sub-spaces called Voronoi cells. Each subgroup 'centroid' is a prototype of a cluster and is characterized by the mean vector μ of the samples in the subgroup. Note that the 'centroids' are often not data samples from the whole population P, although they come from the same space.

The objective of the k-means algorithm is to search for an optimal partition of the whole population P that minimizes the *within-cluster sum of squared error*. Selim et al. [91] formulate this optimization problem as a nonconvex mathematical programming problem, i.e.,

Goal:

$$\min \sum_{j=1}^{k} \sum_{i=1}^{|P|} w_{ij} * d\left(x_i, \mu_j\right) \tag{3.3}$$

subject to:

$$\sum_{j=1}^{k} w_{ij} = 1, \quad 1 \leq i \leq |P| \tag{3.4}$$

$$w_{ij} \in \{0,1\}, \quad 1 \leq i \leq |P|,\ 1 \leq j \leq k \tag{3.5}$$

where μ_j is the *mean* (or centroid, or representative vector) for the cluster c_j, $w_{ij} \in \{0,1\}$ is an element of the membership matrix $W_{|P|\times k}$, and indicates the membership of the sample x_i to the subgroup c_j. The Matching Dissimilarity Measure $d\left(x_i, \mu_j\right)$ is defined by the *squared Euclidean distance* instead of the regular Euclidean distance, i.e.,

$$d(x_i, \mu_j) = \left(x_{i1} - \mu_{j1}\right)^2 + \left(x_{i2} - \mu_{j2}\right)^2 + \cdots + \left(x_{iz} - \mu_{jz}\right)^2 \tag{3.6}$$

Within-cluster sum-of-squares criterion can measure the level of the internal coherence of clusters. However, It has the following drawbacks:

- It assumes that clusters are *convex* and *isotropic*, which is not always true. It performs poorly in irregular-shaped clusters.
- This criterion is not normalized: lower values indicate better results, and zero implies an optimal result. However, Euclidean distances tend to be inflated in high-dimensional spaces. This phenomenon is also known as the "*curse of dimensionality*". We can alleviate this problem by applying a dimensionality reduction technique prior to the k-means clustering algorithm.

Variant V.3.1: Without limiting the computation time, k-means is guaranteed to converge. However, this may result in a local minimum solution. The performance of k-means clustering is highly dependent on the initial centroids. In practice, the algorithm is often executed several times with different centroid initializations. This issue is addressed by the k-means++ initialization paradigm [92], which initializes the centroids to be far away from each other, generally leading to a better output than the random initialization paradigm.

Variant V.3.2: The problem defined by equation 3.3 can also be generalized to a fuzzy (C-means) clustering problem [93]. Fuzzy clustering is a soft version of k-means and allows data samples to be assigned to multiple subgroups with different membership degrees. The objective of fuzzy clustering algorithms is defined as follows:

Goal:

$$\min \sum_{j=1}^{k} \sum_{i=1}^{|P|} \left(w_{ij}\right)^{\alpha} * d\left(x_i, \mu_j\right) \tag{3.7}$$

Subject to:

$$\sum_{j=1}^{k} w_{ij} = 1, \quad 1 \le i \le |P| \tag{3.8}$$

$$w_{ij} \in [0,1], \quad 1 \le i \le |P|, \, 1 \le j \le k \tag{3.9}$$

where $\alpha, \alpha \in [1, \text{inf})$ is the fuzziness coefficient. The variable $w_{ij} \in [0,1]$ denotes the membership degrees of the sample x_i to be in the subgroup c_j.

K-modes clustering

In case all descriptive features of samples are categorical variables, we consider k-modes with 'Cao' initialization [94] as an appropriate clustering approach. Compared to random initialization, 'Cao' initialization considers both the distance between samples and the density of samples, guaranteeing the stability of initialized centroids. Generally, a large density value of a sample x implies a great number of samples around x.

The k-modes algorithm aims to divide the population P into k subgroups.

$$\min \sum_{j=1}^{k} \sum_{i=1}^{|P|} w_{ij} * d\left(x_i, q_j\right) \tag{3.10}$$

where q_j is the *mode* (or centroid, or representative vector) for the cluster c_j, $w_{ij} \in \{0,1\}$ is an element of the membership matrix $W_{|P| \times k}$, and indicates the membership of the sample x_i to the subgroup c_j. The Matching Dissimilarity Measure $d\left(x_i, q_j\right)$ is defined by the *Hamming distance*:

$$d(x_i, q_j) = \sum_{l=1}^{z} \delta\left(x_{il}, q_{jl}\right) \tag{3.11}$$

and

$$\delta\left(x_{il}, q_{jl}\right) = \left\{ \begin{array}{l} 0, x_{il} = q_{jl} \\ 1, x_{il} \neq q_{jl} \end{array} \right\} \tag{3.12}$$

where x_{il} and q_{jl} denote the categorical value of l-th descriptive feature in x_i and q_j respectively.

Definition D.1: the *mode* of a group of samples $c_j = \left\{s_1, s_2, \ldots, s_{|c_j|}\right\}$ is a representative vector $q_j = (q_{j1}, q_{j2}, \ldots, q_{jz})$ that minimizes $D(c, q_j) = \sum_{i=1}^{|c|} d\left(x_i, q_j\right)$, where each sample x_i is described by a z dimensional categorical attribute vector $(A_1, A_2, \ldots, A_z)$.

In other words, each element $q_{jl}(1 \leq l \leq z)$ is the most frequent value of the l-th descriptive feature in the subgroup c_j. Note that the centroid q_j is not necessarily an element of c_j.

k-prototypes clustering

The k-means algorithm is a commonly used clustering technique for numerical variables. However, this algorithm is unsuitable for the dataset containing categorical variables since the similarity measure used in k-means is the conventional Euclidian distance. The Euclidian distance is only valid for numerical feature vectors. In addition, k-modes clustering is only valid for categorical data but not mixed data types, i.e., a dataset containing both categorical and numerical variables.

To solve such a problem, Huang et al. [95] proposed the k-prototypes clustering algorithm, which can handle both numerical and categorical variables (i.e., the mixed data types). The k-prototypes algorithm aims to split the population P into k disjoint subgroups.

$$\min \sum_{j=1}^{k} \sum_{i=1}^{|P|} w_{ij} * d\left(x_i, q_j\right) \tag{3.13}$$

where q_j is the *prototype* (or centroid, or representative vector) for the cluster c_j, $w_{ij} \in \{0,1\}$ is an element of the membership matrix $W_{|P|\times k}$, and indicates the membership of the sample x_i to the subgroup c_j. Given a feature vector x_i, we assume that the first o descriptive features are numerical variables $\{x_{i1}, x_{i2}, \ldots, x_{io}\}$ and the remaining features $\{x_{i\{o+1\}}, x_{\{io+2\}}, \ldots, x_z\}$ are categorical variables. The Matching Dissimilarity Measure $d\left(x_i, q_j\right)$ of the k-prototypes clustering is defined by:

$$d\left(x_i, q_j\right) = \sum_{l=1}^{o} \left(x_{il}^r - q_{jl}^r\right)^2 + \mu_j \sum_{l=o+1}^{z} \delta\left(x_{il}^c, q_{jl}^c\right) \tag{3.14}$$

where the distance consists of two parts, the distance between numerical variables is defined by the squared Euclidean distance, and the distance between categorical variables is defined by the Humming distance.

Application of the clustering algorithms

As recommended by [96], we tilt toward positive samples and ignore potentially existing common characteristics shared by negative samples. In other words, we assume that positive samples consist of multiple subgroups, whereas negative samples are drawn from a single distribution. Intuitively, for the problem of hospital miscoding, positive samples (i.e., miscoded subjects) share specific common characteristics that possibly reflect underlying miscoding behaviors of medical coders and also the underlying causes of miscoding, while negative samples (i.e., correctly coded subjects) behave 'normally' in all descriptive features regarding a potential coding error.

In the fitting phase, we fit the selected clustering algorithm with positive samples and search centroids for a given number of subgroups k, where a centroid represents a representative profile of a subgroup. In the classification phase, we predict the subgroup to which a negative sample belongs. In other words, we project a negative sample into the clustered z dimensional attribute space and put it into its corresponding subgroup according to the distances between the sample and centroids of each subgroup.

In this chapter, we consider the whole population as a homogeneous group, and each subject is an IID sample. Based on the above assumption, we model uncertainties of subject flow. The unknown routing probability that a subject from the whole population flows into a subgroup c_j can be simply defined as:

$$\mathrm{p}\left(c_j\right) = \frac{\sum_{j=1}^{k} 1\left(s \in c_j\right)}{|P|} \tag{3.15}$$

The numerator and the denominator denote the number of subjects in the subgroup c_j and the whole subject population P, respectively. To determine the unknown risk probability related to each subgroup, a method based on Bayesian inference is proposed in the next section.

3.4.2. Bayesian inference

For each subgroup, the unknown subject flow probability can be denoted by $\mathrm{p}(c_j)$ with which a new generic subject flows into the subgroup c_j. The risk probability is defined as the unknown probability $\mathrm{p}\left(\mathrm{r}|c_j\right)$ that a subject in c_j to be at-risk. Since the whole subject population is homogeneous, at-risk or not can be considered as an IID trial with probability $\mathrm{p}\left(\mathrm{r}|c_j\right)$. In other words, the subject sample traverses ACBI with Bernoulli random variable of probability $\mathrm{p}(c_j)$ to decide to which cluster the subject sample belongs and with Bernoulli random variable of probability $\mathrm{p}\left(\mathrm{r}|c_j\right)$ to be at-risk.

With a given subject sample of c_j and its prior distribution of the probability $\mathrm{p}\left(\mathrm{r}|c_j\right)$, the Bayesian inference approach can be used to infer the posterior distribution of the probability $\mathrm{p}\left(\mathrm{r}|c_j\right)$. Bayes rule defines the posterior distribution of a parameter θ given a subject sample $s = (x, y)$ by:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{g(x)} \tag{3.16}$$

where $g(x)$ is the marginal likelihood of the observation x, a constant considering all possible values of θ. $\pi(\theta)$ is the prior distribution representing the prior belief about θ without considering the observation x. $f(x|\theta)$ is the conditional probability of θ given x, called likelihood. The posterior distribution of the parameter θ is inferred as $\pi(\theta|x)$ for the current sample x and is updated when a new sample is available. With Bayes rules, we combine the prior knowledge $\pi(\theta)$ with the knowledge we gained from the

subject samples $f(x|\theta)$. Bayesian inference provides a dynamic inference system and an accurate estimate of the true value of the parameter θ.

Let us assume the most conservative case that we have no prior belief about the parameter θ_j for each subgroup c_j and each subject in subgroup c_j has equal probabilities of being at-risk or not. In this case, the prior distribution follows a uniform distribution. Consequently, the posterior distribution would be a Beta distribution. The prior distribution is defined below. Note that $U(0,1) = \text{Beta}(1,1)$.

$$prior\ distribution: \theta_j \sim U(0,1) \tag{3.17}$$

According to Bayesian theory [97], the Beta distribution is the conjugate prior to the Binomial distribution. Note that Bernoulli distribution $B(p)$ is a particular case of Binomial distribution $B(n,p)$ when the number of trials n equals one. Following the Bayesian rule as defined in (3.16). The posterior distribution can be inferred as follows:

$$posterior\ distribution: \theta_j \sim \text{Beta}\left(1+R_j, 1+N_j\right) \tag{3.18}$$

$$\text{Beta}\left(1+R_j, 1+N_j\right) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} = \frac{x^{R_j}(1-x)^{N_j}}{B(\alpha,\beta)} \tag{3.19}$$

where R_j denotes the number of or positive samples ($y=1$) and N_j the number of negative samples ($y=0$) in the subgroup c_j. The beta function B is a normalization term to guarantee that the sum of probabilities equals one.

For each subgroup c_j with risk posterior distribution of parameter θ_j, the risk probability of a subject belonging to c_j can be determined by $\text{p}\left(\text{r}|c_j\right) = \text{E}\left[\theta_j\right]$. Note that the expectation value of a random variable governed by a beta distribution $\text{Beta}(\alpha,\beta)$ is $\alpha/(\alpha+\beta)$.

In this section, based on Bayesian theory, the risk probability of a given subgroup is defined. Nevertheless, the number k of subgroups used in the previous section to partition the population is a hyper-parameter of the selected clustering algorithm and profoundly influences the outputs. The following section proposes an alternating training process based on clustering and Bayesian inference to determine the unknown number of subgroups.

3.4.3. Alternating clustering and Bayesian inference

To determine the unknown number of subgroups k, two objectives are considered: (i) maximizing the number of at-risk subjects and risk-free subjects, (ii) minimizing the credible interval of predicted risk posterior distributions.

To identify as many at-risk and risk-free subjects as possible, the first sub-objective (3.20) is proposed to maximize the expected distance between the mode of subgroup c_j ($mode_j$) and $mode^*$ the mode of the whole subject population. The equation (3.21)

defines the mode of c_j with a posterior distribution Beta(α, β) and $\alpha, \beta > 1$. Notably, the mode is the most likely value of a Beta distribution and corresponds to the peak in the probability density function (PDF). The weights term $\omega_j = \sum_{j=1}^{k} 1\left(x \in c_j\right) / N$ is defined as the ratio of the subgroup size to the size of the whole population to give more (positive or negative) impact for bigger clusters.

$$\text{objective}_1 : \max \sum_{j=1}^{k} \omega_j \left| mode_j - mode^* \right| \tag{3.20}$$

$$mode_j = \frac{\alpha - 1}{\alpha + \beta - 2} = \frac{R_j}{R_j + N_j} \tag{3.21}$$

The second sub-objective (3.22) is set to guarantee the convergence of predicted posterior distributions. To achieve it, we utilize the HPDI (highest posterior density interval), which is simply the shortest interval $[L, H]$ on a posterior probability density distribution for a given confidence level. From this interval, we extract D, the distance between L, the lower, and H, the higher bound.

$$\text{objective}_2 : \min \sum_{j=1}^{k} \omega_j \left| D_j \right| \tag{3.22}$$

Finally, the alternating training process that takes into account the two sub-objectives mentioned above is defined by:

$$(1 - \gamma) * \text{objective}_1 + \gamma * \text{objective}_2 \tag{3.23}$$

The hyper-parameter γ ranges from zero to one and balances the weights of the expected HPDI distance and the expected mode distance.

3.4.4. Optimal health intervention rationing

This subsection solves the health intervention rationing problem with an integer programming model. Given a set of subgroups C, we minimize the total health intervention budget.

Goal:

$$\min \sum_{j \in C} \mathrm{p}\left(c_j\right) * u_j \tag{3.24}$$

subject to:

$$\sum_{j \in C} \mathrm{p}\left(c_j\right) \mathrm{p}\left(\mathrm{r} | c_j\right) * v_j \leq \lambda \tag{3.25}$$

$$\sum_{j \in C} u_j + v_j = 1 \tag{3.26}$$

$$u_j, v_j \in \{0,1\}, \forall\, j \in C \tag{3.27}$$

where u_j is a binary decision variable equal to one if c_j is an at-risk subgroup, v_j is a binary decision variable equal to one if c_j is a risk-free subgroup. Only at-risk subgroups are given recommended health interventions.

3.4.5. Evaluation metrics

In this subsection, we provide several metrics for the performance evaluation. Given the subject flow probability $\mathrm{p}(c_j)$ and the at-risk probability $\mathrm{p}\left(\mathrm{r}|c_j\right)$, we define the probability of true positive, true negative, false positive and false negative. i.e.

$$\mathrm{p}(\mathrm{TP}) = \sum_{j \in C} \mathrm{p}\left(c_j\right) \mathrm{p}\left(\mathrm{r}|c_j\right) * u_j \tag{3.28}$$

$$\mathrm{p}(\mathrm{FP}) = \sum_{j \in C} \mathrm{p}\left(c_j\right) \left(1 - \mathrm{p}\left(\mathrm{r}|c_j\right)\right) * u_j \tag{3.29}$$

$$\mathrm{p}(\mathrm{TN}) = \sum_{j \in C} \mathrm{p}\left(c_j\right) \left(1 - \mathrm{p}\left(\mathrm{r}|c_j\right)\right) * v_j \tag{3.30}$$

$$\mathrm{p}(\mathrm{FN}) = \sum_{j \in C} \mathrm{p}\left(c_j\right) \mathrm{p}\left(\mathrm{r}|c_j\right) * v_j \tag{3.31}$$

where p(FN) is the probability that a subject is at-risk but is not given health interventions. These metrics are then used to evaluate the performance of ACBI.

3.4.6. Characterizing selected significant subgroups

Given a solution of the ACBI model, each discovered cluster is considered a subtype. In this subselection, we aim to extract frequent patterns for discovered clusters. For each discovered cluster c_j, this subsection aims to determine common characteristics of the patient stays in the cluster. To this end, a frequently used association rule mining algorithm, called FP-Growth [98], is adopted.

Let T be the data set and $I = \{i_1, i_2, \ldots, i_n\}$ be a set of binary descriptive features. An association rule can be expressed as an implication in the form of $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$. X and Y are disjoint item sets, called antecedent and consequent, respectively (i.e., $X \cap Y = \emptyset$). The support indicates the frequency that an item-set presents in T, i.e., $\mathrm{support}(X \Rightarrow Y) = T(X \cup Y)/T$, where $T(X \cup Y)$ denotes the number of records containing both X and Y in T. The confidence of a rule indicates the probability of the rule to be true. i.e. $\mathrm{confidence}(X \Rightarrow Y) = \mathrm{support}\,(X \Rightarrow Y)/\,\mathrm{support}\,(X)$.

3.4.7. Summary

This section defines a transparent and dynamic framework to address the health intervention rationing problem. The approach is designed for miscoding identification

and correction. However, to show that this method is general and can be used in other healthcare prevention applications. The proposed approach is also tested on another problem - the readmission prevention problem. In the following two sections, the proposed approach is applied, evaluated, and validated on two real-life study cases: (i) rationing of reviews of patients' EHRs for the hospital miscoding problem (Section 3.5); (ii) rationing of health interventions for the readmission problem (Section 3.6). More specifically, We evaluate the risk of a given population and give patients with high risk (either risk of readmission or risk of miscoding) a health intervention to solve the given problems (either a health intervention to avoid readmission - post-discharge follow-up or a health intervention to prevent miscoding - review of patient EHRs).

3.5. Case study: risk modeling for miscoding screening

3.5.1. Research context

As assigned medical codes are widely used in various critical fields (e.g., epidemiology, health resource scheduling, health services reimbursement), coding errors possess the potential to produce the butterfly effect and far-reaching consequences. Often the hospital miscoding problem is considered a binary classification problem, where the objective is to predict if a hospital stay will be miscoded or not. However, this ignores the uncertainty of miscoding and thus results in poor prediction performance.

In this study case, we set the problem as a chance-constrained medical review rationing problem: only hospital stays with high miscoding risk are given supplemental medical reviews, while hospital stays at low risk are excluded. The problem is characterized by the unknown subpopulations, the unknown subject flow probability, the unknown miscoding risk, and two contradictory objectives: (i) maximizing the number of significant clusters which have much higher or lower miscoding risk; (ii) maximizing the confidence level of estimated miscoding risks. Experiments on the allocation of medical reviews are conducted and explained. The experimental results show a 46.71% reduction in the miscoding rate. The results imply that the proposed approach is promising and is capable of reducing potentially preventable miscoding and subsequently leads to the reduction of hospital costs.

3.5.2. Data description

In the section, a case study will be undertaken on an undernutrition-related dataset, where each data sample is assigned an undernutrition-related ICD code. We aim to identify and remove coding errors from the dataset by allocating medical reviews to possible miscoded cases. The data is collected from the University Hospital of Saint-Étienne, which is the largest hospital in Saint-Etienne. Compared to the undernutrition data used in Chapter 2, some cases with abnormal feature values (i.e.,

cases with BMI > 120, cases with weight > 150 kg, etc.) are excluded in this study case. Finally, we formed a smaller dataset containing 32,856 data samples, of which 11.18% are miscoded.

Dimensionality reduction and data visualization

Above all, a principle component analysis (PCA) [99] is performed to uncover the underlying structure in the dataset of numerical features (without target factor). In this analysis, numerical features, including age, length of stay (LOS), BMI, weight, height, weight evolution within one month, weight evolution within six months, albumin, pre-albumin, and CRP, are taken into account.

Prior to the PCA analysis, a variable standardization technique is adopted to scale each numerical feature to a given range (i.e., between 0 and 1). The advantages of adopting this standardization technique are (i) it is robust to numerical features with small standard deviation, and (ii) it can preserve zero values in sparse data. The transformation of a given feature f can be described by the following formulas:

$$f_{\text{scaled}} = \mu_f \times (f_{\max} - f_{\min}) + f_{\min}, \text{ and } \mu_f = (f - f_{\min})/(f_{\max} - f_{\min})$$

PCA is a linear dimensional reduction technique and is often used to project a high-dimensional dataset to a lower-dimensional space. Specifically, PCA decomposes a dataset with multiple numerical descriptive features into a set of orthogonal components using the Singular Value Decomposition (SVD) technique. Data is then represented as points in a 3-dimensional Euclidean space. Components are sorted in descending order according to the calculated inertia (or variances). The first component is the most critical dimension, and the second component is the second essential component, and so on.

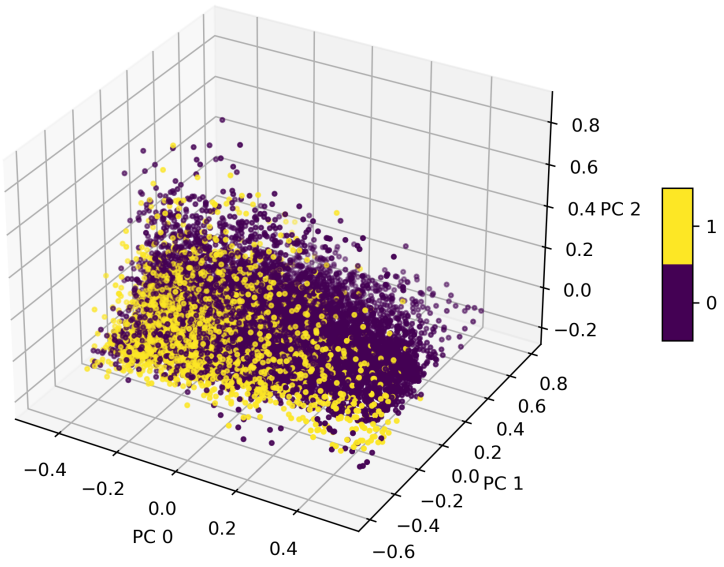


Figure 3.1.: PCA projection of the data set

In figure 3.1, the first principle component, PC 0, holds 66.637% of the information (or explains 66.637% of variance) in the dataset, while the third component, PC 2, contains only 8.086% of information. The results show that, in most cases, it is difficult to distinguish miscoded cases from correctly coded ones. However, there are also a large number of correctly coded cases that are far from the mixed zone and can be easily distinguished. Intuitively, the underlying data structure (i.e., non-uniform distribution of data samples) in the PCA space indicates the possible existence of population subgroups.

3.5.3. Experimental results

First of all, a trade-off should be made. A relatively large γ guarantees the convergence of posterior distributions, while a relatively small γ tends to identify more high-risk and low-risk subjects. In this study case, we set γ to 0.5, which means that subgoal (or equation) 3.20 and subgoal (or equation) 3.22 have the same level of importance.

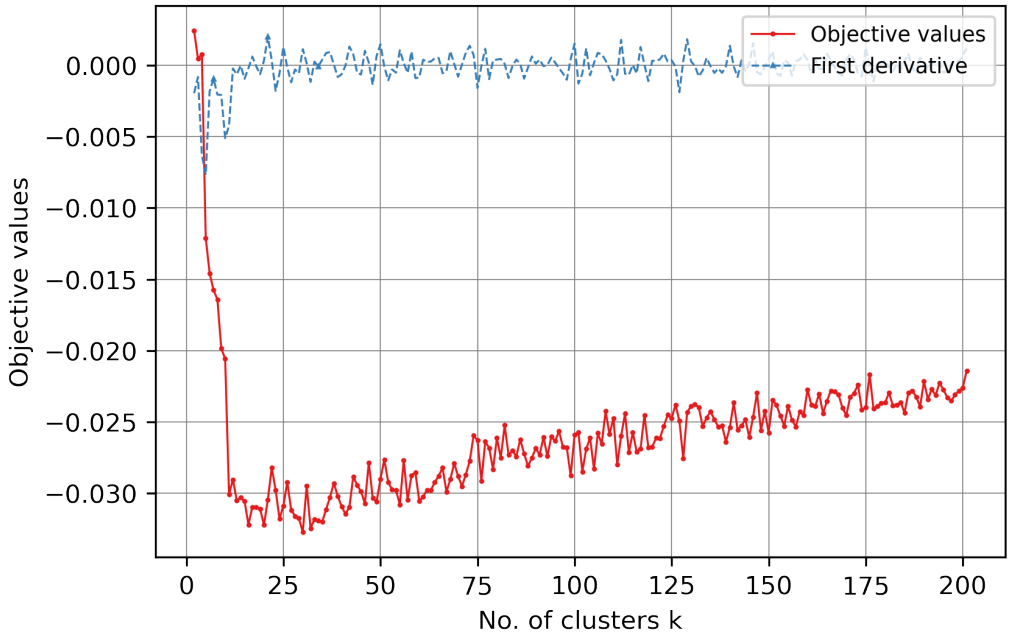


Figure 3.2.: Alternating training process.

After this, we run the alternating training process mentioned in subsection 3.4.3 to determine the optimal number of clusters k. Figure 3.2 demonstrates the relation between the objective values calculated from the equation 3.23 and the number of clusters k during the alternating training phase. In this step, we aim to minimize the objective value of the equation 3.22. The blue curve indicates the first derivative of the red curve. As the number of clusters increases, the objective value rapidly decreases to the optimal (or minimum) value. We did not search for all possible solutions since too many clusters (i.e., for k > 200) do not benefit later interpretable analysis and the

coding staff.

For the most extreme case, where the number of clusters equals the number of input samples, each cluster contains only one sample (i.e., either a miscoded case or a correctly coded case). The miscoding risk for each cluster is either 0 or 1. In this case, only the high-risk clusters (i.e., clusters whose miscoding risk is equal to 1) need to be selected to review.

Based on the results shown in the Figure 3.2, a conservative model with k=16 (and the minimal objective value) is then selected. Table 3.1 shows the estimated subject flow probabilities and risk probabilities for each of the discovered clusters. The estimated risk probability for the whole population is 0.112. On top of this, several clusters of high-risk and low-risk are discovered.

Table 3.1.: Estimated probabilities

Population	No. subjects	No. miscoded subjects	No. negative samples	Subj flow probability	Risk probability
c0	5,784	934	4,850	0.176	0.162
c1	943	128	815	0.029	0.137
c2	503	88	415	0.015	0.176
c3	275	44	231	0.008	0.162
c4	492	196	296	0.015	0.399
c5	1,236	114	1,122	0.038	0.093
c6	951	193	758	0.029	0.204
c7	346	53	293	0.011	0.155
c8	156	31	125	0.005	0.203
c9	910	286	624	0.028	0.315
c10	1,654	93	1,561	0.050	0.057
c11	10,556	461	10,095	0.321	0.044
c12	49	23	26	0.001	0.471
c13	6,009	206	5,803	0.183	0.034
c14	1,078	473	605	0.033	0.439
c15	1,914	353	1,561	0.058	0.185
whole population	32,856	3,676	29,180	1	**0.112**

Table 3.2 shows the results of mixed integer programming (MIP) model. We prescribe medical reviews to predicted positive samples (i.e., TP, FP) and assume medical reviews could completely remove underlying coding errors. Subsequently, the efficiency of the MIP model can be measured by the efficiency score, i.e. EFF = RMR / PMR, EFF $\in [0, \infty)$, where RMR $\in [0, 1]$ and PMR $\in [0, 1]$ denote Reduction of Miscoding Rate and Prescribed Medical Reviews, respectively. The higher the EFF score, the better the efficiency.

Based on the results shown in Table 3.2, there is an evident trade-off between the number of subjects reviewed and the false negative probability p(FN). A false negative

case is a miscoded case that is not assigned a medical review. Intuitively, the more medical reviews are assigned, the fewer false negative cases we retain since prescribing medical reviews to all subjects will ensure to cover all miscoded subjects but at the expense of an increased number of unnecessary medical reviews.

Table 3.2.: Results of the integer programming model

λ	p(TP)	p(FP)	p(TN)	p(FN)	RMR(%)	PMR	EFF
1	0	0	0.88782	0.11217	0	0	0
0.5	0	0	0.88782	0.11217	0	0	0
0.25	0	0	0.88782	0.11217	0	0	0
0.12	0	0	0.88782	0.11217	0	0	0
0.11	0.0023359	0.0096862	0.87814	0.10983	0.23131	0.012022	19.24050
0.10	0.014399	0.018409	0.86941	0.097772	1.43961	0.032809	43.87755
0.09	0.022433	0.035212	0.85261	0.089739	2.24007	0.057645	38.85955
0.08	0.032384	0.059866	0.82796	0.079787	3.23228	0.09225	35.037941
0.07	0.042334	0.10253	0.78528	0.069837	4.21145	0.14487	29.13865
0.06	**0.052396**	**0.14924**	**0.73858**	**0.059776**	**5.22583**	**0.20163**	**25.91698**
0.05	0.062192	0.21446	0.67335	0.049979	6.20891	0.27666	22.44224
0.04	0.072659	0.25866	0.62916	0.039512	7.25590	0.33132	21.89968
0.03	0.082469	0.30650	0.58132	0.029702	8.23289	0.38897	21.16588
0.02	0.092262	0.51721	0.37061	0.019910	9.20684	0.60947	15.10611
0.01	0.10231	0.66296	0.22486	0.0098574	10.20818	0.76527	13.33916
0	0.11217	0.88782	0	0	11.18821	1	11.18821

Finally, with $\lambda = 0.06$, we ration medical reviews to 20.16% of subjects and yield a 46.71% of reduction in miscoding rate, which shows the efficiency of ACBI. The RMR can be further increased at the cost of an increase in the false-positive rate p(FP), i.e., an increase in the number of correctly coded subjects who are given additional medical reviews.

Interpretability enhancement

To interpret the obtained results, we plot centroids of the four subgroups with the highest risk probabilities in Figure 3.3, i.e., $p(r|c4) = 0.399$, $p(r|c9) = 0.315$, $p(r|c12) = 0.471$, $p(r|c14) = 0.439$. Intuitively, these centroids are well separated according to the descriptive features (or descriptive features). For instance, the centroid of subgroup c12 depicts a virtual male subject older than 71 who has a short-term stay (i.e., 1-2 nights) in the medical department.

