

Open Research Online

The Open University's repository of research publications and other research outputs

Health Condition Evolution for Effective Use of Electronic Records: Knowledge Representation, Acquisition, and Reasoning

Thesis

How to cite:

Morales Tirado, Alba Catalina (2023). Health Condition Evolution for Effective Use of Electronic Records: Knowledge Representation, Acquisition, and Reasoning. PhD thesis The Open University.

For guidance on citations see [FAQs](#).

© 2022 Alba Catalina Morales Tirado



<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Version: Version of Record

Link(s) to article on publisher's website:

<http://dx.doi.org/doi:10.21954/ou.ro.00015402>

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

oro.open.ac.uk



HEALTH CONDITION EVOLUTION FOR EFFECTIVE USE OF ELECTRONIC RECORDS: KNOWLEDGE REPRESENTATION, ACQUISITION, AND REASONING

ALBA CATALINA MORALES TIRADO

KNOWLEDGE MEDIA INSTITUTE

THE OPEN UNIVERSITY

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2022

Abstract

Smart City initiatives aim to enhance the effective management of resources while providing quality services to citizens. Central to these initiatives is the use of large-scale datasets that enable intelligent analytics and reasoning components in support of resource optimisation and service provision. Recently, there has been a growing interest in aspects of smart living, particularly due to the increasing adoption and use of Electronic Health Records (EHR).

A Smart City can introduce intelligent systems to support the usage of EHR to improve emergency response services. For instance, data derived from EHR is used in primary emergency care, as a component of emergency decision support systems and for monitoring public health. However, the delivery of healthcare information to emergency bodies must be balanced against the concerns related to citizens' privacy. Besides, emergency services face challenges in interpreting this data; the heterogeneity of sources and the large amount of available information represents a significant barrier.

This thesis investigates the use of EHR for deriving useful information about people requiring assistance during an emergency, focusing on making rich data accessible to emergency services while minimising the amount of exchanged information. To perform this task, an intelligent system needs to estimate the probability that a potentially relevant condition mentioned in a health record is still valid at the time of the emergency. During our research work, we followed a knowledge engineering approach and developed the required knowledge components to support the intelligent delivery of relevant health information about people involved in an emergency situation. These components, which include a knowledge component for representation and reasoning, and a novel knowledge base modelling the evolution of a large number of health conditions, form the basis of CONRAD, a system which is able to support effectively decision-making in an emergency scenario.

**To my beloved parents, Gladys and Hugo,
my brothers and my handsome nephew.**

'It is only with the heart that one can see rightly; what is essential is invisible to the eye.'

The Little Prince (1943)

Antoine de Saint-Exupéry

Acknowledgements

Five years ago, I decided to start a new journey, and if you knew me then, you could tell that I always dreamed of doing a PhD. Besides the academic challenge, part of me always liked to be out of my 'comfort zone'; therefore, it was no surprise when I finally decided to take the first steps.

The start of the journey was not easy; I applied to different places and was rejected immediately. Sometimes, I got a reply, and after a few emails, I never heard back from people. A couple of times, I was invited to interviews and then rejected. After a year, it seemed like the PhD was not for me, but one day out of nothing, I got this email from the OU, followed by an interview!

A few months later, I was flying to the UK to start a PhD. Only when I landed I realised: maybe I went too far with being 'out of my comfort zone'; perhaps being offered a place was not the most challenging part of the journey after all ...

First and foremost, my sincere gratitude goes to my two exceptional supervisors, Enrico Daga and Enrico Motta. I am immensely grateful to Enrico Daga for the time he dedicated to our meetings sharing his precious knowledge and expertise. Words fall short of expressing my gratitude for his guidance, mentorship, support and all the hours he patiently spent reviewing my manuscripts. Enrico Daga's contagious work ethic encouraged me to continue working even when things went wrong or simply when life was hard. I am thankful to Enrico Motta for always trusting our work; his wise point of view on things made our monthly meetings enriching discussions.

Many thanks to my examiners, John Domingue and Dave Robertson, for their enriching comments on my thesis and making the viva an enjoyable, unforgettable discussion. I am grateful to Anna DeLiddo, for chairing the session; she was on top of every aspect of the process while being a reassuring presence on the day.

I also want to highlight the importance of being part of the Knowledge Media Institute - KMi. During the second year of my PhD, we faced the most unlikely situation: a pandemic. Days, months, and years passed, and amid uncertainty, colleagues kept doing their best, published valuable articles, and won well-deserved awards while looking after each other; being part of KMi made life brighter during this period. Of course, a huge thanks to the KMi Admin team for the help accommodating students going back to campus in the middle of the pandemic (your work keeps the KMi machine running seamlessly).

Without the support of amazing colleagues and friends, this journey would not have been the same. A special thank you to Ashling Third and Angelo Salatino, who read the entire

thesis and provided valuable feedback in preparation for the final day. To my colleagues who were there, and I mean physically present at the lab (*because this is the exception nowadays!*), particularly during the writing months: Angelo! and Julian, I cannot thank you enough for all your advice, support and for supplying much-needed brain food (desserts). To the brilliant, inspiring, hard-working doctors: Agnese Chiatti, Venetia Brown and Retno Larasati. I am immensely grateful we had the chance to share this incredible journey and also close this chapter together. It is an honour to have you as my friends and it has been extremely essential to have such strong women to look up to.

A special thought goes to all my friends back home in Ecuador and around the world. You guys kept cheering, offering a word, a smile or simply a hug; you were there through the worst of times and the best: Jennifer, Majo, Magy, Isa and Diego! I feel blessed to have you in my life; you always reminded me that there is more in life than work, and it is pretty great!

Finally and most importantly, this work is dedicated to my beloved family: thank you for your love every step of the way, for getting used to the distance, the time difference and video calls; words are not enough to express how much *los amo*. Mom and dad, thanks for encouraging me to pursue my dreams. I learned from you what hard work can do, what a smile can reach and how a hug can cure almost everything. Thanks to my brothers Josue and Diego, my sister-in-law Alejandra, and my gorgeous nephew Andres for your support; your love kept me going when life was hard. To grandpa: you were at the starting line with me, and I know you are with me at the finish line, somewhere close. Dear family, I hope to honour you with my work every day; you are my blessing!

Yes, in the beginning, I thought this journey was too much for me; now, I can safely say the PhD journey has not been easy. Yet time proved that it was not impossible either! That hard work, consistency, and passion make things a bit easier. All the ups and downs during the PhD journey helped me grow and develop my skills.

Time also confirmed that I landed in the right place. I cannot imagine pursuing a PhD anywhere other than here in KMi. The challenge, the place, the work and the people in my life made the trip enjoyable and a true blessing.

*'No pienses que no pasa nada porque no ves tu crecimiento,
las grandes cosas crecen en silencio.'*

BUDA

Contents

1	Introduction	1
1.1	Motivation	3
1.1.1	Use of electronic health records during fire emergencies	5
1.1.2	Access to health records by ambulance services	7
1.1.3	Exchange of EHR in the Smart City setting	8
1.2	Hypotheses and Research Questions	11
1.3	Research Methodology	16
1.4	Approach	17
1.5	Thesis outline and contributions	19
1.5.1	Contributions	19
1.5.2	Overview of the thesis	20
1.5.3	Publications	20
2	State of the art	23
2.1	Use of healthcare information in the Smart City context	23
2.1.1	Use of Health records in emergency support	25
2.1.2	Intelligent systems to assist Emergency services	28
2.2	Knowledge rep. and reasn. with healthcare data	31
2.2.1	Reasoning on health recovery	32
2.2.2	Ontologies and Semantic Web in healthcare	34
2.3	KG construction from unstructured information	35
2.3.1	Information extraction from natural language sources and supervised text classification	36
2.3.2	Healthcare applications	38
2.4	Semantic Web Technologies	39
2.4.1	RDF/OWL	39
2.4.2	Representation of health records	43

3	Requirements' analysis	55
3.1	Expanded Motivating scenario	55
3.2	Scenario-based analysis - real setting	57
3.3	Requirements elicitation	60
4	Knowledge Representation	65
4.1	Methodology	65
4.2	Abstracting the scenario	67
4.3	Knowledge reasoning – principles	68
4.3.1	Thematic Analysis	69
4.3.2	Competency Questions	72
4.4	Building HECON - Health Condition Evolution Ontology	74
4.4.1	HECON core concepts	75
4.4.2	Provenance representation	76
4.5	Ontology evaluation	79
4.6	Discussion	85
5	Knowledge acquisition and KG construction	87
5.1	Approach overview	88
5.2	Corpus preparation	90
5.3	Knowledge components identification	91
5.3.1	Building a gold standard dataset of HES	92
5.3.2	Training and testing Machine Learning algorithms for the classifica- tion task	93
5.3.3	Application of the machine learning approach.	95
5.3.4	Consistency check	96
5.4	Knowledge completion	98
5.5	Human-in-the-loop	101
5.6	Knowledge Graph	102
5.6.1	Data statistics	103
5.7	Conclusions	104
6	Reasoning with Health Condition Evolution	107
6.1	Context: the Smart City scenario	108
6.2	Dataset of health records	110
6.2.1	FHRI representation and annotation	111

6.3	Reasoning on health records validity	112
6.3.1	Severity score	116
6.3.2	Supporting interpretation	117
6.4	Intelligent System architecture	119
6.5	CONRAD - Health Condition Radar	119
6.6	Conclusions	121
7	Evaluation	123
7.1	User study - Towards a KG of health condition evolution	124
7.1.1	User study methodology	125
7.1.2	Precision analysis	130
7.1.3	Evaluating the knowledge completion task	133
7.1.4	Sustainability	136
7.1.5	Discussion	137
7.2	Intelligent System Evaluation	138
7.2.1	Gold standard dataset	138
7.2.2	CONRAD overview	139
7.2.3	Experiments design	141
7.2.4	Results and discussion	141
7.3	Chapter conclusions	144
8	Discussion and Conclusions	147
8.1	Discussion	148
8.2	Limitations of the study and Future work	153
8.3	Conclusions	156
A	Appendix - Glossary	173
B	Appendix - Summary software	179
C	Appendix - Emergency response	181
C.1	Emergency management	181
D	Appendix - Requirements' analysis	183
D.1	Fire related documentation	183
D.2	Personal Emergency Evacuation Plan - PEEP	186
D.3	Poster	187

E	Appendix - Intelligent System Design	189
E.1	FHIR specification	189
F	Appendix - Building HECON Ontology - resources	193
F.1	Text snippets	193

List of Figures

2.1	RDF IRI typical components	40
2.2	RDF Literal typical components	41
2.3	RDF triple	41
2.4	RDF triple example	41
2.5	Using FHIR resources to represent an Electronic Health Records	47
2.6	SNOMED CT structure and top-level concepts	49
2.7	SNOMED CT components explained	50
2.8	SNOMED CT Expression Constraint Language example	51
3.1	Summary of Health and Safety data flow for fire planning	59
4.1	Health Condition Evolution Statement (HES) representation	72
4.2	HECON Ontology core concepts.	76
4.3	HECON Ontology provenance representation.	77
4.4	Query outcome for CQs5-7.	83
5.1	Knowledge Acquisition pipeline	89
5.2	Binary classification training	94
5.3	ML Training process	95
5.4	Prediction process	96
5.5	HES KG representation	103
6.1	CONRAD - data flow and implementation setting	110
6.2	Synthetic EHR representation using FHIR resources	111
6.3	Reasoning on HES: Direction interpretation.	113
6.4	Reasoning on HES: Time range interpretation.	114
6.5	Reasoning on HES: Severity score.	116
6.6	CONRAD system architecture.	120

7.1	User study: interface task one.	127
7.2	User study: interface task two.	128
7.3	Precision @ k	131
7.4	Comparison total number of annotations and correct answers	135
7.5	CONRAD's output - People in need of assistance and type of vulnerability .	140
C.1	Emergency management cycle	182
D.1	Personal Emergency Evacuation Plan - PEEP form	186
D.2	Poster: Towards privacy-aware intelligent systems for emergency response .	187
E.1	FHIR mapping. Part 1/4	190
E.2	FHIR mapping. Part 2/4	191
E.3	FHIR mapping. Part 3/4	192
E.4	FHIR mapping. Part 4/4	192

List of Tables

2.1	List of data generated by Synthea	54
4.1	Examples - text describing health evolution.	70
4.2	Examples of text snippets describing health evolution	70
4.3	Summary of expressions used abstract health condition evolution features .	71
4.4	Examples - sentences and Health Evolution Statement representation. . . .	73
4.5	Competency Questions - requirements on health evolution information . . .	79
4.6	Results execution SPARQL	81
4.7	Competency Questions - time range information	81
4.8	Competency Questions - provenance information	83
4.9	Results - provenance information	85
5.1	Summary of data collected from web sources	91
5.2	Finding matching SNOMED CT concepts using Levenshtein distance	92
5.3	Number of sentences per HES in the training dataset	93
5.4	ML training results: Accuracy per algorithm & HES features	95
5.5	Best performing ML algorithms hyper-parameters configuration	95
5.6	HES best confidence value	98
5.7	Propagation rules details.	100
5.8	Core. A number of instances per class.	104
5.9	Provenance. Number of instances per class.	104
6.1	Reasoning on EHR - examples	115
6.2	Types of disabilities and correspondent Key Concept	117
6.3	Ranked top 3 reasons for assistance	118
7.1	Participants by level of expertise	128
7.2	First task: number of annotated SNOMED concepts per participant	131
7.3	First task: Precision@k per participant	132

7.4	First task: Total annotations by familiarity	133
7.5	Second task: Number of annotated SNOMED concepts per participant . . .	134
7.6	Second task: Total annotations grouped by rule	135
7.7	Second task: Total Correct annotations per participant	136
7.8	Experiments results	142
7.9	Examples severity score	143
7.10	Precision @ 3 categories of disability	144
A.1	Glossary	173
B.1	Contributions and repository description	179
F.1	List of text snippets used to build a subset of sentences.	193

Listings

2.1	RDF graph in N-Triples syntax	42
2.2	RDF graph in N-Triples, Turtle syntax	42
2.3	OWL syntax example	44
4.1	HES: type, pace and time range. E.g., Fracture of ankle	80
4.2	Competency questions one to four SPARQL.	80
4.3	Fracture of ankle - time range information.	81
4.4	Fracture of ankle health condition evolution information time range.	82
4.5	Fracture of ankle health HES provenance information.	84
4.6	Query provenance information.	84
5.1	KG data - triple example	103

Chapter 1

Introduction

Smart Cities have emerged as a socio-technical environment in which powerful platforms are designed, built and maintained to improve public services and create better urban spaces for citizens [Pellicer et al. 2013; Cassandras 2016]. As part of the Smart City vision, Intelligent Systems (IS) collect and use data from a wide variety of sources in order to optimise services such as healthcare, transportation, energy distribution, emergency response, and commerce, among others. Smart solutions' prime objective is to ensure sustainable use of resources (economically, environmentally, and socially), facilitating the management of services in the cities.