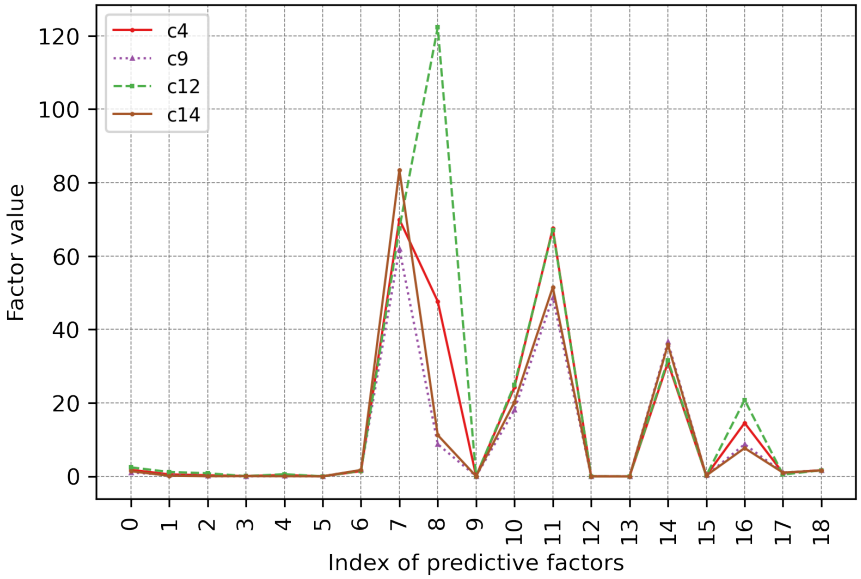


Figure 3.3.: cluster centroids

Table 3.3 shows the top three frequent patterns for subgroup c14 - the cluster containing the most data samples among the four clusters c4, c9, c12, c14. The generated rules are consistent with the profile provided by the subgroup c14 in Figure 3.3. Furthermore, subgroups whose true positive probability is significantly higher than others usually contain several strong rules (i.e., rules with a high confidence value) related to miscoding. Otherwise, the generated patterns are usually less significant. This implies that at-risk subgroups may share common characteristics. Interestingly, ACBI identifies some meaningful and frequent patterns of subgroups, even if it has no medical knowledge.

Table 3.3.: Top 3 frequent patterns for $Y = \{c_{12}\}$

ant.	cons.	sup.	conf.
{surgery_1_2, albumin_35_inf, wgt_evol_1mo_-5%_inf}	{miscoded}	0.0123	0.724
{missing_value_existed, albumin_35_inf, wgt_evol_1mo_-5%_inf}	{miscoded}	0.0121	0.716
{obstet_1_2, albumin_35_inf, wgt_evol_6mos_-15%_inf}	{miscoded}	0.0119	0.719

3.5.4. Summary

This section focuses on the modeling of miscoding risk for miscoding screening and elimination. We explore interpretable approaches such as clustering, Bayesian inference, and association rule mining to increase interpretability. Miscoding correction

involves not only a procedure to identify at-risk subjects but also provides pieces of evidence that possibly reflect underlying reasons resulting in miscoding. The cause of miscoding is related to not only shared medical factors but also shared demographics that probably reflect several interactions between the subjects and the medical units or physicians.

The numerical results show a 46.71% reduction in the miscoding rate. The results imply that the proposed approach is promising and is capable of reducing potentially preventable miscoding and subsequently leading to the reduction of hospital costs.

3.6. Case study: risk modeling for readmission prevention

3.6.1. Research context

Unplanned readmission to the emergency department (ED) is an inherent and increasing problem in the healthcare domain. It has many negative impacts, such as: (i) a negative effect on subjects' health, (ii) a strain on medical resources allocation, and (iii) a contribution to potentially avoidable costs. Verhaegh et al. [100] show that short-term (30 days or less) readmissions are challenging to prevent. Only high-intensive interventions show effectiveness for the readmission rate reduction. In order to eliminate hospital readmissions, some high-intensity interventions are suggested by Verhaegh et al. [100], including (i) frequent communications between the reception department and the primary healthcare provider, (ii) healthcare coordination by a registered nurse, (iii) follow-up visits within 72 hours of discharge. Therefore, to reduce the readmission rate, the main challenge is determining the appropriate subpopulation to which supplemental health interventions should be given.

A study [101] in The U.S. indicates that the readmission rate is related to hospital quality but independent of patients. Two studies [102, 103] reveal that readmission risk prediction is an institutional-specific problem and point out that the hospital is a significant discriminant factor for the prediction on a multi-hospital dataset. Hebert et al. [104] performed readmission prediction on historical data of a single hospital and yielded poor accuracy, which suggests that risk factors and at-risk subjects may vary over time. Practitioners need to continuously adjust and update the prediction model to react to the variation of local trends. In this case, the dynamic nature of Bayesian inference shows unique advantages when dealing with problems in continuous time.

In [80], various techniques for predicting early hospital readmission are benchmarked. The deep neural networks and ensemble methods yield the highest accuracies. These sophisticated models usually outstrip interpretable models such as Logistic Regression (LR), Generalized Linear Models (GLMs), Generalized Additive Models (GAMs), Naïve Bayes, single decision trees, and rule-based models. Numerous techniques to model readmission risk were also investigated. Tonkikg et al. [105] employ the Malnutrition Universal Screening Tool ('MUST') and some well-known

risk factors of readmission for readmission risk identification. Caruana [25] proposes a generalized linear model (GAM) for readmission risk modeling and prediction and emphasis the importance of model transparency.

The readmission problem is often modeled as a binary classification problem whose objective is to predict whether a subject will be readmitted or not. Nevertheless, it ignores the uncertainty of readmission and usually results in poor prediction quality. As preliminary studies presented in section 3.6.3, we initially treat the problem as a binary classification problem. The regression and probability-based models generalize well on the test set, while the classification models are usually overfitting and perform worse than the former. The results demonstrate the value of modeling the uncertainty of readmission. Therefore, in this study case, we aim to model the readmission risk to solve the health intervention rationing problem instead of the straightforward prediction of the readmission.

Specifically, the problem is defined as a chance-constrained health intervention rationing problem: at-risk subjects are targeted and given supplemental health interventions [100], while the remaining subjects are treated as outpatients. The objective is to profile subjects, identify at-risk subjects, and select specific groups of subjects to which additional health interventions are recommended while addressing the unknown subject flow probabilities and the unknown risk probabilities. We leverage the approach proposed in section 3.4 and investigate its efficiency on a real-life readmission data set. Results are promising and show the method could lead up to a 34.42% reduction in the readmission rate.

3.6.2. Data description

In this study case, we utilize a real-life data set extracted from the University Hospital of Saint-Étienne in France. This study focuses on subjects with emergency stays in the ED. After filtering any records with at least one missing value, a data set with 90,996 de-identified records was formed. The data set is highly imbalanced and only includes 4.678% readmitted subjects.

The feature engineering step is constructed and conducted to organize the available information for all subjects uniformly.

- **Handling missing value**: in order to ensure the reliability of the results, records with any missing value are removed.
- **Removing redundant features**: some coarse-grained features such as 'arrived from home (binary)' and 'arrived from medical units (binary)' are removed since the information in these features is already contained in the fine-grained feature 'admission mode (categorical)'.
- **Forming the complex features**: original binary features with a sequential or order relationship (i.e., 'age_1_5 (binary)', 'age_6_15 (binary)', etc.) are integrated into a single complex feature such as 'age (ordinal)' and 'length of stay (ordinal)'.

Original ordinal features such as 'admission mode (categorical)', and 'discharge mode (categorical)' without any sequential relationship are then encoded as multiple binary features by using the one-hot encode method.

- **Splitting the dataset**: to facilitate assessment, 50% of data of the dataset form the training set, and the remaining 50% of data forms the test set, which is exclusively used for the performance evaluation.

After performing the feature engineering step mentioned above, we retain 34 categorical descriptive features describing each subject. Each subject is assigned a binary label (next72present): one if the subject is readmitted to the ED within 72 hours of discharge and zero otherwise. The descriptive features, along with statistics and detailed descriptions corresponding to each feature of the pre-processed dataset, are shown in table 3.4. In the table, ICD-10-FR (international classification of diseases, 10-th revision in French) [5] is a commonly used diagnostic coding system for epidemiology and healthcare management in French. The column 'Values (distribution on P)' shows the different values of a feature and the distribution of these values across the population, while 'Values (distribution on P^+)' denotes different values of a feature and the distribution of these values among positive (or readmitted) samples.

The database of CHU Saint-Etienne also contains medical records and lab test results. However, these features are usually unavailable in the initial stage when a subject arrives at the ED of CHU Saint-Etienne, but rather after medical treatment and consultation. Thus, using these posterior features are not appropriate for prevention purpose. Therefore, these features are not taken into account during the data extraction process.

According to the statistics marked in bold in Table 3.4, intuitively, there are some significant characteristics. For instance, in the data set, 58.49% of readmitted subjects are males. Subjects with a main diagnosis of '21' seem more likely to be readmitted (10.01%) compared to the readmission rate of the whole population (4.31%). Traumatic injury and poisoning is the most frequent (44.6%) ICD principal diagnosis. Besides, most subjects (97.16%) arrived in the ED from home. In addition, after medical treatment at the ED, most subjects (78.09%) return home. A more in-depth study of the interactions between descriptive features leads to developing algorithms for at-risk subjects detection, which is introduced in the next subsection.

Table 3.4.: Descriptive features and statistics of the readmission dataset

Feature	Type	Values (distribution on P)	Values (distribution on P^+)	Description
gender	binary	1 (**54.70%**), 2 (45.30%)	1 (**58.49%**), 2 (41.51%)	The gender of subjects. 1 (male), 2 (female)

Table 3.4 continued from previous page

Feature	Type	Values (distribution on P)	Values (distribution on P+)	Description
age	ordinal	0 (1.68%), 1 (7.37%), 2 (9.91%), 3 (18.46%), 4 (30.19%), 5 (16.47%), 6 (7.19%), 7 (8.37%).	0 (2.51%), 1 (8.15%), 2 (7.85%), 3 (18.23%), 4 (33.05%), 5 (16.98%), 6 (6.55%), 7 (6.67%).	The age of subjects. 0 (newborn), 1 (1-5 year-old), 2 (6-15 year-old), 3 (16-25 year-old), 4 (26-50 year-old), 5 (51-70 year-old), 6 (71-80 year-old), 7 (80 and above).
ICD main disgnosis	binary	0 (9.83%), 1 (1.61%), 2 (0.23%), 3 (0.32%), 4 (0.54%), 5 (2.84%), 6 (1.42%), 7 (0.48%), 8 (0.66%), 9 (3.19%), 10 (3.39%), 11 (3.89%), 12 (2.11%), 13 (2.76%), 14 (3.08%), 15 (0.45%), 16 (0.038%), 17 (0.035%), 18 (14.07%), 19 (**44.62%**), 20 (0.085%), 21 (**4.31%**), 22 (0.024%).	0 (10.85%), 1 (1.88%), 2 (0.71%), 3 (0.14%), 4 (0.40%), 5 (3.74%), 6 (1.10%), 7 (0.23%), 8 (0.35%), 9 (1.22%), 10 (2.47%), 11 (2.70%). 12 (4.67%), 13 (2.28%), 14 (3.64%), 15 (0.80%), 16 (0.047%), 17 (0.047%), 18 (15.18%), 19 (**37.89%**). 20 (0.14%), 21 (**10.01%**), 22 (0.14%).	The 22 main diagnosis categories of ICD-10-FR 0 (no data), 1-22 (ICD category codes)

Table 3.4 continued from previous page

Feature	Type	Values (distribution on P)	Values (distribution on P+)	Description
admission mode	binary	1 (1.33%), 2 (1.55%), 3 (**97.13%**).	1 (2.63%), 2 (0.92%), 3 (**96.45%**).	The department that a subject comes from. 1 (another medical unit of the same legal entity), 2 (another medical unit of another legal entity), 3 (domicile, substitube such as a social medical institution).
discharge mode	binary	1 (19.70%), 2 (2.19%), 3 (**78.09%**), 4 (0.014%).	1 (12.99%), 2 (1.93%), 3 (**85.08%**), 4 (0%).	The department to which a subject is transferred after the medical treatment. 1 (a medical unit in the same legal entity) 2 (a medical unit in another legal entity) 3 domicile (including discharge without an agreement, home care, social medical institution, accommodation center, etc.)
severity	ordinal	1 (16.26%), 2 (69.91%), 3 (12.51%), 4 (0.94%), 5 (0.35%).	1 (18.51%), 2 (70.26%), 3 (10.85%), 4 (0.33%), 5 (0.047%).	The severity of illness range from 1 (lowest) to 5 (highest)
LOS	ordinal	0 (21.41%), 1 (25.81%), 2 (15.95%), 3 (10.22%), 4 (6.77%), 5 (4.46%), 6 (15.39%).	0 (21.73%), 1 (25.96%), 2 (16.51%), 3 (10.55%), 4 (7.33%), 5 (4.25%), 6 (13.67%).	Length of stay of an admitted subject (in hours) 0 (less than 1 hour), 1 (1 hour), 2 (2 hours), 3 (3 hours), 4 (4 hours), 5 (5 hours), 6 (6 hours and above).

Table 3.4 continued from previous page

Feature	Type	Values (distribution on P)	Values (distribution on P+)	Description
previous72present	binary	0 (95.22%), 1 (**4.78%**).	0 (88.28%), 1 (**11.72%**).	A feature indicating whether a subject was present in the ED in the previous 72 hours.
next72present	binary	0 (95.32%), 1 (4.68%).	0 (0%), 1 (100%).	The target variable indicating whether a subject readmitted to the ED within 72 hours of discharge.

3.6.3. Experimental results

Analysis of preliminary experimental results

In Table 3.5, we demonstrate the results of various supervised machine-learning techniques on the data set. In the experiments, only training data is balanced using a random oversampling strategy; the testing data remain imbalanced. Hyper-parameters are determined using a randomized search method with 5-fold cross-validation. Although the supervised techniques yield only moderate accuracy, the results imply the possible existence of subgroups, i.e., (i) the SLR model yields higher scores (i.e., AUC and F1-score) than the LR model, (ii) a simple ensemble method that trains a local classifier per subgroup performs better than a single global classifier. When using the simple ensemble method, the scores of multiple classifiers are calculated by the sum of the weighted scores of each cluster, where the weights are the ratio of the cluster size to the size of the entire test set. In this case study, local classifiers are trained on four predicted sub-clusters separately. Accordingly, the hyper-parameters of each local classier are totally different.

Table 3.5.: Prediction accuracy of supervised algorithms

Method	Accuracy	Precision	Recall	AU-ROC	F1-score
LR	0.683	0.0575	0.382	0.538	0.0978
SLR	0.684	0.0608	0.400	0.547	0.103
Simple ensemble of SLRs (c=4)	0.684	0.0660	0.450	0.567	0.115
GAM	0.687	0.0608	0.396	0.547	0.103
Naive Bayes	0.390	0.0540	0.722	0.548	0.101
Decision tree	0.725	0.0604	0.332	0.583	0.102

When a prediction model is used in an industrial application, optimizing the model either for precision (which allows the model independently and accurately to determine readmission on a small fraction of subjects) or for recall (which allows the model

redundantly predicts possible at-risk subjects as positive samples to cover subjects who are actually readmitted (True positive) as many as possible) is an inevitable step. A model with a high recall can be used in a setup to assist doctors in making decisions for specific subjects. In this case study, a relatively high recall is preferred. The highest recall score of 0.772 is achieved by the Naïve Bayes classifier.

Although these models are not capable of accurately identifying at-risk subjects, it allows us to explore interactions between descriptive features. For instance, the SLR model shows that ICD principal diagnosis, admission, and discharge mode are the most critical discriminant features. To increase the prediction interpretability and accuracy, these preliminary studies are complemented by experiments with ACBI. The application of this method to the readmission dataset is explained in the next section.

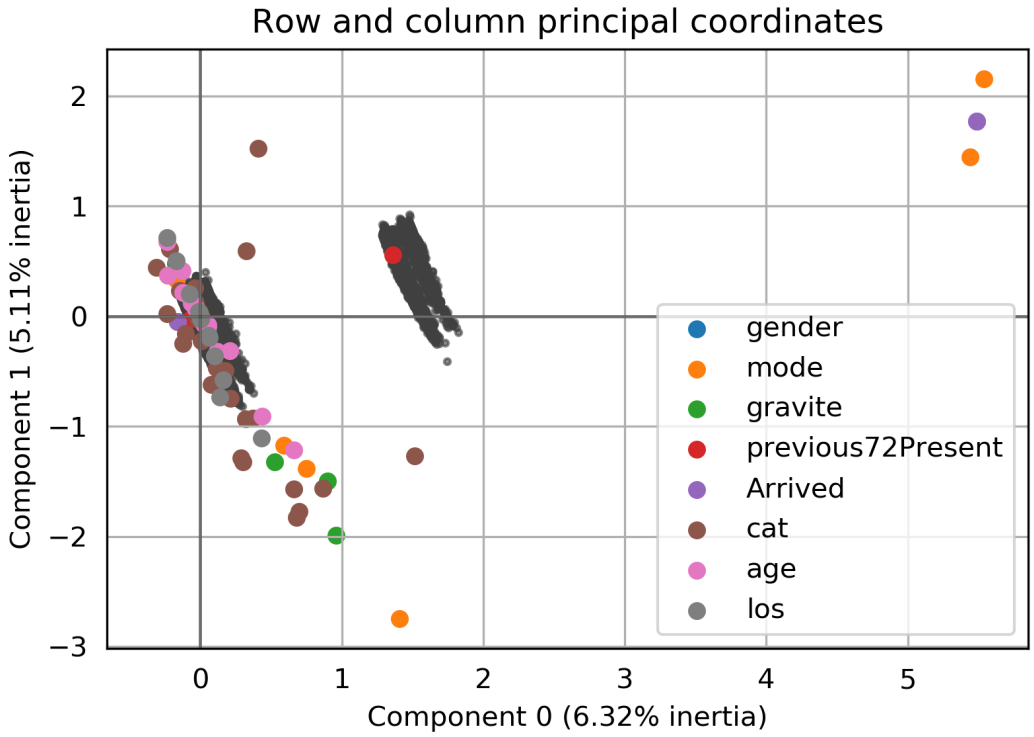


Figure 3.4.: 2D visualization of the data set

Since each record is composed of categorical features, we first apply the Multiple Correspondence Analysis (MCA) [106], a dimensionality reduction method, to uncover the underlying structure of 34-dimensional categorical dependent features (without target variable) of the dataset. First of all, All ordinal features are one-hot encoded to generate an indicator matrix. Based on the generated index matrix, MCA calculates the chi-square distances between one-hot encoded descriptive features (columns) and between subjects (rows) to discover associations. Data are represented as points in a 2-dimensional Euclidean space called 'map'. Components are sorted in descending order according to calculated variances. The first component is the most critical dimension. The second component is the second essential component, and so on.

In Figure 3.4, one-hot encoded descriptive features are marked by big colored points and subjects by small gray points. The distance between any subject points or feature points measures their similarity (or dissimilarity). At the top right of the figure, since the feature 'mode' (especially admission mode) is a superset of the feature 'arrived' (including 'arrived from home' and 'arrived from medical unit'), the 'mode' feature is close to the feature 'arrived'. Two red points with a significant distance show that subjects who were present previously and those who were not present had a big difference. Intuitively, the underlying data structure of subjects (gray points) implies the possible existence of population subgroups.

We draw a hyper-parameter curve for the alternating training process. In the curve, the number of subgroups k decreases as hyper-parameter γ increases until k reaches one. The key points where the first derivative value changes abruptly are A, B, and C. The corresponding k are 4, 37, and 89, respectively. The γ is then selected from these points. As mentioned in 3.4.3, a trade-off should be made, a relatively large γ guarantees the convergence of posterior distributions, while a relatively small γ tends to identify more at-risk and risk-free subjects. We apply the integer programming model to these 3 cases, and a relatively accurate and conservative model at point B is selected.

Table 3.6.: Results of integer programming model

λ	P(TP)	P(FP)	P(TN)	P(FN)	RRR (%)	PHI (%)	Eff
0.047	0	0	0.95	0.046	0	0	0
0.046	6.3e-4	1.9e-3	0.95	0.046	1.33	0.27	4.99
0.040	6.4e-3	5.9e-2	0.89	0.040	13.56	6.45	2.10
0.030	0.016	0.24	0.72	0.030	34.42	25.18	1.37
0.020	0.026	0.40	0.55	0.020	56.43	42.63	1.32
0.010	0.036	0.61	0.35	9.9e-3	77.62	64.00	1.21
0.005	0.041	0.76	0.19	4.9e-3	88.38	80.14	1.10
0.001	0.045	0.87	0.082	9.8e-4	96.68	91.71	1.05
0	0.046	0.95	0	0	100	100	1.00

Table 3.6 shows the results of integer programming on point B. We prescribe health interventions to predicted positive samples (i.e., TP, FP) and assume health interventions could completely prevent potential readmissions (i.e., TP). Subsequently, the efficiency of the health intervention rationing problem can be measured by the efficiency score, i.e. Eff = RRR / PHI, Eff $\in [0, \infty)$, where RRR $\in [0, 1]$ and PHI $\in [0, 1]$ denote Readmission Rate Reduction and Prescribed Health Intervention in percentage, respectively. The higher the Eff score, the better the efficiency.

With $\lambda = 0.03$, we ration health interventions to 25.18% subjects and yield a 34.42% reduction in readmission rate, which shows the efficiency of ACBI. The RRR can be further increased at the cost of an increase in the false-positive rate, i.e., an increase in the number of risk-free subjects who are given additional health interventions.

Interpretability of results

To interpret the obtained results, we plot the centroids of the first three subgroups in Figure 3.5. Intuitively, these centroids are well separated according to descriptive features. For instance, the centroid of subgroup three depicts a virtual male subject with an ICD principal diagnosis of 19^{th} class, which indicates an injury, a poisoning, or other consequences of external causes. He has a short stay in the ED. He comes from another medical unit and then returns to his home after discharge.

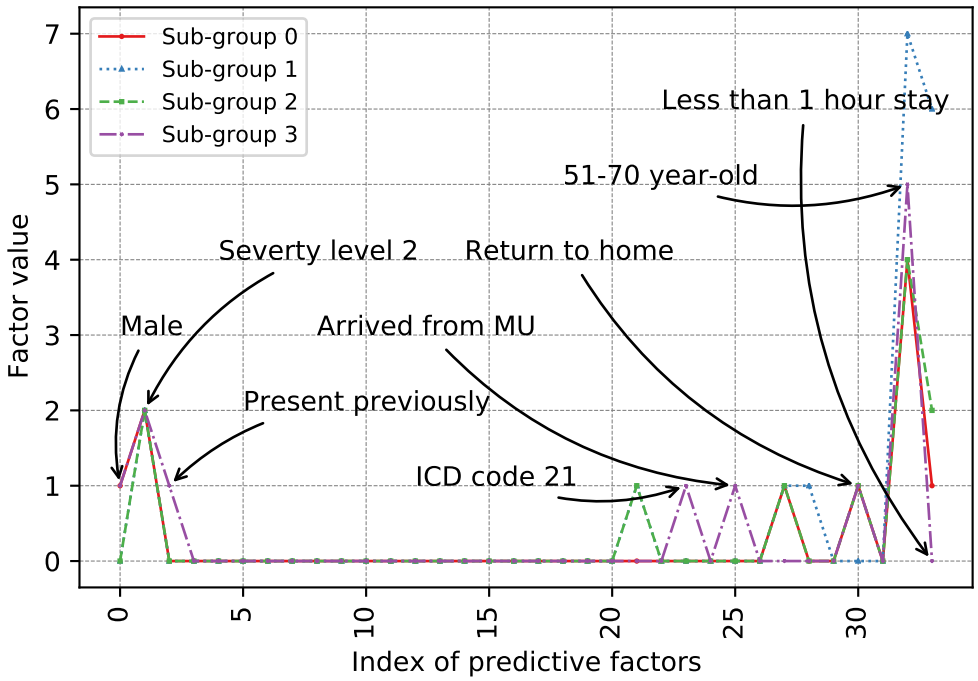


Figure 3.5.: Centroids of the first three subgroups

Intuitively, the centroids are well separated according to descriptive features. However, to extend the interpretability of the results, we aim to detect association rules within each subgroup that impact the readmission risk. To this end, an association rule mining algorithm named FP-Growth [98] is leveraged. FP-Growth is frequently used to identify patterns in data sets and is used in this case study to identify risk factors of the discovered subgroups. FP-Growth is applied to the training set and test set separately. Only patterns that are present in both sets are accepted as final association rules. The generated patterns can be provided to physicians for more interpretability.

Table 3.7 shows the top four frequent patterns for subgroup three. The first one indicates that 62% of males between 51 and 70, with a severity 2 code from chapter 21 of the ICD coding convention, have been readmitted. Furthermore, subgroups whose true positive probability is significantly higher than others usually contain several strong rules (i.e., rules with a high confidence value) related to readmission.

Otherwise, the generated patterns are usually less significant. This implies that at-risk subgroups may share common characteristics. Interestingly, ACBI identifies some meaningful and frequent patterns of subgroups, even if it completely has no medical knowledge.

3.6.4. Summary

This section focuses on the modeling of readmission risk for readmission prevention. We explore interpretable approaches such as clustering, Bayesian inference, and association rule mining to increase interpretability. Readmission prevention involves not only a procedure to identify at-risk subjects but also providing pieces of evidence that possibly reflect underlying pathologies and reasons resulting in a readmission. The cause of readmission is related to not only shared medical factors that possibly reflect specific underlying pathologies but also shared demographics that probably reflect several interactions between the subjects and the medical units or physicians.

Also, the limited number of descriptive features leads to moderate results for most classification methods. One of the future works would be data augmentation, especially the increase in the number of descriptive features by taking into account more subject characteristics related to readmissions. Ablation studies should be conducted to test the contribution of each added descriptive feature to the accuracy score.

Table 3.7.: Top four decision rules for subgroup 3

Antecedent	Consequent	Antecedent support	Consequent support	Support	Confidence	Lift	Leverage	Conviction
male, icd_21, age_51_70, severity_2, discharge_home.	next72Present	0.0197	0.172	0.0123	0625	3.634	0.00890	2.208
male, icd_21, age_51_70, discharge_home, previous72Present.	next72Present	0.0172	0.172	0.00983	0.571	3.322	0.00687	1.932
male, icd_21, age_51_70, severity_2, previous72Present.	next72Present	0.0172	0.172	0.00983	0.571	3.322	0.00687	1.932
male, icd_21, age_51_70, severity_2, previous72Present, discharge_home.	next72Present	0.0172	0.172	0.00983	0.571	3.322	0.00687	1.932

3.7. Conclusion and perspectives

This chapter defines a transparent and dynamic framework to address the health intervention rationing problem. In addition, the proposed approach is applied, evaluated, and validated on two different real-life study cases. For instance, proper rationing of medical reviews of patients' EHRs can solve the hospital miscoding problem (Section 3.5), and appropriate allocation of post-discharge follow-ups can prevent the hospital readmission problem (Section 3.6). More specifically, We evaluate the risk of a given population and give at-risk patients (either risk of readmission or risk of miscoding) a health intervention (either a health intervention to avoid readmission - post-discharge follow-up, or a health intervention to prevent miscoding - review of patient EHRs).

Our main contribution is the proposition of a novel approach that discovers patterns, identifies at-risk subjects, and determines the appropriate set of health interventions for at-risk subjects. Our research highlights certain essential insights to discover patterns or clues for at-risk subjects. This approach could be easily ported to the healthcare management system and is valuable when dealing with large-scale databases.

The proposed approach faces some limits. In future work, there is an interesting perspective on embedding the alternating training process and integer programming model into one single process. This way, not only the integer programming model could define the subgroups to which health interventions should be given, but it could do it while shaping the subgroups at the same time, i.e., by optimally defining k the number of subgroups to cluster with the k-modes algorithm. This could lead to a significant increase in the robustness of ACBI.

So far, Part I includes two semi-automated methods for profiling and identifying hospital miscoding for coding error correction. In part II, we will introduce two automated approaches for the hospital miscoding problem accounting for the experience and knowledge gained from Part I. The two automated approaches presented in Part II can identify and correct coding errors without user intervention.

Part II.

Hospital miscoding correction budget allocation

4. A two stage approach for optimizing miscoding correction budget

Summary

Abstract

In this chapter, we develop a two-stage optimization approach for the hospital miscoding problem by profiling and allocating medical reviews to miscoded cases. The primary goal is to assess the efficiency of the proposed approach, mainly through the overall savings observed on unnecessary medical reviews. The secondary goal is to provide counterfactual explanations to medical coding staff and improve the quality of their coding practices. In the process of profiling miscoding subtypes (see Chapter 2), the miscoding subtype is characterized by a set of similar subjects instead of a set of features. However, it is better to introduce a set of features (called correction set) to characterize miscoding behaviors and provide causes of coding errors to medical coders. To deal with this shortcoming, this chapter introduces the correction set for profiling miscoding behaviors and the concept of decision lists for coding error correction. A new mathematical formulation is introduced, and a two-stage approach is presented. The proposed approach combines clustering for subject profiling, mixed integer programming (MIP) for medical review allocation, and counterfactual explanation for more interpretability. A case study is presented to deal with the hospital miscoding problem and evaluate the performance of the proposed approach.

Keywords: Hospital miscoding, clustering, coding review budget optimization, mathematical programming.

Résumé du chapitre

Dans ce chapitre, nous développons une approche d'optimisation pour le problème de mauvais codage hospitalier en établissant le profil de mauvais codage et en allouant les examens médicaux aux cas mal codés. L'objectif principal est d'évaluer l'efficacité de l'approche proposée, notamment via les économies globales réalisées sur les examens médicaux inutiles . L'objectif secondaire est de fournir des explications contrefactuelles au personnel du codage médical et d'améliorer la qualité de leurs pratiques de codage. Dans le processus de profilage des sous-types de mauvais codage (voir chapitre 2), le sous-type de mauvais codage est caractérisé par un ensemble de sujets similaires au lieu d'un ensemble de caractéristiques. Or, il est préférable d'introduire un ensemble de caractéristiques (appelé ensemble de correction) pour caractériser les comportements de mauvais codage et fournir les causes des erreurs de codage aux codeurs médicaux. Pour remédier à cette lacune, ce chapitre présente l'ensemble de correction pour le profilage des comportements de mauvais codage et le concept de listes de décision pour la correction des erreurs de codage. Une nouvelle formulation mathématique est introduite, et une approche en deux étapes est présentée. L'approche proposée combine le clustering pour le profilage des sujets, la programmation mixte en nombres entiers (MIP) pour l'allocation des examens médicaux et l'explication contrefactuelle pour une meilleure interprétabilité. Une étude de cas est présentée pour traiter le problème de l'erreur de codage dans un hôpital et évaluer la performance de l'approche proposée.

Mots-clés: Miscodage hospitalier, algorithme de regroupement, optimisation du budget de la révision du codage, programmation mathématique.

4.1. Introduction

Hospital coding is defined as the process of translating relevant information stored in patient Electronic Health Records (EHRs) into pre-defined medical codes defined in standardized coding guidelines such as ICD-10-FR (diagnosis) and CCAM (medical acts or procedures). In a French teaching hospital (i.e., the University Hospital of Saint-Etienne, CHU-SE), the medical coding task is mainly completed by hand by a professional coding team from the Unit of public health and medical information (SSPIM). Nevertheless, due to the complexity of the coding process, the coding practices in SSPIM are often subjective and error-prone.

This chapter is our collaboration with CHU-SE. In France, the provision of appropriate reimbursement of health services via the activity-based funding model (T2A) depends on accurate medical coding. To improve the accuracy of hospital coding, we proposed a code correction approach consisting of four steps: (i) identifying discrepancies in medical code assignment; (ii) constructing homogeneous miscoding subgroups in terms of patient characteristics; (iii) assigning a set of medical reviews to miscoded subjects for eliminating coding errors presented; (iv) detecting changes in feature values for correcting discovered miscoded cases. It is worth noting that the proposed approach is also concerned with contrastive and counterfactual explanations of medical miscoding. While other methods aim to answer the question, "Why is a given subject miscoded?". Our approach seeks to answer the question, *"what does a medical coder need to do to correct a given miscoded case?"*.

One reason why CHU-SE undertakes medical coding audits is because of observed mismatches between the assigned medical codes and as documented in patient EHRs. In CHU-SE, the coding audit task is periodically performed by several medical coders in SSPIM. Given the intrinsic complexity of the audit task and the heterogeneity of the subpopulation to audit, the current practice of SSPIM has been applying a simple heuristic to select a subpopulation and check all the contents of patient EHRs for each subject in this subpopulation. Such a practice is inefficient and raises challenges to the problem of the coding audit. Many miscoded cases are not taken into account, while many unnecessary medical reviews are allocated to properly coded cases. In addition, even if a subject is miscoded, reviewing all content of his/her EHRs is not necessary. Such a conservative practice leads to unnecessary waste of efforts and time of medical coders and subsequently results in an increase in the cost of the entire healthcare system.

The intuition behind our approach is that the current practice of SSPIM is leveraging a coarse-grained heuristic (consisting of a few rules of thumb) to target a general subpopulation; but, with appropriate profiling of coding errors, there is a great chance to decrease the miscoding rate without increasing the efforts of medical reviewers. In other words, *we assume that the optimal allocation of medical reviews depends on the personalized profiling of coding errors.*

We aim to propose an optimization approach to allocating the medical review budget by profiling coding errors. By incorporating some practical considerations,

we make the following assumptions about the proposed approach: (i) all raw data and relevant information used to assign medical codes are complete and included in patients' EHRs; (ii) for each miscoded subject, there exists a finite set of features (called, a correction set) in the patient's EHRs that can remove and explain the miscoding; (iii) each correction set corresponds to specific miscoding behavior and is considered the underlying cause of the corresponding miscoding behavior. In this study, hypotheses (ii) and (iii) are based on hypothesis (i). The presence of missing variables would violate our assumptions.

This chapter is organized as follows. The problem definition is presented in Section 4.2. Section 4.3 introduces the proposed approach and Section 4.4 presents our experimental results on a real-life study case. In Section 4.5, we conclude our paper with a brief conclusion.

4.2. Problem definition

This chapter addresses the problem of hospital miscoding correction for a given set of inpatient stays $P = \{p_i\}_{i=1,2,\ldots,|P|}$. Each subject $p_i = (x_i, y_i)$ is characterized by a $z, z \in \mathbb{Z}^+$ dimensional feature vector, i.e., $x_i \in \mathbb{R}^z$ and $\mathbb{R}^z = \left(f_1, f_2, \ldots, f_z\right)$, and is labeled with a medical code $y_i \in Y$, $Y = \{\text{code}_1, \text{code}_2, \ldots, \text{code}_{|Y|}\}$. The feature set is indicated by $\phi = \{f_1, f_2, \ldots, f_z\}$.

The assigned labels y are medical codes confirmed by the PMSI (i.e., Le Programme de Médicalisation des Systèmes d'Information, in French) and are considered the gold standard for hospital service reimbursement. In this study, the given code y_i is checked against a coding guideline [5] issued by the HAS (i.e., Haute Autorité de Santé, in French). Such a coding guideline is defined on a feature subspace $F = \{f_1.f_2, \ldots, f_{z'}\} \subset \phi, z' \leq z, z' \in \mathbb{Z}^+$ and is denoted by the function: $\text{rec} : \{f_1.f_2, \ldots, f_{z'}\} \rightarrow Y$. Given a subject $p_i = (x_i, y_i)$, the mismatch between the reassigned code $\text{rec}(x_i)$ and the raw code y_i indicates the occurrence of a coding error, i.e., $\text{rec}(x_i) \neq y_i$. In this study, we concentrate on the miscoding subpopulation $I \subseteq P$ and devote ourselves to eliminating and explaining coding errors presented in the population I.

We assume that changes in the values of certain features lead to desired outcomes. In order to provide counterfactual explanation for each miscoded subject, the feature set $c \subseteq F$ is considered a **correction set** of subject p_i if $\text{rec}(x_i|_{\phi-c}) = y_i$. The term $x_i|_{\phi-c}$ projects vector x_i onto a reduced subspace $\phi - c$. By changing the values of features contained in the correction set c, we can make the reassigned code the same as the raw code, i.e., $\text{rec}(x_i|_{\phi-c}) = y_i$. As a result, for a given miscoded subject, the correction set provides coding staff with features $\{f : \forall\, f \in c\}$ that contain coding errors and need to be reviewed.

Specifically, we consider the problem of screening a set of correction sets for a heterogeneous population, where the goal is to eliminate medical coding errors, e.g., eliminating the mismatch between a re-coded code $\text{rec}(x_i)$ and the original code y_i for a given subject $p_i = (x_i, y_i)$, i.e., $\text{rec}(x_i) \neq y_i$. Given a subject population P and

a set F of descriptive features $F = \{f_1, f_2, \ldots, f_{z'}\}$, we seek to maximize the number of mismatches eliminated $\sum_{i=1}^{|P|} 1\left(\text{rec}(x_i) \neq y_i\right)$ by selecting correction sets from the feature set $c \subseteq F$, such that the number of selected features does not exceed λ, i.e., $\sum |c| \leq \lambda$.

The problem solving is in need of three consecutive steps: subject profiling, correction set evaluation, and medical review rationing. After filtering out properly coded cases, the miscoding population I can be divided into multiple disjoint subpopulations $I_i, i = \{1, 2, \ldots, k\}$, i.e., $\{I_1 \cup I_2, \ldots, \cup I_k\} = I$. The second step is to iterate over all possible correction sets $C \subseteq \{f_1, f_2, \ldots, f_{z'}\}$ for each subpopulation and compute the reduction of the coding errors s (called the score) for them. In the final step, the proposed optimization model selects a set of correction sets in order to maximize the number of coding errors reduced, i.e.,

$$\max E_s \left[\sum_{i \in I} \sum_{j \in C} s_{ij} * u_{ij} \right] \tag{4.1}$$

such that the percentage of features to review is at most λ, $\lambda \in (0, 1]$, i.e.,

$$E_s \left[\sum_{i \in I} \sum_{j \in C} |I_i| * \left|c_j\right| * u_{ij} \right] < \lambda \tag{4.2}$$

Where u_{ij} is a binary decision variable equal to 1 if the correction set c_j of the subpopulation I_i is selected for review. This is known as a medical review allocation problem under chance constraints (Equation 4.2), in which only selected features are given a supplemental review.

4.3. Methodology

4.3.1. Overview of the methodology

In this section, we aim to propose a practical solution approach for the hospital miscoding problem. To achieve this goal, we have to ensure that the proposed approach has a set of highly desirable properties in medical coding practices, i.e., accuracy and transparency. As illustrated in Figure 4.1, we first introduce a clustering algorithm for subject profiling. Subjects with similar characteristics are grouped together. We also provide a representative portrait for each discovered subpopulation for more transparency. Next, the medical coding rules in unstructured text format are converted into a set of structured decision lists. These decision lists, together with the possible correction sets, are used to formulate a score function for evaluating the goodness of correction sets. For the resource allocation optimization, with a limited number of medical reviews, we establish a tailored Mixed Integer Programming (MIP) model to select and correct coding errors to the maximum extent possible.

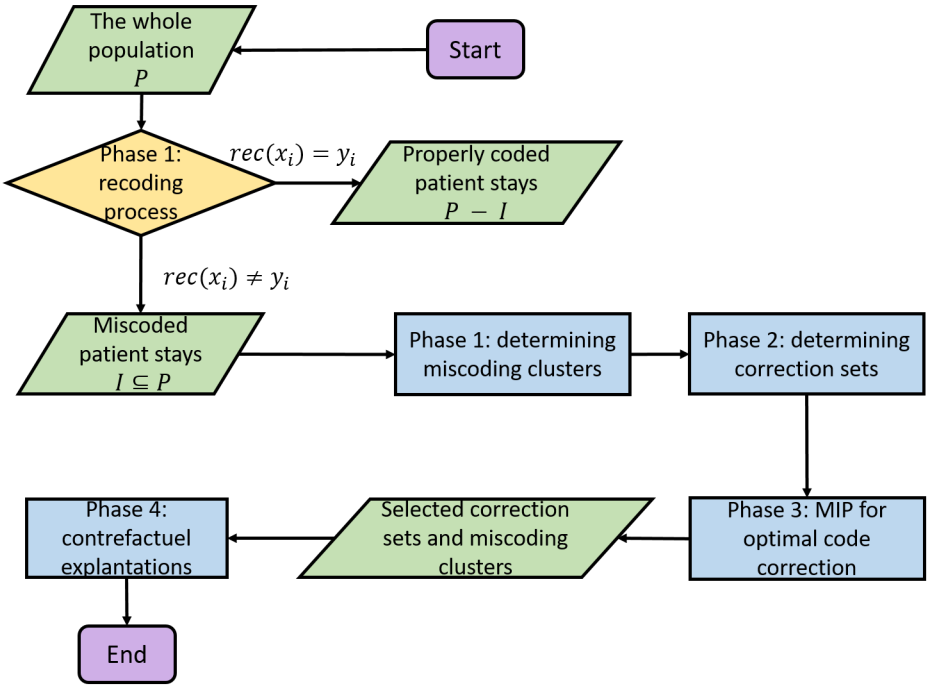


Figure 4.1.: An overview of the proposed approach

In phase 4, we measure coding errors and provide counterfactual explanations to medical coders for each miscoded case. The overall framework of the proposed approach is illustrated in Figure 4.1, and the technical details are described in the following subsections.

4.3.2. Population partitioning and subject profiling

4.3.2.1. K-mode clustering

The purpose of population partitioning is mainly to isolate mutually exclusive subgroups that possibly reflect the underlying subtypes of coding errors that are often neglected in real-time coding practices. Our dataset is extracted from the University Hospital of Saint Etienne, from which each medical coder is responsible for one or more medical units. Based on this fact, it is reasonable to assume that subjects from the same medical unit are coded by a single coder and thus may reflect certain coding behaviors of that coder. A good population partitioning is the one in which subjects in the same subgroup are similar in terms of coding (or miscoding) behaviors w.r.t their medical unit.

In case all descriptive features of samples are categorical variables, it is reasonable to leverage the k-mode clustering algorithm. We aim to partition the miscoding population I into k disjoint subgroups $I_i, i = \{1,2,\ldots,k\}$ of equal variance, i.e., $I = \{I_1 \cup I_2 \cup,\ldots,\cup I_k\}$, in which each sample belongs to the subgroup with the nearest mode. Compared to random initialization, "Cao" initialization considers both the

distance between samples and the density of samples, guaranteeing the stability of initialized centroids. Specifically, the subgroups are constructed by k-mode clustering of the demographic data $\{f_{z'}, f_{z'+1}, \ldots, f_z\}$ since the demographic data are ordinal variables or binary variables. You can refer to the Chapter 3 Section 3.4.1 for a detailed description of the k-mode clustering technique.

4.3.2.2. Determining the optimal number of clusters

In this study, silhouette coefficient s [107] is used to assess the natural (or optimal) number of clusters k in a given dataset P. The silhouette coefficient is between the worst value -1 and the best value 1, i.e., $s \in [0,1]$. The silhouette coefficient score is high if clusters are dense and well separated, which corresponds to the standard definition of a cluster. The model with the highest silhouette score is considered the best. Silhouette scores of 0 indicate that clusters are overlapped. A negative silhouette score generally indicates that some samples are assigned to wrong clusters, as assigning them to a different cluster is more appropriate. Note that the silhouette coefficient is defined on $k \in \{2,3,\ldots,|P|-1\}$, where $|P|$ denotes the number of samples in the given dataset P. The silhouette score s for a given data sample $p_i \in I_A$ (a data sample p_i of the cluster I_A) is defined as:

$$s(p_i) = \begin{cases} 1 - a(p_i)/b(p_i), & \text{if } a(p_i) < b(p_i) \\ 0, & \text{if } a(p_i) = b(p_i) \\ b(p_i)/a(p_i) - 1, & \text{if } a(p_i) > b(p_i) \end{cases} \tag{4.3}$$

where

$$a(p_i) = \frac{1}{|I_A| - 1} \sum_{p_j \in I_A, i \neq j} d(p_i, p_j) \tag{4.4}$$

$$b(p_i) = \min_{B \neq A} \frac{1}{|I_B|} \sum_{p_j \in I_B} d(p_i, p_j) \tag{4.5}$$

As defined in the equation 4.4, the mean intra-cluster distance $a(p_i)$ measures the mean distance between the sample p_i and all other samples in the same cluster I_A. $|I_A|$ denotes the number of samples in cluster I_A. $d(p_i, p_j)$ indicates the distance between data samples $p_i, p_i \in I_A$ and $p_j, p_j \in I_B$. Besides, $|I_A| - 1$ is placed in the denominator of the Equation 4.4 since we do not need to include the distance $d(p_i, p_i)$ in the sum operation $\sum$. In sample words, $a(i)$ measures the goodness of assigning the sample p_i to cluster I_A. The smaller the $a(p_i)$, the better the assignment.

As defined in the equation 4.5, the mean nearest-cluster distance $b(p_i)$ measures the mean of the distance from the sample p_i to all samples in the "neighboring" cluster I_B, $I_B \neq I_A$. The cluster I_B is considered the next nearest cluster of p_i since it is the best fit cluster for the sample p_i (i.e., considering the min operator in the above formula) except the cluster I_A. In simple term, $b(p_i)$ is defined as the smallest mean distance

between the sample p_i and all samples in the next nearest cluster I_B, of which p_i is not a member, i.e., $p_i \notin I_B$.

Overall, such an optimization process can be simply defined by:

$$s = \max_k \tilde{s}(P,k) \tag{4.6}$$

Where $\tilde{s}(P,k)$ denotes the mean value of $s(p_i,k)$ over the entire dataset P for a given number of clusters k.

4.3.3. Correction set traversal and evaluation

According to the ICD coding manual [5], each ICD code corresponds to a set of unstructured coding rules expressed in textual format. We first convert the ICD coding rule into the form of a decision rule. On top of that, we introduce the concepts of correction set c and decision list r. To evaluate the number of coding errors eliminated by a given correction set, we construct a score function s(p, c) for the downstream optimization task presented in the next section.

Definition D.4.1. Given a non-empty set F of z' distinct features, a k-combination of F is defined as all possible subsets of k distinct features of F. The number of items in k-combination can be denoted as $|C_{z'}^k| = z'!/\left[k!\left(z'-k\right)!\right], k \leq z', K \in \mathbb{N}$. The set C is defined as k-combinations for all possible k excluding the empty set, and the number of combinations contained in C is $|C| = \sum_{1 \leq k \leq z'} C_{z'}^k = 2^{z'} - 1$.

Definition D.4.2. An atomic rule a on an input domain $R^{z'}$ is a Boolean function consisting of a descriptive feature, a relation operator, and a threshold value that returns true or false. Given an input sample $p = \left(x, y\right),\ x \in R^{z'}$, we define that x satisfies the atomic rule a if $a(x)$ is assessed as true. For instance, the relational expression "age < 70" in Table 4.1 is an atomic rule.

Table 4.1.: An excerpt of the decision list r_{E44} for the code E44

	precursor		**successor**
IF	age < 70 ⋀ wgt evol in 1mo ≤ -0.05 ⋀ wgt evol in 1mo > -0.1 ⋀ wgt evol in 6mos > -0.15 ⋀ BMI > 17 ⋀ albumin > 30	THEN assigning the code	E44
ELSE IF	age ≥ 70 ⋀ CRP < 15 ⋀ wgt evol in 1mo > -0.1 ⋀ wgt evol in 1mo ≤ -0.05 ⋀ wgt evol in 6mos > -0.15 ⋀ BMI ≥ 18 ⋀ albumin ≥ 30	THEN assigning the code	E44
ELSE IF	age ≥ 70 ⋀ CRP ≥ 15 ⋀ wgt evol in 1mo > -0.1 ⋀ wgt evol in 6mos ≤ -0.1 ⋀ wgt evol in 6mos > -0.15 ⋀ BMI ≥ 18 ⋀ albumin ≥ 30	THEN assigning the code	E44
…	…	…	…
ELSE	–	do not assign the code	E44

Definition D.4.3. A decision rule $d = (\alpha, e)$ is a Boolean expression of the following form: "IF x fulfills all requirements of the precursor α, THEN the successor e.". A precursor α corresponds to the front part between the "IF" and "THEN" keywords consisting of multiple atomic rules linked by the logical operator AND, i.e.,

$\alpha = \left(a_1 \wedge a_2, \ldots, \wedge a_{|\alpha|}\right), |\alpha| \in \mathbb{N}$. A successor e is the evaluation statement after the "THEN" keyword of a given decision rule d. The first row of Table 4.1 shows an example of decision rule.

Definition D.4.4. A decision list $r : R^{z'} \rightarrow [0,1]$ on an input domain $R^{z'}$ consists of a list of decision rules with the following form: "IF x fulfills all requirements of the precursor α_1, THEN e_1; ELSE IF x fulfills all requirements of the precursor α_2, THEN e_2; ...; ELSE IF x fulfills all requirements of the precursor $\alpha_{|r|-1}$, THEN $e_{|r|-1}$; ELSE $e_{|r|}$;" where $|r|$ indicates the length of the decision list, including the last else clause. In a simple term, a decision list r can be expressed in the following mathematical form: $r = (\alpha_1, e_1), (\alpha_2, e_2), \ldots, \left(\alpha_{|r|-1}, e_{|r|-1}\right), e_{|r|}$. The expression r_{code} denotes the decision list describing the coding rules of the given "code". For instance, a decision list r_{E44} of the ICD code E44 is given in Table 4.1.

Definition D.4.5. For convenience, a prefix *pre* of a decision rule is the front part of the decision rule without the last else clause, while a suffix *suf* is the last part of a decision rule representing the last else clause. For a given rule list r, *pre* $= \left(\alpha_1, p_1\right), \left(\alpha_2, p_2\right), \ldots, \left(\alpha_{|r|-1}, p_{|r|-1}\right)$ and *suf* $= p_(|r|)$.

Definition D.4.6. The space of a rule list r on $R^{z'}$ is defined by $R^{z'}_{rec}$, and $R^{z'} \subseteq R^{z'}_{rec}$). In this study, the whole space consists of multiple subspaces, i.e., $R^{z'} = R^{z'}_{r1} \cup R^{z'}_{r2}, \ldots, \cup R^{z'}_{r_{|Y|}}$. In addition, the subspaces of decision rules are mutually exclusive. In other words, the intersection of any two subspaces is an empty set, i.e., $R^{z'}_{r_i} \cap R^{z'}_{r_j} = \emptyset, i \neq j, i, j \in \{1, 2, \ldots, |Y|\}$.

Definition D.4.7. Given a rule list r, we define that an input sample $x \in R^{z'}$ is captured by the i-th precursor ($i \in \{0, 1, \ldots, |r|-1\}$) of the rule list if x does not satisfy precursors $\alpha_1, \alpha_2, \ldots, \alpha_{i-1}$, but fulfills all requirements of the precursor α_i. In other words, α_i is the first precursor that x is trapped in. We say an input sample x is captured by a rule list r if the sample can be captured by one precursor of the rule list. The capture function of a rule list r_{code} is defined by the mapping $\text{capt}(x, r_{\text{code}})$, which returns 1 if the sample x is captured by the rule list r_{code}, otherwise returns 0.

Definition D.4.8. Given a rule list r and a correction set $c \in C$, the **masked decision list** r^c_{code} is defined by masking out (setting to true) all atomic rules of r containing features listed in the correction set c. For instance, given a correction set $c = \{\text{age}, \text{CRP}\}$ and a decision list r_{E44}: "IF age $< 70 \wedge$ BMI > 17 THEN assigning E44; ELSE IF albumin $\geq 30 \wedge$ CRP $< 15 \wedge$ wgt evol in 1mo ≤ 0.05, THEN assigning E44;... ; ELSE do not assign E44", the masked decision list r^c_{E44} has the following form: "IF BMI > 17 THEN assigning E44; ELSE IF albumin $\geq 30 \wedge$ wgt evol in 1mo ≤ 0.05, THEN assigning E44;... ; ELSE do not assign E44".

Definition D.4.9. Given a subject $p = (x, y), p \in I$ and a correction set $c \in C$, the score function is defined by the following formula:

$$\begin{aligned} \text{s(p,c)} &= s_{\text{total}}(x, y, c) + s_{\text{partial}}(x, y, c) \\ &= \text{capt}\left(\text{x}, r^c_y\right) + \sum_{x \in \{x | \text{capt}\left(\text{x}, r^c_y\right) = 0\}} \text{partial}(x, c, A(y), a) \end{aligned} \tag{4.7}$$

where

$$\text{parial}(x, c, A(y), a) = \begin{cases} \gamma^a, \text{ if capt}\left(\text{x}, r^c_{A_a(y)}\right) = 1 \\ \text{partial } (x, c, A(y), a+1), \text{ otherwise} \end{cases} \tag{4.8}$$

The score function is defined as the sum of a complete correction score and a partial correction score. The term $A(y)$ indicates a list of medical codes that are close to the target code y. $A(y)$ is a list of similar codes in descending order defined by the coding staff. The higher the ranking (a) of a code, the more similar it is to the target code y.

Specifically, given a subject p, if the masked coding rule r^c_y can produce the same code as y, i.e., $\text{capt}(x, r^c_y) = 1$, then we say that the coding error is explained and removed by the correction set C.

The recursive function $\text{parial}(x, c, A(y), a)$ compute a score for subjects whose coding errors cannot be completely corrected and removed. This implies that if the correction set, c, is not able to completely eliminate the coding error from a sample x, then the approximate substitution can partially eliminate the negative effects of the miscoding.

The discount factor $\gamma \in [0, 1]$ defines the horizon of the miscoding correction, which represents how much importance is given to relevant codes compared to the target code y. In addition, for a given population I_i, we consider that a complete correction is always better than the partial correction, $\gamma \times |I_i| \leq 1, \forall I_i \in \text{I}$

Ultimately, the score of a correction set c_j for a given subpopulation I_i is simply defined as the sum of scores of each individual in the population, i.e., $s_{ij} = \sum_{p \in I_i} s\left(p, c_j\right)$.

4.3.4. Optimal medical review allocation

This subsection defines the medical review allocation problem by a mixed-integer programming model. The relevant notations are given below:

Notations:

- F: the set of features for miscoding correction, i.e., $F = \{f_1, f_2, \ldots, f_{z'}\}$;
- I: the set of miscoding clusters, $I_i \in I, i \in \{1, 2, \ldots, k\}$, i.e., $I = \{I_1 \cup I_2 \cup, \ldots, \cup I_k\}$;
- C: correction sets, i.e., $\{c : c \subseteq F, c \neq \emptyset\}$

Given a set of miscoding subgroups $I_i, i = \{1, 2, \ldots, k\}$ determined in Section 4.3.2, a feature set F and the correction sets C for each subgroup I_i, we seek to maximize the number of coding errors reduced, i.e.,

Goal:

$$\max \sum_{i \in I} \sum_{j \in C} s_{ij} * u_{ij} \tag{4.9}$$

Subject to:

$$\sum_{i\in I}\sum_{j\in C}|I_i| * \left|c_j\right| * u_{ij} \leq \lambda \tag{4.10}$$

$$0 \leq \sum_{j\in C} u_{ij} \leq 1,\ \forall i \in I \tag{4.11}$$

$$\sum_{f\in F} v_f \leq \beta,\ \beta \in \{1,2\ldots,|z'|\} \tag{4.12}$$

$$\delta_{ij}^{f} * u_{ij} \leq v_f,\ \forall i \in I, j \in C, f \in F \tag{4.13}$$

$$\sum_{i\in I}\sum_{j\in C} \delta_{ij}^{f} * u_{ij} \geq v_f,\ \forall f \in F \tag{4.14}$$

$$u_{ij} \in \{0,1\},\ \forall i \in I, j \in C \tag{4.15}$$

$$v_f \in \{0,1\},\ \forall f \in F \tag{4.16}$$

Where u_{ij} is a binary decision variable equal to one if the j-th correction set c_j of the subpopulation I_i is selected for review. The constraint 4.11 indicates that the MIP model can select at most one correction set for each subpopulation I_i. The binary decision variable v_f equal to one if the feature $f \in \{1,2,\ldots,z'\}$ appears in one of the selected correction sets. The constrain 4.12 points out that the selected unique features cannot exceed the upper bound β, (i.e., the maximum cardinality of correction sets). In addition, δ_{ij}^{f} is a pre-computed matrix that equals to one if the feature f appears in the correction set c_j of the subgroup I_i.

4.3.5. Counterfactual explanation of coding errors

Consider a selected correction set c and the set I_c of miscoded patient stays $p_i = (x_i, y_i), p_i \in I$ assigned to it. It is possible to match the reassigned code $\text{rec}^c(x_i|_{\phi-c})$ and the given code y_i simply by changing the value of features in c. The purpose of this section is to show such changes and provide counterfactual explanations to medical coders. By doing so, we can tell medical coders what to do to correct given miscoded cases.

Considering first a patient $p_i \in I_c$ with feature vector x_i and given code y_i. Let ρ be the first atomic rule in alphabetical order of coding recommendation for code y_i such that $\rho|_{F-c}(x_i) = 1$. We call ρ the coding error atomic rule of patient p_i.

Definition D.4.10: Let ρ be the coding error atomic rule of a given patient p_i defined by a subset $S \subset F$ and intervals $[\text{LB}_f, \text{UB}_f]_{f\in S}$. We define the **minimal correction** MCN_{if} for all feature $f \in c$ as follows:

$$\mathrm{MCN}_{if} = \begin{cases} 0, \text{if} f \notin S \vee x_{if} \in [\mathrm{LB}_f, \mathrm{UB}_f] \\ x_{if} - \mathrm{LB}_f, \text{if} f \in S \wedge x_{if} < \mathrm{LB}_f \\ x_{if} - \mathrm{UB}_f, \text{if} f \in S \wedge x_{if} > \mathrm{UB}_f \end{cases}$$

In the above definition, $f \in S \wedge x_{if} < \mathrm{LB}_f$ implies a negative error of feature f and an upward correction is needed for code y_i, whereas $f \in S \wedge x_{if} > \mathrm{UB}_f$ implies a positive error of feature f and a downward correction is needed.

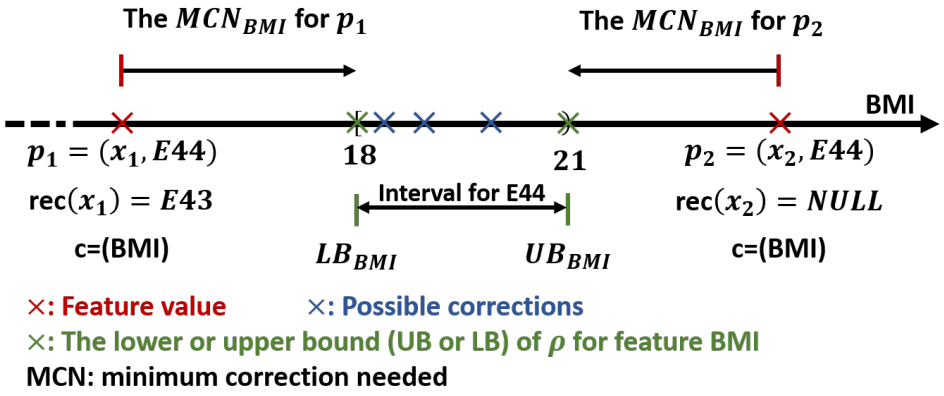


Figure 4.2.: The minimal correction needed

For instance, in Figure 4.2, the patient stay p_1 is miscoded. To reproduce the code E44, we have to change the value of the feature f = "BMI" by at least $|x_{if} - \mathrm{LB}_f|$.

4.4. Case study: medical review rationing for miscoding correction

As part of the hospital coding task, the screening and correction of miscoded cases in hospitals typically involve the review of Electronic Health Records (EHRs). Many medical reviews are dedicated to correcting miscoded cases but are not efficient for the general population. The problem of over-rationing of medical reviews is often present in French healthcare institutions.

In this section, we focus on the rationing of medical reviews for miscoded subjects that are not useful for correctly coded subjects. In the remaining parts, we present the application of the proposed approach on the undernutrition dataset, which is extracted from the University Hospital of Saint Etienne, CHU-SE. Detailed description of this dataset is provided in Chapter 2, Section 2.5.1 and Chapter 3, Section 3.5. The dataset contains 32,856 EHRs (i.e., $|P| = 32,856$) in total, of which 3,676 EHRs are miscoded (i.e., $|I| = 3,676$). All possible codes are from the ICD-10 coding system, i.e., $\text{code}_{\text{HAS}}, \text{code}_{\text{PMSI}} \in \text{ICD}$.

4.4.1. The current practice of the DIM

For malnutrition-related ICD codes, the review process is handled by a medical coder from the medical information department (DIM). However, currently, the medical coder does not have direct access to the features used in the guideline to determine the level of malnutrition. Therefore, the malnutrition-related coding review consists of a fairly simple heuristic: patients who have been assigned a nutritionist visit during their stay and for whom no malnutrition has been coded are reviewed. The underlying idea is that a nutritionist usually visits a patient who is strongly suspected of being malnourished.