In this context, healthcare information is a significant and valuable resource; its use has received increased attention in recent years [Pramanik et al. 2017; Zahid et al. 2021]. However, using healthcare data has its challenges. First, a person's health record contains a large amount of information, so finding a way to detect relevant data is essential. For example, information extracted from health records could reveal recent medical events or chronic health conditions indicating a particular need for support during an emergency. Second, as health records contain very detailed information, it could be time-consuming and hard to interpret for the untrained eye, risking that vital information could be overlooked. Third, sensitive data should be processed lawfully. Any data processing, including its collection, storage and exchange, should be treated fairly and responsibly, according to established data regulations, lifting any data privacy concerns. If this data is misused or disclosed accidentally, it could affect city services and crucially cause damage to citizens' privacy. Moreover, adequate handling of personal data is a legal requirement for organisations that collect, hold or process data. Regulations such as the General Data Protection Regulation (GDPR) establish administrative fines of *'up to 4% of annual global turnover or €20 million -whichever is greatest'* [European Parliament 2016] as a result of infringements of the regulation. As more cities

are adopting the Smart City vision, it is imperative to address privacy issues and protect against any misuse of healthcare data. Ultimately, providing condensed and readily available information on up-to-date health data constitutes a fundamental step to better use of health records.

An area of application in Smart Cities pertains to improving emergency services response [Rego et al. 2018; Moreira, Sinderen, and Pires 2019]. Typically, planning and dealing with emergencies such as fire, flooding, and accidents, among others, is challenging for emergency responders, especially when their activities are not coordinated or supported with accurate and most recent information about the event. Emergency Managers require data that could help them act promptly and make use of appropriate resources [Phillips, Neal, and G. Webb 2016].

In this thesis, we investigate the use of health records data with the purpose of supporting first responders' activities during an emergency. The main aim is to address the problem of extracting timely and valuable information from Electronic Health Records (EHR), which could help decision-makers draw a picture of people's recent medical events, potentially indicating that a person suffers some issue and hence requires assistance. In particular, we research methods and approaches for extracting and representing relevant information from health records at a given point in time and support the detection of vulnerable people during an emergency. This research investigates how to process and represent the extracted data to facilitate health records interpretation while minimising the amount of information delivered; this is a fundamental requirement to tackle privacy concerns when dealing with healthcare data. By adhering to data protection principles such as data minimisation, we intend to incorporate Privacy-by-Design (PbD) [European Parliament 2016] at the data level. Our approach is based on knowledge engineering and applies Semantic Web technologies into a solution that effectively accesses, represents, and extracts relevant information from health records and makes it available to the relevant professionals during an emergency.

The work in this thesis is the first that studies the representation of the evolution of health conditions in the context of supporting emergency services. Furthermore, it supports the extraction and adaptation of information in Smart City data exchanges, particularly minimising personal data processing in the dynamic context of emergency response.

1.1 Motivation

With the increasing number of people moving to urban spaces, efficiency and sustainable use of resources are becoming pressing issues for local governments. One approach to tackle this problem has been adopting the Smart City vision as a socio-technical solution to enhance resource management and build more sustainable cities.

An area of increasing interest is related to emergency response activities. As described in [Bartoli et al. 2015], public safety services are under pressure to respond more efficiently to emergencies (for example, local incidents, fires, and flooding). The response to incidents should be timely, and the right information can leverage the emergency response activities of decision-makers such as firefighters, police officers, and paramedics.

During emergency response operations, information about people involved in the emergency, along with casualties, is of paramount importance [Prasanna 2010; Nunavath, Prinz, and Comes 2016]. For example, during a fire emergency, responders' goals include ensuring the health and safety of the identified casualties [Prasanna 2010]. Access to consistent, up-to-date and relevant information that allows emergency services (for instance, firefighters and police) to assess the nature and extent of the event becomes crucial. They need to familiarise themselves with the situation and *gather information such as the type of people involved (male, female, children, disabled people), their medical condition, and special support needs*.

In addition, the value of having information about people with special needs is highly acknowledged by Health and Safety personnel and risk assessment good practices. For instance, educational premises are obliged to perform risk assessments, identify people in a vulnerable situation, discuss their health issues or impediments, and prepare tailored evacuation plans [Britain, Communities, and Government 2006; UK Government 2007b]. Emergency services may gather this data by approaching security personnel, fire wardens, or people present at the moment of the emergency. Despite this information coming from reliable sources, emergency services should use it with caution, reasonably challenging its *completeness and accuracy* (for example, information about employees requiring assistance because of a temporary disability might have changed in a matter of weeks).

Since the risk assessments are performed using medical information, Smart Cities provide the perfect conditions to leverage this process and improve services for citizens' benefit. In an interconnected city, this information can be gathered on demand, and health records provide an excellent opportunity to access up-to-date and accurate data about peoples' latest health issues. Specifically, health records include details of medical events that can be used

to identify vulnerable people or people who require special assistance in an emergency. For example, a record of a permanent condition such as ‘Fracture of vertebral column’ clearly identifies a person that has mobility difficulties; what is more, this information can be helpful in different scenarios. For instance, during a fire event, people with mobility issues might need special support to mobilise in an environment where there are stairs or other obstacles. This information could assist first responders in planning and prioritising efforts to evacuate this person. In a like manner, this information can be helpful for paramedics who can provide appropriate treatment and coordinate the transportation of a patient to a care site (an ambulatory care clinic or emergency department) according to their health circumstances.

During an emergency event, first responders should act immediately. In this scenario, analysing a large and detailed amount of information is not a sustainable practice for emergency responders. For instance, identifying helpful information can be a time-consuming task for emergency managers, and relevant information could be overlooked. Therefore, it is imperative to make relevant information available on time and avoid the overload of information to decision-makers. In addition, an intelligent system automatically analysing this data will require some knowledge to *identify valuable information* from people’s current health events [Morales Tirado, Daga, and Motta 2020]. Furthermore, the exchange and use of highly sensitive data also raise concerns about privacy and security. Organisations are reluctant to share personal information amidst growing concerns about breaches of regulations [UK Government 2007a], such as the EU General Data Protection Regulation (GDPR)¹ and the UK Data Protection Act². *Minimising the amount of data exchanged* to what is helpful not only facilitates the use of electronic health records but also safeguards citizens’ privacy.

Therefore, in a Smart City ecosystem, it is imperative to facilitate the interaction between healthcare providers and emergency services in a way that healthcare providers are able to deliver useful information while minimising privacy concerns. Crucially, this data exchange can assist emergency services’ needs for meaningful data on a person’s recent or current health issues. In the subsequent sections, we will detail the importance of the use of electronic health records in emergency settings using three paradigmatic use cases. Additionally, we will describe some challenges that become crucial when providing information to emergency services and detail the key motivations underlying our research on the use of health records to support the information needs of emergency services.

¹GDPR: <https://gdpr-info.eu/>

²UK Data Protection Act: <https://www.gov.uk/data-protection>

1.1.1 Use of electronic health records during fire emergencies

Emergency planning comprises activities that help organisations identify risks, prioritise planning development, and review results. Having processes and plans in place should prevent emergencies, and if they occur, good planning should reduce or mitigate the effect of the emergency [UK Government 2013b]. Vulnerable people are considered a key group as they may be less able to help themselves during an emergency [UK Government 2008]. This includes people with disabilities, mental health issues and children [UK Government 2008; UK Government 2013b]. Planning should look carefully into the special arrangements or assistance these groups may need.

Organisations are bound to follow governmental policies and procedures that guide the elaboration of action plans that should be followed during emergencies. For example, the UK's fire safety guidelines [UK Government 2007b] state that all employees should notify their line managers and the Health and Safety Department if they require assistance executing a fire evacuation plan. The assistance could be due to a disability or a temporary or chronic health condition. The organisation should ensure the design of a Personal Emergency Evacuation Plan (PEEP) tailored to employees' special needs. To develop the PEEP, the organisation collects employees' health information, such as the type of disability and discusses the support needed according to the person's capabilities and health issues. The PEEP is a form that records a detailed step-by-step plan of the agreed evacuation route the employee should follow in case of evacuation; it also records employees' personal information, the description of the type of disability and health-related data. Once the PEEP form is complete, it must be communicated to the relevant parties: Fire Wardens, all Appointed Helpers, and the relevant staff at the Health and Safety Department. The prime objective of elaborating the PEEP is to identify employees in vulnerable situations (for instance, people with mobility issues or mental health difficulties) and provide appropriate support in case of a fire evacuation.

When a fire emergency happens, the person responsible (fire wardens, security personnel and members of the health and safety department) should ensure that first responders are aware of the information contained in the PEEP. Typically, first responders attending emergencies gather this type of information and use it to assess the nature and extent of the event, make timely decisions, guide their personnel and distribute resources accordingly [Nunavath, Prinz, and Comes 2016]. The PEEP constitutes an excellent case in which healthcare data is stored and used to support emergency planning and response activities.

Clearly, having the latest health information allows organisations to assess, prepare and

adopt action plans (such as the PEEP) to follow in case of an emergency. More importantly, the latest information on ongoing health issues supports emergency services activities since it allows for the identification of vulnerable people's needs. Even though the PEEP provides means to collect this valuable information, health data management reveals other challenges organisations should face.

- Although procedures indicate a regular update of the PEEP form, employees' most recent health events might take some time to be recorded, particularly in large organisations. This could lead to *incomplete or inaccurate information*.
- It is possible that employees or visitors could consider their health issues irrelevant, particularly if they are experimenting with temporal disabilities. For example, most 'Sprain of ankle' problems recover after some time; however, if recent, it might result in some mobility impediment. This particular piece of information could draw the attention of firefighters planning the evacuation; information otherwise, not considered.
- Employees and visitors might choose not to disclose their current health status. For instance, sharing sensitive information about their disabilities might make them feel uncomfortable, affecting their willingness to *disclose personal information*. Furthermore, knowing that health information will be shared across different instances (within the organisation) may raise data privacy and security concerns.
- Access to the PEEP forms might be *time-consuming* and, in the case of emergencies where physical or digital infrastructure is affected, even unfeasible.

The collection and processing of health data prove to be challenging for organisations, even though its collection is intended to support emergency planning and responding operations.

The smart city ecosystem expedites the identification of people in an emergency; for instance, location information can be retrieved from Access Control Systems (ACS), building monitoring systems, and tracking devices. Crucially, the Smart City perspective opens opportunities to leverage the use of Electronic Health Records (EHR) and facilitate its management. Intelligent systems interacting in a Smart City environment could facilitate the retrieval of up-to-date EHR from health providers, its subsequent analysis and use to identify people requiring assistance, making this type of information accessible to emergency services. However, there are significant challenges when managing health records. First, a person's health record contains extensive and fine-grained information, which could be

overwhelming for firefighters and fire wardens; a large amount of data makes it difficult to find relevant information, and important health events could be overlooked. Second, health records contain very sensitive information. Preventing the disclosure of personal data while providing emergency services with usable information is an important and difficult problem [Morales Tirado, Daga, and Motta 2020]. In principle, an intelligent system could act as a mediator between the healthcare data provider and the emergency services. The intelligent system could help the health service provider's data managers to deliver useful information while minimising personal data exchange (following the GDPR principle of data minimisation for adequate, relevant and limited processing of data [Information Commissioner's Office 2016]).

Importantly, such a solution would relieve the organisation from the management of complex and sensitive data. Furthermore, it would make useful and up-to-date information readily available during an emergency.

1.1.2 Access to health records by ambulance services

Prehospital care is an essential part of the emergency health care system. In the UK, pre-hospital care providers responded to approximately 11.7 million calls between 2018 and 2019, and the demand increased rapidly with an average year-on-year increase of 6% [NHS England 2019]. Around 90% of calls are classified as 'urgent care', whereas 10% are for life-threatening emergencies requiring transfer to an emergency department [Porter et al. 2020]. Specifically, ambulance services are vital in delivering better health care outside the hospital when it is more suitable for patients. In other words, ambulance services can act as an enabler for safe non-conveyance of patients and shift to out-of-hospital treatment. In recent years, the UK has carried out major policy reviews [Keogh 2013; McClelland 2013] in the emergency and urgent care system. As a result, the Department of Health and Social Care strategy encourages ambulance services to integrate practices that safely provide alternatives to transport to the hospital (non-conveyance). The ultimate goal is to provide better healthcare services, reduce hospital waiting times, and efficiently reduce expensive treatments.