However, this heuristic has two significant limits. Firstly, it is only an approximation of the guideline violation. It does not ensure that the resulting set of patients to review consists only of miscoded patients. Secondly, even under the hypothesis that the resulting set is only composed of miscoded patients, the current practice would only target a subset of miscoded patients. Indeed, the heuristic is designed in such a way that selects and reviews only a subpart of under-coded patient stays, i.e., patient $p_i \in P$ for whom (i) $y_i = \text{NULL}$, $\text{rec}(x_i) = \text{E44}$ or E43, and (ii) $y_i = \text{E44}$, $\text{rec}(x_i) = \text{E43}$.

On a miscoding review conducted on patients in February 2021, the heuristic extracted 123 patients' EHRs to review (corresponding to $123 \times 6 = 738$ descriptive features). Out of this set of patients, nine were actually miscoded. This shows that the current practice is highly underperforming since the estimated miscoding rate is only 7.31%. Some coding errors are not considered and taken into account for review.

4.4.2. Optimal number of clusters

In practice, clusters are constructed on health pathways-related data. The variables related to health pathways are ordinal or binary variables. For simplification purposes, all ordinal features are discretized into binary variables. For example, the feature 'number of visits to medical unit' $\in \{0,1,\ldots,10\}$ is discretized into three binary variables: (i) 'medical_0' $\in \{0,1\}$ (i.e., has the patient visited the medical unit in the past?), (ii) 'medical_1_2' $\in \{0,1\}$ (i.e., is the number of visits to medical unit between 1 and 2 (or not)?), (iii) medical_3_over' $\in \{0,1\}$ (i.e., is the number of visits to medical unit greater than 3 (or not)?).

To determine the optimal number of clusters, the silhouette coefficient described in subsection 4.3.2 is used to determine the optimal number of clusters k. In other words, the objective of this step is to optimize the objective defined in Equation 4.6. Figure 4.3 demonstrates the relation between the objective values calculated from the equation 4.3 and the number of clusters k. The blue curve indicates the first derivative of the red curve. As the number of clusters increases, the target value rapidly decreases to the optimal (or minimum) value. We did not search for all possible solutions since too many clusters (i.e., for k > 100) do not benefit the coding correction task and subsequent interpretable analysis.

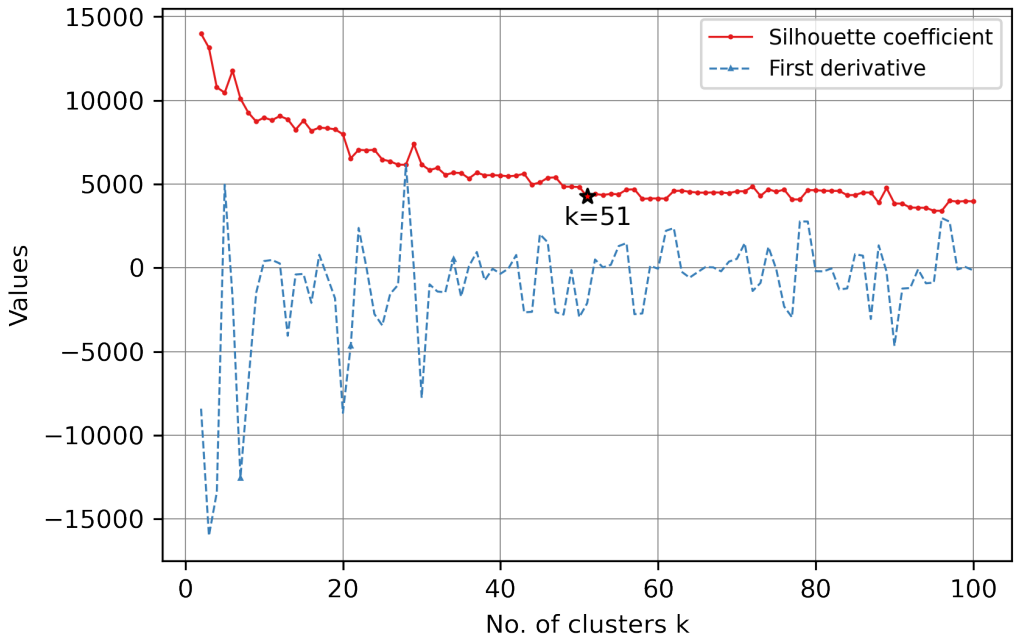


Figure 4.3.: Determining the optimal number of clusters.

Based on the results shown in the Figure 4.3, a conservative model with k=51 is then selected. Table 3.1 shows the cluster centroids for the first 15 clusters. As illustrated by cluster centroids, several clusters with different healthcare pathways are discovered.

Table 4.2.: Cluster centroids

cluster_ID	cluster_size	medical_0	medical_1_2	medical_3_over	surgery_0	surgery_1_2	surgery_3_over	rehab_0	rehab_1_2	rehab_3_over	icu_0	icu_1_2
0	5	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
1	190	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
2	75	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
3	161	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
4	24	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
5	1	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
6	16	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
7	19	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
8	46	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
9	66	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
10	364	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
11	67	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
12	55	0.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0
13	40	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
14	19	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0

Table 4.3.: Continued from the above table

cluster_ID	icu_3_over	CM_0	CM_1_2	CM_3_over	obstet_0	obstet_1_2	obstet_3_over	gender_male	gender_female	los_2_7	los_8_21	los_22_over
0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0
1	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
2	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
3	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
4	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
5	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
6	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0
7	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
8	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
9	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
10	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
11	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0
12	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
13	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
14	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0

4.4.3. Efficient medical review allocation

First of all, the hyper-parameter γ should be select. The γ reveals the importance of a partial correction. As defined in Equation 4.7, a γ value of 1 indicates that a partial correction corresponds to a complete correction, while a γ value of 0 means that partial corrections are useless. We set γ to 0.5, meaning that a complete correction equals two partial corrections.

According to the coding manual [5], we select a set of relevant features F to construct several recommendation functions $\text{rec}(x) = \text{NULL}, \text{rec}(x) = \text{E43}$, and $\text{rec}(x) = \text{E44}$. Once the recommendation functions are constructed, we are able to compute a set of correction sets C for each subpopulation I_i. After that, we set the maximum cardinality of corrections to β in order to limit the size of the problem.

With the above pre-computed constants, the optimization model proposed in Section 4.3.4 is well defined. The decision variable u_{ic} either assigns or does not assign a medical review c to a subgroup I_i. The decision variable v_f either makes use of or does not makes use of the feature f for miscoding correction. The goal is to maximize the total number of coding errors reduced such that the cumulative number of features to review is below a pre-defined upper bound λ.

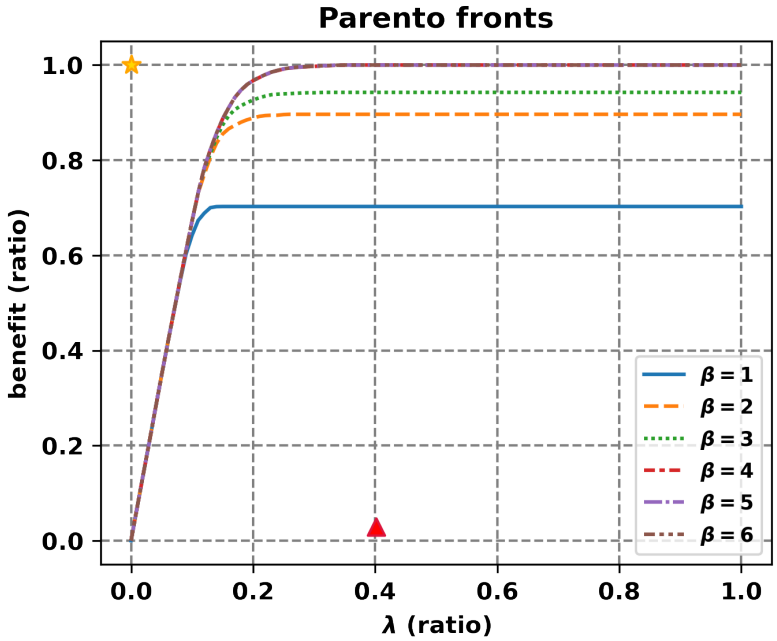


Figure 4.4.: Pareto fronts for the MIP model with different correction efforts and sizes of correction sets.

For this study case, an optimization model is defined by a tuple of (β, λ), where β is the maximum cardinality of correction sets, and λ denotes the maximum number of features to review. Figure 4.4 shows the Pareto fronts of the resulting MIP models, where a benefit (ratio) of 1 corresponds to the maximum benefit reachable, i.e., 3676

coding errors for the current dataset. A λ (ratio) of 1 corresponds to reviewing all $|I| \times |F| = 3676 * 6 = 22056$ features. The yellow star denotes the optimal solution. The red triangle represents the current practice of the DIM, which is calculated according to the figures given in subsection 4.4.1.

Table 4.4 shows results of the MIP models ($\beta = 4$). In simple terms, the term *"λ"* denotes the maximum number of features allowed for review. The term *"Objective values"* indicates the objective values of MIP models obtained by Equation 4.9. In addition, the term *"Error eliminated (%)"* shows the percentage of errors eliminated by MIP models. By varying the pre-defined upper bound λ, code correction solutions corresponding to different levels of the tradeoff between the medical review budget and the increased benefit are obtained and provided to the decision-maker.

Table 4.4.: Experimental results of the MIP models ($\beta = 4$)

λ	Objective values	Errors eliminated (%)	λ	Objective values	Errors eliminated (%)
0.02	512.5	0.1394	0.32	3670.5	0.9985
0.04	1022	0.2780	0.33	3672.5	0.9990
0.06	1524	0.4145	0.34	3674.5	0.9995
0.08	2004.5	0.5452	0.35	3675	0.9997
0.1	2471	0.6721	0.36	3676	1
0.12	2881.5	0.7838	0.37	3676	1
0.14	3148	0.8563	0.38	3676	1
0.16	3346	0.9102	0.39	3676	1
0.18	3480.5	0.9468	0.4	3676	1
0.2	3555	0.9670	0.5	3676	1
0.22	3601	0.9795	0.6	3676	1
0.24	3633	0.9883	0.7	3676	1
0.26	3652	0.9934	0.8	3676	1
0.28	3660.5	0.9957	0.9	3676	1
0.3	3665.5	0.9971	1	3676	1
0.31	3668	0.9978			

As mentioned in section 4.3.4, a tradeoff should be made. A relatively large value of β guarantees the relatively fast convergence of the objective values, while a relatively small value of λ tends to reduce the computational cost. By varying the pre-defined upper bound λ, code correction solutions corresponding to different levels of the tradeoff between the medical review budget and the increased benefit are obtained and provided to the decision-maker. Based on the experimental results presented in Table 4.4, the model with the parameters ($\beta = 4, \lambda = 0.36$) is therefore selected, which corresponds to review 1,323 out of 22,056 descriptive features, and eliminates all coding errors presented in the entire population P.

4.4.4. Insights from counterfactual explanation

In this subsection, we present the counterfactual explanation (CE) for each corrected coding error $p_i = (x_i, y_i) \in I$, $\text{rec}(x_i) \neq y_i$, i.e., the set of features $\{f : f \in c\}$ to review, and for each feature, the feature values to change in order to get the desired code y_i, i.e., $\text{rec}^c(x_i) = y_i$.

Table 4.5 shows several examples of counterfactual explanations. For instance, given the miscoded case p_{245} from the cluster c_5, we can assign the code "E43" to the subject (or the hospital stay) p_{245} only if we review the feature "BMI" and decrease 'BMI' by 5.40. Another example is the miscoded case p_{1275} from cluster c_4. In this case, instead of reviewing only one feature, we have to review three features and make some modifications to them for code correction. Note that, in this study, the maximum cardinality of a correction set is 6. These counterfactual explanations are intuitive and can be provided to medical coders for improvement in coding practice.

The coding error presented in the given dataset may be eliminated by providing professional coding training to medical coders. We suggest the decision-maker develop management measures to improve coders' coding practices, i.e., (i) providing standardized training regularly to medical coders to help them form proper and consistent coding habits and (ii) designing an effective work handover plan to ensure continuity of medical coding tasks and improve the efficiency of the coding task.

Table 4.5.: The counterfactual explanations (excerpt)

subj_ID	y_i	rec(x_i)	cluster_ID	correction set: c	correction
p245	E43	Null	c5	('BMI')	BMI-5.40
p15692	E44	Null	c5	('BMI')	BMI-4.34
p356	E44	E43	c5	('BMI')	BMI+3.43
p152	E44	Null	c1	('wet_evol_in_6_mo')	wet_evol_in_6_mo+0.012
p2514	E43	Null	c1	('wgt_evol_in_6mo')	wet_evol_in_6_mo+0.045
p7756	Null	E43	c6	('BMI', 'albumine')	BMI+2.32, albumin+5.81
p59	E44	E43	c6	('BMI', 'albumine')	BMI+3.21, albumin+3.54
p132	E44	Null	c34	('BMI', 'albumine')	BMI-7.43, albumine-6.99
p6358	E43	Null	c34	('BMI', 'albumine')	BMI-3.28, albumin-4.27
p7783	E43	E44	c16	('wgt_evol_in_6mo', 'albumine')	wet_evol_in_6_mo-0.025, albumin-5.97
p4	E43	E44	c16	('wgt_evol_in_6mo', 'albumine')	wet_evol_in_6_mo+0.015, albumin+7.42
p15	Null	E44	c4	('BMI', 'wgt_evol_in_1mo', 'albumine')	BMI+3.25, wgt_evol_in_1mo+0.026, albumin-1.34
p1275	E44	Null	c4	('BMI', 'wgt_evol_in_1mo', 'albumine')	BMI-1.46, wgt_evol_in_1mo-0.088, albumin-7.53
p112	Null	E44	c10	('BMI', 'wgt_evol_in_1mo', 'albumine')	BMI+7.93, wgt_evol_in_1mo+0.012, albumin+9.25
p12567	Null	E43	c10	('BMI', 'wgt_evol_in_1mo', 'albumine')	BMI+3.24, wgt_evol_in_1mo+0.0057, albumin+3.23

4.5. Discussion and conclusion

This chapter addressed the problem of miscoding correction in healthcare institutions subject to constraints of public health services. We propose a two-stage optimization approach by a combination of clustering for subject profiling, mixed integer programming (MIP) for medical review allocation, and counterfactual explanation for interpretability enhancement.

The proposed approach is tested in a real case study to allocate medical reviews. The application of the proposed approach reduces unnecessary medical reviews by $(8856 - 1323)/8856 = 85\%$ (i.e., only 1,323 descriptive features are reviewed compared to the current practice in the SSPIM (i.e., reviewing about $738 \times 12 = 8{,}856$ descriptive features in one year, and can not reduce all underlying coding error) while reducing all underlying coding errors. We assume that some miscoding subtypes exist in the given dataset. The use of clustering algorithms to determine similar profiles in terms of miscoding subtypes ameliorate the efficiency of the miscoding correction.

The proposed approach offers an effective and practical way to deal with the hospital miscoding problem. We emphasize two main contributions from the experimental results: (i) the optimization program offers a solution to save unnecessary medical reviews while removing all underlying coding errors; (ii) the provided counterfactual explanations can be provided to medical coders to prevent repetitive miscoding behaviors. The experimental results ensure the validity and applicability of the proposed approach.

In this chapter, we proposed a two-stage approach for automated correction of hospital miscoding. A clustering algorithm is applied prior to the application of the proposed optimization model. A complete set of correction sets is also assigned with each discovered cluster. In addition to the concept of correction-set proposed, in the next chapter, an important element, called hypercube, is proposed in Chapter 5 to refine the miscoding profiling approach and generalize the proposed approach to a wider range of applications. On top of that, a novel optimization model is proposed to adapt to the ameliorated profiling approach. Please continue reading the next chapter for a more detailed explanation.

5. An integrated approach for optimizing miscoding correction budget

Summary

Abstract of the chapter

As coded health data are widely adopted in various critical areas (i.e., health services reimbursement, hospital resource scheduling, epidemiology research), medical miscoding has the potential to produce widespread and far-reaching negative impacts. In France, the appropriate reimbursement of hospital fiscal revenue in the activity-based funding system (T2A) relies on accurate, complete, and periodical medical coding. To raise the health services reimbursement, we set the problem as a chance-constrained medical review rationing problem. Tiny coding errors with high financial benefits are preferentially targeted and given supplemental medical reviews, while the remainder is excluded. A novel clustering-based optimization approach [1] is proposed to allocate medical reviews based on profiles of patient stays. A case study on patient stays associated with malnutrition-related ICD codes is presented, and the performance of the proposed methodology is assessed. A significant increase in health services reimbursement is achieved with a limited number of subjects' features reviewed.

Keywords: **Keywords:** Hospital miscoding, clustering, coding review budget optimization, mathematical programming.

[1] This chapter is based on our previous work submitted to IEEE Transactions on Automation Science and Engineering as "A clustering-based optimization approach for hospital miscoding correction" (under revision) by Chen HE, Benjamin DALMAS, Cedric BOUSQUET, Beatrice TROMBERT-PAVIOT, and Xiaolan XIE.

Résumé du chapitre

Les codes hospitaliers étant largement utilisés dans divers domaines critiques (remboursement des services de santé, planification des ressources hospitalières, recherche épidémiologique), les erreurs de codage médical peuvent avoir des répercussions négatives étendues et de grande ampleur. En France, le remboursement approprié des recettes fiscales des hôpitaux dans le cadre du système de financement par activité (T2A) repose sur un codage médical précis, complet et périodique. Pour augmenter le remboursement des services de santé, nous avons défini le problème comme un problème de rationnement de la révision médicale sous contrainte de chance. Les petites erreurs de codage présentant des avantages financiers élevés sont ciblées de manière préférentielle et font l'objet de révisions médicales supplémentaires, tandis que les autres sont exclues. Une nouvelle approche d'optimisation basée sur le clustering [1] est proposée pour allouer les examens médicaux en fonction des profils de séjours des patients. Une étude de cas sur les séjours de patients associés à des codes CIM liés à la malnutrition est présentée, et la performance de la méthodologie proposée est évaluée. Une augmentation significative du remboursement des services de santé est obtenue avec un nombre limité de caractéristiques des sujets examinés.

Mots-clés: Miscodage hospitalier, algorithme de regroupement, optimisation du budget de la révision du codage, programmation mathématique.

[1] Ce chapitre est basé sur notre travail précédent soumis à IEEE Transactions on Automation Science and Engineering sous le titre "A clustering-based optimization approach for hospital miscoding correction" (en cours de révision) par Chen HE, Benjamin DALMAS, Cedric BOUSQUET, Beatrice TROMBERT-PAVIOT, et Xiaolan XIE.

5.1. Introduction

In the previous chapter, i.e., chapter 4, we proposed a patient-oriented methodology, which is based on the assumption that similar patients might share a similar miscoding reason. On the contrary, chapter 5 aims to propose a feature-oriented approach. We assume that the miscoding reasons lie outside of the patients, more in the coding habits, personal experience, willingness of medical coders, etc. Based on the above assumption, we make use of the correction set (i.e., a set of features or a feature combination) to characterize hospital miscoding instead of using well-known clustering techniques.

This chapter is part of our collaboration with a French territory hospital - CHU-SE. In France, all public and private hospitals have to periodically report their medical activities to obtain health services refunds. For a given patient stay, the assigned medical codes are used to compute a DRG (Disease-Related Group) code. The resulting DRG is a summary code that serves as the basis for allocating the budget to the hospital. Therefore, the completeness and accuracy of the assigned medical codes are critical factors in the process of hospital reimbursement.

In this chapter, we define the hospital benefit as a combination of hospital reimbursement generated by the correction of under-coded patients (i.e., patients whose code is less severe than it should be, thus resulting in an inferior reimbursement) and the financial penalties avoided by correcting over-coded patients (i.e., patients whose code is more severe than it should be, thus resulting in a superior reimbursement). To increase the financial benefit of CHU-SE, we propose an optimization approach consisting of the following steps: (i) identifying inconsistencies in medical code assignment; (ii) allocating a limited number of medical reviews to miscoded patient stays to correct corresponding coding errors; (iii) detecting changes of the resulting DRG; (iv) measuring revenue variation associated with DRG changes;

The review of assigned medical codes raises challenges. In CHU-SE, the miscoding review task is regularly performed by hand by the coding staff from the DIM. Given that the population to review is heterogeneous and the intrinsic difficulty of the review task, the current practice of the DIM has been adopting a heuristic to target a specific subpopulation and allocate a complete set of medical reviews to all patient stays in the subpopulation. Such a practice is inefficient and conservative. Many unnecessary medical reviews are allocated to patient stays that do not lead to financial benefits (i.e., correctly coded patient stays). In addition, for a miscoded patient stay, not all relevant descriptive features are worth giving a review. This, unfortunately, results in an unnecessary waste of time and efforts of medical coders and subsequently increases the cost and burden of the entire French healthcare system.

The fundamental problem that motivates this study is whether a medical review is necessary for all descriptive features of a miscoded subject. Intuitively, the optimal allocation of medical reviews varies according to the characteristics of patient stays and types of coding errors. With appropriate profiling of patient stays and miscoding behaviors, it is possible to significantly reduce the number of allocated medical reviews

without decreasing the financial benefit obtained. In a preliminary study, [34], a topological space projection approach, called Mapper, is used for defining profiling rules to determine high-risk subpopulations for which medical review is needed and those for which medical review is unnecessary.

In addition, the methodology proposed in chapter 4 has the limitation of not being able to take into account the heterogeneity of miscoding behaviors. In chapter 4, subjects in a subgroup are similar in the sense of subject portraits or descriptive features, and each subgroup is homogeneous regarding descriptive features. However, subjects with similar feature values can not reflect a specific miscoding behavior of medical coders. In this chapter, we define medical miscoding behaviors (or subtypes) in advance and assign miscoded subjects to subgroups according to subject attributes and portraits, where each subgroup represents a miscoding behavior or a miscoding subtype.

This study aims to develop an optimization approach to ration the medical review budget by profiling miscoding behaviors. The proposed approach is based on the following hypotheses: (i) all information and data related to assigned medical codes are complete and available in subjects' EHRs; (ii) for each miscoded hospital stay, there exists a finite number of features (a correction set) in the subject's EHR that can explain and eliminate the corresponding coding error; (iii) each correction set corresponds to a specific miscoding behavior, and a group of miscoded patient stays corrected by the correction set. Therefore, in this chapter, "correction set" and "correction cluster" are two interchangeable terms. The correction set is also considered the underlying cause of such a miscoding behavior. In this study, hypotheses (ii) and (iii) are valid only if hypothesis (i) holds. Care should be taken if the data are not complete.

More specifically, based on the official coding manual [5] and the concepts presented in section 5.3.2, we translate coding rules in text format to several recommendation functions. With the provided recommendation functions, each hospital stay is recorded without looking at the original code. Given a subject x_i and the assigned code y_i, a coding error is identified if a re-coded code $\text{rec}(x_i)$ does not match its corresponding initial code y_i, i.e., $\text{rec}(x_i) \neq y_i$.

Coding errors are identified. Next, we address the cost-efficiency problem of the medical review task. Such a medical review allocation is subject to the chance constraint, i.e., coding errors associated with high fiscal gains are preferentially targeted and have their recommended medical reviews. In addition, miscoded patient stays are grouped into homogeneous clusters in terms of miscoding behaviors, and underlying factors are identified for each discovered cluster.

A novel clustering-based optimization approach is proposed. Our contribution is multifold: (i) a general framework to deal with the miscoding censoring task; (ii) a novel clustering-based optimization approach to allocating medical reviews by profiling patient stays (section 5.3.2 and 5.3.3); (iii) in combination with statistical analysis techniques presented in section 5.3.4 and 5.3.4, the proposed approach provides causes (i.e., the estimated statistics of discovered miscoding behaviors)

and consequences (i.e., variation in hospital fiscal revenue) for identified miscoding behaviors. This approach has the ability to identify, correct and quantify miscoding behaviors for a complete data set. An increase in hospital revenue of 6,992,489.69 € is presented with the application of the proposed approach.

The remaining of this chapter is organized as follows. Section 5.2 defines the code correction problem and gives an overview of the proposed general framework. The proposed methodology is presented in sections 5.3.2, 5.3.3, and 5.3.4. Piratical considerations related to the implementation of the proposed approach are presented in 5.4. In section 5.3.5, we provide several model variants to meet the personalized requirements of a wide variety of practitioners. Finally, the conclusion and independent perspectives are presented in section 5.5.

5.2. Problem definition

This chapter considers the problem of coding error correction for a given population $P = \{p_i\}_{i=\{1,2,\ldots,|P|\}}$. Each patient stay $p_i = (x_i, y_i)$ is characterized by a z dimensional feature vector $x_i = (f_1, f_2, \ldots, f_z) \in \mathbb{R}^z$, and $z \in \mathbb{Z}^+$, and is labeled with a medical code $y_i \in Y = \{1, 2, \ldots, |Y|\}$. The feature set is denoted by $\Phi = \{f_1, f_2, \ldots, f_z\}$. The given code y_i is checked against a given coding recommendation $\text{rec} : \{f_1, f_2, \ldots, f_{z'}\} \rightarrow Y$ defined on a reduced feature space $F = \{f_1, f_2, \ldots f_{z'}\} \subset \Phi$, $z' \leq z$ and $z' \in \mathbb{Z}^+$. A patient stay p_i is said miscoded if $\text{rec}(x_i) \neq y_i$, is said over-coded if $\text{rec}(x_i) < y_i$ and under-coded if $\text{rec}(x_i) > y_i$. In this chapter, we concentrate on the miscoded patient stays $I \subseteq P$ and devote ourselves to explaining and efficiently eliminating miscoding behaviors presented in the population I.

We assume that coding errors are **observational errors** and are caused by wrong observed values of certain features. In order to identify features with wrong observed values, a set of features $c \subseteq F$ is considered a **correction set** of p_i if $\exists\, x \in \mathbb{R}^z : \text{rec}^c(x_i) = y_i,\ x|_{\Phi-c} = x_i|_{\Phi-c}$ holds. The term $x|_{\Phi-c}$ denotes the projection of vector x to space $\Phi - c$, which implies that, by changing the value of features included in the correction set c, we can make the recommendation code the same as the given code, i.e., $\text{rec}^c(x_i|_{\Phi-c}) = y_i$. As a result, for a given miscoded patient stay, the correction set provides medical coders with features $\{f : \forall\, f \in c\}$ that contains observational errors and need to be checked.

The cost of code correction is related to three factors: (i) the complexity of code correction depending on the number of features to review for each patient stay, i.e., the cardinality of the correction set; (ii) the diversity of verification, i.e., the number of correction sets; (iii) the total workload depending on the cumulative number of features to check across all miscoded patient stays. Based on this, we define **a code correction solution** as a partial assignment of miscoded patient stays to a finite set of clusters, each associated with a correction set.

A code correction solution is feasible if (i) all correction sets are of cardinality no more than a pre-defined upper bound m, i.e., $|c| \leq m$, (ii) the number of clusters is no

more than a pre-defined upper bound K, and (iii) the cumulative number of features to check is no more than the maximum workload maxIF.

The goal of the code correction optimization is to maximize health service reimbursement. To this end, we assume that each patient stay p_i is associated with a DRG code and a fiscal revenue, i.e., $\text{gain}_i : Y \to \mathbb{R}^+$. As a result, fiscal revenue is $\text{gain}_i(y_i)$ according to the given code y_i and $\text{gain}_i(\text{rec}(x_i))$ according to the given coding convention. We assume the financial benefit of code correction to be $\left|\text{gain}_i\left(\text{rec}\left(x_i\right)\right) - \text{gain}_i\left(y_i\right)\right|$. Although a patient is corrected, the gain can be zero if the correction does not change the patient DRG, i.e $\text{gain}_i(\text{rec}(x_i)) = \text{gain}_i\left(y_i\right)$ with $y_i \neq rec(x_i)$.

To summarize, the problem of code correction consists in determining a feasible code correction solution in order to maximize the total financial benefit. Apart from the above primary goal, we set the following secondary goals: (i) characterizing in terms of features discovered clusters (or miscoding behaviors) and (ii) quantifying the error distribution of each feature in a correction set.

Apart from the above primary goal, to provide valuable statistics for the final decision-making, we set the following secondary goals: (i) characterizing in terms of features discovered clusters (or miscoding behaviors) and (ii) quantifying the error distribution of each feature in a correction set.

Remark. In the above definition, each feature is a real number. Extensions to binary, integer, or qualitative features are straightforward. For qualitative features defined on a finite set, we assume that the finite set is strictly ordered. We also assume that the code set Y is strictly ordered. Nevertheless, the strict order of the code set is only needed in the analysis of over-coding or under-coding.

Remark. The financial impact $\left|\text{gain}_i\left(\text{rec}\left(x_i\right)\right) - \text{gain}_i\left(y_i\right)\right|$ can also be extended to a more general form: $\phi_i(\text{rec}(x_i), y_i)$. In particular, we can include some fixed penalty costs for miscoding. Note that the dependence of the financial impact to subject index i allows us to take into account the dependency of the financial impact on other relevant treatments, which is common for co-morbid or multi-morbid patients. In our case study, the revenue of malnutrition coding depends on the main DRG code of the subject. Further, by restricting miscoding to under-coding or over-coding, the financial impact $\left|\text{gain}_i\left(\text{rec}\left(x_i\right)\right) - \text{gain}_i\left(y_i\right)\right|$ reduces to under-coding gain or over-coding penalty. The former is the main concern of our partner hospital.

5.3. Methodology

5.3.1. Overview of the methodology

To solve the problem of code correction, we leverage techniques from three research fields: optimization, data mining, and statistics. We propose a formal mathematical representation of medical coding conventions for coding error identification, correction sets to characterize various miscoding behaviors, a mixed integer programming

(MIP) model to solve the optimal code correction problem, a data mining algorithm to characterize discovered correction clusters, and finally, histogram statistics to quantify error distributions. The aerial view of the proposed method is shown in Figure 5.1.