However, integrating practices to safely avoid conveyance has been difficult, primarily due to the lack or minimal access to up-to-date and useful information about the patient's health [Clark et al. 2019; Patterson et al. 2019]. Often paramedics can only rely on patients' symptoms at the moment of the emergency or scarce information provided by a dispatch team, family members or carers. As a consequence, paramedics have limited options but

to convey patients to hospitals, which is not always the most appropriate option or even desired by the patient. Having access to health records, crucially containing information on ongoing health issues or the latest health events (e.g., conditions and medications), can support paramedics' activities.

In what follows, we describe some challenges faced by first responders:

- Information such as latest medications, allergies, details of chronic or permanent conditions, care and crisis plans are crucial when assessing the type of care required [Zorab, Robinson, and Endacott 2015; NHS England 2019]. Ambulance staff typically have almost no access to health records, and when they can access them, *the amount of data could be overwhelming*. Paramedics can benefit from health records information; however, the data should be *relevant* in order to make well-informed and timely decisions.
- Understanding patients' pre-existing medical conditions and recent treatments becomes a critical data asset for ambulance personnel with no prior knowledge of patients' conditions. Health records can *provide up-to-date information* [Patterson et al. 2019] of paramount importance when deciding the type of treatment required.
- Concerns about *privacy and security* represent a significant barrier to exchanging electronic health records among health organisations. Particularly when a patient's health records are managed by ambulance services and primary care separately [Smith 2017]. Reducing the amount of information exchanged by health services to the most relevant can minimise concerns about data disclosure while providing useful information.

In summary, access to *relevant* electronic health records is still a pressing issue for paramedics. Emergency services need relevant information about patients' health situation, for instance, details of latest medications, allergies, details of chronic or permanent conditions, care and crisis plans. Although all this information is contained in the EHR, access to the medical history alone does not guarantee a swift use of the data. Therefore, paramedics acting in a pressing environment require some kind of support to carefully examine up-to-date and recent health conditions. Such access to data indicating ongoing health issues has the potential to leverage out-of-hospital care and assist paramedics decision-making process.

1.1.3 Exchange of EHR in the Smart City setting

One key component of Smart Cities is data. As might be expected, the shift to electronic health records has led to a collection of large-scale datasets, which opens opportunities for

implementing Smart City systems to improve citizens' lives, in particular, by assisting health care services [Xu et al. 2014; Majumder et al. 2017]. In what follows, we will explain how this information could be used in different aspects of daily life to benefit citizens in the Smart City context.

Access to EHR for pharmacy dispensing. As mentioned previously, electronic health records hold historical information about prescriptions and medications patients are taking. EHR can be used to monitor patients' treatment, identify drug and allergies interactions, and assess medication appropriateness [VanLangen and Wellman 2018]. This evaluation is quite an important task, as any mistake in the prescription or medication could result in health episodes ranging from mild discomfort to severe scenarios of life-threatening allergic reactions [Allen and Sequist 2012].

For instance, when a person suffers a mild issue, he/she might initially contact a pharmacist. Usually, pharmacists will ask about the symptoms, current medication, and if the patient has allergies; they will base their assessment and following prescription on little information provided by the patient or carer. Additionally, pharmacists use this information to identify and resolve possible medication discrepancies. Although the process of initiating medication prescriptions is well established, without an appropriate verification of the patient's medication details, allergies or drug-related problems may go unnoticed and therefore increase the risk of an error.

Hence, particularly for medications provided over the counter or in a care home scenario, access to a patient's recent health problems and current medication history becomes a significant source of information to avoid errors or unwanted health reactions. Up-to-date information about current prescriptions and especially details of chronic issues can support the decision-making process for health professionals assisting ill patients. Although EHR contain this type of information, accessing health records is not enough. There is a need to obtain information that reveals conditions that are affecting the patient at the current time and that an intelligent system could derive from electronic healthcare records.

EHR for monitoring cities' health care systems. Cities across the world are experiencing a rapid rate of urban growth. A growing population brings particular challenges for local governments regarding the provision of public services and the management of resources. For instance, the demand from governments, public services and patients to have reliable access to health information was never as crucial as with the ongoing COVID-19³ pandemic. Smart healthcare systems are gaining momentum in the Smart City paradigm. Such intelli-

³COVID-19: <https://en.wikipedia.org/wiki/COVID-19>

gent systems should be able to monitor disease outbreaks, and the latest patients' diagnoses continuously and efficiently exchange healthcare data to decision-makers [J.-H. Kim and J.-Y. Kim 2022; Verma 2022].

Early detection and diagnosis are keys to addressing any disease outbreak. Smart Cities should be able to take advantage of the different systems in place to obtain crucial information for managing a health outbreak. In this context, electronic health records constitute crucial data assets, as they contain the latest information on citizens' health. Intelligent systems interacting in a Smart City environment could *facilitate the retrieval of up-to-date electronic health records, its subsequent analysis* and, therefore, better monitoring and management of a disease outbreak. The retrieval and exchange of electronic health records should be focused on the latest or ongoing health issues, thus minimising the amount of information to be analysed, the time used to process such information and concerns about privacy and security breaches. Information about the number of people with certain diseases, medication, and hospitalisation statistics is the type of data that can be used to monitor and supervise disease outbreaks; furthermore, decision-makers can use it to plan and distribute resources accordingly. Although, large data sources, such as EHR, have become accessible not all data is useful in a scenario where the decision-makers should identify and analyse current conditions.

Clearly, accessing and extracting recent events from electronic health records is not sufficient in order to provide information that supports decision-makers during an emergency. Events (conditions, procedures, among others) recorded in EHR evolve differently in a period of time. Therefore, in order to get an account of the *current health status of a person*, a system should *reason on the evolution of a given condition* (event) in other words, if the event is still ongoing at a certain point in time. The work presented in this thesis addresses the problem of identifying relevant information, particularly the ongoing health issues of citizens involved in an emergency event, which can support decision-makers during the early response. Specifically, this thesis examines how to facilitate the identification of useful data extracted from EHR, which can then be used to indicate an ongoing health problem and consequently reveal if the EHR owner requires assistance or needs support during an emergency. Consequently, this process will reduce the amount of information that first responders or organisations have to process. However, analysing ongoing or chronic health events registered in a person's health records is not a straightforward process, and it is a challenging study area. As reported by [Alfattni, Peek, and Nenadic 2020; Li et al. 2020] significant research has been conducted to extract medical information from natural language sources. However,

researchers have not treated *health evolution* or the analysis of EHR to evaluate ongoing or current health issues in much detail. To the best of our knowledge, none of the existing approaches studies the representation of *health evolution* for supporting the detection of people requiring assistance in the context of supporting emergency services.

1.2 Hypotheses and Research Questions

The main research question investigated in this thesis is:

How to identify current health issues from EHR, to support the accurate identification of people requiring assistance during an emergency?

Our central hypothesis is that it is possible to provide emergency services with additional information about people requiring assistance during an emergency by automatically analysing people's health records. Health records can be used to draw a picture of people's current health status and derive the health issues that indicate a vulnerable situation or need for special assistance.

Therefore, the main focus of this research is to provide decision-makers with relevant information about current health conditions extracted from EHR, so that it could help them assess if a person has an ongoing health issue and requires assistance during an emergency. As discussed earlier, EHR contains a very detailed and large amount of information about a patient's health history. In order to facilitate its use by first responders, a system should be able to identify the events that are ongoing at the moment of the emergency. In this way, a system performing this automatic analysis facilitates the task of assessing if a person requires special assistance, which in turn reduces the information delivered to decision-makers and minimises concerns about privacy and security breaches.

In what follows, we describe the specific research questions we address and the related hypotheses:

Research question 1. *How to formally represent health evolution to support the identification of current/ongoing health issues?*

In an ideal scenario, for a doctor, it takes some time to assess a patient's current health status; the doctor would ask the patient about current symptoms or issues and read and analyse the patient's historical health records. The doctor's knowledge allows him to make an initial diagnosis of the ongoing health issues of the patient and advise the patient on what to expect and how long will it take to recover.

In an emergency, decisions should be taken on expeditiously. Although a system can provide emergency managers with access to electronic health records, they still have to analyse historical medical records and detect ongoing issues promptly.

Ideally, a system should be able to automatically analyse if a medical event is improving or worsening over a certain period. However, a system requires some knowledge about medical events' evolution over time. A formal representation of 'health evolution' is needed to guide our system and support the evaluation of current health issues. For example, medical events typically have a temporal span associated with their evolution, e.g., improving (a broken leg) or worsening (Alzheimer) over a certain period or becoming chronic (asthma). Therefore, it is imperative to formalise a model for representing health evolution over time in the context of emergency response. All this information should be represented in a computer-readable form.

Our hypothesis for knowledge representation of ongoing health issues is:

Hypothesis 1. *A model for representing the evolution of health events over time can support the detection of ongoing health issues in electronic health records.*

As presented previously, analysing health conditions is not a straightforward process. We hypothesise that by examining in detail how medical events' evolution is described in natural language, we can abstract a formal model that solves the problem of representing health condition evolution in a computer-readable format. From the analysis, we expect to derive the characteristics that define health condition evolution and generalise a model that represents the different types of condition evolution and the estimated recovery time. Therefore, a system using this model can infer whether a medical event improves or declines, how long it takes for a person to recover, or if a medical event is chronic. Ideally, we want to identify what are the characteristics that define the health condition evolution. How elements such as the type of condition (improvement, deterioration, chronic), duration (maximum and minimum convalesce time) and pace (if a condition recovery changes fast or slow) can be used to represent the evolution of health conditions over time.

Research question 2. *How to build a database of health condition evolution?*

This question is closely related to the previous one. A system that has the means to represent health condition evolution now requires specific data, for example, about convalescence time, data indicating the time a fracture usually improves and eventually recovers.

An essential aspect of EHR data handling and exchange is the standardised representa-

tion of clinical terminology. SNOMED CT⁴ [SNOMED International 2017], ICD-10⁵ and LOINC⁶ are among the most used schemes; these schemes comprise medical terms (such as disorders, procedures, among others), support reporting and monitoring. However, these terminology systems do not include definitions or structured information describing health evolution which reveals the need for a reliable database that describes medical events' convalescence time and recovery. In contrast, unstructured information regarding known health issues' convalescence time and recovery can be collected from reliable public sources such as NHS England⁷ and MAYO Clinic⁸.

Knowledge acquisition and text classification techniques can automatically extract and build a database of health evolution information. In addition, it is essential to envision how to expand the collection of such information either in an automatised manner or by providing tools to build this database with the collaboration of domain experts.

It is then clear that a database of health evolution constitutes an essential part of the health evolution representation.

Thus, our hypothesis for this question is the following:

Hypothesis 2. *It is possible to build a structured database of health evolution information using open, public and authoritative data sources.*

We hypothesise that information about health evolution can be collected and extracted from public sources. Knowledge acquisition techniques such as Natural Language Processing (NLP) and Machine Learning (ML) can be used to build a semi-automatic supervised classification pipeline and develop a structured database of health evolution information. The resulting database can be published and made available using Semantic Web technologies and linked to well-known medical terminology standards such as SNOMED CT. In this way, it can be extended by aggregating information on the recovery time of medical events.

Research question 3. *How to automatically reason on EHR to identify ongoing health issues?*

We envision a system that automatically estimates whether a specific condition stored in a health record holds at a certain point in time and, therefore, can be considered valid. The idea is to use the database and the elements representing the health condition evolution to

⁴SNOMED CT: Systematised Nomenclature of Medicine Clinical Terms <https://www.snomed.org/>

⁵ICD-10: International Classification of Diseases

⁶LOINC: Logical Observation Identifiers, Names, and Codes

⁷NHS England: <https://www.nhs.uk/conditions/>

⁸MAYO Clinic: <https://www.mayoclinic.org/diseases-conditions>

identify only relevant records. Specifically, a health condition is relevant if the condition is still ongoing (an individual has not recovered from it) when the emergency occurs (e.g. the fire started in the building). For example, a certain condition that generally improves in two weeks and occurs three days before the emergency is likely to impact a person's health. This person then might require special assistance to evacuate.

Our hypothesis here is:

Hypothesis 3. *It is possible to formalise the rules for estimating automatically health condition evolution.*

Typically, a health condition is still ongoing if: the recovery time has not passed yet, the condition is chronic, or the condition deteriorates in time. The idea here is to use the characteristics identified during the formalisation of the health condition evolution model to design a set of reasoning rules that will allow an intelligent system to estimate automatically if a health condition is ongoing at a specific moment.

Research question 4. *How to leverage the developed knowledge components to build an intelligent system capable of identifying people requiring assistance during an emergency?*

The design of a solution for extracting valuable information from significantly large data sources, such as health records, covers different aspects of the data management process. We envision a solution that reduces the amount of data to be disclosed by identifying helpful information for a given task (in this case, the latest health issues) and finally delivers accessible and easy-to-interpret information to final users, in our case for a fire emergency scenario.

Our objective is to identify all the requirements that these components (software or knowledge bases) should fulfil to achieve the primary goal of providing emergency services with information about people in need of help during a fire evacuation.

The idea is to analyse the data requirements and use the knowledge components developed previously to identify ongoing health issues from EHR. We also explore regulations for fire evacuation and types of assistance or disabilities regarding an impediment to performing an evacuation plan [UK Government 2007b], which could assist the emergency team in identifying people that require attention and the associated impediment. The ultimate goal is to contextualise the provided information without overwhelming emergency services.

Our hypothesis for this question is:

Hypothesis 4. *By relying on the components developed before, it is possible to design an intelligent system that uses health records data to deliver relevant information to emergency*

services.

To solve the problem of identifying useful information from EHR, a system should be able to estimate the duration of a given health condition, emulating the process that an emergency responder would follow. This analysis can be performed by an intelligent system that brings together the knowledge components developed before. Ideally, the system should be able to receive as input people's EHR, process them and finally deliver as an outcome a list of people that have an ongoing health condition and, therefore, require assistance.

The process of reasoning on health evolution can be supported by the representation of the health condition evolution component and the information collected from structured data about health condition evolution. This analysis can significantly reduce the amount of sensitive data to be delivered to emergency services, as only the records indicating a condition is in progress will be considered. Furthermore, the system could be capable of processing the data and, ideally, classifying the relevant health events according to UK governmental guidelines on types of disabilities [UK Government 2007b].

The proposed intelligent system architecture should follow best knowledge engineering practices and use Semantic Web technologies that support the development of the required knowledge components.

Ideally, we consider the instantiation of our solution in an environment where the interaction with other intelligent systems supports access to EHR swiftly:

Assumption 1. The designed solution relies on the assumption that in a Smart City environment, its information technology infrastructure allows determining the location of a person and, therefore, knowing if this person is involved in an emergency. For example, an intelligent building has an Access Control System (ACS) that automatically registers when people enter or leave the premises. By accessing the registers of the ACS, the software solution can identify the people in the building and communicate with the health provider and, therefore, use this information to retrieve occupants' health records.

Assumption 2. Electronic health records are represented using well-established standards such as SNOMED CT and FHIR⁹. The exchange of EHR is supported by a communication infrastructure between the EHR provider and the emergency services already in place.

⁹Fast Healthcare Interoperability Resources: <https://www.hl7.org/fhir/overview.html>

1.3 Research Methodology

The main goal of this research is to make health records information accessible to emergency services, first by recognising ongoing health issues that identify people requiring assistance and second, by processing these records in order to minimise the amount of information exchanged with first responders. To attain our objective and address the research questions stated previously, we focus on designing and developing a solution at the knowledge level [Newell 1982], thus adopting a knowledge engineering point of view. Although critical in a real scenario, this research work does not focus on designing and developing a complete software platform that implements security policies or uses techniques such as cryptography, control access or anonymisation to access EHR. Instead, we focus on designing artefacts that could serve as components of an intelligent system to make an automatic assessment of relevant EHR.

To this end, our research is built upon Design Science [Johannesson and Perjons 2021] as a general methodology that guides our research. This method framework includes five main activities: (1) problem investigation and (2) requirements definition, (3) artefact design and development, and finally, (4) demonstration and (5) evaluation.

Our research is structured to reflect this methodology. It starts by *describing the main problem* and the motivations behind our research. Next, we investigate work in the context of Smart Cities and the use of health records to support emergency services activities. We also summarise the challenges and issues organisations and emergency services face when using healthcare data. Then, we *define the requirements* of the components/artefacts to be developed to answer each research question proposed in previous sections. In order to *design and develop the artefacts* we work in three main areas: Knowledge acquisition, representation and reasoning for identifying ongoing health issues and using a common-sense knowledge base to transform data.

In our methodology, the *demonstration and evaluation* of each artefact function is grounded in the experimental evaluation [Dodig-Crnkovic 2002; Schiaffonati and Verdicchio 2014] to assess whether each component of our solution is producing the expected results. Health-related data is considered highly sensitive information. Hence, to prevent any disclosure of private information, in our experiments, we make use of synthetic healthcare data encoded using FHIR¹⁰. Specifically, we use Synthea [Walonoski et al. 2018], an open-source software that generates synthetic electronic health records. Each patient's health record is generated

¹⁰Fast Healthcare Interoperability Resources is a standard describing data formats and elements and an application programming interface for exchanging electronic health records.

independently and simulates the health registers from birth to death through modular representations of various diseases. The final prototype system aims to validate our research hypothesis using a synthetic health records dataset to identify people with ongoing health issues and compare the results with a manually generated gold standard of people requiring assistance.

1.4 Approach

The use of rich data sources to support emergency services activities is an area that has brought attention lately, particularly in the context of Smart Cities where heterogeneous data sources and devices interact to bring updated information to final users. Our research focuses on developing a pipeline that uses EHR data as input to identify and extract useful information that emergency services can use during emergencies. In this context, our proposed approach addresses the problem from the perspective of Knowledge Engineering with particular attention to Knowledge representation and reasoning and Semantic technologies.

In order to obtain updated information about people with ongoing health issues, we identify Electronic Health Records (EHR) as the main source of information. An advantage of using EHR is the use of well-established standards such as FHIR (Fast Healthcare Interoperability Resources) standard specification for exchanging healthcare information and SNOMED CT for describing medical terms. Next, we identify the knowledge requirements according to the emergency and formalise them in the form of research questions. The research questions guide the identification of the different knowledge components that are part of our solution. We can list four components:

- An ontology that abstracts the definition of health condition evolution
- A Knowledge Graph of structured information about health evolution
- A knowledge component that performs the reasoning on health evolution
- An intelligent system that uses all the components listed above to extract and deliver information to emergency services

The process of recovering from a health situation is not limited to stating the convalescence time, typically described as having a minimum and maximum duration. There are situations such as chronic conditions that could deteriorate over time. Therefore, for our

first component, we analyse how the evolution of health events is described in natural language; we use two main sources: NHS England and MAYO Clinic. The Health Evolution Statement (HES) is the resulting representation of the health recovery process once we have analysed the aforementioned sources. We build a model based on these concepts, resulting in HECON, the Health Condition Evolution Ontology. Our ontology aims to support the identification of ongoing health events at a given point in time by representing medical health events' evolution information. In particular, we link the health evolution data to the SNOMED CT taxonomy and extend it by aggregating information on the recovery time of medical events.

For our second component, since there is no existing structured data about health evolution available to reuse, we develop a data pipeline for constructing a database of health evolution annotations linked to SNOMED CT concepts. We rely on knowledge acquisition techniques such as Natural Language Processing (NLP) and Machine Learning (ML) and build a semi-automatic supervised classification pipeline to extract condition evolution information from unstructured text collected from websites such as NHS England and MAYO Clinic. We train different ML algorithms and classify sentences according to the different components of health evolution using HECON. The result is at least one HES for each health condition in the data source. We follow a Linked Data approach to publish the newly obtained database as a Knowledge Graph of health evolution, creating an accessible resource. To evaluate the results of the text classification process, we designed a user study directed to health domain experts. The objective was threefold, validate the overall applicability of the knowledge extraction approach, further populate the HES and capture domain experts' knowledge so that we obtain a high-quality curated database of HES.

We explore how to better characterise the information delivered to emergency services. We use UK guidelines for fire evacuation [UK Government 2007b], which provide a list of categories that represent disabilities.

Finally, to address our general research question, we developed a system which takes as input an EHR. CONRAD, an Intelligent System for Emergency Support, processes the data following our proposed methodology to identify people requiring assistance and generate a summary of the type of help they require. We evaluated our system in a simulated environment, where access to the identification of people in the building is possible using the building's Access Control System. Similarly, we assume that access to the public health provider is available. Since health-related data is considered highly sensitive, we use synthetic healthcare data to evaluate our proposed approach. We use the Synthea open-source

software [Walonoski et al. 2018], which generates synthetic electronic health records. To evaluate the overall approach, we designed a set of experiments following different hypotheses created to test the use of HECON descriptions of health condition evolution.

1.5 Thesis outline and contributions

In what follows, we list the contributions of our work and summarise the content of this thesis and the publications that are part of this research.

1.5.1 Contributions

This thesis aims at contributing with a process that automatically identifies relevant information from rich and makes it accessible to emergency managers. Our approach relied heavily on Knowledge representation and reasoning, Knowledge acquisition, and Semantic Web technologies as fundamental building blocks of a knowledge-based system. From a Knowledge representation and reasoning perspective, our main contribution is that we built a model for describing the evolution of health events, formally represented as HECON Ontology, and provide the means to reason on this model. From a Knowledge acquisition perspective, we developed a data pipeline to automatically collect and extract information on health evolution from public web sources. Following, we describe some of the contributions (all contributions are published, see Appendix B for the list of software, tools, and datasets developed):

- A model for representing and reasoning about the evolution of health events over time: HECON - Health Condition Evolution Ontology (Chapter 4).
- A pipeline that allows the *semi-automatic* extraction of information about health evolution (Chapter 5) and includes humans in the loop to curate the generated data.
- A Knowledge Graph (KG) that defines an abstraction of the available data about health evolution. The KG includes information that extends the descriptions of the clinical concepts provided by SNOMED CT (Chapter 5).
- We present CONRAD - Health Condition Radar, the intelligent system that implements the proposed methodology. It uses as data input a sample of randomly selected synthetic health records [Walonoski et al. 2018]. The final output is a list of people with ongoing health issues that potentially require assistance to evacuate during a fire emergency (Chapter 6).

1.5.2 Overview of the thesis

The material of this thesis is distributed in individual chapters as follows:

The next chapter is dedicated to the state of the art. First, we review the adoption of healthcare information to assist emergency services in the context of Smart Cities and the use of health records to support emergency responders; it also includes a brief review of the emergency response process. Next, we introduce concepts and literature about Knowledge representation and reasoning with healthcare data, followed by Knowledge acquisition methods and knowledge graph construction from unstructured data. We close this chapter by reviewing the Semantic Web technologies and standards used to represent and exchange Electronic Health Records (EHR), querying RDF data, and using synthetic health records.

The following four Chapters provide details of our framework. In Chapter 3, we describe the motivating scenario and the survey of findings on data management practices for emergencies. This includes a description of the intended stakeholders. Chapter 4 presents our approach for building an ontology to represent health events evolution, the requirements and the sources used. In Chapter 5 we describe the pipeline for building the database of health evolution and the construction of the Knowledge Graph Knowledge. Chapter 6 draws together the elements studied during our research into CONRAD - Health Condition Radar, an Intelligent System to provide emergency services with information about vulnerable people. We describe all the components of our proposed architecture and its evaluation.

In the last two Chapters, we present the evaluation of the elements we developed during the research, the user study results, and the final discussions. Chapter 7 is divided into three sections. The first section describes the evaluation of the Knowledge Graph; the second section presents the results of the user study conducted with domain experts to validate the database of health event evolution. The last section describes the results obtained by the CONRAD system, the software that applies this thesis proposed approach. Chapter 8 closes this thesis with some final remarks, including a discussion of the results achieved by our approach, some limitations and future directions.

1.5.3 Publications

Below we list the publications covering this work and we associate them with the relevant chapters.

Chapter 4

- Morales Tirado, Alba; Daga, Enrico and Motta, Enrico (2022). HECON: Health Con-

dition Evolution Ontology. In: Proceedings of 5th Workshop on Semantic Web solutions for large-scale biomedical data analytics, co-event with The ESWC 2022: Extended Semantic Web Conference. SeWeBMeDA-2022, Hersonissos, Greece.

Chapter 5

- Morales Tirado, Alba; Daga, Enrico and Motta, Enrico (2022). Towards a Knowledge Graph of Health Evolution. In: Knowledge Engineering and Knowledge Management. EKAW 2022, Bolzano, Italy.
- Morales Tirado, Alba; Daga, Enrico and Motta, Enrico (2021). Reasoning on Health Condition Evolution for Enhanced Detection of Vulnerable People in Emergency Settings. In: Proceedings of the 11th Knowledge Capture Conference. K-CAP 2021, Virtual Event, USA.

Chapter 6

- Morales Tirado, Alba; Daga, Enrico and Motta, Enrico (2022). (*Demo paper*) CONRAD Health Condition Radar - an Intelligent System for Emergency Support. In: Proceedings of 5th Workshop on Semantic Web solutions for large-scale biomedical data analytics, co-event with The ESWC 2022: Extended Semantic Web Conference. SeWeBMeDA-2022, Hersonissos, Greece.
- Morales Tirado, Alba; Daga, Enrico and Motta, Enrico (2021). Reasoning on Health Condition Evolution for Enhanced Detection of Vulnerable People in Emergency Settings. In: Proceedings of the 11th Knowledge Capture Conference. K-CAP 2021, Virtual Event, USA.
- Morales Tirado, Alba; Daga, Enrico and Motta, Enrico (2020). Effective use of personal health records to support emergency services. In: Knowledge Engineering and Knowledge Management. EKAW 2020, Virtual Event, Bolzano, Italy.
- Morales Tirado, Alba (2019). Towards a privacy-aware information system for emergency response. International Semantic Web Research Summer School (ISWS), Bertinoro, Italy. (*Best Poster Award*)¹¹
- Morales Tirado, Alba; Daga, Enrico and Motta, Enrico (2019). Towards a privacy-aware information system for emergency response. Proceedings of 16th International

¹¹See Appendix D.3

Conference on Information Systems for Crisis Response and Management (ISCRAM 2019), Valencia, Spain.

Chapter 7

- Morales Tirado, Alba; Daga, Enrico and Motta, Enrico (2022). Towards a Knowledge Graph of Health Evolution. In: Knowledge Engineering and Knowledge Management. EKAW 2022, Bolzano, Italy.
- Morales Tirado, Alba; Daga, Enrico and Motta, Enrico (2021). Reasoning on Health Condition Evolution for Enhanced Detection of Vulnerable People in Emergency Settings. In: Proceedings of the 11th Knowledge Capture Conference. K-CAP 2021, Virtual Event, USA.