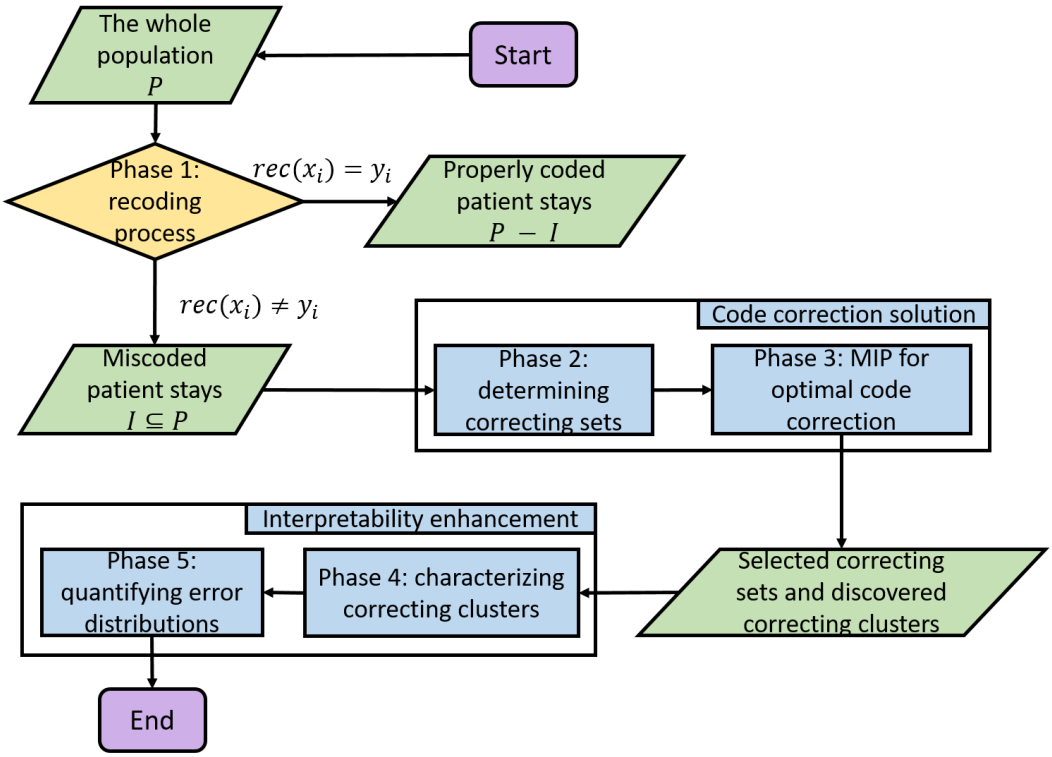


Figure 5.1.: A clustering-based optimization approach

More specifically, we use the mathematical representation of medical coding convention to identifying coding errors (i.e., patient stays for whom rec$(x_i) \neq y_i$. We assume that the population I is heterogeneous and contains various miscoding behaviors. In this study, each miscoding behavior is characterized by a correction set. For each miscoded case p_i, we then determine a collection of correction sets to characterize underlying miscoding behaviors associated with the case p_i. These inferred correction sets are what we need for the code correction problem presented in phase 3. In this study, we formulate the optimal code correction problem as a MIP problem. Technically, the proposed MIP model screens a correction set c for each selected miscoded case $p_i \in P$ to eliminate coding errors presented in the heterogeneous population I (called medical review rationing and code correction) while putting miscoded cases which are similar in the sense of miscoding behavior into a homogeneous cluster (called error profiling and population partitioning).

To characterize discovered correction clusters, we leverage an association rule mining algorithm to extract frequent patterns for each discovered cluster. The generated patterns may reveal the interaction between descriptive factors and miscoding behaviors. Extracted patterns can be provided to the decision-makers to evaluate and

improve the coding practices of medical coders. In phase 5, we construct histogram statistics for each discovered cluster to quantify the error distribution of each feature presented in the corresponding correction set. Apart from this, the histogram intersection technique is used to quantify the similarity of two discovered miscoding behaviors.

In the following sections, we give a detailed description of each component of the proposed methodology.

5.3.2. Coding recommendation and correction sets

This section proposes a formal representation of coding recommendations and addresses the determination of correction sets for each miscoded case.

Formal representation of coding recommendation

This subsection proposes a formal mathematical representation of coding recommendation by the union ($\cup$) of hypercubes and disjunction ($\vee$) of atomic rules.

Definition D.5.1:
A hypercube $\theta \subset \mathbb{R}^{z'}$ is defined by a subset $S \subset F$ and a set of closed intervals $[\mathrm{LB}_f, \mathrm{UB}_f]$ associated with features in S such that $\theta = \left\{ x \in \mathbb{R}^{z'} : x_{if} \in \left[\mathrm{LB}_f, \mathrm{UB}_f\right], \forall f \in S \right\}$. By convention, $\mathrm{LB}_f = -\infty$ ($\mathrm{UB}_f = \infty$) implies no lower (upper) bound.

Definition D.5.2:
An atomic rule ρ on an input domain $R^{z'}$ is a Boolean function: $\rho : F \rightarrow \{0,1\}$ associated with a subset $S \subset F$ and a set of closed intervals $[\mathrm{LB}_f, \mathrm{UB}_f]$ such that $\rho(x_i) = \wedge_{f \in S}\left(x_{if} \in \left[\mathrm{LB}_f, \mathrm{UB}_f\right]\right)$ where $\wedge$ stands for the logical operator AND.

From above, there is a one-to-one correspondence between hypercubes and atomic rules, and we will use both interchangeably.

Assumption A.5.1:
For any code $y \in Y$, the preimage $\mathrm{rec}^{-1}(y)$ of the recommendation function is a union of a finite number of hypercubes ($\mathrm{rec}^{-1} = \cup_{i=1}^{|r|}(\theta_i)$) in the input feature space F. Equivalently, the **recommendation function** $\mathrm{rec}(x) = y$ is the disjunction of atomic rules ($\mathrm{rec} = \vee_{i=1}^{|r|}(\rho_i)$) and is defined by a Boolean function $\mathrm{rec} : \mathbb{R}^{z'} \rightarrow [0,1]$.

Table 5.1.: An excerpt of the recommendation function rec(x) = $E44$ for the code E44

	precursor		**successor**
IF	age ∈ (0,70) ∧ wgt evol in 1mo ∈ [−10%,−5%) ∧ wgt evol in 6mos ∈ (−15%,∞) ∧ BMI ∈ (17,∞) ∧ albumin ∈ (30,∞)	THEN assigning the code	E44
ELSE IF	age ∈ [70,∞) ∧ C-reactive protein ∈ [0,15) ∧ wgt evol in 1mo ∈ (−10%,−5%] ∧ wgt evol in 6mos ∈ (−15%,∞) ∧ BMI ∈ [18,∞) ∧ albumin ∈ [30,∞)	THEN assigning the code	E44
ELSE IF	age ∈ [70,∞) ∧ C-reactive protein ∈ [15,∞) ∧ wgt evol in 1mo ∈ (−10%,∞) ∧ wgt evol in 6mos ∈ (−15%,−10%] ∧ BMI ∈ [18,∞) ∧ albumin ∈ [30,∞)	THEN assigning the code	E44
…	…	…	…
ELSE	–	do not assign the code	E44

Examples:
For illustration purposes, a recommendation function rec(x) = E44 is given in Table 5.1, each row of the table represents an atomic rule. The coding recommendation for code E44 can also be represented by the union of the following hypercubes (with both open and closed intervals).

$$\begin{array}{c} \text{age} \\ \text{C-reactive protein} \\ \text{wgt evol in 1mo (\%)} \\ \text{wgt evol in 6mos (\%)} \\ \text{BMI} \\ \text{albumin} \end{array} \begin{bmatrix} (0,70) \\ (0,\infty) \\ (-10,-5] \\ (-15,\infty) \\ (17,\infty) \\ (30,\infty) \end{bmatrix} \cup \begin{bmatrix} [70,\infty) \\ [0,15) \\ (-10,-5] \\ (-15,\infty) \\ [18,\infty) \\ [30,\infty) \end{bmatrix} \cup \begin{bmatrix} [70,\infty) \\ (15,\infty) \\ (-10,\infty] \\ (-15,-10) \\ [18,\infty) \\ [30,\infty) \end{bmatrix} \cup \cdots$$

In the above expression, each column, excluding the leftmost column, represents a hypercube.

Remark:
The **Assumption A.5.1** is not restrictive for features that are binary, integer, or qualitative. For real features, the hypercube cannot represents exactly constraints such as $x_{if} > \text{LB}_f$ or $x_{if} < \text{UB}_{if}$. However, they can be approximated by $x_{if} \geq \text{LB}'_f$ or $x_{if} \leq \text{UB}'_{if}$ for realistic applications without loss of generality.

Assumption A.5.2:
For each hypercube θ, there exists $x \notin \theta$ for each feature $f \in S$ such that $x_f \notin [\text{LB}_f, \text{UB}_f]$. In other words, none of the intervals is redundant.

Remark:
The **Assumption A.5.1** can be relaxed with the hypercube replaced by $\times_{f\in S} W_f$ where W_f is any non-empty subset of the domain of feature f and $\times_{f\in S} W_f$ is any direct product of the sets W_f. All results of the chapter still hold if sub-domains W_f are all (closed or open) intervals. For more general W_f, all results except the error quantification presented in section 5.3.4 still hold. $W_f = \mathbb{R} - [\text{LB}_f, \text{UB}_f]$ is an example of general W_f.

Correction sets of a miscoded patient stay

This subsection addresses the determination of the correction sets of a miscoded patient stay $p_i = (x_i, y_i)$, $\text{rec}(x_i) \neq y_i$. It relies on the concept of atomic rule projection, defined as follows.

Definition D.5.3:
For any subset $c \in F$ and an atomic rule ρ defined by a subset $S \in F$ and intervals $\left[\text{LB}_f, \text{UB}_f\right]_{f \in S}$, the partial projection of ρ is defined by the following expression: $\rho\big|_{F-c}(x) = \left(\exists\, x' \in \mathbb{R}^{z'} : \rho\left(x'\right) = 1,\, x'\big|_{F-c} = x|_{F-c}\right)$.

Intuitively, the **masked recommendation function** $\text{rec}^c = \vee_{i=1}^{|r|}(\rho_i|_{F-c})$ is defined by removing intervals associated with features f listed in the correction set c, i.e., $\{f : \forall f \in c\}$.

By definition, $\rho\big|_{F-c}(x) = \wedge_{f \in \{S-c\}}\left(x_f \in \left[\text{LB}_f, \text{UB}_f\right]\right)$. As a result, c is a correction set of a patient stay p_i if and only if $\rho|_{F-c}(x_i) = 1$ for at least one atomic rule ρ of the masked recommendation function rec^c.

By definition, if c is a correction set, c' is a correction set as well for all supersets c' of c, i.e., $c' \supset c$. The concept defined in **Definition D.5.4** allows reducing the number of correction sets to consider.

Definition D.5.4:
A correction set c of a miscoded patient stay p is said minimal if it is covered by any other correction sets c' of the patient stay.

The algorithm 1 summarizes the process of determining the set of minimal correction sets C_i for a given miscoded patient stay $p_i \in I$.

Algorithm 1 Computation of minimal correction sets

Input: A miscoded patient stay $p_i = (x_i, y_i)$, $p_i \in I$;
Output: A set C_i of minimal correction sets for p_i;
Step 1: Determine the atomic rule $\rho_k, k \in \{1, 2, \ldots, |r|\}$ of the Boolean function $\text{rec}(x) = y_i$;
Step 2: Initialization: $C_i \leftarrow \emptyset$, correction set size $j \leftarrow 1$;
Step 3: For all subsets $c \subset F, |c| = j$ that are not supersets of any set in C_i

- **3.1**: Check whether c is a correction set, i.e., $\exists\, k \in \{1, 2, \ldots, |r|\} : \rho_k\big|_{F-c}(x_i) = 1$;
- **3.2**: If yes, add c to C_i.

Step 4: If $j = |F|$, stop; otherwise, $j \leftarrow j + 1$ and go to step 3.

Based on the concepts proposed in this subsection, we defined and computed minimal correction sets C_i, which will be used in the next section for optimal code correction.

5.3.3. Optimal code correction

This section proposes a Mixed Integer Programming (MIP) model for optimal code correction and also considers the linear relaxation problem of the proposed model.

Chance-constrained valuation model (CCVM)

This subsection addresses the problem of optimal code correction, i.e., determining a feasible code correction solution in order to maximize the total financial benefit. Specifically, a code correction solution is a partial assignment of miscoded cases to a finite set of correction clusters, each associated with a correction set. We propose a MIP model for the problem. The relevant notations and variables are given below:

Notations

- F: set of relevant features for code recommendation;
- I: set of miscoded subjects $p_i \in I, i \in \{1,2,\ldots,|I|\}$;
- m: maximum size of correction set, $m \in \mathbb{Z}^+$;
- K: maximum number of correction clusters, $K \in \mathbb{Z}^+$;
- $C^{(m)}$: set of feature subsets of size at most m, i.e., $C^{(m)} = \{c \subset F : 0 < |c| \leq m\}$;
- C_i: set of minimal correction sets for subjects $p_i \in I$;
- C_i^+: set of all correction sets of subjects $p_i \in I$, i.e., $C_i^+ = \left\{c \subset F : \exists\, c' \in C_i, c' \subset c\right\}$;

Decision variables

- u_c: a binary variable equal to 1 if $c \in C^{(m)}$ is selected as a correction cluster, i.e., the correction cluster c captures at least one miscoded case $p_i \in I$ and becomes a non-empty cluster $c \neq \emptyset$.
- v_{ic}: a binary variable equal to 1 if the subject p_i is assigned to the correction cluster of correction set c, i.e., the set of features in correction set ($f \in c$) of the miscoded case p_i is selected for review.

With these predefined constants and variables, the optimal code correction problem is defined by the following MIP model:

Goal:

$$G^{\text{CCVM}} = \max G(u,v) =$$

$$\max \sum_{i \in I} \sum_{c \in \{C^{(m)} \cap C_i^+\}} \left|\text{gain}_i\left(\text{rec}(x_i)\right) - \text{gain}_i\left(y_i\right)\right| \times v_{ic} \tag{5.1}$$

Subject to:

$$\sum_{c \in C^{(m)}} u_c \leq \mathrm{K} \tag{5.2}$$

$$v_{ic} \leq u_c, \ \forall \ (i,c) \in I \times \left\{C^{(m)} \cap C_i^+\right\} \tag{5.3}$$

$$\sum_{c \in \{C^{(m)} \cap C_i\}} v_{ic} \leq 1, \forall \ i \in I \tag{5.4}$$

$$\sum_{i \in I} \sum_{c \in \{C^{(m)} \cap C_i\}} |c| \times v_{ic} \leq \mathrm{maxIF} \tag{5.5}$$

$$v_{ic} \in \{0,1\}, \ \forall \ (i,c) \in I \times \left\{C^{(m)} \cap C_i^+\right\} \tag{5.6}$$

$$u_c \in \{0,1\}, \ \forall \ c \in C^{(m)} \tag{5.7}$$

where CCVM stands for Chance-Constrained Valuation Model. The objective function 5.1 maximizes the total financial benefit, constraint 5.7 defines the set of correction clusters, and 5.6 defines the subjects and correction sets to review, i.e., only a subset of miscoded cases are selected and reviewed. On top of this, the constraint (5.2) limits the maximum number of clusters by K. The constraint (5.4) allows the CCVM model to assign a miscoded case p_i to at most one correction cluster. After that, a selected miscoded case p_i is placed in one of the correction clusters based on the constraint (5.3). The formula (5.5) is the chance constraint, which limits the total correction workload to at most MaxIF.

Constraint 5.5 is a knapsack constraint. CCVM reduces to a classification knapsack problem for the special case of independent feature sets of different subjects. As a result, the CCVM is NP-hard. Nevertheless, the theorem presented in the next subsection shows that the integer constraint 5.6 of assignment variable v_{ic} is nearly relaxable, and the number of cluster selection variables u_c is limited. Hence, the CCVM is not extremely hard.

Linear relaxation problem (LP)

In this subsection, we consider the linear relaxation problem (LP) of the proposed CCVM model (5.1)-(5.7). In the LP, the integrity constraint $v_{ic} \in \{0,1\}$ is replaced by the continuous variable constraint $v_{ic} \in [0,1]$.

Theorem T.5.1:
$G^{\mathrm{LP(CCVM)}} \geq G^{\mathrm{CCVM}}$, where $G^{\mathrm{LP(CCVM)}}$ and G^{CCVM} are the optimal criterion values of LP(CCVM) and CCVM. Further, for any feasible solution (u,v) of LP(CCVM), there exists a feasible solution (u,v') of CCVM such that $G(u,v) < G(u,v') - A$ where $A = \max_{i \in I} \left|\mathrm{gain}_i\left(\mathrm{rec}\left(x_i\right) - \mathrm{gain}_i\left(y_i\right)\right)\right|$.

The theorem T.5.1 implies that the linear relaxation optimum deviates from the true optimum by at most A, i.e., the maximum benefit of a code correction.

Proof of theorem T.5.1:
the proof of $G^{\text{LP(CCVM)}} \geq G^{\text{CCVM}}$ is trivial. For any feasible solution (u, v) of the LP(CCVM), let IF $= \sum_{i \in I} \sum_{c \in \{C^{(m)} \cap C_i^+\}} |c| \times v_{ic}$. v' is derived as follows: (i) determining the optimal open cluster c_i of minimal cardinality for each patient stay $p_i \in I$, i.e., $c_i = \arg\min_{c \in C_i^+ : u_c = 1} |c|$; (ii) ranking patient stays I in non-increasing order of $|\text{gain}_i(\text{rec}(x_i) - \text{gain}_i(y_i)|/|c_i|$; (iii) determining a solution v'' such that $v''_{ic_i} = \min\left(1, \frac{1}{|c_j|}\left(\text{IF} - \sum_{j=1}^{i-1} |c_j| \times v''_{jc_j}\right)\right)$; (iv) $v'_{ic_i} = \left\lfloor v''_{ic_i} \right\rfloor$.

By construction, v'' is the optimal solution of the relaxed knapsack problem with object weight $|c_i|$, object profit $|\text{gain}_i(\text{rec}(x_i) - \text{gain}_i(y_i)|$ and sack capacity IF. As a result, $G(u, v'') \geq V(u, v)$. By construction, v'' has at most one real feasible solution for CCVM. Hence, $G(u, v) - G(u, v') \leq G(u, v'') - G(u, v') < A$. Q.E.D.

Remark:
In the CCVM defined above, the objective is set to maximize fiscal revenue. The workload is defined as the cumulative number of features to check. Other definition of workload and objective, and variants of the CCVM model (5.1)-(5.7) are considered in section 5.3.5.

In this section, with limited human resources, we established an accurate and transparent MIP model to allocate the medical review budget and correct coding errors to the maximum extent possible while generating homogeneous clusters for error profiling. In the next section, with a given code correction solution, we provide more detailed statistical information and evidence for decision support. Strategies for characterizing selected clusters and quantifying error distribution are presented in section 5.3.4 and 5.3.4, respectively.

5.3.4. Miscoding explanation

For any given code correction solution, this section addresses the second objective of this study: (i) to characterize in terms of features discovered correction clusters and (ii) to quantify the error distribution of each feature presented in a given correction set.

Characterizing selected correction clusters

Given a code correction solution, each discovered correction cluster is considered a miscoding behavior. In this subselection, we aim to extract frequent patterns for discovered miscoding behaviors. For each cluster c selected in the code correction solution, this subsection aims at determining common characteristics of the patient stays in cluster c, i.e., the set of patient stays p_i such that $v_{ic} = 1$. To this end, a frequently used association rule mining algorithm, called FP-Growth [98], is adopted.

More specifically, consider the complete set P of patient stays (transactions in association rule terminologies) and two sets of binary features (items): a set B of binary features obtained by discretizing features of the complete feature set Φ and

another set B' of binary features including $c, c \subset F : u_c = 1$, under-coding, over-coding. Each patient is defined by a triplet (i, t, t') where i is its ID, $t \subset B$ and $t' \subset B$ are two itemsets with $c \in t'$ if patient stay p_i is miscoded and belongs to cluster c and under-coding $\in t'$ if patient i is under-coded.

A frequent pattern is an implication of the form $X \Rightarrow Y$, where antecedent $X \subset B$ and consequent $Y \subset B'$ are disjoint binary feature sets. The support measures the frequency that an implication presents in the whole population P, i.e., $\sup(X \Rightarrow Y) = \left|\left\{\left(i, t, t'\right) \in P : X \subset t, Y \subset t'\right\}\right| / |P|$. The confidence of an implication indicates the probability of the implication to be true, i.e., $\text{conf}(X \Rightarrow Y) = \sup(X \Rightarrow Y) / \sup(X)$, where $\sup(X)$ indicates the number of itemsets X presented in P.

Quantifying error distribution

Consider a selected correction set c and the set P_c of miscoded patient stays $p_i = (x_i, y_i), p_i \in I$ assigned to it. It is possible to match the masked code recommendation rec^c and the given code y_i simply by changing the value of features in c. The purpose of this section is to show the distribution of such changes or, equivalently, the distribution of feature value errors. By doing so, we can check whether a coding error is due to a minor error of feature values, in which case the coding error is acceptable.

Consider first a patient $p_i \in P_c$ with feature vector x_i and given code y_i. Let ρ be the first atomic rule in alphabetical order of coding recommendation for code y_i such that $\rho|_{F-c}(x_i) = 1$. We call ρ the coding error atomic rule of patient p_i.

Definition D.5.5:
Let ρ be the coding error atomic rule of a given patient p_i defined by a subset $S \subset F$ and intervals $[\text{LB}_f, \text{UB}_f]_{f \in S}$. We define the **minimal observational error** MOE_{if} for all feature $f \in c$ as follows:

$$\text{MOE}_{if} = \begin{cases} 0, \text{if} f \notin S \vee x_{if} \in \left[\text{LB}_f, \text{UB}_f\right] \\ x_{if} - \text{LB}_f, \text{if} f \in S \wedge x_{if} < \text{LB}_f \\ x_{if} - \text{UB}_f, \text{if} f \in S \wedge x_{if} > \text{UB}_f \end{cases}$$

In the above definition, $f \in S \wedge x_{if} < \text{LB}_f$ implies a negative error of feature f and an upward correction is needed for code y_i, whereas $f \in S \wedge x_{if} > \text{UB}_f$ implies a positive error of feature f and a downward correction is needed. For instance, in Figure 5.2, the patient stay p_1 is miscoded. To reproduce the code E44, we have to change the value of the feature "BMI" by at least $|x_{if} - \text{LB}_f|$.

Given a correction set $c = (f_1, f_2, \ldots, f_{|c|})$ and a group of patient stays $\{p_1, p_2, \ldots, p_n\}$ in the corresponding cluster, the error distribution of feature $f, f \in c$ is characterized by $D_f = \{\text{MOE}_f^1, \text{MOE}_f^2, \ldots, \text{MOE}_f^n\}$. After this, we propose to represent the error distribution of feature f by computing its normalized histogram H_f containing s bins. Based on this, the error distribution of a cluster c can be denoted as $D_c = \{H_{f_1}(\text{D}_{f_1}), H_{f_2}(\text{D}_{f_2}), \ldots, H_{f_{|c|}}(\text{D}_{f_{|c|}})\}$.

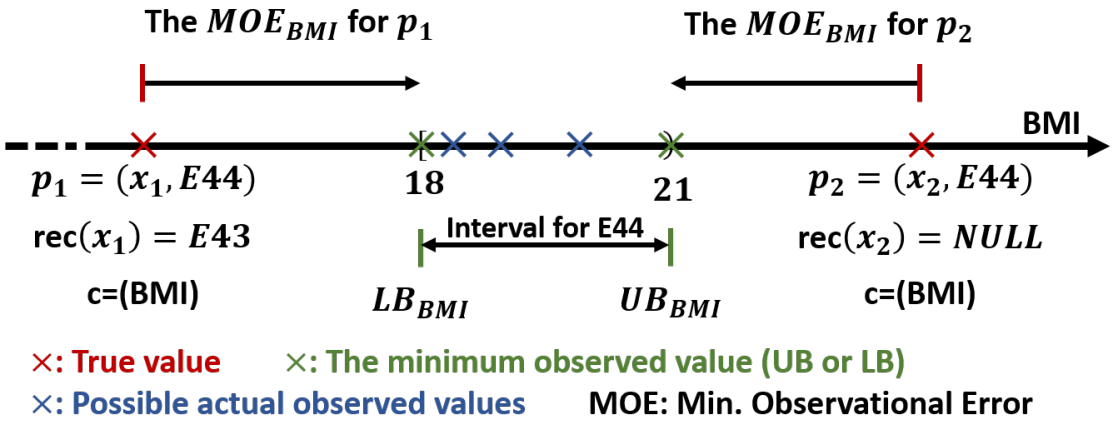


Figure 5.2.: The minimal observational error

We also propose to compare error distributions on a given feature f using histogram intersection, which is frequently used for image similarity evaluation [108]. Given two normalized histograms H_f and H'_f of feature f, each containing s bins, the similarity between two histograms is defined by the formula $\text{Sim}(H_f, H'_f) = H_f \cap H'_f = \sum_{i=1}^{s} \min(H_f(i), H'_f(i))$.

5.3.5. Model variants

This section extends the optimal code correction model (CCVM) to address the balance of financial benefit and the number of miscoded cases to review and the balance of coding correction workload and the number of miscoded cases to review. We address these balances by variants of the CCVM model and use the same notation.

5.3.5.1. Financial benefit vs. the number of miscoded cases to review

This subsection considers the balance of financial benefits and the number of cases to review. This model variant is similar to the CCVM model but with maximal coding workload maxIF replaced by the maximal number maxI of miscoded patient stays to review. This is equivalent to representing the code correction workload by the number of cases to review instead of the number of features to check, i.e.,

Variant V.5.1:

$$\max G(u, v) = \sum_{i \in I} \sum_{c \in C^{(m)}) \cap C_i^+} \left| \text{gain}_i \left(\text{rec}(x_i) - \text{gain}_i \left(y_i \right) \right) \right| \times v_{ic} \tag{5.8}$$

subject to constraints (5.2)-(5.4), (5.6)-(5.7) and

$$\sum_{i \in I} \sum_{c \in \{C^{(m)} \cap C_i^+\}} v_{ic} \leq \text{maxI} \tag{5.9}$$

In contrast to CCVM, the integrity constraint of binary variables of Variant 1 can be relaxed. In this sense, Variant 1 is easier to solve than CCVM.

Theorem T.5.2:
Let LP(Variant V.5.1) be the linear relaxation of Variant V.5.1, i.e. with $v_{ic} \in 0,1$ replaced by $v_{ic} \in [0,1]$. Then $G^{\text{LP}}(\text{Variant V.5.1}) = G^{\text{Variant V.5.1}}$, where $G^{\text{LP}}(\text{Variant V.5.1})$ and $G^{\text{Variant V.5.1}}$ are the optimal criterion value of LP(Variant V.5.1) and Variant V.5.1.

Proof of Theorem T.5.2:
For any feasible solution u, the optimal v that is feasible for both LP(Variant V.5.1) and Variant V.5.1 is determined as follows: : (i) determine the set of patient stays p_i with at least one correction set selected, i.e., $\sum_{c \in C_i^+} u_c \geq 1$, (ii) rank the patient stays in non-increasing order of $\left|\text{gain}_i\left(\text{rec}\left(x_i\right) - \text{gain}_i\left(y_i\right)\right)\right|$, (iii) assign the first maxI patient stays to any of their selected correction sets. Q.E.D.

5.3.5.2. Coding correction workload vs. the number of miscoded cases to review

This subsection departs from the previous financial benefit approaches and considers the problem of how to balance the coding error corrections and the workload represented by the cumulative number of features to review. This model variant is useful when financial information is missing. The problem is modeled as follows:

Variant V.5.2:

$$\min G(u,v) = \sum_{i \in I} \sum_{c \in C^{(m)} \cap C_i^+} |c| \times v_{ic} \tag{5.10}$$

subject to constraints (5.2)-(5.4), (5.6)-(5.7) and

$$\sum_{i \in I} \sum_{c \in C^{(m)} \cap C_i^+} v_{ic} = \text{maxI} \tag{5.11}$$

Again, the integrity constraint of binary variables of Variant V.5.2 can be relaxed. In this sense, Variant V.5.2 is easier to solve than CCVM.

Theorem T.5.3:
Let LP(Variant V.5.2) be the linear relaxation of Variant V.5.2, i.e. with $v_{ic} \in 0,1$ replaced by $v_{ic} \in [0,1]$. Then $G^{\text{LP}}(\text{Variant V.5.2}) = G^{\text{Variant V.5.2}}$, where $G^{\text{LP}}(\text{Variant V.5.2})$ and $G^{\text{Variant V.5.2}}$ are the optimal criterion value of LP(Variant V.5.2) and Variant V.5.2.

The proof is similar to that of Theorem T.5.2 with $\left|\text{gain}_i\left(\text{rec}\left(x_i\right) - \text{gain}_i\left(y_i\right)\right)\right|$ replaced by $|C|$ and hence omitted.

Of course, the balance of coding workload and coding error corrections can also be equivalently addressed by the following model variant:

Variant V.5.2′:

$$\max G(u,v) = \sum_{i \in I} \sum_{c \in C^{(m)} \cap C_i^+} v_{ic} \tag{5.12}$$

subject to constraints (5.2)-(5.4), (5.6)-(5.7) and

$$\sum_{i \in I} \sum_{c \in \{C^{(m)} \cap C_i^+\}} |c| \times v_{ic} \leq \text{maxIF} \tag{5.13}$$

However, as CCVM, the integrity constraint of binary variables of Variant V.5.2' cannot be relaxed, and hence, Variant V.5.2' is harder to solve than Variant V.5.2. For this reason, Variant V.5.2 is better for studying the balance of coding workload and coding error corrections.

5.3.5.3. Other variants

Variant V.5.3: To limit the number of subjects to review, we propose an additional constraint (5.14) coupled with the CCVM model (5.1)-(5.7), i.e,

$$\sum_{i \in I} \sum_{c \in \{C^{(m)} \cap C_i^+\}} v_{ic} \leq \text{maxI} \tag{5.14}$$

where the number of subjects to review can not exceed a predefined upper bound maxI.

Variant V.5.4: To avoid generating small clusters, we propose a variant of CCVM model (5.1)-(5.7) incorporating three additional constraints (5.15)-(5.17), i.e.,

$$\text{minSize} * v_{ic} \leq w_c, \ \forall \ (i,c) \in I \times \left\{C^{(m)} \cap C_i^+\right\} \tag{5.15}$$

$$\sum_{i \in I} v_{ic} = w_c, \ \forall \ c \in \{C^{(m)} \cap C_i^+\} \tag{5.16}$$

$$w_c \in \{0,1,\ldots,|I|\}, \ \forall \ c \in C^{(m)} \tag{5.17}$$

where w_c is an integer decision variable indicating the number of patient stays contained in the cluster c, the predefined lower bound minSize specifies the minimum size of clusters.

5.4. Implementation of the proposed approach in practice

In practice, the implementation of the proposed approach starts with the identification of various (numerical or categorical) descriptive features relevant to the coding

procedure of a given set of medical codes. After that, we define a workflow to extract a dataset of patient stays with relevant features. Finally, we have to define a personalized goal (i.e., maximizing the financial gains, maximizing the number of coding errors reduced, etc.) and implement one of the models presented in this chapter.

For practical implementation, the proposed approach is flexible enough. So that we can take into account some piratical considerations. For instance, we can add the constraints (5.15)-(5.17) to our model. This can limit the minimum size of discovered clusters.