Chapter 2

State of the art

In this chapter, we cover background information and provide the reader with an understanding of the context in which we develop our research. We deal with the problem of identifying relevant data from EHR from the Knowledge engineering perspective. In order to tackle this problem, we start (Section 2.1) by giving an overview of the use of EHR and describing the considerations we have to take into account when managing sensitive information such as health records. We also review the latest research on Smart City applications to assist emergency response activities and identify the opportunities and gaps in the literature regarding intelligent systems support for processing EHR. Then we move to introducing the areas in which we place our contributions, covering current research on Knowledge representation and reasoning on health condition evolution (Section 2.2), with particular emphasis on the identification of ongoing medical issues-oriented to support emergency response activities. Next, we describe approaches and technologies used to build Knowledge Graphs from unstructured data sources, particularly natural language (Section 2.3). The last Section 2.4 is dedicated to giving an overview of the Semantic Web tools that are the base of the knowledge components presented in this thesis.

2.1 Use of healthcare information in the Smart City context

In 2018, 55% of the world population was living in urban areas, and by 2050, it is expected that this number will reach 68% [United Nations 2022]. The rapid growth of the population in urban spaces is challenging local governments to deliver better services (e.g. transportation, energy and water supply, environment, education, and emergency management, among others), improve resource management, ensure sustainability and address social needs. The Smart City concept has emerged as a solution to create better and sustainable urban envi-

ronments, taking into account the needs of citizens, the information generated within the city, and using Information and Communication Technologies (ICT) to enhance the services [Bowerman et al. 2000].

Although the meaning of Smart City is not strictly defined, it has attracted the attention of many local governments as an answer to serve the urban population better. London [Greater London Authority 2018], Milton Keynes [Caird, Hudson, and Kortuem 2016], and Barcelona [Bakıcı, Almirall, and Wareham 2013] are examples of successful case studies. Smart City initiatives try to enhance the use of resources and build sustainable cities by taking into consideration crucial areas, such as urban planning (water, energy, food, air pollution, street lighting, health care, emergencies), use of existing infrastructure (buildings, roads, housing), business support, mobility and transportation, technology, education, environment, and governance [Lombardi et al. 2012]

Broadly speaking, Smart City systems use information from a wide variety of sources in order to optimise services and, therefore, enhance the management of cities. A set of heterogeneous systems and infrastructures collect, exchange, store and process real-time and on-demand information from different elements, including sensors, communication networks, Internet of Things (IoT) devices, specialised databases and state-of-the-art data centres, and so forth. Additionally, information also may come from other sources, for example, from local governments, the private sector and citizens' interactions with social media platforms [Zygiaris 2013; Gharaibeh et al. 2017] and even distributed under specific policies [Daga, d'Aquin, et al. 2015].

An area of growing interest is the provision of better emergency and healthcare services. The significant increase in urban population density calls for local governments to implement smart public management solutions to respond quickly to emergencies; besides, calls for better health management after the latest COVID-19 outbreak have driven the attention of research on smart solutions [Moreira, Sinderen, and Pires 2019; Allam and Jones 2020]. In the case of an emergency (such as fire, earthquakes, floods, terrorist attacks, and hazardous material incidents, among others), detailed information could represent a life-saving resource [Chehade, Matta, Jean-Baptiste Pothin, et al. 2018] for emergency responders such as firefighters, police, health bodies, or local authorities. For example, in the case of a fire event, an intelligent building equipped with indoor tracking and motion detection systems can provide emergency services with information about the location of people on each floor of the building [R. Srinivasan, Mohan, and P. Srinivasan 2016]. This detailed information is critical for emergency responders and can help plan and execute rescue operations.

It is clear that enormous volumes of data drive many activities in a Smart City, such as expansion planning and use of resources, among others. On the one hand, the promise of smarter management of cities leads to continuous monitoring and collection of granular and incredibly heterogeneous data. Furthermore, the adoption of advanced and ubiquitous technology has made the constant growth of large amounts of data every day [Hashem et al. 2016] even easier. On the other hand, dealing with such an enormous amount of data, particularly personal and sensitive data, its collection, processing, and storage, but overall, the exchange of this information within Smart City systems has raised concerns about the security and privacy of the citizens [Curzon, Almeahmadi, and El-Khatib 2019]. According to the GDPR legislation, personal data is defined as “*any information relating to an identified or identifiable natural person*”; additionally, it considers a particular category defined as “sensitive data”, which is entitled to higher protection. For instance, information regarding health conditions, disabilities, sexual orientation, and location are considered sensitive data. Crucially, issues about dissemination or disclosure appear when sharing information with external agents: once the information is shared, the owner has no control over how their data will be used, by whom, and for what purposes.

In addition, providing access to a large and detailed amount of information is not always helpful in the context of a Smart City environment. Traditional approaches, largely through human iteration, are unable to analyse such vast amounts of data, particularly in emergency scenarios. Identifying helpful information can be time-consuming for emergency managers, and untrained eyes could overlook relevant details.

The following section describes the context in which EHR is being used by emergency services and provides a thorough analysis of the challenges and opportunities when managing electronic health records. Next, section 2.1.2 reviews the solutions proposed to support emergency services activities and how they tackle some previously described problems.

2.1.1 Use of Health records in emergency support

Electronic health records (EHR) were proposed as a ‘digital tool’ that helps health service providers manage patients’ care and track patients’ historical conditions. Furthermore, the use of EHR data has seen a significant adoption thanks to the development of standard technologies for the exchange and representation of healthcare. Two components can be mentioned; first, the Fast Healthcare Interoperability Resources (FHIR) standard is used to facilitate the exchange of EHR between heterogeneous systems. Second, the adoption of established and controlled terminology to represent clinical terminology such as SNOMED

CT enable the representation of accurate medical facts¹.

The shift from paper-based records to the implementation of information technologies has leveraged the development of health-related smart applications and intelligent systems, opening up opportunities for patients to manage their care options. For instance, health care organisations are becoming aware that EHR can support other activities, including improving patients' experience, facilitating care for patients with chronic conditions, audit and financial planning, and ensuring the continuity of care outside hospitals boundaries [Wiljer et al. 2008; P. B. Jensen, L. J. Jensen, and Brunak 2012; Shah and Khan 2020].

Furthermore, the exponential gathering of information has made it possible not only to collect health data related to patients' demographics and current conditions, but it has changed how medical professionals manage (filter and interpret) information about prescriptions, laboratory test results, and medical imaging data, among others. The data generated by different healthcare organisations and recorded by clinical professionals in different areas [McGraw and Mandl 2021] is growing significantly [Stanford Medicine 2017].

Clearly, access to health data is not only crucial for activities related to the delivery of care, but there is also a growing trend for its exploitation outside the primary health care domain. In fact, the information contained in health records represents a life-saving resource for first responders [Nunavath, Prinz, and Comes 2016; McNeill et al. 2019]. Moreover, emergency responders value swift and transparent access to data contained in EHR [Ben-Assuli and Leshno 2016] which crucially supports their emergency operations (reviewed in detail in Section 1.1). For example, firefighters and organisations benefit from having relevant information about disabilities that affect people's capacity to evacuate [UK Government 2008; Phillips, Neal, and G. Webb 2016]. Similarly, paramedics can provide better treatment when accessing the latest medication or type of care required by patients with chronic conditions [Clark et al. 2019; Patterson et al. 2019].

However, emergency response is exceptional in nature. The provision of data to emergency services, particularly electronic health records, differs from other areas of healthcare because emergency responders typically perform their activities in a high-stress and accelerated decision-making environment [Rosenfield, Harvey, and Jessa 2019]. These particular circumstances should be considered to avoid problems regarding the processing and interpretation of rich information sources such as electronic health records.

In this scenario, identifying relevant information could become a challenging task when EHR contribute a large amount of fine-grained data because:

¹Section 2.4.2 gives a thorough description of both standards.

- Healthcare data is highly specialised and may be difficult to interpret by the personnel involved in supporting the evacuation (e.g., firefighters).
- A large amount of data makes it difficult to find relevant information.
- Exchange of sensitive information might put citizens' privacy at risk.

One of the main obstacles is the problem related to the dissemination of personal data. It seems to be a barrier to appropriate information sharing within emergency services and healthcare service providers (such as hospitals), crucially impacting the delivery of emergency services and becoming a real issue for emergency responders. A report from the UK government referring to the London bombing attacks (7 July 2005) points out that the *'Limitation on the initial collection and subsequent sharing of data...'* was due to concerns about the exchange of personal data [UK Government 2006; UK Government 2007a]. Consequently, in 2007, the UK Cabinet Office issued a Practical Guide document to help those faced with making decisions on using personal information; the guideline clarifies the framework of the Data Protection Act 1998 [UK Government 2007a]. Nevertheless, the decision to share personal data is based on the evaluation of the public interest benefit and still has to be made by emergency responders at the moment of the emergency, thus adding tasks to their normal activities. The issues related to privacy that hamper the effective reuse of data can be summarised as follows:

- Disclosure or dissemination of sensitive information (such as health conditions, disabilities, sexual orientation, location, among others).
- Use data for purposes other than initially stated (such as advertising).
- The exchange/dissemination of personal data with third parties (insurance companies, the government, including emergency bodies) [Bertino 2016].
- Of particular concern is compliance and breaches of regulations² (e.g., GDPR³, UK DPA⁴, HIPPA⁵), leading to unlawful personal data exchange during emergency response.

²These regulations establish data protection principles, responsibilities and obligations that apply to organisations that process personal data; they describe the framework under which personal data can be 'processed', providing that it is lawful to do so.

³General Data Protection Regulation (GDPR) [European Parliament 2016].

⁴UK Data Protection Act [UK Government 2007a].

⁵Health Insurance Portability and Accountability Act (HIPPA) [HHS 1996]

Therefore, finding a solution that can access healthcare data, filter out the relevant information and process it to deliver meaningful summaries while preserving citizens' privacy is imperative. Measures must be taken to balance the trade-off between utility (and its potential for secondary use) and privacy.

The following section starts with an overview of the emergency response process, the actors involved, and the beneficiaries of using EHR. Then, we focus on reviewing current approaches developed to tackle these concerns and identify opportunities and areas of exploration.

2.1.2 Intelligent systems to assist Emergency services

The success of emergency response services depends on coordinating various agents required to attend emergency events. Broadly speaking, an emergency is defined as an event that threatens severe damage to life, properties, health, or the environment in a determined place [UK Government 2013a]. The term *emergency* is defined as a local incident, part of everyday life, while the term *disaster* refers to extreme and devastating events [Al-Dahash, Thayaparan, and Kulatunga 2016].

We refer to *emergency management*⁶ as the discipline that applies science, technology, planning and management processes to handle emergencies [Wilson and Oyola-Yemaiel 2001]. Managing emergencies involves accomplishing complex monitoring, planning, response and recovery tasks. In this thesis, we focus on the *response phase* of the emergency management cycle. We use *emergency services* as a generic term to group police, health, and fire and rescue agencies [UK Government 2013a]; it includes (and is not limited to) firefighters, paramedics, ambulance services, emergency departments⁷.

Designing and building intelligent systems for emergency response involves attaining communication and data needs on the different levels of emergency management (from decision-makers at a higher level to first responders at the primary execution level) and across various organisations [Turoff and Chumer 2004]. On an operation level, planning and dealing with emergencies such as fires should be coordinated and supported with useful information about the event. Access to consistent, up-to-date and relevant data allows emergency services to assess the nature and extent of the event and represents a crucial step. Previous research by [Prasanna 2010; Nunavath, Prinz, and Comes 2016] has identified that emergency services, specifically firefighters, require information about people involved in

⁶See Appendix C.1 for a concise description of Emergency Management

⁷For a complete list of definitions see Appendix A

the emergency along with casualties. They need to familiarise themselves with the situation, and *gather information such as the type of people involved (male, female, children, disabled people), their medical condition, and special support needs*. Other data requirements include but are not limited to the type of event (fire, car accident, among others), resources available, hazard materials and emergency crew location.

To address this need for information, intelligent systems dedicated to emergency response take advantage of the Smart City paradigm to gather data from heterogeneous systems and infrastructures. These elements manage data in real-time and on-demand, including mobile devices, data repositories, sensors, Internet of Things (IoT) devices, specialised databases and other smart systems. Moreover, in a Smart City environment, this data can be swiftly accessed and dynamically updated; therefore, the latest and most accurate information can reach rapidly emergency services and emergency managers, allowing them to make life-saving decisions. In this respect, ‘response to emergencies’ was identified as one of the most relevant application domains in the context of healthcare provision in Smart Cities⁸ [Rocha et al. 2019].

There are a variety of smart systems designed to provide solutions to specific city problems. For instance, lengthy ambulance response time [Rego et al. 2018], discontinuous monitoring for potential natural disasters [Boukerche and Coutinho 2018; Lung, Buchman, and Sabou 2018], lack of information and coordination among authorities and emergency services [Turoff and Chumer 2004; Shibuya and Tanaka 2019; Chehade, Matta, Jean-Baptiste Pothin, et al. 2020] among other problems. In this review, we focus on solutions that address firefighters’ data requirements and the approaches that deal with challenges posed by using electronic health records, described in detail in the previous section.

For instance, plenty of research focuses on *providing a fast emergency response* by monitoring the health of the disabled and elderly’s using smart home solutions [Hussain et al. 2015; Loukil et al. 2017; Majumder et al. 2017]. Although these solutions gather healthcare-related data, in this case, physiological signs (e.g., heart rate, body temperature) from patients [Hussain et al. 2015], *the support to emergency response services is oriented to reduce the time it takes for them to reach the scene*, the system generates automatic alerts based on changes in vital signs, thus making data immediately available to hospitals and emergency medical services [Gharaibeh et al. 2017].