In addition, as coded data is continuously generated, it is necessary to execute the selected model periodically. We do not suggest running the model for each new data generated due to the expensive computation power needed. On the other hand, the provided histogram statistics and frequent patterns will no longer be reliable If the amount of data is insufficient. Instead, we recommend modifying the selected model according to the new version of coding conventions and executing the model regularly (i.e., quarterly or annually).

5.5. Conclusion and perspectives

This chapter addressed the problem of optimal code correction subject to stakeholders' requirements of a French territory hospital. We propose a data-driven approach combining correcting sets for profiling miscoding behaviors and mixed integer programming for medical review allocation.

In real-life applications, the proposed approach is flexible enough to fit stakeholders' objectives. The model variants presented in section 5.3.5 can be used as a reference, and additional constraints can be taken into account. In addition, as data is continuously generated, it is necessary to apply the proposed approach periodically. We do not recommend applying the proposed approach for each new data point obtained due to the necessary computation cost. We recommend applying the proposed approach regularly (i.e., quarterly, annually) and adjusting the model based on updates of coding conventions and the health reimbursement policy.

This chapter presents the proposed approach and provides some model variants for more flexibility. Practical considerations are also taken into account. In the next chapter, we present a real-life case study on the CHU-SE by applying the proposed approach. Experimental results show that the proposed approach is efficient and reliable.

6. In-site operation, evaluation and validation of the integrated optimization model in the University Hospital of Saint-Etienne

Summary

Abstract of the chapter

Medical miscoding has a significant negative impact on hospitals, with a financial loss for under-coding and a penalty for over-coding. Whether a medical review is necessary for all descriptive features of a miscoded subject? Is it possible to reduce unnecessary medical reviews without compromising the goal of increasing hospital financial benefits? This chapter attempts to answer these questions by practically applying the data-driven optimization approach presented in Chapter 5 on a real-life dataset to determine a limited number of miscoding clusters and the set of features to review for each in order to best balance the financial benefits and the medical review workload. In the dataset, around 11.18.% of the 33143 cases audited reflected an ICD change and a subsequent DRG (Diagnosis Related Group) change. The application of the proposed approach leads to a significant increase in hospital fiscal revenue of nearly 6,992,489.69 € while reviewing only a small number of descriptive features (5293 out of 22056 features, or 24% of features). Results show that the proposed approach is able to remove all coding errors present in the given dataset. Causes are also provided for each discovered coding error subtype to ameliorate medical coders' coding practices. Furthermore, The proposed approach allows the decision-maker to balance the cost-benefit and the requirement of public health institutions (i.e., miscoding rate).

Résumé du chapitre

Est-il nécessaire de effectuer un contrôle médical pour toutes les variables descriptives d'un sujet mal codé ? Est-il possible de réduire les contrôles médicaux inutiles sans compromettre l'objectif d'augmenter les bénéfices financiers des hôpitaux ? Cet article tente de répondre à ces questions en appliquant de manière pratique l'approche d'optimisation basée sur les données présentée au chapitre 5 sur un ensemble de données réel. Dans cet ensemble de données, environ 11,18 % des 33143 cas audités reflétaient un changement de code CIM et un changement ultérieur de GHM (groupes homogènes de malades). L'application de l'approche proposée conduit à une augmentation significative des recettes fiscales des hôpitaux de près de 6 992 489,69 €, tout en n'examinant qu'un petit nombre de variables descriptives (5293 sur 22056 variables, soit 24% des variables). Les résultats montrent que l'approche proposée est capable de supprimer toutes les erreurs de codage présentes dans l'ensemble de données donné. Les causes sont également fournies pour chaque sous-type d'erreur de codage découvert afin d'améliorer les pratiques de codage des codeurs médicaux. En outre, l'approche proposée permet au décideur d'équilibrer le rapport coût-bénéfice et les exigences des institutions de santé publique (c'est-à-dire le taux de mauvais codage).

6.1. The protocol for method evaluation and validation

6.1.1. Context

In hospitals, a discharge summary (RSS) is generated for each patient by a physician at the time of discharge. The discharge summary often contains a short description of the patient's history and health condition, as well as some coded medical information (ICD codes (diagnosis), CCAM codes (medical procedure), etc.). Note that CCAM codes are not involved in this study (the coding of undernutrition).

In general, 1 RSS contains 1 DP (ICD primary diagnosis code) and 0+ DAS (zero or more than zero ICD significant associated diagnoses). The undernutrition-related ICD codes are coded in DAS and have different severity levels, i.e., (i) mild or moderate undernutrition (E44): level 2, (ii) severe undernutrition (E43): level 3. In the SSPIM, the coding task is done using an integrated software (a keyword search tool, the default editor) in the Web100T platform.

ICD codes are used in many aspects of health services, such as coding quality control, health insurance control, and health services reimbursement, and therefore have a financial impact on society. More specifically, under-coding results in financial loss to the hospital, while over-coding often leads to health insurance control (inspection and verification by insurance companies). At the SSPIM, CQT (transversal quality control) on undernutrition is based on the passage of dieticians (supported by the prescription of food supplements) and on the BMI value.

A grouping heuristic (algorithm) is used to generate health service reimbursements. Given a patient p_i and his or her standardized discharge summary (RSS) x_i, the heuristic algorithm generates the corresponding health service reimbursement (or the tariff) through the following workflow: $x_i \Rightarrow (\text{GHM}_{i1}, \text{GHM}_{i2}, \ldots) \Rightarrow \text{GHS}_i \Rightarrow \text{tariff}_i$.

The severity of a patient is defined by various factors, including undernutrition-related factors and others. The code E44 (moderate under-nutrition) is at a severity level of 2. In other words, if a patient is diagnosed with moderate undernutrition, his or her severity is at least 2, i.e., severity level >= 2. The final severity is determined by taking all assigned ICD diagnosis codes, i.e., DP, DAS, DR. Note that the severity level of a patient is also reflected in the last character of his/her GHM code and GHS code.

In SSPIM, a regular CQT process is as follows: given a patient at level 1 with moderate under-nutrition (E44), then the medical coder has to change the severity level from 1 to 2 and thus get a higher tariff. This is so-called the correction of under-coded subjects. In this case, we get a positive tariff difference ($\text{tariff}(\text{rec}(x_i)) - \text{tariff}(y_i) > 0$) and can therefore increase the hospital's financial revenue. Next, if the patient is at level 3 and has moderate malnutrition (E44), then we stay at level 2 (no change in the tariff). Obviously, all the over-coded cases were ignored, and the finical impact of overcoding thus cannot be assessed. The severity of a patient and the health services

reimbursement are positively correlated. The more severe a patient's condition is, the more health reimbursement the hospital receives.

In this section, we describe the protocol for evaluating the optimization method proposed in section 5. Given a population $P = \{p_i\}_{i=\{1,2,...,|P|\}}$, each patient stay $p_i = (x_i, y_i)$ is characterized by a feature vector x_i (extracted from patient discharge summary and EHRs) and is labeled with a ICD code y_i. Under this context, the PMSI code y_i (the ICD code assigned by PMSI) is considered the gold standard. The optimization approach proposed in section 5 is leveraged for detecting and correcting undernutrition coding errors in the PMSI. Specifically, the proposed approach is used to detect as well as explain the differences between the re-coded ICD codes $\text{rec}(x_i)$ and the PMSI codes y_i. The assignment of re-coded codes is based on the HAS recommendations.

There are various causes of coding errors: (i) data entry error in patient discharge summary (RSS); (ii) missing data in patient EHRs and discharge summary (RSS); (iii) the HAS guidelines are used for clinical purposes and code quality control (QC) but can not describe coding process of the PMSI code y_i; (iv) the coding process of the PMSI takes into account the medical knowledge of medical coders. Instructions for coding are integrated into the methodological guideline (CoCoA [6]), and the last venison of the HAS guideline. To avoid coding errors caused by reasons (i) and (ii), We encourage physicians and nurses to enter related information during hospitalization at least twice, as it is critical for under-nutrition coding and is an indicator in the national-level quality control.

6.1.2. Objective

The objective of this chapter is to evaluate and validate the optimization approach proposed in chapter 5 in a real situation in the SSPIM. Specifically, the following workflow is designed to identify the origin of a coding error: characterizing the coding error ⇒ identifying the reasons for the miscoding ⇒ recommending what should be changed in the future (prevention is better than remedy) = identifying the miscoding habits (or behaviors).

In which medical services, under which conditions, and in which areas of the patient space (the particular subpopulations) that patient stays are subject to miscoding risk (operational assessment)? Which types of coding errors should be targeted and corrected? We prioritize targeting significant or frequent coding errors first. Is there a simple and efficient way to identify under-coded patient stays? What is the minimum amount of data (i.e., the number of features or feature vectors) to be checked for a given number of miscoded patient stays? What is the time spent on each record (file or feature vector), i.e., the cost of acquiring and verifying each variable and each file? In this retrospective study, we assume that features have the same acquisition difficulty. In future prospective studies, we recommend measuring the acquisition difficulty of each feature and weighting each feature to fit the real situation in the SSPIM.

In this chapter, we conduct a retrospective study and evaluate the proposed method

on historical data for 2018. In addition, the SSPIM plans to conduct a prospective study in the next year (2023) to verify the effectiveness of the proposed method.

6.1.3. Methodology

6.1.3.1. Optimization algorithm deployment and application

Deployment of the optimization algorithm

EMSE deploys the algorithm to SSPIM. Specifically, the different tasks are assigned as follows:

- SSPIM proposes to integrate the algorithm coded in Python into a web application with the Django server in the form of an API (application programming interface), sending POST requests, inputting data, retrieving the responses from the server, managing the display (front end web pages), and making the experimental results available to TIMs.
- EMSE is required to provide a Python script that can be encapsulated in the form of functions that can be called remotely. EMSE also needs to provide a technical document describing the dependencies of the Python project (i.e., python libraries used in the project), each function's input parameters, the functionality each function achieves, and the response as well.
- CPLEX is a solver for optimization algorithms written in C that can be called from java, Python, c++, etc. It is adopted by EMSE for solving the optimization problem. CPLEX is proprietary software and is free for projects with research or educational purposes. In addition, installation and use of CPLEX require large storage space and high computational costs. For reasons of data security and the cost of software subscription, the PuLP library (for the python language) is considered. The PuLP can connect to a solver that is installed on PCs, including (i) proprietary solvers such as CPLEX and GUROBI and (ii) open-source solvers such as GLPK, PULP-CBC, COIN-OR, and SciPy.

Data extraction

The SSPIM performs the following steps to extract relevant data from the hospital's database (WEB-100T),

1. Conducting a simulation on the PMSI data to see what happens when undernutrition ICD codes are removed from the discharge summary (RSS). Performing a grouping simulation (the grouping heuristic) to see the impact on the level of severity and the tariff.
2. Identifying the hospitalization ID using the PMSI data under the following given conditions.

a) The patient's first admission date is within a given time interval (usually one month) - month M. For example, patients with an admission date between Feb. 1 and Feb. 28, 2023;

b) Patients aged 18 years and older;

c) The patient is an inpatient, and the length of stay (LOS) is at least one night;

3. Extracting relevant data (variables) from the EHRs (provided by the Information Systems Department of the CHU-SE).

   a) Biological data: albumin, pre-albumin, CRP (only used in the old HAS recommendation);

   b) Weight (at first admission, at month M), weight (history at month M-3), weight (history at month M-6), BMI;

   c) Filtering out abnormal data records by identifying inconsistencies (e.g., some BMI data are outliers, e.g., BMI < 5 or BMI > 120);

4. Calculating weight change: weight evaluation within three months (in percentage), weight evaluation within six months (in percentage).

5. Retrieving data on dietitian interventions (the CQT process).

6. The cross join of PMSI data and the data provided by the dietitian. The SSPIM can thus get more relevant data (variables).

   a) Date of birth (used to calculate patient age);

   b) Gender;

   c) GHM (with severity level);

   d) Length of stay (LOS)

   e) number of visits to medical, surgical, obstetrical, continuing care, rehabilitation units;

   f) Deaths;

7. Anonymizing data in SSPIM before method evaluation.

Application of the proposed algorithm

The optimization algorithm identifies a fixed number of files (or variables) to review by employing a threshold, e.g., the 200 most error-prone files or variables. At worst, the effort (the number of files checked) will be the same as the benefit (the number of errors removed). We conduct two separate experiments: (i) a control group - the files selected by the SSPIM, and (ii) an experimental group: the files identified by the optimization algorithm. We merge the two groups of data samples by identifying the common records in both groups (using hospitalization ID). Next, we analyze data samples that do not belong to the intersection of the two groups. Does SSPIM remove records from its list to remain a constant effort? i.e., the number of records checked is equal to the number of records checked in the initial SSPIM list.

The outputs of the optimization algorithm are shown as follows:

1. Subpopulations
   a) The optimization algorithm generates multiple subgroups, each representing a specific miscoding behavior;
   b) With limited resources (i.e., the number of variables or files to review), the optimization algorithm selects subgroups to review and corrects the coding errors in the selected subgroups;
   c) For each of these subgroups, the optimization algorithm identifies the variables to be checked. For each variable to be checked, the algorithm shows the contradiction with the HAS recommendation;
2. All coding errors (both under-coding and over-coding) modify the severity level of the GHM.
   - Re-identifying the groups that are mostly responsible, taking into account the severity of the GHM;
3. Deviation between PMSI codes y_i and HAS recommendations $\text{rec}(x_i)$.
   - Identifying the coding rule that has been violated. A bit-mask allows us to indicate the violated coding rule (one bit per HAS atomic rule). This information should be displayed on the front end of the web application and should be available to medical coders;

6.1.3.2. Web application development and deployment

The web application, as illustrated in Figure 6.1, can be divided into three parts, which include the optimization algorithm, the development and deployment of the web application, and the evaluation of the optimization algorithm. The optimization algorithm is provided by EMSE in the form of Python code. As a starting point, the optimization algorithm takes the under-nutrition data from the WEB100T database as an input and generates experimental results. The experimental results are then transferred to the back-end database via the web front end.

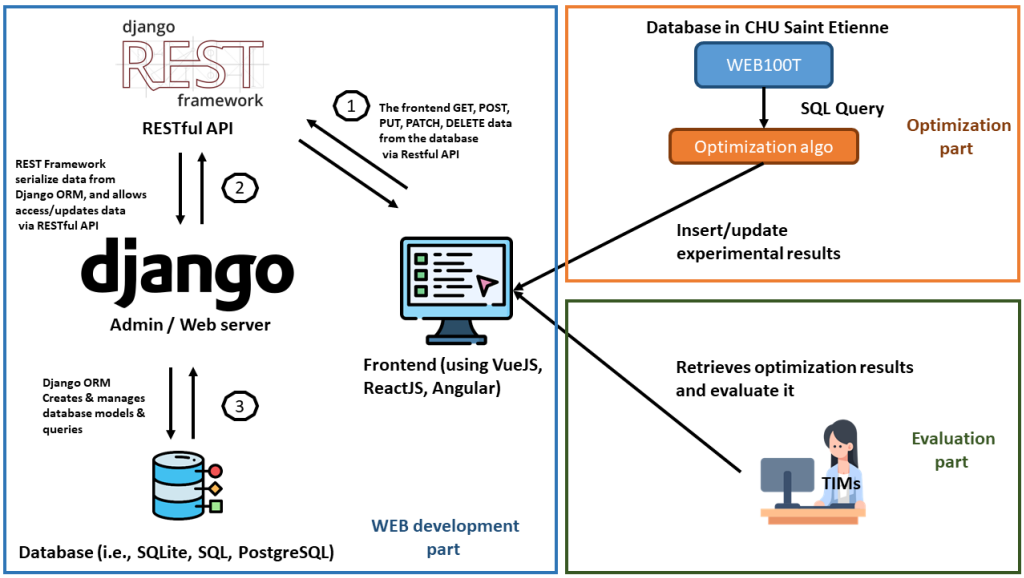


Figure 6.1.: Standard workflow of the web application.

The second part is the development of the web application. We plan to construct a web application using the Django web framework and then expose the data stored in the back-end database to other applications via RESTful APIs and front-end display. Finally, a front-end web view (constructed by VueJS, ReactJS, or Angular) is made available for the TIMs. Since the optimization part is already done (or implemented), in the next step, the SSPIM is required to implement the web application and evaluate the optimization model proposed in chapter 5. In general, the development of the Django web application consists of multiple tasks, i.e.,

- The design and construction of Django internal database (SQLite, SQL, etc.);
- The design and construction of the Django web application. This tutorial [1] provides a comprehensive and complete tutorial for this topic;
- The design and construction of Django RESTful APIs. APIs are used for information transfer;
- The design and construction of web front-end web pages using VueJs, ReactJs, Angular, etc.
- The rental of a web server;
- Deploying the Django web application to the web server;

[1] Django Web Framework (Python): https://developer.mozilla.org/en-US/docs/Learn/Server-side/Django

6.1.3.3. The evaluation and validation of the optimization algorithm

Every month, at the end of month M, the SSPIM works on the data for the month M-1. For example, at the end of February, data for January are processed. At the beginning of each CQT process, an upper bound on how many cases to review is to be discussed with the TSH in SSPIM.

1. Evaluation metrics
    a) The number of coding errors corrected, the number of files (or variables) reviewed;
    b) Financial impact (including the potential gain from correcting under-coded samples and the avoidable loss or penalty from correcting over-coded samples). The TIM or MIM, who reviews a record and modify the corresponding ICD code, will be informed of the tariff change in Web100T. The change in tariff is then recorded and can be used to compute the final financial impact.
    c) The EMSE provides indicators to track the changes: a column (binary variable) indicating whether a data sample is selected and corrected or not, a column (variable) indicating the selected variables to review, a column (variable) indicating the change in value for each selected variable, a column for adding free comments, which is useful for the TIM but not for this study, one column indicating the new ICD code assigned, one column indicating the new GHM assigned, one column indicating the new GHS assigned, one column indicating the new tariff by changing the old ICD code;
    d) The cost: Number of cases reviewed;
    e) Effort to review files (the time used to review a specific variable, but it is difficult to measure.);
    f) The benefit: percentage of coding errors corrected by the optimization algorithm, the percentage of coding errors corrected by the manual method (CQT process);
    g) percentage of data samples that led to a change in tariff;
2. Qualitative analysis: Identification of subgroups at risk of miscoding;
3. Quantitative analysis: how much benefit can be gained by applying the optimization algorithm vs. how much benefit can be gained by using the "classic" method (CQT process);

Note that the most time-consuming step is when DSI [1] extracts the data from the DPI [2].

[1]The information systems department (DSI) is the department responsible for an institution's information system (IS). It is in charge of defining the IS architecture, designing, installing, deploying, and operating the IS. (Direction des systèmes d'information in French

[2]An electronic health record (EHR) is a collection of patient history and health information in digital

6.1.3.4. Ethics and expected results

Ethics: This study is not interventional, and will not interfere with the patient's treatment. We do not expect benefits from the patients. We work on anonymized data. However, for EMSE, the data is not anonymous to SSPIM. Relevant data from the EHR are licensed to EMSE for relevant studies. However, EMSE has to submit an authorization request to the CHU-SE ethics committee before using the data.

Expected results: Evaluating the benefits of this approach for the SSPIM: The time saved and the financial benefits generated for the CHU-SE.

6.2. Retrospective evaluation and validation of the optimization model

Malnutrition affects more than 2 million patients across France. With such a high disease base, even a tiny error in code assignment can lead to a significant impact on the operational efficiency of healthcare institutions. This chapter presents the application of the proposed approach to a real-life study case. The dataset used is come from the DIM of the CHU-SE in France. Medical information centers such as the DIM of CHU-SE are institutions connected to the French national health information system, where medical coding is an essential task to ensure the operation of hospitals.

The medical code censoring task in the DIM is to check the correctness and completeness of medical codes prior to submitting coded health information to the French Health Authority (HAS). A number of medical reviews are dedicated to the screening and correction of inaccurate medical codes. However, the current practices in the DIM rely on a heuristic of reviewing all EHRs of selected hospital stays to raise financial reimbursement from the HAS. This strategy results in an increase in labor costs that, at the national level, lead to a decrease in the operational efficiency of the national healthcare system.

In the remainder of this chapter, we first present the optimal code correction solution for the miscoded cases, including detailed experimental results of each component of the approach proposed in chapter 5. At the end of this chapter, we also analyze the financial impact on the hospital's fiscal revenue.

6.2.1. Study population

The dataset contains patients admitted to the CHU-SE in 2018, for whom we have EHRs of their hospitalization, as well as malnutrition-related information. Subjects under eighteen years old were filtered out from the dataset. Eventually, a data set with 32856 ($|P| = 32856$) anonymized EHRs was formed. A preliminary analysis was conducted to select a set of relevant features to include in this study case. Relevant

formats. (dossier patient informatisé (DPI) in French)

descriptive features are presented in Table 6.1. In total, 17 distinct features are taken into account.

The dataset is highly imbalanced and includes 11.18% miscoded patient stays ($|I| = 3676$). Among the entire population, 2.92% of subjects died during their hospitalization. Moreover, about 29.2% cells of the dataset are missing. These missing values are filled using an imputation method based on weighted k-Nearest Neighbors (KNN) [62], as it has been proven to be the most efficient imputation strategy on this dataset [34]. More descriptive statistics on the whole population P can be found in Table 6.2.

In this particular study case, we are interested in malnutrition-related ICD codes, including without malnutrition (NULL), protein-calorie malnutrition of moderate and mild degree (E44), and severe protein-calorie malnutrition (E43), i.e., $\text{rec}(x), y \in Y, Y = \{\text{NULL, E43, E44}\}$. The term $|Y|$ indicates the number of unique medical codes presented in set Y. The space of a recommendation function rec on $\mathbb{R}^{z'}$ is defined by $\mathbb{R}_r^{z'}$ and $\mathbb{R}_r^{z'} \subseteq \mathbb{R}^{z'}$. In this study, the whole space consists of multiple subspaces, i.e., $\mathbb{R}^{z'} = \mathbb{R}_{r_1}^{z'} \cup \mathbb{R}_{r_2}^{z'}, \ldots, \cup \mathbb{R}_{r_{|Y|}}^{z'} = \mathbb{R}_{NULL}^{z'} \cup \mathbb{R}_{E43}^{z'} \cup \mathbb{R}_{E44}^{z'}$. In addition, the subspaces of recommendation functions are mutually exclusive, i.e., $\mathbb{R}_{r_i}^{Z'} \cap \mathbb{R}_{r_j}^{Z'} = \emptyset, i \neq j, i, j \in \{1, 2, \ldots, |Y|\}$.

Table 6.1.: dataset content

Features related to the medical coding task $F = \{f_1, f_2, \ldots, f_{z'}\}, z' \leq z$	age, body mass index (BMI), weight evolution in 1 month (in percent), weight evolution in 6 months (in percent), albumin, C-reactive protein (CRP)
Healthcare-pathway related features $T = \{f_{z'}, f_{z'+1}, \ldots, f_z\}$	gender, body weight, length of stay (LOS), death occurred, containing missing values, number of visits to various medical departments including the medical department(med), surgical department(surgery), rehabilitation ward(rehab), intensive care unit (ICU), continuous monitoring department(CM), department of obstetrics(obstet)

Table 6.2.: Statistics of main variables for the whole population P

Variable	Min	Max	Mean	Std Dev	Skew.	Kurt.
Age	18.0	104.0	62.54	20.022	-0.47	-0.80
Weight	25.50	125.0	73.53	10.14	0.51	5.31
BMI	11.020	59.45	26.42	3.49	1.25	8.75
wgt evol in 1 mo	-0.65	0.32	6.8e-4	0.015	-8.81	261.31
wgt evol in 6 mos	-0.38	0.21	6.8e-3	0.017	-4.96	42.57
Albumin	9.10	57.90	36.83	3.17	-2.83	14.21
CRP	0.30	519.59	22.0	42.16	4.19	22.36

Table 6.3.: Statistics of main variables for under-coded cases I^+

Variable	Min	Max	Mean	Std Dev	Skewness.	kurtosis.
Age	18.0	99.5	71.16	17.62	-1.21	0.93
Weight	25.50	122.0	59.84	14.44	0.61	0.51
BMI	11.020	52	21.79	4.95	0.94	1.92
wgt evol in 1 mo	-0.65	0.20	-0.021	0.051	-3.90	27.45
wgt evol in 6 mos	-0.31	0.10	-0.049	0.044	-0.60	1.58
Albumin	9.80	54.59	34.12	5.42	-0.68	1.48
CRP	0.30	356.89	18.14	34.85	4.83	30.089

In addition, a subject $p_i = (x_i, y_i)$ is said over coded if $\text{rec}(x_i) < y_i$, and is said under coded if $\text{rec}(x_i) > y_i$. Table 6.2 shows the basic statistics of the medical codes included in the dataset. Correction of under coded subjects leads to financial gains, i.e., $\text{gain}_i(\text{rec}(x_i)) > \text{gain}_i(y_i)$, while correcting over coded subjects can avoid potential financial penalties from French Health Care Authority (HAS), i.e., $\text{gain}_i(\text{rec}(x_i)) < \text{gain}_i(y_i)$. The population I^+ and I^- ($|I| = |I^+| + |I^-|$) consist of 1545 under-coded subjects and 2131 over-coded subjects, respectively. Table 6.3 and 6.4 show some basic statistics for the population I^+ and I^-, respectively.

Table 6.4.: Statistics of main variables for over-coded cases I^-

Variable	Min	Max	Mean	Std Dev	Skewness	kurtosis
Age	18.0	104	76.12	15.63	-1.15	1.19
Weight	33.70	124.5	71.085	9.57	0.11	4.091
BMI	17.069	45.409	25.73	3.28	1.16	7.15
wgt evol in 1 mo	-0.085	0.16	1.4e-3	0.016	-3.45	34.43
wgt evol in 6 mos	-0.13	0.21	-5.9e-3	0.018	-0.86	27.35
Albumin	9.10	51.90	37.74	5.26	-1.63	2.54
CRP	0.30	458.29	37.31	56.94	2.91	10.17

In addition, to simulate the process of health service reimbursement, we generated some synthetic DRG codes based on a heuristic D provided by the DIM. The heuristic D is defined by a function $D : Y \rightarrow \sigma$, where σ is a set of relevant DRG codes, i.e., $\sigma = \{\text{DRG}_1, \text{DRG}_2, \ldots, \text{DRG}_{|\sigma|}\}$. For instance, given a patient stay $p_i = (x_i, y_i), p_i \in P$, the DRG codes are generated by $\text{DRG}_i = D(y_i)$, and $\text{DRG}'_i = D(\text{rec}(x_i))$. Therefore, the financial benefit of correcting the patient stay p_i can also be expressed by $\left|\text{gain}_i\left(D\left(\text{rec}\left(x_i\right)\right)\right) - \text{gain}_i\left(D\left(y_i\right)\right)\right|$.

6.2.2. The current practice of the medical information department (DIM)

For malnutrition-related ICD codes, the quality control (QC) process is handled by a medical coder from the DIM. The current practice of the nutritionist has been

correcting only coding errors for the following two special cases: (i) patient stays $p_i \in P$ that are classified as without undernutrition, i.e., y_i = NULL, rec(x_i) = E43, and y_i = NULL, rec(x_i) = E44; (ii) patient stays $p_i \in P$ for whom a visit to nutritionist is appointed. Such a practice is conservative and considers only a small fraction of coding errors.

Table 6.5.: One-year estimation of the QC process

No. patient stays reviewed	No. features reviewed	No. errors corrected	Financial benefits
1476	8856	108	205,674.36 €

For the quality control conducted in February 2021, the medical coders reviewed EHRs of 123 patient stays (corresponding to $123 \times 6 = 738$ descriptive features) and corrected 9 coding errors. The financial gains obtained from this QC process are 17,139.53 €. Based on this, the table 6.5 estimates the financial benefits of quality control for one year (twelve months).

6.2.3. Optimal code correction and medical review rationing

According to the coding manual [5], we select a set of relevant features F to construct several recommendation functions rec(x) = NULL, rec(x) = E43, and rec(x) = E44. Once the recommendation functions are constructed, we are able to compute a set of minimal correcting sets C_i for the population I based on the proposed algorithm 1. After that, with the feature set F, we compute and put $C^{(m)}$ into the CCVM model to limit the size of the problem.

With the above pre-computed constants, the CCVM proposed in section 5.3.3 is well defined. The decision variable v_{ic} either assigns or does not assign a medical review c to a subject i. The decision variable u_c either puts or does not put at least one patient stay into a correction cluster c. The goal is to maximize the total hospital reimbursement such that the cumulative number of features to review is below a pre-defined upper bound maxIF, i.e., to raise financial gains of under-coding and avoid financial penalties of over-coding.

For this study case, an optimization model is defined by a tuple of (m, K), where m is the maximum size of correcting sets, and K denotes the maximum number of clusters. Figure 6.2 shows the Pareto fronts of the resulting CCVM models, where a reimbursement (ratio) of 1 corresponds to the maximum benefit reachable, i.e, 6,992,490 € for the current dataset. A maxIF (ratio) of 1 corresponds to reviewing all $|I| \times |F| = 22056$ features. The yellow star denotes the optimal solution. The red triangle represents the current practice of the DIM, which is calculated according to the figures given in Table 6.5. As mentioned in section 5.2, a tradeoff should be made. A relatively large value of m guarantees the relatively fast convergence of the CCVM models, while a relatively small value of m tends to reduce the number of decision

variables as well as the computational cost. Besides, we can further speed up the convergence of Pareto efficiency as the number of clusters K increases. However, when the number of clusters K exceeds 4, the effect of K on Pareto efficiency starts to diminish. This phenomenon is also known as diminishing marginal utility. Based on the experimental results presented in Figure 6.2, the model with the parameters $(m = 4, K = 4)$ is therefore selected.

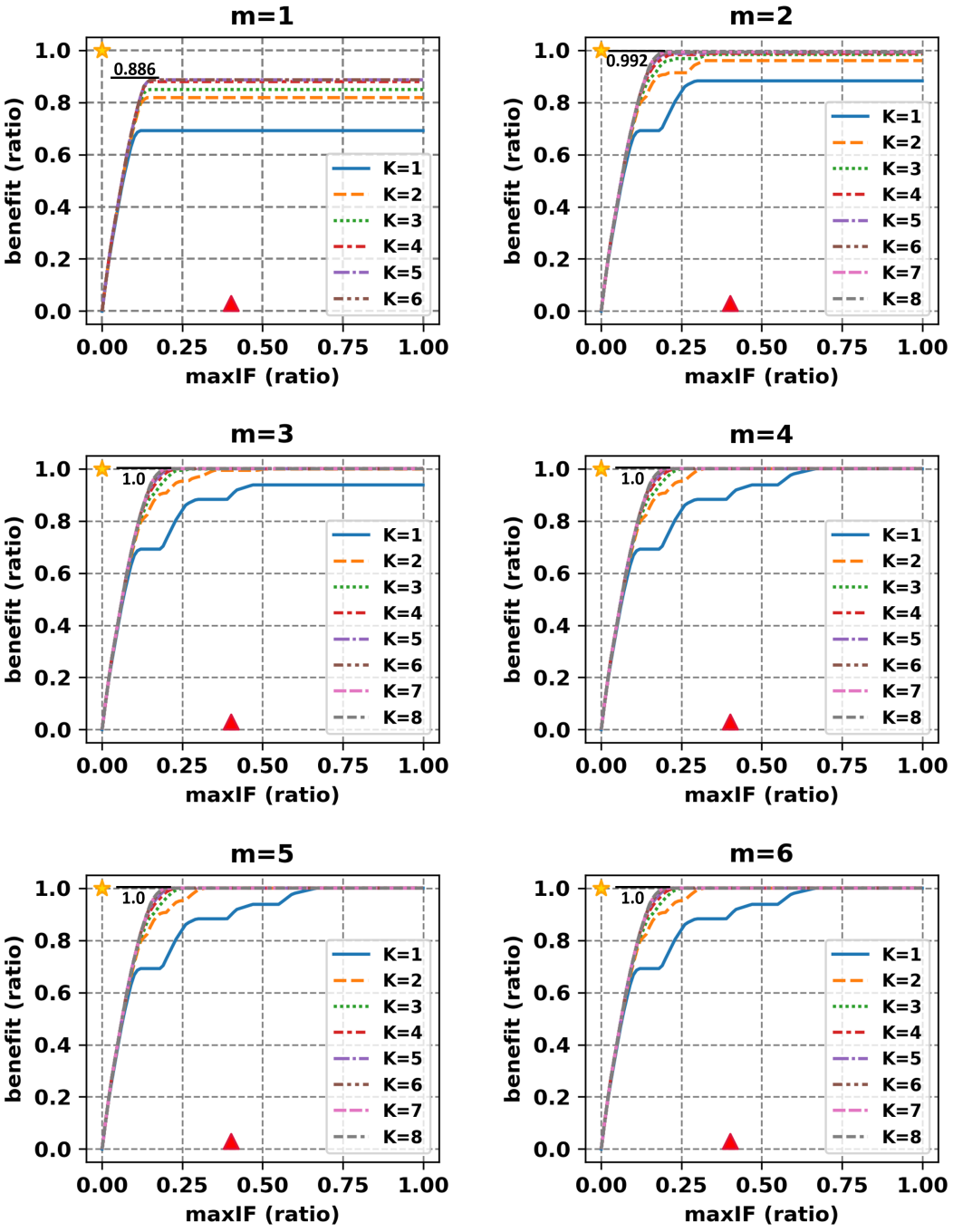


Figure 6.2.: Pareto fronts for the population I with different sizes (m) of correcting sets.

Table 6.6 shows the results of the CCVM model ($m = 4, K = 4$). Several evaluation metrics, including false-positive rate (FP(%)) and false-negative rate (FN(%)), are provided for performance evaluation. In simple terms, the FN reflects the current miscoding rate, while the FP indicates the reduction of the miscoding rate. The terms IF′ and I′ are the number of features reviewed and the number of subjects reviewed, respectively. By varying the pre-defined upper bound maxIF, code correction solutions corresponding to different levels of the tradeoff between the medical review budget and the increased reimbursement are obtained and provided to the decision-maker.

Table 6.6.: Experimental results of the CCVM model (m=4, K=4)

maxIF (ratio)	maxIF	IF′	I′	FP(%)	FN(%)	Revenue(€)
0	0	0	0	0	11.188	0
0.02	441.12	441	441	1.342	9.845	1316819
0.04	882.24	882	882	2.684	8.503	2401032
0.06	1323.36	1323	1323	4.0266	7.161	3381407
0.08	1764.48	1764	1764	5.368	5.819	4328575
0.10	2205.6	2205	2205	6.711	4.477	5090740
0.12	2646.72	2646	2646	8.0533	3.134	5739781
0.14	3087.84	3087	2932	8.923	2.264	6196202
0.16	3528.96	3528	3103	9.444	1.743	6520863
0.18	3970.08	3970	3204	9.751	1.436	6718737
0.20	4411.2	4411	3509	10.679	0.508	6887553
0.22	4852.32	4852	3623	11.0269	0.161	6875101
0.23	5072.88	5072	3675	11.185	0.00304	6990274
0.24	**5293.44**	**5293**	**3676**	**11.188**	**0**	**6992490**
0.25	5514	5514	3676	11.188	0	6992490
0.5	11028	11028	3676	11.188	0	6992490
0.75	16542	14704	3676	11.188	0	6992490
1	22056	14704	3676	11.188	0	6992490

From Figure 6.3, there is an evident trade-off between the workload allocated and the obtained false-negative rate. In this study, a false negative is a patient stay that is miscoded but is not allocated with a medical review. Intuitively, the more the medical reviews are allocated, the fewer false negatives we obtain since allocating medical reviews to all possible features will ensure to cover all miscoding sources (or features) but at the expense of an increased number of unnecessary medical reviews. On the other hand, the fewer the medical reviews we allocate, the less the unnecessary expenses (see Table 6.6), however, at the cost of failing to identify certain coding errors that might be critical for the increase in the hospital fiscal revenue. Notably, the objective of the CCVM model is to increase financial gains by eliminating coding errors rather than directly eliminating coding errors. A variant is provided to eliminate coding errors instead of increasing the financial gain in section 5.3.5.

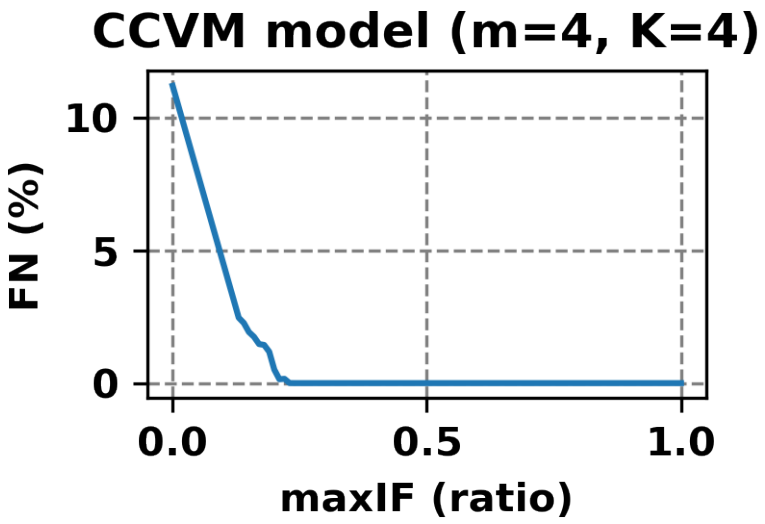


Figure 6.3.: Percentage of FN by number of features reviewed

With a maxIF(ratio) of 0.24, we audit 5293 descriptive features for the given population I. The CCVM model with parameters (m=4, K=4, maxIF=5293.44) eliminates all coding errors presented in the whole population P and yields a significant increase in fiscal revenue of 6,992,490 €, which shows the efficiency of the proposed approach. The Table 6.7 shows the optimal solution obtained by such a CCVM model. The most significant factor of underlying miscoding behaviors is the BMI entered.

Table 6.7.: The optimal code correction solution

Correcting cluster	Correcting set	No. of subjects	No. of overcoding	No. of undercoding
c_2	{BMI}	2541	1947	594
c_17	{wgt_evol_1mo, albumin}	779	184	595
c_26	{age, wgt_evol_1mo, wgt_evol_6mos}	230	0	230
c_52	{BMI, wgt_evol_1mo, albumin, wgt_evol_6mos}	126	0	126
-	-	3676	2131	1545

6.2.4. Characteristics of the selected clusters

In this subsection, we focus on extracting frequent patterns $X \Rightarrow Y$ from patient stays for each discovered cluster. As a data preprocessing strategy, binary variables are retained, and each integer variable is discretized into multiple binary variables. In Table 6.8 and 6.9, "obstet_0" means that subjects never visited the department of obstetrics during their hospital stay, and "surgeyr_1_2" means that subjects visited the surgical department once or twice during their hospitalization.

Table 6.8.: Top 3 frequent patterns for $Y = \{c_2\}$

ant.	cons.	sup.	conf.
{surgery_1_2, albumin_35_inf, wgt_evol_1mo_-5%_inf}	{c_2}	0.0630	0.919
{missing_value_existed, albumin_35_inf, wgt_evol_1mo_-5%_inf}	{c_2}	0.0628	0.918
{obstet_0, albumin_35_inf, wgt_evol_6mos_-15%_inf}	{c_2}	0.0620	0.955
{icu_0, wgt_evol_1m_-5%_inf, wgt_evol_1m_-10%_inf}	{c_2}	0.0573	0.709
{rehab_0, wgt_evol_1m_-5%_inf, wgt_evol_1m_-10%_inf}	{c_2}	0.0562	0.712

Table 6.9.: Top 3 frequent patterns for $Y = \{\text{over\_coding}, c_2\}$

ant.	cons.	sup.	conf.
{wgt_evol_1mo_-5%_inf, wgt_evol_6mos_-10%_inf}	{over_coding, c_2}	0.0585	0.707
{obstet_0, wgt_evol_1mo_-5%_inf, wgt_evol_6mos_-10%_inf}	{over_coding, c_2}	0.0584	0.706
{surgery_1_2, wgt_evol_1mo_-5%_inf, wgt_evol_6mos_-10%_inf}	{over_coding, c_2}	0.0580	0.706
{icu_0, wgt_evol_1m_-5%_inf, wgt_evol_1m_-10%_inf}	{over_coding, c_2}	0.0573	0.709
{rehab_0, wgt_evol_1m_-5%_inf, wgt_evol_1m_-10%_inf}	{over_coding, c_2}	0.0562	0.712

Table 6.8 and 6.9 show several frequent patterns for the largest cluster $c_2 = \{\text{BMI}\}$, which accounts for about 69.12% (2541) of miscoded patient stays $|I| = 3676$. The support values are relatively low. This is reasonable since the support value for cluster

c_2 can not exceed $2541/|P| = 0.0773$. We find several strong rules with high confidence values, implying that the miscoding behavior $c_2 = \{\text{BMI}\}$ is also associated with the healthcare pathway-related factors. In the process of coding malnutrition-related patient stays, subjects with the above profiles have a higher probability of being miscoded (or over-coded). Medical coders should be aware of patient stays with such profiles and double-check subjects' BMI values to avoid this type of coding error.

6.2.5. Error distribution of the selected clusters

In this subsection, we present error distributions of each feature in discovered correcting clusters.

Figure 6.4 (a) presents an example of error distribution on the cluster $c_2 = \{\text{BMI}\}$. This error distribution has a relatively large standard deviation (SD), i.e., SD=4.471, which indicates that the MOE_{BMI} varies widely over time. This type of coding error is called **random error** and is mainly due to one or more uncontrollable underlying factors: (i) changes in the coding habits of the responsible coder over time, (ii) the update of the coding guideline, (iii) the adoption of a novel health information system, and (iv) the turnover of medical coders. We suggest the decision-maker develop management measures to improve coders' coding practices, i.e., (i) providing regular standardized training to medical coders to help them form proper and consistent coding habits and (ii) designing an effective work handover plan to ensure continuity of medical coding tasks and improve the efficiency of the reimbursement procedure.

Figure 6.4 (b) and (c) show error distributions on the second-largest cluster $c_{17} = \{\text{wgt\_evol\_in\_1mo}, \text{albumin}\}$. Among the two error distributions, MOE for feature "wgt_evol_in_1mo" is concentrated around the mean value and varies slightly, i.e., SD=0.038. This type of coding error (presented in Figure 6.4 (b)) is most likely due to certain stable and repetitive miscoding habits of the responsible coder (**systematic error**) and may be reduced with standardized procedures such as providing professional coding training to the coder. In Figure 6.4 (c), the $\text{MOE}_{albumin}$ is mainly concentrated in a small negative range, i.e., albumin $\in [-20, 0]$. This significant characteristic of miscoding behaviors should be taken into account when developing a standardized training program. Apart from that, error distributions for the cluster $c_5 2$ are also provided in Figure 6.5.

In both Figure 6.4 and Figure 6.5, Kernel density estimation (KDE) with Gaussian kernel is used to estimate the probability density function (PDF) of each given feature (variable). We use a rule of thumb, called the Scott Rule [109], to determine the band-with. The Scott Rule is defined by the following formula:

$$n^{(-1/(d+4))} \tag{6.1}$$

where the term n is the number of samples and d indicates the number of dimensions.

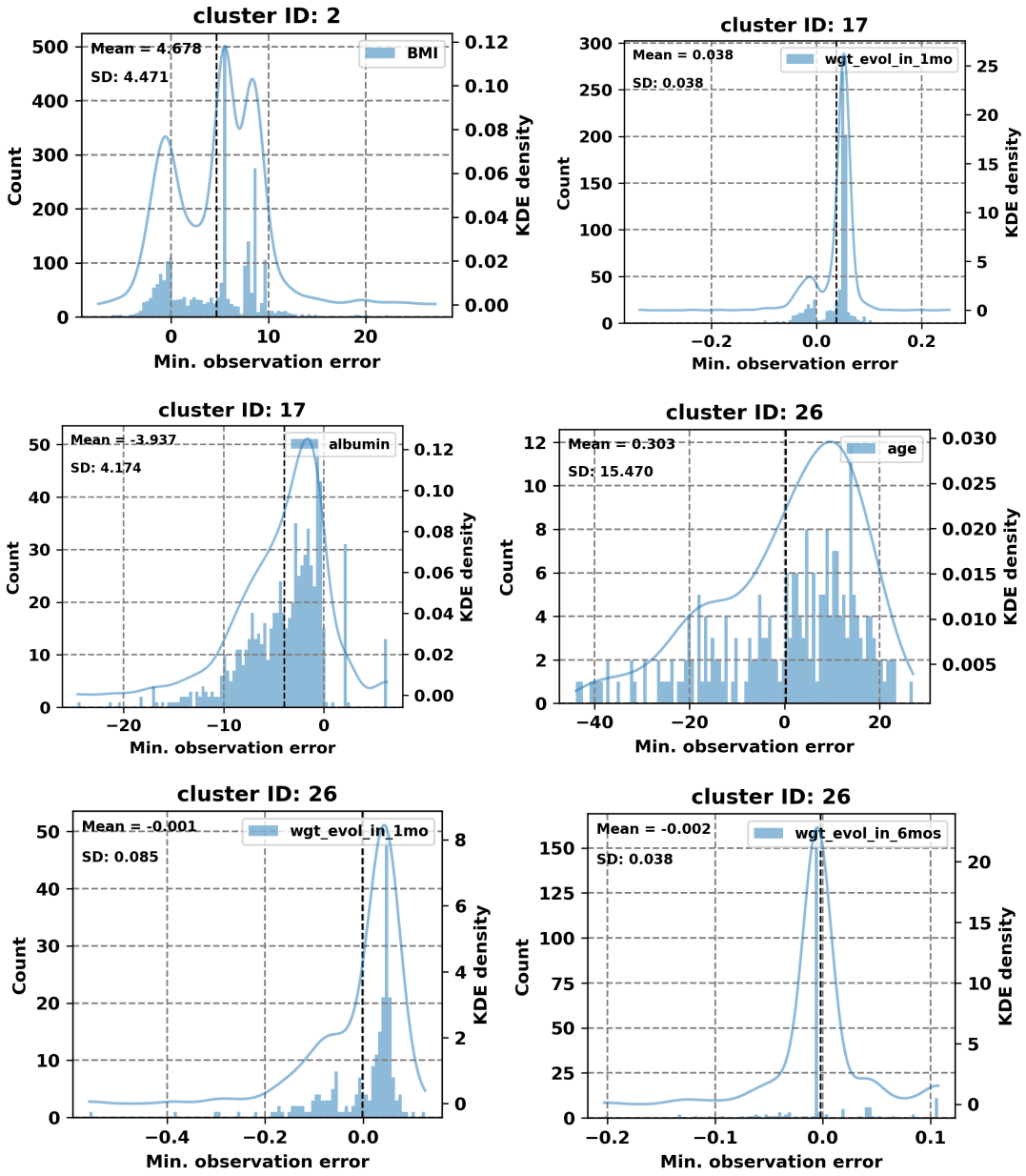


Figure 6.4.: Error distribution on BMI (a) for the for the largest cluster c_2, on wgt evol in 1mo (b) and albumin (c) for the second-largest cluster c_{17}. In addition, figure (d), (e), and (f) show error distributions on different features presented in the correcting set of the cluster c_{26}.

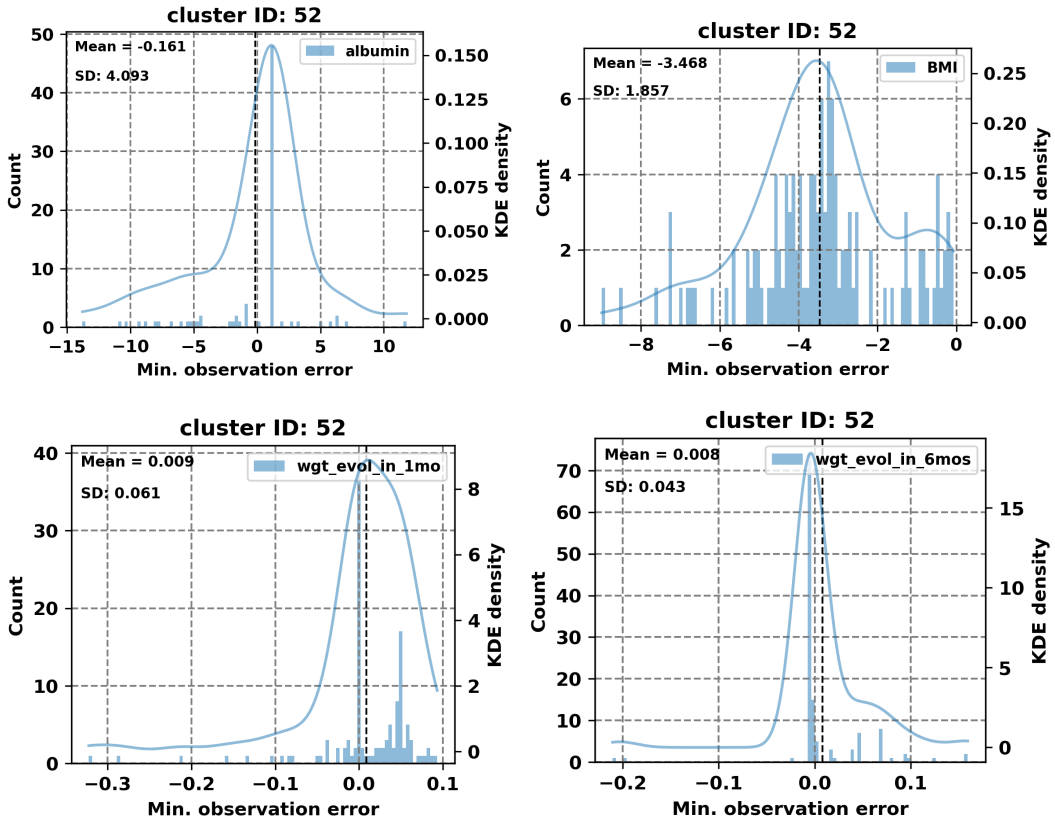


Figure 6.5.: Error distribution on albumin (a), on BMI (b), on wgt evol in 1mo (c) and on wgt evol in 6mos (d) for the cluster c_{52}.

6.2.6. Financial impacts on the hospital

Given that the primary concern of our partner hospital (CHU-SE) is to increase its annual financial revenue, it is crucial to study financial gains generated by correcting under-coded subjects and financial penalties caused by over-coding separately. For this purpose, independent studies are conducted on population I^+ and population I^-, respectively.

We first apply the proposed approach to the population I^+. For a given maximum size of correcting sets m=4, Figure 6.6 gives the Pareto fronts for different cluster sizes. The maximum financial gain generated by our CCVM model is 2,906,047.40 €, which accounts for about 41.56% of the total financial benefit (6,992,489.69 €).

After that, we work on the population I^-, and determine the optimal code correction solution for them. The resulting Pareto fronts for a given parameter m=4 is shown in Figure 6.6. The potential financial penalty avoided by applying the proposed approach is 4,086,442.29 €. This accounts for about 58.44% of the total financial benefit (6,992,489.69 €).

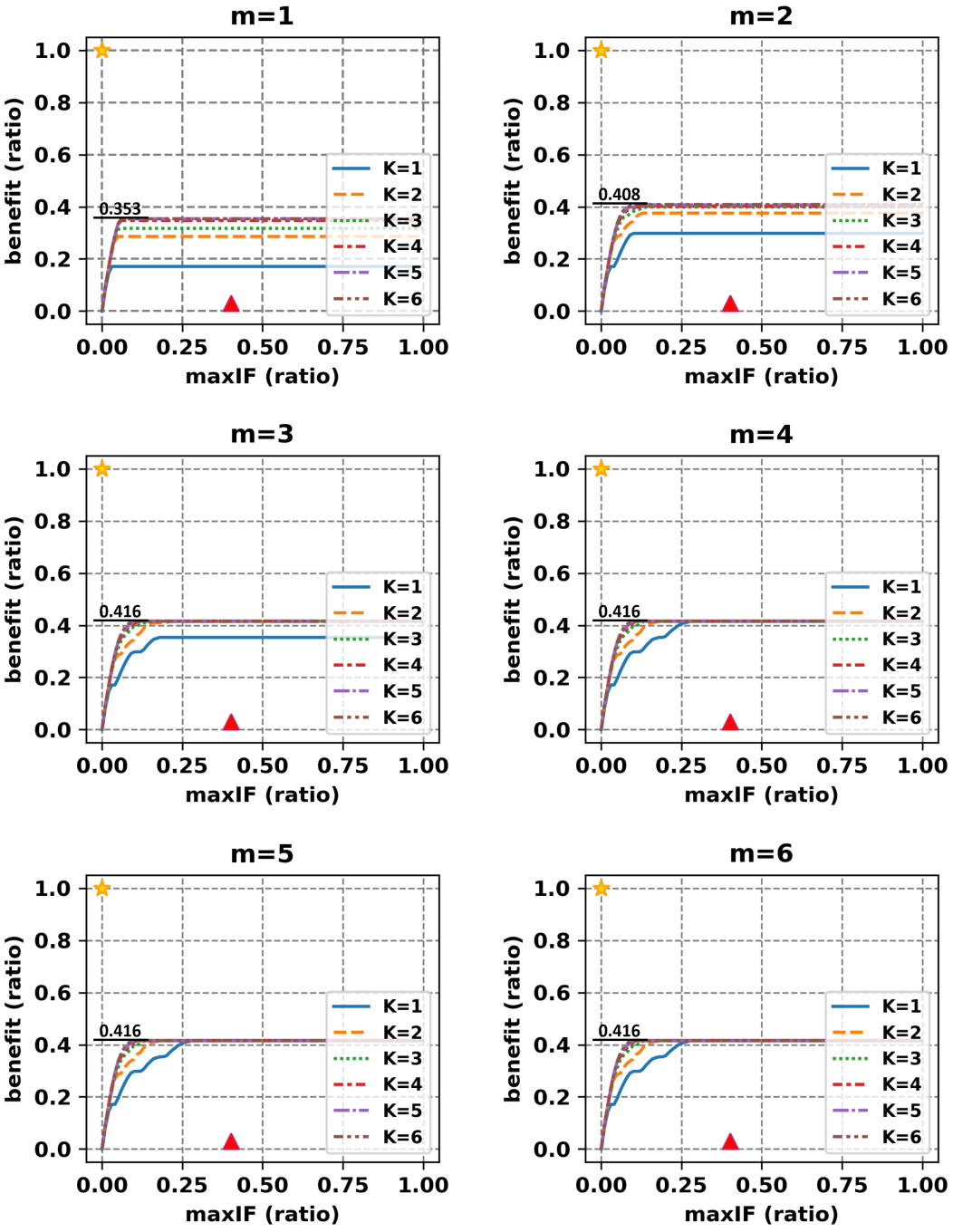


Figure 6.6.: Pareto fronts for the population I^+

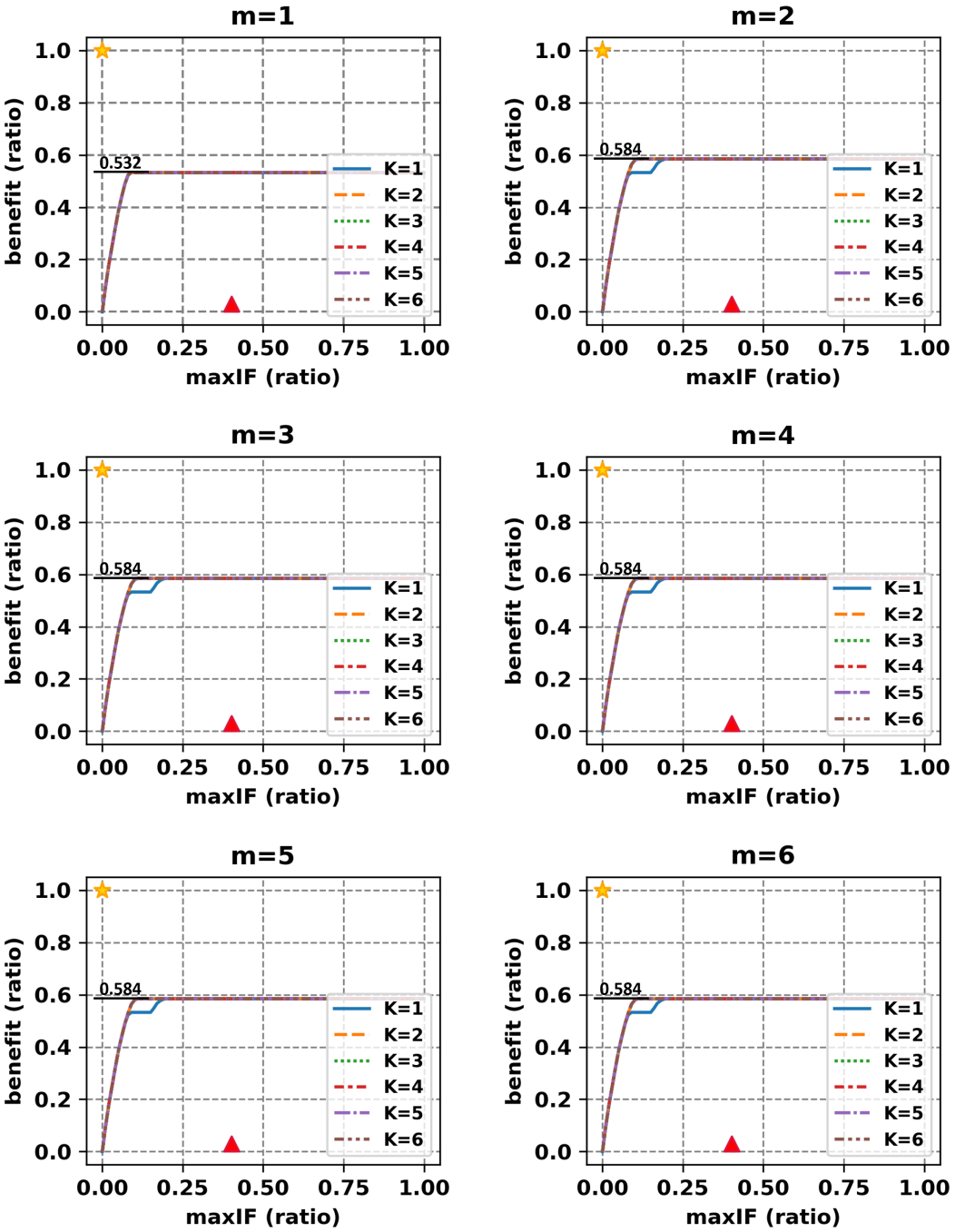


Figure 6.7.: Pareto fronts for the population I^{-}

6.3. Conclusion and perspectives

The approach proposed in chapter 5 is evaluated on a real-life dataset to allocate medical reviews for increasing health services reimbursement. The application of the proposed approach reduces unnecessary medical reviews up to 2933 (i.e., only 5293 descriptive features are reviewed) compared to the current practice in the DIM (i.e., reviewing 8856 descriptive features) while increasing the expected financial benefits by 6,786,815.33 €.

The proposed clustering-based optimization approach offers an effective and practical way to deal with the optimal code correction problem. From the experimental results, we highlight three critical contributions. First, the optimization program offers a solution to cut unnecessary medical reviews. Second, recommendation functions and correcting sets provide a strategy to automate the miscoding review task, where coding rules are expressed in text format. Third, miscoding behaviors (or subtypes) and relevant statistics are also addressed and provided to the decision-maker. It was concluded that the proposed approach plays a critical role in reducing the workload of medical coders, identifying causes of coding errors, and, hence, improving coding practice in routine medical codes assignment.

As a future research direction, the proposed approach can be applied to a larger set of medical codes by incorporating relevant features and medical codes. In this case, the maximum cardinality m should be considered and used to limit the size of the CCVM model, i.e., the number of decision variables. The time and space complexity should also be considered and evaluated for large CCVM models. Meta-heuristic algorithms should be considered to reduce the computation time.

Another research perspective lies in the definition of miscoding behavior. In addition to identifying mismatches between original codes y_i and re-assigned codes $\text{rec}(x_i)$, it would be interesting to understand the reason for the miscoding beyond the difference in feature values. We can not exclude that the medical coders had good reasons to deviate from the guidelines and miscode on purpose. A possible research direction would be to investigate how "spread" is the miscoding behaviors among the coders, the services that admitted the patients, etc. The more "shared" the miscoding behavior, the more it can reveal the limits of coding education as well as the coding guidelines.

6.4. Appendix

In this section, we present the code recommendation for malnutrition-related ICD codes. Overall, malnutrition diagnosis can be divided into two consecutive steps: (i) determining the presence of malnutrition and (ii) determining the severity. The following subsections introduce the code recommendation for adults and seniors, respectively.