Research concerning emergency management in Smart Cities initially focused on emer-

⁸The other application domains are: population surveillance, active ageing, healthy lifestyle, support to disabled people, care services and socialisation.

gency communications in emergencies [Bartoli et al. 2015]. A great deal of research also focuses on *diagnosis and decision support systems*. In this context, information management to support emergency services embraces diverse areas. For example, plenty of research focuses on decision support systems for data integration and utilisation of provenance data [McNeill et al. 2019] and the use of semantics for the integration of heterogeneous knowledge [Shi et al. 2017], all with the crucial aim of providing enough information that supports decision-makers tasks. One hot topic for big cities is the use of hardware solutions such as smart sensors or IoT (Internet of Things) that cover larger city areas, particularly for monitoring. The aim is to develop smart solutions to predict, prevent, and react faster and efficiently to dangerous situations [Lung, Buchman, and Sabou 2018; Boukerche and Coutinho 2018]. Data used by these types of smart solutions vary from collaborative systems to crowd-sourcing information [Abu-Elkheir, Hassanein, and Oteafy 2016] to the collection of data from personal devices such as smartphones [Garcia et al. 2018] and smart devices.

From the literature, we identify that plenty of research is dedicated to developing better decision support systems and making accessible information available thanks to the Smart City ecosystem; however, the use of electronic health records proves minimal.

Crisis management is another significant area of interest in smart cities and emergency response. The object is to provide real-time data availability and support for incident control and crisis management intelligence. Smart solutions are oriented to collecting, integrating and processing all possible data from a crisis scenario to support first responders and different organisations [Palmieri et al. 2016].

However, for emergency services, the use and exchange of such information still face considerable barriers in terms of access, data requirements and interpretation, and technological adoption [Rosenfield, Harvey, and Jessa 2019]. Although the use of EHR is still limited in all these cases (crises management and decision support systems), there is a clear understanding of how much emergency response operations benefit from having access to up-to-date information.

There is also growing research that focuses on *solutions facilitating confidential access to health records during emergency events* [Thummavet and Vasupongayya 2013; Yu et al. 2017; Romanou 2018]. For instance, [Thummavet and Vasupongayya 2013] propose a framework for granting access to health records by emergency services when patients are unconscious. Their approach relies on access control procedures, and it assumes EHR owners share the exact definition of *valuable information* as emergency responders. Emergency care staff are granted access to a patient's health records according to one of the three con-

fidentiality levels (secure, restricted and exclusive). Patients are free to define what type of information is part of each of the three levels of confidentiality. In other words, there is an great risk that patients' judgement of valuable information differs from what emergency services require in order to perform their duties. Similar approaches use blockchain technology [Rajput et al. 2019] for granting granular access to EHR. Other solutions propose protocols to enable anonymous data exchange between stakeholders in cloud environments [Rahman et al. 2016].

As described, much of the available literature on emergency support concentrates on gathering real-time health-related data or developing systems that securely provide access to EHR within a health organisation. However, we can identify the difficulty in exchanging medical data outside health service providers' boundaries and emergency responders [Porter et al. 2020]. Therefore, a limited number of intelligent systems deal with electronic health records and process them to provide emergency services with much-needed data [Morales Tirado, Daga, and Motta 2022a].

Well-known methods to protect data privacy, for instance, encryption, anonymisation, pseudonymisation and other technological tools [Xiang and Cai 2021], proved to be part of the solution to protect patients' privacy. However, other methods should be devised in an emergency scenario where a person requiring assistance should be identified to retrieve accurate and precise health information.

Although many approaches and techniques are available to secure and anonymise the information, these are not infallible. In recent years, however, a different approach has aimed at minimising personal information disclosure: the **privacy-utility trade-off** [Cook et al. 2018; Morales Tirado, Daga, and Motta 2020; Morales Tirado, Daga, and Motta 2021]. The crucial strategy is to balance both issues, reusing health data under the premise of protecting privacy. To the best of our knowledge, there is an opportunity to design an approach that undertakes the problem of optimising the trade-off between sensitivity and utility while accessing health records during emergency events, hence minimising data sensitivity before it is exchanged.

2.2 Knowledge representation and reasoning with health-care data

Knowledge representation is a sub-field of artificial intelligence that deals with the problem of representing, maintaining and manipulating knowledge about an application domain

[Lakemeyer and Nebel 1994]. In this section, we focus on knowledge representation and reasoning on electronic health records, particularly how to represent the evolution of health conditions to derive information for emergency response use.

In their daily activities, healthcare professionals require knowledge of the structure and functions of the human body to diagnose and manage disorders; they also should consider the knowledge of methods and means to treat them. Furthermore, it is crucial that all this knowledge is organised for the correct management of patients' diseases and overall health [Hommersom and Lucas 2015].

Early systems dealing with healthcare knowledge were developed following researchers' representation methods and were usually oriented to solve particular problems. This close collaboration with experts gave rise to the phrase *knowledge-based systems*, employed to denote information systems that use a formal representation of domain experts' knowledge intended to approximate human reasoning to solve problems in a specific domain. However, knowledge can also be extracted from heterogeneous sources; a clear example is the smart city ecosystem where emergency services systems benefit from accessing up-to-date data of people involved in an emergency [Morales Tirado, Daga, and Motta 2020; Morales Tirado, Daga, and Motta 2021].

In what follows, we present an overview of how knowledge representation and reasoning approaches are implemented in order to use healthcare data, particularly to benefit from electronic health records rich information.

2.2.1 Reasoning on health recovery

The increasing adoption of electronic health records (EHR) has leveraged the collection of medical information at all levels of healthcare provision. Electronic health records comprise information in structured and unstructured format. Typically structured data consists of coded and time-stamped information about a patient's demographics, appointments, procedures, test results, and treatments, among other details. Unstructured data, on the other hand, include notes and free text used by healthcare professionals to record explanations of symptoms and details of treatments expressed in natural language.

Clearly, the information registered by EHR is growing considerably [Stanford Medicine 2017], generating a need for a quick and thorough analysis of patient records. This analysis has tended to focus on the knowledge representation and reasoning of unstructured data obtained from electronic health records, for instance, clinical text, notes written by doctors, nurses, paramedics and test images [Alfattni, Peek, and Nenadic 2020; Tayefi et al. 2021].

This type of data is often richer, this is, details of symptoms and decisions are taken based on this information. Of particular attention is the extraction of the temporal relations between medical events [Alfattni, Peek, and Nenadic 2020; Olex and McInnes 2021] registered as unstructured data. These efforts are oriented to provide *detailed* information about patient's medical history facilitating treatment and diagnosis [Zhou and Hripcsak 2007]. However, the literature does not explore areas such as emergency response and domains where this analysis should be performed under time, resource and data access constraints.

Electronic health records are widely used for forecasting or the detection of health situations or diseases that could happen in the future [Soyiri and Reidpath 2013]. For instance, research work is dedicated to the analysis and clinical mining of EHR data oriented to solve problems such as the detection and prediction of healthcare-associated infections, detection of adverse drug events, suicide prevention, detection of cancer symptoms and disease outbreaks [Dalianis 2018]. A clear example is the demand from governments, public services and patients to have reliable access to health information, crucially of great importance with the ongoing COVID-19 pandemic⁹. Smart healthcare platforms are becoming essential in this context, as they could take advantage of different systems managing healthcare data in order to obtain crucial information for health outbreak management [J.-H. Kim and J.-Y. Kim 2022; Verma 2022]. Intelligent systems interacting in the Smart City ecosystem can facilitate the retrieval of up-to-date information [Allam and Jones 2020; W. Webb and Toh 2020]. Health forecast and disease prediction rely heavily on machine learning techniques [Daoud et al. N.D.] and identifying features that characterise a specific disease.

As described previously, research on the representation and reasoning with healthcare data has been focused mainly on the use of unstructured data stored in EHR and the representation of temporal relations [Olex and McInnes 2021] of health events. Furthermore, there is an important amount of work on clinical data mining that relies on the use of Machine Learning (ML) and Natural Processing Language (NLP), mainly oriented to the prediction of diseases. Areas such as clinical research, long healthcare provision and prediction of diseases benefit largely from this research. However, we identify an opportunity in areas such as emergency response where information about patients' relevant health events could indicate the presence of disabilities and the need for support.

⁹COVID-19 pandemic: <https://en.wikipedia.org/wiki/COVID-19>

2.2.2 Ontologies and Semantic Web in healthcare

Semantic Web technologies and ontologies are widely used in many real-world applications [Healy, Kameas, and Poli 2010; Earley 2016]. Semantic Web technologies and ontologies have been used in healthcare data and biomedical domains since the early days. Healthcare ontologies facilitate the processing of large datasets while providing practical solutions to support healthcare management and integration of knowledge and data. It is safe to say that healthcare ontologies are pivotal for knowledge representation and data management.

Ontologies are used in different areas of the healthcare domain, for instance, to organise knowledge about medical terminology, encode health records (e.g., lab results, diagnoses), and integration and interoperability of data in order to share and reuse clinical information.

Regarding medical terminology, ontologies are used to represent large and standard medical terms. For instance, well recognised and established terminologies include: SNOMED CT [SNOMED International 2017], LOINC¹⁰ [A. Srinivasan et al. 2006], IC10¹¹ [Möller, Manuel and Sonntag, Daniel and Ernst, Patrick 2013]. Also, one of the most commonly used standards for EHR exchange and interoperability is the Fast Healthcare Interoperability Resources-HL7 (FHIR-HL7) [HL7 2019]. This family of standards describes the Resource Description Framework (RDF) representation of FHIR resources. We will explore more details about medical terminology and EHR representation standards in Section 2.4.2.

As indicated previously, EHR is used for detecting health situations or predicting diseases [Soyiri and Reidpath 2013]. Ontologies are used to model the medical knowledge available to patients and the treatment of specific medical problems, such as the care of chronically ill patients [Riaño et al. 2012] or the representation of symptoms and diagnosis of diabetes [Mekruksavanich 2016]. Data from electronic health records, particularly unstructured data, is also used to formalise ontologies that support medical reasoning for breast cancer diagnosis [Oyelade et al. 2021]. These are just a few examples of work using ontologies to represent and reason healthcare data to detect future diseases or ongoing health issues [Zahid et al. 2021]. However, during our review, we identified that most ontologies are focused on the specific detection of health issues rather than the general representation of health evolution, particularly to support emergency response data requirements as described in [Morales Tirado, Daga, and Motta 2022b].

Within the Semantic Web community, it is strongly encouraged to reuse existing ontology

¹⁰Logical Observation Identifiers Names and Codes terminology (LOINC) is a nomenclature for clinical laboratory tests.

¹¹ICD10: International Classification of Diseases and Related Health Problems.

models. Ontology guidelines and frameworks like NeOn Methodology [Suarez-Figueroa, A. Gomez-Perez, and Fernandez-Lopez 2012] propose flexible approaches for building ontologies, with particular attention on ‘ontology engineering by reuse’. Therefore in our work, we reuse ontologies in domains such as Time Ontology [Hobbs and Feng 2006] and PROV Ontology [Lebo et al. 2013] for provenance information.

The OWL-Time ontology represents temporal concepts and describes the temporal properties of resources. Time Ontology provides a vocabulary that expresses facts about instants and intervals, along with information about durations, including specific date-time information [Hobbs and Feng 2006]. The dedicated namespace for Time Ontology terms is <http://www.w3.org/2006/time#> and the suggested prefix is `time`.

The PROV Ontology (PROV-O) provides a set of classes, properties, and restrictions that represent provenance information generated by different systems. PROV-O can be used to extend or create specialised new classes or properties that adjust to specific domain requirements to represent provenance. The dedicated namespace for PROV-O terms is <http://www.w3.org/ns/prov#> and the suggested prefix is `prov`.

2.3 Knowledge graph construction from unstructured information

In recent years, KG has been widely used to represent information extracted from natural language and computer vision. Additionally, domain knowledge expressed in KG is used as input data for machine learning models to improve predictions [Chaudhri, Chittar, and Genesereth 2021]. The incorporation of human knowledge is one of the areas of interest in artificial intelligence (AI), the objective is that intelligent systems can use the knowledge represented in Knowledge Graphs (KG) to execute tasks otherwise complex.

A precise definition of KG has proved elusive [Ehrlinger and Wöß 2016; Bergman 2019; Fensel et al. 2020]. However, Hogan et al. (2021) have provided an inclusive definition where the authors conceived a knowledge graph as ‘*a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities.*’. This definition helps distinguish the two elements of a KG structure: entities and relationships. Where entities can be real-world objects or facts that describe abstract concepts, and relationships represent the connection between these objects in a specific domain. Both elements have semantic descriptions meaning they have well-defined types and properties [Ji et al. 2022]. The formal representation is

defined using a machine-readable schema or ontology. The standard language used to represent ontologies is OWL (Web Ontology Language) [J. M. Gomez-Perez et al. 2017]. There are various knowledge representation models available to store KG (for instance, Resource Description Framework Schema (RDFS), JavaScript Object Notation (JSON)). Section 2.4 is dedicated to describing RDF and OWL in more detail.

In general, knowledge graphs can be grouped into two categories: open knowledge graphs and enterprise knowledge graphs. Open knowledge graphs are accessible, and their content has been made available online for public use; a well-known KG is Wikidata¹². Enterprise knowledge graphs, on the other hand, are typically developed for commercial uses and their access is restricted; a very well-known example is Google Knowledge Graph.

Knowledge Acquisition (KA) groups the different methods and techniques to acquire knowledge, build KGs and store them using KG. Broadly speaking, KA tasks are divided into three main categories: knowledge graph completion, relation extraction and entity discovery [Ji et al. 2022]. There are also other tasks related to knowledge capture, for instance, the capture of domain experts' knowledge and automatic or semi-automatic approaches created from heuristics. Furthermore, knowledge can be captured from heterogeneous and large-scale data sources [Hogan et al. 2021].