6.4.1. Diagnosis of malnutrition in adults (18 <= age < 70))

To determine the presence of malnutrition in adults, the coding guideline [5] requires the subjects to satisfy at least one of the phenotypic criteria listed below:

- BMI < 18.5 kg/m^2;
- Weight loss ≥ 5% in 1 month or ≥ 10% in 6 months;

Next, in the process of determining the severity of malnutrition, a single criterion of severe malnutrition (E43) out-weights one or more criteria of moderate malnutrition (E44). One can assign the code severe malnutrition (E43) to an adult if at least one of the criteria listed below is satisfied:

- BMI ≤ 17 kg/m^2;
- Weight loss ≥ 10% in 1 month or ≥ 15% in 6 months;
- Albumin ≤ 30 g/L;

After that, adults need to meet at least one of the following criteria to confirm the diagnosis E44, i.e.,

- 17 < BMI < 18.5 kg/m^2;
- Weight loss ≥ 5% in 1 month or ≥ 10% in 6 months;
- 30 < albumin < 35 g/L;

6.4.2. Diagnosis of malnutrition in seniors (age >= 70)

Seniors need to meet at least one of the following criteria to confirm the presence of malnutrition:

- BMI < 22 kg/m^2;
- Weight loss ≥ 5% in 1 month or ≥ 10% in 6 months;
- Albumin* < 35 g/L;

Interpretation of the "albumin*" must consider the inflammatory state, assessed with the c-reactive protein. In other words, the criteria Albumin* < 35 g/L should be removed from the above list if the patient stay has an inflammation response (C-reactive protein ≥ 15)

Next, one can assign the code severe malnutrition (E43) to a senior if at least one of the criteria listed below is satisfied:

- BMI $< 20\ kg/m^2$;
- Weight loss $\geq$ 10% in 1 month or $\geq$ 15% in 6 months;
- Albumin $\leq 30\ g/L$;

In the event that none of the above conditions are satisfied, the criteria listed below can be used to confirm the diagnosis of moderate malnutrition (E44):

- $20 \leq$ BMI $< 22\ kg/m^2$;
- Weight loss $\geq$ 5% in 1 month or $\geq$ 10% in 6 months;
- Albumin $\geq 30\ g/L$;

Note that the "albumin" and "albumin*" are all measured by immunonephelometry or immunoturbidimetry.

General conclusion

Main contributions

In this thesis, we have explored a series of optimization-based approaches to automating the Quality Control (QC) process for hospital miscoding correction. In the literature, hospital miscoding audits are often seen as a manual review problem. Such a practice requires medical coders and physicians to review medico-administrative data, identify code inconsistencies and data omissions, and correct coding errors. This is far from an efficient practice since information is stored in different forms (i.e., structured tabular data, semi-structured chart data, unstructured textual data) and in different locations (i.e., different databases, different tables, and different columns) in hospital databases. In addition, the information in hospital databases differs in quality for each medical specialty and each patient stay. Previous QC practice depends on several coding specialists (each one comes from a particular medical specialty), who are expected to be professionally trained and be able to identify coding errors from patient histories efficiently.

Instead, we rely on collecting all available information from patient histories, detecting and correcting coding errors on the extracted information. Such a strategy allows us to have a global view of the hospital miscoding problem and optimize the miscoding correction budget. This requires a holistic optimization framework for the hospital miscoding problem, in which profiling of miscoding behaviors can have a large influence on the performance of the entire optimization framework. We have considered multiple perspectives for miscoding profiling and attempt to characterize and model miscoding behaviors appropriately. Our main contributions are in the holistic framework of miscoding profiling, and miscoding correction budget optimization, which also provides complementary information (i.e., miscoding explanation) for coding practice improvement.

For all proposed approaches, we select the integrated optimization approach presented in Chapter 5, for which we validate the results on a de-identified dataset that is directly extracted from a real-life database. We have no assumptions about how the dataset should be, which leads to the proposed approach being general.

Outlook

Research is a never-ending process that raises novel research questions every time a possible solution is provided to a problem.

In chapter 6, we conducted a retrospective validation of the integrated optimization approach on a real-life dataset. For the miscoding correction problem, a natural next step is the integration of the proposed approaches into the healthcare information system of the University Hospital of Saint Etienne (CHU-SE). On top of that, we can investigate the usability of the proposed approach, either as (i) clinical decision support (CDS: suggestion of top-N miscoding correction solutions in real-time) or (ii) automation (automated correction of coding errors). We can also evaluate whether the provided approach results in an improvement of the miscoding quality control process by (i) a reduction in average time for reviewing a patient stay, (ii) a reduced number of documents reviewed for a given population, (iii) a decrease in hospital miscoding rate, and (iv) an increase in hospital fiscal revenue. This can help the miscoding quality control process and validate the proposed approaches in a real-life environment.

Further research into the applicability of the proposed approaches to large populations is also worth being considered. The proposed approaches might not be implementable for large populations, where a large number of medical codes and patient features are included. To deal with this problem, novel meta-heuristics should be proposed to generate high-quality solutions and derive optimal solutions from them. The application of the proposed approaches to large populations requires linking medico-administrative data from multiple sources. Linking these data provides opportunities on the integration of heterogeneous data.

Besides, another research direction might be proving NLP-based approaches (i.e., Natural Language Processing-based approaches) to automate the hospital coding process since recent studies on Artificial Intelligence shows high performance on text classification tasks. These are only a part of the possibilities and research directions for new solution approaches that can tackle these challenges.

Bibliography

[1] G. R. Bramer, "International statistical classification of diseases and related health problems. tenth revision.", *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales*, vol. 41, no. 1, pp. 32–36, 1988 (cit. on p. 1).

[2] F. Sahraoui, D. Jolly, *et al.*, "La révolution de la classification commune des actes médicaux: ccam", *Dossier "la Réforme de l'assurance maladie". ADSP*, vol. 53, no. 54, pp. 54–5, 2006 (cit. on p. 1).

[3] J.-P. Domin, "Le programme de médicalisation des systèmes d'information (pmsi). de l'indicateur de comptabilité hospitalière au mode de tarification (1982-2012)", *Histoire, médecine et santé*, no. 4, pp. 69–87, 2013 (cit. on p. 1).

[4] S. Naran, A. Hudovsky, J. Antscherl, S. Howells, and S. Nouraei, "Audit of accuracy of clinical coding in oral surgery", *British Journal of Oral and Maxillofacial Surgery*, vol. 52, no. 8, pp. 735–739, 2014 (cit. on pp. 13, 20, 25).

[5] V. Delahaye-Guillocheau and C. Baude, *Le Guide méthodologique de production des informations relatives à l'activité médicale et à sa facturation en médecine, chirurgie, obstétrique et odontologie*. 2022, Available at: https://www.atih.sante.fr/guide-methodologique-mco-2021, Accessed on June, 2022 (cit. on pp. 18, 46, 76, 91, 95, 103, 112, 141, 153).

[6] L. C. des Codeurs Anonymes, *CoCoA 2021 Vademecum des TIM et des DIM pour le codage des maladies et des problèmes de santé connexes dans le cadre du PMSI-MCO et du PMSI-SSR*. 2022, Available at: http://docs.collectif-cocoa.org/index.php?title=CIM-10_CoCoA, Accessed on June, 2022 (cit. on pp. 18, 22, 132).

[7] S. Selvadurai *et al.*, "Cholangiocarcinoma miscoding in hepatobiliary centres", *European Journal of Surgical Oncology*, vol. 47, no. 3, pp. 635–639, 2021 (cit. on pp. 20, 25, 27).

[8] M.-T. Hsieh *et al.*, "Validation of icd-10-cm diagnosis codes for identification of patients with acute hemorrhagic stroke in a national health insurance claims database", *Clinical Epidemiology*, vol. 13, p. 43, 2021 (cit. on pp. 20, 27).

[9] S. Nouraei *et al.*, "A study of clinical coding accuracy in surgery: implications for the use of administrative big data for outcomes management", *Annals of surgery*, vol. 261, no. 6, pp. 1096–1107, 2015 (cit. on pp. 20, 24, 25).

[10] N. A. Heywood *et al.*, "Improving accuracy of clinical coding in surgery: collaboration is key", *Journal of Surgical Research*, vol. 204, no. 2, pp. 490–495, 2016 (cit. on p. 20).

[11] D. Lorence and L. Chen, "Disparities in health information quality across the rural–urban continuum: where is coded data more reliable?", *Journal of medical systems*, vol. 32, no. 1, pp. 1–8, 2008 (cit. on p. 23).

[12] C. Moje, T. J. Jackson, and P. McNair, "Adverse events in victorian admissions for elective surgery", *Australian Health Review*, vol. 30, no. 3, pp. 333–343, 2006 (cit. on pp. 23, 26).

[13] H. Pervez, A. Bhargwa, and M. J. Parker, "Accuracy and reliability of the clinical indicators related to hip fractures", *Injury*, vol. 34, no. 7, pp. 522–524, 2003 (cit. on p. 23).

[14] J. Horsky, E. A. Drucker, and H. Z. Ramelson, "Accuracy and completeness of clinical coding using icd-10 for ambulatory visits", in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2017, 2017, p. 912 (cit. on pp. 24, 25).

[15] B. Reid, C. Allen, and J. McIntosh, "Investigation of leukaemia and lymphoma ar-drgs at a sydney teaching hospital", *Health Information Management*, vol. 34, no. 2, pp. 34–39, 2005 (cit. on p. 25).

[16] J. Marshall and D. Adema, "Reinventing radiology reimbursement.", *Radiology Management*, vol. 27, no. 2, pp. 36–44, 2005 (cit. on p. 25).

[17] R. française, *Ordonnance n° 96-346 du 24 avril 1996 portant réforme de l'hospitalisation publique et privée*. 2022, Available at: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000742206/, Accessed on June, 2021 (cit. on p. 25).

[18] S. Santos, G. Murphy, K. Baxter, and K. M. Robinson, "Organisational factors affecting the quality of hospital clinical coding", *Health Information Management Journal*, vol. 37, no. 1, pp. 25–37, 2008 (cit. on p. 25).

[19] T.-H. Lu, M.-C. Lee, and M.-C. Chou, "Accuracy of cause-of-death coding in taiwan: types of miscoding and effects on mortality statistics", *International journal of epidemiology*, vol. 29, no. 2, pp. 336–343, 2000 (cit. on p. 25).

[20] A. Ballaro, S. Oliver, and M. Emberton, "Do we do what they say we do? coding errors in urology", *BJU international*, vol. 85, no. 4, pp. 389–391, 2000 (cit. on p. 25).

[21] S. Jameson and M. Reed, "Payment by results and coding practice in the national health service: the importance for orthopaedic surgeons", *The Journal of Bone and Joint Surgery. British volume*, vol. 89, no. 11, pp. 1427–1430, 2007 (cit. on p. 25).

[22] J. A. Schoenman, J. P. Sutton, A. Elixhauser, and D. Love, "Understanding and enhancing the value of hospital discharge data", *Medical Care Research and Review*, vol. 64, no. 4, pp. 449–468, 2007, PMID: 17684112. DOI: 10.1177/1077558707301963. eprint: https://doi.org/10.1177/1077558707301963. [Online]. Available: https://doi.org/10.1177/1077558707301963 (cit. on p. 26).

[23] E. A. S. Nelson *et al.*, "Surveillance of childhood diarrhoeal disease in hong kong, using standardized hospital discharge data", *Epidemiology and Infection*, vol. 132, no. 4, pp. 619–626, 2004, ISSN: 09502688, 14694409. [Online]. Available: http://www.jstor.org/stable/3865380 (visited on 09/23/2022) (cit. on p. 26).

[24] A. Ramalho, J. Souza, and A. Freitas, "The use of artificial intelligence for clinical coding automation: a bibliometric analysis", in *International Symposium on Distributed Computing and Artificial Intelligence*, Springer, 2020, pp. 274–283 (cit. on p. 26).

[25] R. Caruana *et al.*, "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission", in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721–1730 (cit. on pp. 26, 58, 75).

[26] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, and W. R. Hersh, "A systematic literature review of automated clinical coding and classification systems", *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 646–651, 2010 (cit. on pp. 26, 35).

[27] S. N. Payrovnaziri *et al.*, "Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review", *Journal of the American Medical Informatics Association*, vol. 27, no. 7, pp. 1173–1185, 2020 (cit. on p. 26).

[28] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text", *arXiv preprint arXiv:1802.05695*, 2018 (cit. on p. 26).

[29] F. Teng *et al.*, "A review on deep neural networks for icd coding", *IEEE Transactions on Knowledge and Data Engineering*, 2022 (cit. on p. 27).

[30] C. Rudin *et al.*, "Interpretable machine learning: fundamental principles and 10 grand challenges", *Statistics Surveys*, vol. 16, pp. 1–85, 2022 (cit. on p. 27).

[31] W. Kamal, S. Björnsdottir, O. Kämpe, and Y. T. Lagerros, "Concordance between icd-10 codes and clinical diagnosis of hypoparathyroidism in sweden", *Clinical Epidemiology*, vol. 12, p. 327, 2020 (cit. on p. 27).

[32] S. De Lusignan *et al.*, "A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data", *Diabetic Medicine*, vol. 27, no. 2, pp. 203–209, 2010 (cit. on p. 27).

[33] S. De Lusignan *et al.*, "Miscoding, misclassification and misdiagnosis of diabetes in primary care", *Diabetic medicine*, vol. 29, no. 2, pp. 181–189, 2012 (cit. on p. 27).

[34] C. He, B. Dalmas, C. Bousquet, B. T. Pavior, and X. Xie, "A topological and optimization based methodology to identify and correct icd miscoding behaviors", in *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2021, pp. 1382–1387 (cit. on pp. 33, 34, 112, 139).

[35] K. J. O'malley *et al.*, "Measuring diagnoses: icd code accuracy", *Health services research*, vol. 40, no. 5p2, pp. 1620–1639, 2005 (cit. on p. 35).

[36] L. Euler, "Solutio problematis ad geometriam situs pertinentis", *Commentarii academiae scientiarum Petropolitanae*, pp. 128–140, 1741 (cit. on p. 36).

[37] G. Carlsson, "Topology and data", *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009 (cit. on p. 36).

[38] G. Singh, F. Mémoli, and G. E. Carlsson, "Topological methods for the analysis of high dimensional data sets and 3d object recognition.", *SPBG*, vol. 91, p. 100, 2007 (cit. on pp. 36–38).

[39] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification", in *Proceedings 41st annual symposium on foundations of computer science*, IEEE, 2000, pp. 454–463 (cit. on p. 36).

[40] H. Abdi, "Metric multidimensional scaling (mds): analyzing distance matrices", *Encyclopedia of measurement and statistics*, pp. 1–13, 2007 (cit. on p. 36).

[41] H. Abdi and L. J. Williams, "Principal component analysis", *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010 (cit. on p. 36).

[42] E. Munch, "A user's guide to topological data analysis", *Journal of Learning Analytics,*, pp. 47–61, 2017 (cit. on p. 36).

[43] P. Y. Lum *et al.*, "Extracting insights from the shape of complex data using topology", *Scientific reports*, vol. 3, no. 1, pp. 1–8, 2013 (cit. on p. 36).

[44] M. Alagappan, *From 5 to 13: redefining the positions in basketball*, Lecture, 2012 (cit. on p. 37).

[45] J. Liu, J. Wang, Q. Zheng, W. Zhang, and L. Jiang, "Uncovering precision phenotype-biomarker associations in traumatic brain injury using topological data analysis", *Science of the Total Environment*, pp. 260–267, 2012 (cit. on p. 37).

[46] M. Rucco *et al.*, "Using topological data analysis for diagnosis pulmonary embolism", *arXiv preprint arXiv:1409.5020*, 2014 (cit. on p. 37).

[47] A. Savic, G. Toth, and L. Duponchel, "Topological data analysis (tda) applied to reveal pedogenetic principles of european topsoil system", *Science of the Total Environment*, pp. 1091–1100, 2017 (cit. on p. 37).

[48] V. Deshmukh, T. E. Berger, E. Bradley, and J. D. Meiss, "Leveraging the mathematics of shape for solar magnetic eruption prediction", *Journal of Space Weather and Space Climate*, vol. 10, p. 13, 2020 (cit. on p. 37).

[49] L. Wasserman, "Topological data analysis", *Annual Review of Statistics and Its Application*, vol. 5, pp. 501–532, 2018 (cit. on p. 37).

[50] R. Kraft, *Illustrations of data analysis using the mapper algorithm and persistent homology*, 2016 (cit. on p. 37).

[51] F. Chazal and B. Michel, "An introduction to topological data analysis: fundamental and practical aspects for data scientists", *Frontiers in artificial intelligence*, vol. 4, 2021 (cit. on pp. 37, 40).

[52] M. M. Nicolau, A. J. Levine, and G. E. Carlsson, "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.", *Proceedings of the National Academy of Sciences of the United States of America*, pp. 7265–70, 2011 (cit. on p. 37).

[53] L. Li *et al.*, "Identification of type 2 diabetes subgroups through topological analysis of patient similarity", *Science Translational Medicine*, 2015 (cit. on p. 37).

[54] B. Y. Torres *et al.*, "Tracking resilience to infections by mapping disease space", *PLoS biology*, 2016 (cit. on p. 37).

[55] J. L. Nielson *et al.*, "Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury", *Nature communications*, vol. 6, no. 1, pp. 1–12, 2015 (cit. on p. 38).

[56] J. L. Nielson *et al.*, "Uncovering precision phenotype-biomarker associations in traumatic brain injury using topological data analysis", *PLOS ONE*, pp. 1–19, 2017 (cit. on p. 38).

[57] D. Steinberg and P. Colla, "Cart: classification and regression trees", *The top ten algorithms in data mining*, vol. 9, p. 179, 2009 (cit. on pp. 38, 48, 50).

[58] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications", *biometrics*, vol. 21, pp. 768–769, 1965 (cit. on pp. 38, 60).

[59] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.", in *Kdd*, vol. 96, 1996, pp. 226–231 (cit. on pp. 38, 41).

[60] M. Berthold and B. Wiswedel, "Learning in parallel universes", *Data Mining and Knowledge Discovery*, vol. 21, Jul. 2010. DOI: 10.1007/s10618-010-0170-1 (cit. on p. 40).

[61] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan", *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017 (cit. on p. 41).

[62] O. Troyanskaya *et al.*, "Missing value estimation methods for dna microarrays", *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001 (cit. on pp. 45, 139).

[63] S. v. Buuren and K. Groothuis-Oudshoorn, "Mice: multivariate imputation by chained equations in r", *Journal of statistical software*, pp. 1–68, 2010 (cit. on p. 45).

[64] C. He, B. Dalmas, and X. Xie, "Acbi: an alternating clustering and bayesian inference approach for optimizing medical intervention budget under chance constraints", in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2020, pp. 55–60 (cit. on pp. 55, 56).

[65] Z. Landsman and M. Sherris, "Risk measures and insurance premium principles", *Insurance: Mathematics and Economics*, vol. 29, no. 1, pp. 103–115, 2001 (cit. on p. 57).

[66] A.-M. Kuhn and B. J. Youngberg, "The need for risk management to evolve to assure a culture of safety", *BMJ Quality & Safety*, vol. 11, no. 2, pp. 158–162, 2002 (cit. on p. 57).

[67] S. M. Grundy, R. Pasternak, P. Greenland, S. Smith, and V. Fuster, "Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the american heart association and the american college of cardiology", *Journal of the American College of Cardiology*, vol. 34, no. 4, pp. 1348–1359, 1999 (cit. on p. 57).

[68] A. Zehrouni, V. Augusto, T. Garaix, R. Phan, and X. Xie, "Health care emergency plan modeling and simulation in case of major flood", in *2017 Winter Simulation Conference (WSC)*, IEEE, 2017, pp. 2994–3005 (cit. on p. 57).

[69] G. M. Breakwell, *The psychology of risk*. Cambridge University Press, 2014 (cit. on p. 57).

[70] A. Artetxe, A. Beristain, and M. Grana, "Predictive models for hospital readmission risk: a systematic review of methods", *Computer methods and programs in biomedicine*, vol. 164, pp. 49–64, 2018 (cit. on p. 58).

[71] C. Van Walraven *et al.*, "Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community", *Cmaj*, vol. 182, no. 6, pp. 551–557, 2010 (cit. on p. 58).

[72] K. Zolfaghar *et al.*, "Big data solutions for predicting risk-of-readmission for congestive heart failure patients", in *2013 IEEE International Conference on Big Data*, IEEE, 2013, pp. 64–71 (cit. on p. 58).

[73] B. Strack *et al.*, "Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records", *BioMed research international*, vol. 2014, 2014 (cit. on p. 58).

[74] J. Billings, J. Dixon, T. Mijanovich, and D. Wennberg, "Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients", *Bmj*, vol. 333, no. 7563, p. 327, 2006 (cit. on p. 58).

[75] J. Billings *et al.*, "Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (parr-30)", *BMJ open*, vol. 2, no. 4, e001667, 2012 (cit. on p. 58).

[76] J. Jiang, S. Hewner, and V. Chandola, "Hospital readmission prediction-applying hierarchical sparsity norms for interpretable models", *arXiv preprint arXiv:1804.01188*, 2018 (cit. on p. 58).

[77] N. E. Breslow, "Analysis of survival data under the proportional hazards model", *International Statistical Review/Revue Internationale de Statistique*, pp. 45–57, 1975 (cit. on p. 58).

[78] A. Hosseinzadeh, M. Izadi, A. Verma, D. Precup, and D. Buckeridge, "Assessing the predictability of hospital readmission using machine learning", in *Twenty-fifth IAAI conference*, 2013 (cit. on p. 58).

[79] X. Min, B. Yu, and F. Wang, "Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on copd", *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019 (cit. on p. 59).

[80] J. Futoma, J. Morris, and J. Lucas, "A comparison of models for predicting early hospital readmissions", *Journal of biomedical informatics*, vol. 56, pp. 229–238, 2015 (cit. on pp. 59, 74).

[81] S. Cui, D. Wang, Y. Wang, P.-W. Yu, and Y. Jin, "An improved support vector machine-based diabetic readmission prediction", *Computer Methods and Programs in Biomedicine*, vol. 166, pp. 123–135, 2018, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2018.10.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260718308083 (cit. on p. 59).

[82] A. Z. A. Hazzouri *et al.*, "Cardiovascular risk score, cognitive decline, and dementia in older m exican a mericans: the role of sex and education", *Journal of the American Heart Association*, vol. 2, no. 2, e004978, 2013 (cit. on p. 59).

[83] M. Kivipelto *et al.*, "Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study", *The Lancet Neurology*, vol. 5, no. 9, pp. 735–741, 2006 (cit. on p. 59).

[84] C. Reitz *et al.*, "A summary risk score for the prediction of alzheimer disease in elderly persons", *Archives of neurology*, vol. 67, no. 7, pp. 835–841, 2010 (cit. on p. 59).

[85] L. G. Exalto *et al.*, "Midlife risk score for the prediction of dementia four decades later", *Alzheimer's & Dementia*, vol. 10, no. 5, pp. 562–570, 2014 (cit. on p. 59).

[86] S. J. Andrews *et al.*, "Validating the role of the australian national university alzheimer's disease risk index (anu-adri) and a genetic risk score in progression to cognitive impairment in a population-based cohort of older adults followed for 12 years", *Alzheimer's research & therapy*, vol. 9, no. 1, pp. 1–12, 2017 (cit. on p. 59).

[87] E. Y. Tang *et al.*, "Current developments in dementia risk prediction modelling: an updated systematic review", *PloS one*, vol. 10, no. 9, e0136181, 2015 (cit. on p. 59).

[88] G. C. Siontis, I. Tzoulaki, K. C. Siontis, and J. P. Ioannidis, "Comparisons of established risk prediction models for cardiovascular disease: systematic review", *Bmj*, vol. 344, 2012 (cit. on p. 59).

[89] S. L. Harrison *et al.*, "Cardiovascular disease risk models and longitudinal changes in cognition: a systematic review", *PloS one*, vol. 9, no. 12, e114431, 2014 (cit. on p. 59).

[90] C. C. Petersen, "Simulation of alternative designs of a cardiovascular risk reduction program for the u.s. air force", *Journal of Medical Systems*, vol. 6, pp. 149–164, 1982 (cit. on p. 59).

[91] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: a generalized convergence theorem and characterization of local optimality", *IEEE Transactions on pattern analysis and machine intelligence*, no. 1, pp. 81–87, 1984 (cit. on p. 60).

[92] S. Vassilvitskii and D. Arthur, "K-means++: the advantages of careful seeding", in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2006, pp. 1027–1035 (cit. on p. 61).

[93] L. Bobrowski and J. C. Bezdek, "C-means clustering with the l/sub l/and l/sub infinity/norms", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 545–554, 1991 (cit. on p. 61).

[94] F. Cao, J. Liang, and L. Bai, "A new initialization method for categorical data clustering", *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 223–10 228, 2009 (cit. on p. 62).

[95] Z. Huang, "Clustering large data sets with mixed numeric and categorical values", in *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, Citeseer, 1997, pp. 21–34 (cit. on p. 63).

[96] W. Dai *et al.*, "Prediction of hospitalization due to heart diseases by supervised learning methods", *International journal of medical informatics*, vol. 84, no. 3, pp. 189–197, 2015 (cit. on p. 63).

[97] J. J. Chen and M. R. Novick, "Bayesian analysis for binomial models with generalized beta prior distributions", *Journal of Educational Statistics*, vol. 9, no. 2, pp. 163–175, 1984 (cit. on p. 65).

[98] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: a frequent-pattern tree approach", *Data mining and knowledge discovery*, vol. 8, no. 1, pp. 53–87, 2004 (cit. on pp. 67, 82, 121).

[99] A. Szlam, Y. Kluger, and M. Tygert, "An implementation of a randomized algorithm for principal component analysis", *arXiv preprint arXiv:1412.3510*, 2014 (cit. on p. 69).

[100] K. J. Verhaegh *et al.*, "Transitional care interventions prevent hospital readmissions for adults with chronic illnesses", *Health affairs*, vol. 33, no. 9, pp. 1531–1539, 2014 (cit. on pp. 74, 75).

[101] H. M. Krumholz *et al.*, "Hospital-readmission risk—isolating hospital effects from patient effects", *New England Journal of Medicine*, vol. 377, no. 11, pp. 1055–1064, 2017 (cit. on p. 74).

[102] A. Bottle, P. Aylin, and A. Majeed, "Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis", *Journal of the Royal Society of Medicine*, vol. 99, no. 8, pp. 406–414, 2006 (cit. on p. 74).

[103] S. Yu *et al.*, "Predicting readmission risk with institution-specific prediction models", *Artificial intelligence in medicine*, vol. 65, no. 2, pp. 89–96, 2015 (cit. on p. 74).

[104] C. Hebert *et al.*, "Diagnosis-specific readmission risk prediction using electronic health data: a retrospective cohort study", *BMC medical informatics and decision making*, vol. 14, no. 1, pp. 1–9, 2014 (cit. on p. 74).

[105] O. Tonkikh *et al.*, "Functional status before and during acute hospitalization and readmission risk identification", *Journal of hospital medicine*, vol. 11, no. 9, pp. 636–641, 2016 (cit. on p. 74).

[106] H. Abdi and D. Valentin, "Multiple correspondence analysis", *Encyclopedia of measurement and statistics*, vol. 2, no. 4, pp. 651–657, 2007 (cit. on p. 80).

[107] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987 (cit. on p. 94).

[108] S. Lee, J. H. Xin, and S. Westland, "Evaluation of image similarity by histogram intersection", *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, vol. 30, no. 4, pp. 265–274, 2005 (cit. on p. 123).

[109] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015 (cit. on p. 147).

List of publications

Journal articles

1. He, Chen, Benjamin Dalmas, Cedric Bousquet, Beatrice Trombert Pavior, and Xiaolan Xie. "A clustering-based optimization approach for hospital miscoding correction." In IEEE Transactions on Automation Science and Engineering (SCI indexed journal). **Conditionally accepted**.

Conference papers

1. He, Chen, Benjamin Dalmas, Cedric Bousquet, Beatrice Trombert Pavior, and Xiaolan Xie. "A topological and optimization based methodology to identify and correct ICD miscoding behaviors." In proceedings of 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), pp. 1382-1387. IEEE, 2021, (EI indexed). DOI: 10.1109/CASE49439.2021.9551661. **Oral presentation**.

2. He, Chen, Benjamin Dalmas, and Xiaolan Xie. "ACBI: An Alternating Clustering and Bayesian Inference approach for optimizing medical intervention budget under chance constraints." In proceedings of 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), pp. 55-60. IEEE, 2020, (EI indexed). DOI: 10.1109/CASE48305.2020.9216791. **Oral presentation**.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT: 2023EMSEM002

Author: Chen HE

Title: Analyzing, optimizing, and explaining hospital miscoding for coding practice improvement

Speciality: Industrial Engineering

Keywords: PMSI program, hospital miscoding, miscoding review budget optimization, machine learning, data mining.

Abstract

The primary objective of this thesis is to develop a series of methodologies based on optimization, machine learning, and data mining techniques, which are then adopted in the context of PMSI (i.e., Information Systems Medicalization Program) for the hospital miscoding problem. A series of data-driven optimization approaches are developed to identify hospital miscoding behaviors, analyze them, model them, and correct them in order to improve the operational efficiency of the University Hospital of Saint Etienne (CHU-SE). This thesis puts a particular emphasis on model transparency and interpretability since they are critical elements to guarantee the fairness of developed applications. An evaluation of the proposed approaches and an analysis of the benefits of the optimization approaches are performed. Experimental results show that the proposed approaches are able to decrease the hospital miscoding rate without increasing the workload of medical coding staff in the hospital. The results are promising and reveal that potential benefits for all private clinics and public hospitals in France are achievable.

MINES Saint-Étienne

Une école de l'IMT

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2023EMSEM002

Auteur : Chen HE

Titre : Analyser, optimiser et expliquer le mauvais codage hospitalier pour l'amélioration des pratiques de codage

Spécialité : Génie industriel

Mots-Clefs : Programme PMSI, erreur de codage hospitalier, optimisation du budget de révision de l'erreur de codage, apprentissage automatique, exploration de données.

Résumé

L'objectif principal de cette thèse est de développer une série de méthodologies basées sur des techniques d'optimisation, d'apprentissage automatique et de fouille de données. Ces méthodes sont ensuite appliquées dans le contexte du PMSI (i.e., Programme de Médicalisation des Systèmes d'Information) pour répondre au problème de mauvais codage hospitalier. Une série d'approches d'optimisation guidées par les données est développée pour identifier les comportements de mauvais codage hospitalier, les analyser, les modéliser et les corriger afin d'améliorer l'efficacité opérationnelle du CHU de Saint Etienne (CHU-SE). Cette thèse met un accent particulier sur la transparence et l'interprétabilité des modèles car ce sont des éléments critiques pour garantir l'équité des applications développées. Une évaluation des approches proposées et une analyse des avantages des approches d'optimisation sont réalisées. Les résultats expérimentaux montrent que les approches proposées sont capables de diminuer le taux de mauvais codage dans les hôpitaux sans nécessairement augmenter la charge de travail du personnel de codage médical dans l'hôpital. Les résultats sont prometteurs et révèlent que des bénéfices potentiels pour toutes les cliniques privées et les hôpitaux publics en France sont réalisables.

ÉCOLE DES MINES DE SAINT-ÉTIENNE
158, cours Fauriel – CS 62362
42023 Saint-Étienne Cedex 2 – France
Tél. +33(0)4 77 42 01 23
www.mines-stetienne.fr

Institut Mines-Télécom

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE L'ÉCONOMIE DE L'INDUSTRIE ET DU NUMÉRIQUE

INSPIRING INNOVATION SINCE 1816