In the following sections, we review the latest literature on knowledge graph construction, particularly in the area of information extraction from natural language sources and the techniques used to transform unstructured data into structured information ready for machine consumption.

2.3.1 Information extraction from natural language sources and supervised text classification

Knowledge extraction or information extraction refers to the analysis of natural language in order to automatically extract structured information and make it explicit for machine and systems consumption [Cunningham 2005; Grishman 2015]. In detail, information extraction tasks include identifying entities, their relationships and the inference of additional information that represents real-world elements; this information can be stored in databases and, importantly, as Knowledge Graphs to be consumed on the Web.

In order to create knowledge graphs that reflect real-world entities, some elements should be considered, for instance, the type of data sources, typically unstructured data in the form of natural language.

¹²Wikidata: <https://wikimediafoundation.org/>

Information extraction is an area of great value in domains where the volume of data, specifically textual data, grows exponentially, and it is almost impossible to process it manually [Grishman 2015]. For instance, electronic health records are a type of data source that nowadays stores information not only generated by healthcare providers but also from health tracker devices and smart monitoring applications [Stanford Medicine 2017] and benefits from swift analysis of information.

Precisely to construct a KG that contains semantically correct data and reflects real-world entities, several methods and approaches are used. In their work, Abu-Salih et al. (2022) review the latest state-of-the-art KG generation approaches and techniques in the healthcare domain. They highlight that knowledge graph construction activities vary depending on the type of knowledge graph, the data sources used to build the KG (for instance, unstructured or structured data), and the data extraction technique (for instance, entity extraction, machine learning, collaborative or manual). Importantly they identify three main tasks of knowledge acquisition [Qu 2022], these include:

- Relation extraction: the purpose is to identify relationships among different entities. It is divided into two tasks: hierarchical relationship extraction (for similar medical entities) and relation extraction of different types of medical entities (for instance, diseases and symptoms).
- Entity discovery: the purpose is to extract known entity names, places, temporal expressions, places names using existing knowledge of the domain or data extracted from other sources. It comprises three tasks (a) Named Entity Recognition (NER), (b) Named Entity Disambiguation (NED) and (c) Named Entity Linking (NEL) [Ji et al. 2022].
- Knowledge graph completion: this task is oriented to identify entities and relationships using the elements that are already part of the KG or exploiting ontological features of other KGs.

The task of extracting information from a text can be achieved by using Natural Language Processing techniques, for instance, Named Entity Recognition (NER). In what follows, we describe the application of text processing, particularly supervised text classification.

Supervised text classification is the preferred machine learning (ML) technique that is used to classify text automatically into specific categories. It relies on tasks such as data retrieval, classification and machine learning (ML) algorithms [Kadhim 2019].

In broad terms, a classification task objective is to organise text according to established categories. Automatic text classification is treated as a supervised machine learning technique. The goal of this technique is to determine whether a given document belongs to the given category or not by looking at the words or terms of that category [Kowsari et al. 2019]. It is called a ‘supervised’ task in the sense that a human should feed the classification algorithm with data already annotated with correct results.

Typically, the process of applying a supervise text classification can be structured into a series of steps: data identification, model selection, training models, evaluating results and prediction.

2.3.2 Healthcare applications

Knowledge graph (KG) construction has attracted considerable attention in recent years. In the healthcare domain, efforts have been directed to encode medical knowledge using Semantic Web technologies such as RDF and OWL; as a result, well-known medical terminology taxonomies and ontologies such as SNOMED CT or ICD10 are available. Additionally, examples of the application of knowledge graphs in the healthcare domain include IBM’s Watson Health, Ali Health’s medical think tank, and Sogou’s AI medical knowledge graph APGC.

Although these efforts have been crucial in terms of the representation and interpretability of EHR, there is still plenty of work to be done to capture knowledge that connects such information among entities [Juric et al. 2020]. For instance, there are currently no comprehensive knowledge graphs containing data on health conditions’ recovery time and SNOMED CT concepts [Morales Tirado, Daga, and Motta 2022b; Morales Tirado, Daga, and Motta 2022c].

In their work, [Abu-Salih et al. 2022] review the latest state-of-the-art KG generation approaches and techniques, along with possible applications in the healthcare domain. Their review indicates that attention has been centred on drug discovery, adverse reactions, diseases and disorders and other healthcare applications. This reflects the move from knowledge bases manually compiled to knowledge bases built using an automated process. Other examples related to medical data are knowledge graphs that reflect relations between diseases and symptoms [Mhadhbi and Akaichi 2017] or connections with other concepts such as risk factors, diseases, and test results, among others. We pay particular attention to Knowledge Graphs in healthcare related to estimating health condition evolution or close areas. In their work Rotmensch et al. (2017) analysed electronic health records to identify and cap-

ture diseases and symptoms entities; their main contribution is a KG that describes diseases, symptoms and their relationships. A specific KG representing depression disorder was developed by [Huang et al. 2017]. In particular, they produced a KG that makes use of medical data sources such as PubMed, and the Unified Medical Language System (UMLS).

Knowledge graphs offer the healthcare and medical domain the opportunities to implement technical solutions that support the analysis of large datasets and derive meaningful insights. The construction and use of knowledge graphs have expanded in the last years, and there is a growing interest in the medical domain; furthermore, there are plenty of opportunities to explore areas such as using KG in emergency support and detecting ongoing health conditions.

2.4 Semantic Web Technologies

This section aims to introduce fundamental concepts of Semantic Web technologies and describe the set of technologies used extensively in our work.

The term Semantic Web was coined by Tim Berners-Lee two decades ago and refers to the basic idea of describing the meaning of Web content in a way that can be interpreted by computers [Berners-Lee and Fischetti 1999]. Intelligent systems and knowledge-base applications can then exploit Web knowledge and execute tasks using human knowledge. For the Semantic Web to work, the first step to take is the cast such knowledge into machine-readable form. These formalisms are called *ontologies* and are based on ontology languages such as Web Ontology Language (OWL). Additionally, computers must have access to structured descriptions and collections of information as well as sets of inference rules that can support their reasoning.

In the following sections, we will review the Semantic Web technologies that comprise tools and technologies for electronic health records representation, use of ontologies and information exchange.

2.4.1 RDF/OWL

RDF stands for Resource Description Framework, a standard data model for the Web which aims to facilitate data exchange. RDF define a graph-structured data model that is flexible and features a global naming scheme that uses native Web identifiers.

RDF is a standard framework for describing resources. Resources are defined as anything that could represent data, such as digital data (e.g., files, web pages); or concrete entities (e.g.,

furniture, people).

There are three types of RDF terms:

- **Internationalised Resource Identifier (IRIs).** As mentioned previously, RDF is a standard for describing and identifying resources. To identify each resource, it is imperative to avoid ambiguity, particularly when considering natural language for the task. For instance, the resource ‘Paris’, which refers to a city, could also represent a person’s name or a street name. For this reason, RDF reuses the native Web naming scheme. The first scheme considered was *URLs (Uniform Resource Locators)*; however, URLs are used exclusively to identify the location of documents on the Web; therefore, it is not suitable when describing generic resources such as cities, and people, among others. A generalisation of URLs, in this case, *URIs (Uniform Resource Identifiers)* was proposed to serve as identifiers for generic resources. However, URIs have a limitation in terms of representation; they are limited to a subset of ASCII¹³ characters [Hogan 2020]. RDF 1.1 introduced the use of *IRIs (Internationalised Resource Identifier)* based on the use of URIs but supporting the use of special characters. Typically IRIs is composed of various parts: *the scheme, host, path and fragment*. The scheme denotes the protocol used to locate the resource (if available), the host provides the location of the server, the path refers to the file in which information about the resource is contained, and the fragment refers to something contained in the file. Figure 2.1 exemplifies the different components of the IRI [Hogan 2020].



Figure 2.1: RDF IRI typical components

- **Literals.** Literals are used to provide human-readable information (for instance, represent descriptions, names, dates, etc.) and represent concrete data values.

In the last published version of RDF, a literal consists of three parts (two optional): a lexical form (a Unicode string), a datatype IRI (which can take a value type defined by XML Schema¹⁴.) and a language tag (indicates the human language of the lexical form). Figure 2.2 displays examples of the definition of literals and their different

¹³ASCII: American Standard Code for Information Interchange <https://en.wikipedia.org/wiki/ASCII>

¹⁴XML Schema: <http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/datatypes.html>

components. Also, in RDF 1.1, all literals without an explicit datatype are assumed to have a datatype IRI `xsd:string`.

Lexical form	"Permanent"
Language tag	"Permanent"@en
Data type IRI	"0.0016"^^xsd:float

Figure 2.2: RDF Literal typical components

- **Blank nodes.** RDF supports the representation of unknown values or resources that otherwise do not have a fixed IRI using blank nodes. Blank nodes describe the existence of *a resource or some resource* but are not interpreted as global identifiers. Typically, blank nodes are represented with an underscore prefix, for instance `_:bnode1`.

RDF terms are used to identify resources, and *triples* are used to make statements about resources and to represent data. A triple consists of three linked data pieces: subject, predicate and object. A set of such triples is called an RDF graph and can be visualised as a node joined by a directed arc (see Figure 2.3).

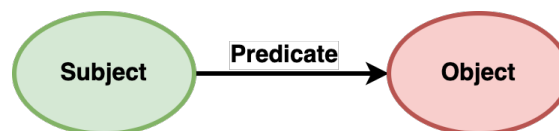


Figure 2.3: RDF triple

Each one of the components of a triple is an element of an RDF statement. The *subject* is a resource, the *predicate* is a relation, and the *object* is either another resource or a literal value. For example, Figure 2.4 illustrates the triple representing the statement (*Alba, was born, Ecuador*)

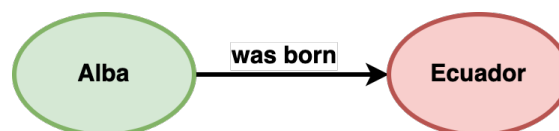


Figure 2.4: RDF triple example

Triples have some restrictions regarding the use of RDF terms to identify resources [Hogan 2020]:

- Subject can only contain IRIs or blank nodes

- Predicates can only contain IRIs
- Blank nodes can appear in the subject or object position

Listing 2.1 displays a few examples of RDF graphs as triples using N-Triple format [World Wide Web Consortium (W3C) 2014a]. In the example, we use RDF to identify a patient's gender and name (lines 1 and 2, respectively). The resource name has two components: family and given name. Lines 3 and 4, use blank nodes to link this information to the patient. For instance, a resource `_:B9c8aef` (a blank node) has as a given name (an IRI, `fhir:HumanName.given`) *Catalina* (the literal, `Catalina^^xsd:string`).

Listing 2.1: RDF graph in N-Triples syntax

```

1 <http://kmi.open.ac.uk/emergency/hr/62f11095> <http://hl7.org/fhir/Patient.gender> "F"
   ^^<http://www.w3.org/2001/XMLSchema#string> .
2 <http://kmi.open.ac.uk/emergency/hr/62f11095> <http://hl7.org/fhir/Patient.humanName>
   _:B9c8aef .
3 _:B9c8aef <http://hl7.org/fhir/HumanName.family> "Baeza"^^<http://www.w3.org/2001/
   XMLSchema#string> .
4 _:B9c8aef <http://hl7.org/fhir/HumanName.given> "Catalina"^^<http://www.w3.org/2001/
   XMLSchema#string> .

```

An easier way for representing RDF data is the use of Turtle format [World Wide Web Consortium (W3C) 2014b]. Turtle is a more convenient version of N-Triples as it makes it easier to define prefixes and reduce repeating subjects; Listing 2.2 illustrates the use of Turtle format facilitating the interpretation of the RDF triples. In this example, the first three lines define the prefixes that will be used throughout the document. Line 9, describes the patient's name including the family and given name.

Listing 2.2: RDF graph in N-Triples, Turtle syntax

```

1 @prefix fhir: <http://hl7.org/fhir/> .
2 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
3 @prefix epo: <http://kmi.open.ac.uk/emergency/EPrOnto/ontology/> .
4
5 ### http://kmi.open.ac.uk/emergency/hr/62f11095
6 <http://kmi.open.ac.uk/emergency/hr/62f11095>
7   rdf:type owl:NamedIndividual ,
8     epo:Patient ;
9   fhir:Patient.humanName [ fhir:HumanName.family "Baeza"^^xsd:string ;
10                           fhir:HumanName.given "Catalina"^^xsd:string ;
11                           fhir:HumanName.maiden "Sandoval"^^xsd:string ;
12                           ] ;

```

RDF provides limited expressive means and does not support the representation of more complex knowledge. The **Web Ontology Language (OWL)** [World Wide Web Consortium (W3C) 2012] is the standard for modelling complex knowledge about things and their relationships. OWL supports the logical reasoning of this knowledge, and its interpretation by computers, making implicit knowledge explicit. OWL is part of the W3C's Semantic Web technology stack, which includes RDF [World Wide Web Consortium (W3C) 2014a] and SPARQL [World Wide Web Consortium (W3C) 2008].

The basic building blocks of OWL are classes, properties and individuals. Classes are used to group *individuals* with something in common to refer to them. Hence, classes essentially represent sets of individuals. Listing 2.3 illustrate an example of class `HealthEvolutionStatement`, and properties `ruleSyntax` and `hasMaxDuration`.

Individuals in OWL are related by properties. There are two types of properties in OWL:

- *Object properties* (`owl:ObjectProperty`) relates individuals (instances) of two OWL classes.
- *Datatype properties* (`owl:DatatypeProperty`) relate individuals (instances) of OWL classes to literal values.

2.4.2 Representation of health records

The Electronic Health Record (EHR) is at the heart of digital healthcare. Access to large and high-quality healthcare repositories is essential for clinical and medical implications such as individual patient care, and clinical decision support, among others. EHR are equally important for non-clinical related uses, including software development, testing, clinical training, and insurance reports, to name a few [Lamprinakos et al. 2014; Leroux, Metke-Jimenez, and Lawley 2017; Bodenreider, Cornet, and Vreeman 2018].

To provide secure and reliable data, it is imperative to ensure interoperability, meaning that different systems or components can exchange and use EHR in a straightforward manner [Benson and Grieve 2021]. However, achieving interoperability has its own challenges. For example, EHR requires adequate data representation that supports understanding of clinical terms among different providers, and this includes disambiguation of health descriptions [Mello et al. 2022]. Representation should also cover means for external sharing; for instance, in the UK, delayed adoption of EHR is attributed to the development of proprietary and tailored EHR by vendors that have not considered the exchange of data among

Listing 2.3: OWL syntax example

```

1 ### http://kmi.open.ac.uk/conrad/HECON#HealthEvolutionStatement
2 hecon:HealthEvolutionStatement rdf:type owl:Class ;
3     owl:disjointUnionOf ( hecon:Decline
4                             hecon:Improvement
5                             hecon:Permanent
6                             hecon:Unaffected
7                         ) ;
8     rdfs:comment "Health Evolution Statement is an abstraction of the
9 components that describe the health recovery process"@en ;
10    rdfs:label "Health Evolution Statement (HES)"@en ;
11    skos:prefLabel "Health Evolution Statement"@en .
12
13 ### http://kmi.open.ac.uk/conrad/HECON#ruleSyntax
14 hecon:ruleSyntax rdf:type owl:DatatypeProperty ;
15     rdfs:range xsd:string ;
16     rdfs:comment "This property indicate the syntax used to build the
17 propagation rule when using a Rule execution Activity"@en ;
18     rdfs:label "rule syntax"@en ;
19     skos:prefLabel "Rule Syntax"@en .
20
21 ### http://kmi.open.ac.uk/conrad/HECON#hasMaxDuration
22 hecon:hasMaxDuration rdf:type owl:ObjectProperty ;
23     rdfs:domain hecon:Progress ;
24     rdfs:range <http://www.w3.org/2006/time#Duration> ;
25     rdfs:comment "Expresses time extend, and represents the maximum duration of
26 convalescence time."@en ;
27     rdfs:label "Has maximum duration"@en ;
28     skos:prefLabel "Has Max Duration"@en .

```

trusts [Zhang et al. 2020]. Thus, interoperability aims to enable diversity while ensuring that systems can work and understand each other.

Benson and Grieve (2021) and the Healthcare Information and Management Systems Society (HIMSS)¹⁵ [HIMSS 2022], identify layers and levels of interoperability, with different structures. Although both structures are different, they coincide in one layer, the semantic interoperability layer. The semantic layer is also considered a data layer and provides systems with a shared understanding of information. It requires the use of standardised definitions, including terminologies and vocabularies.

Standards are a technological solution to describe complex information, such as healthcare data, in order to make it accessible for different systems or organisations. Mello et al. (2022) performed an extensive review of semantic interoperability in EHR; their findings indicate that although there is no general consensus, there is a trend in health standards

¹⁵HIMSS - Healthcare Information and Management Systems Society: <https://www.himss.org/who-we-are>

choice. Regarding data exchange, openEHR¹⁶, ISO/CEN 13606¹⁷ and HL7¹⁸ (including FHIR) formats are the most used. In terms of terminologies, SNOMED CT is the most cited terminology along with ICD (International Classification of Diseases¹⁹). This trend is also reflected in the policies and requirements adopted by public organisations. For instance, the National Health Service (NHS) in the UK has adopted SNOMED CT as a standard vocabulary for recording health records²⁰ [NHS Digital services 2022b], and FHIR to improve the exchange of information²¹ [NHS Digital services 2022a].

Zhang et al. (2020) research argues that the adoption of standards cannot only accelerate the adoption of EHR, but it also aids in the advancements of achieving interoperability. For instance, it allows the creation of high-quality research datasets, eliminates the manual collection of data for support processes such as audits, and overall benefits patients by improving healthcare services. In addition to achieving interoperability, the adoption of EHR and related standards has leveraged the development of intelligent systems and ontologies for emergency support [Galton and Worboys 2011].

Taking all these considerations into account, we adopted two of the most well-known and established standards: FHIR and SNOMED CT. Crucially, we also adhere to industry trends and national requirements regarding EHR adoption, which in turn could anticipate a swift implementation of our solution in a real scenario. In what follows, we will give a detailed description of the aforementioned standards.

FHIR - standard for exchange of EHR

Information exchange is a key factor in achieving functional and useful health data communication. However, the diversity of medical data and the actors involved in the management could turn the information exchange process into a challenge. Lack of communication and incomplete or incomprehensible data could lead to unreliable medical services affecting people's lives and ultimately representing large financial costs for governments and health providers [Lamprinakos et al. 2014].

Therefore, adopting a communication standard is a key component of solving semantic interoperability among heterogeneous systems. Although there is no global consensus on the

¹⁶openEHR - <https://www.openehr.org/>

¹⁷ISO/CEN 13606 - <http://www.en13606.org/information.html>

¹⁸HL7 <http://www.hl7.org/implement/standards/>

¹⁹International Classification of Diseases - ICD: <https://www.who.int/standards/classifications/classification-of-diseases>

²⁰SNOMED CT UK: <https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct>

²¹FHIR UK Core: <https://digital.nhs.uk/services/fhir-uk-core>

use of a unique standard that covers all aspects of health communication, HL7 CDA and the Fast Healthcare Interoperability Resources (FHIR) standards are among the most used and well-known formats [Mello et al. 2022]. For its flexibility and adaptability to stakeholders' requirements, FHIR is used in a large number of domains, such as handling electronic health records (EHR), cloud communications, clinical decision support systems, data analytics, mobile health applications and wearable devices.

In 2011, the Australian Health Level Seven (HL7) organisation published a standard for interoperability called Resources for Healthcare (RFH). This new standard was designed for web technology and based on extensible markup language (XML). Later RFH standard was renamed Fast Health Interoperability Resources (FHIR) and extended previous HL7 standard specifications (HL7 version 2 and version 3) [Ayaz et al. 2021].

FHIR's scope is broad; it covers human and veterinary, administration and financial aspects, clinical care and trials, and public health. The FHIR specification takes a modular approach and defines two components: a representation of healthcare data named 'resources' and a REST API for manipulation of resources (create, read, update, delete, among others²²) [Semenov et al. 2019]. FHIR uses resources as a generic term to refer to common healthcare concepts such as patient, observation, medication, device, and condition, among others; it is the smaller possible unit of a transaction. Each resource is an entity, and it is described in a way that provides meaningful data and supports medical data exchange. FHIR defines more than 150 different types of resources to date (according to version R4 and continues to grow gradually) and uses them to access and perform operations on patient data [HL7 2019].

The latest version of FHIR categorises resources into five groups: foundation (e.g., Provenance, Composition, etc.), base (Patient, Organization, Encounter, etc.), clinical (e.g., Observation, Medication, CarePlan, Communication, etc.), financial (e.g., Coverage, Claim, Account, etc.) and specialised (e.g., ResearchStudy, Measure, Evidence, etc.).

FHIR standard is flexible and adaptable to users data requirements [HL7 2019; Lehne, Luijten, and Thun 2019; Saripalle, Runyan, and Russell 2019]. The specification of a business requirement can be described using a set of resources; for instance, Figure 2.5 illustrates the representation of an electronic health record using FHIR resources Patient, Encounter, Condition, CarePlan, Immunisation, and Medication. Different business requirements could add other resources, such as information about insurance.

HL7 and SNOMED International collaborate closely to ensure that it is clear how to use SNOMED CT in FHIR standards. Both organisations manage the documentation web

²²RESTful API - <http://hl7.org/fhir/http.html>

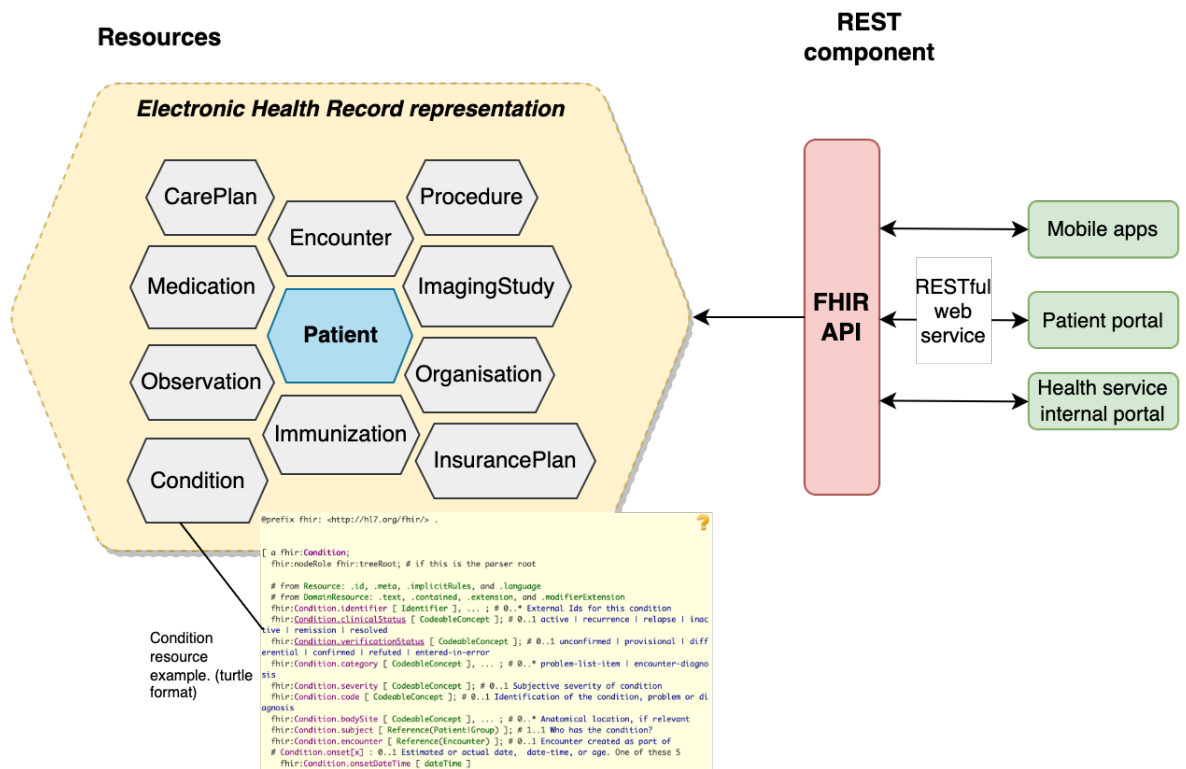


Figure 2.5: Using FHIR resources to represent an Electronic Health Record

page ‘Using SNOMED CT with FHIR’²³ that describes how to identify SNOMED CT code system and identify health conditions using both standards.

SNOMED CT - standard clinical terminology

Historically, the formalisation of clinical terminology has been seen as an issue in the medical domain. This happens for several reasons; for instance, clinical terms often can have one or more concepts used to refer to the same term (synonyms). There are cases where one term can be used to express different concepts depending on the context (homonyms). And there are always terms using acronyms, abbreviations or named after people (eponyms), so the meaning of the concept cannot be directly inferred [Benson and Grieve 2021].

The use of electronic health records (EHR) leverages communication among health service providers and facilitates its storage. An agreed use of clinical terminology aids practitioners in communicating and exchanging medical data plainly, reducing time and effort while ensuring patient safety [SNOMED International 2022d]. Therefore, standardised representation of clinical terminology is an important aspect of interoperability. We can summarise the benefits of having a standardised vocabulary as follows:

- It enables a consistent, shared and understandable representation of clinical and health

²³Using SNOMED CT with FHIR: <http://hl7.org/fhir/snomedct.html>

information.

- It enables other entities, such as decision support systems, to perform data analysis or monitoring on an individual level (for instance, people's current health status) or a community level (for instance, emerging health issues of the population).
- It supports the exchange of appropriate information among organisations delivering healthcare services.
- It supports targeted access to relevant information; for instance, systems can extract information on the latest health issues or related to a particular disease.

The systematic review of semantic interoperability in health records standards reported by Mello et al. (2022) identifies SNOMED CT [SNOMED International 2022c] as the most used terminology. It is worth mentioning other schemes often cited, for instance, ICD-10 (International Classification of Diseases)[World Health Organization 2022] and LOINC (Logical Observation Identifiers, Names, and Codes) [Bodenreider, Cornet, and Vreeman 2018], they focus on terminology for reporting and monitoring diseases, and clinical measurement terms, respectively.

As mentioned previously, SNOMED CT, the Systematised Nomenclature of Medicine Clinical Terms, is the preferred clinical terminology scheme for use in electronic health records [Chang and Mostafa 2021; Mello et al. 2022]. Furthermore, it constitutes an extensive collection of clinical content that has been scientifically validated [SNOMED International 2022c], and it consists of 3535,567 active concepts (as of January 2020 release).

Since 2007, the International Health Terminology Standards Development Organization (IHTSDO, a not-for-profit organisation) has owned and managed SNOMED CT. Since its inception, SNOMED CT has expanded to 40 member countries and has been translated into 11 languages. In what follows, we will describe SNOMED CT structure and its characteristics while providing examples of use in the context of EHR.

SNOMED CT concepts are organised in a hierarchical structure. In this structure, a concept may have *parents* (immediate supertypes) and *ancestors*, as well as *children* (immediate subtypes) and descendants. General concepts are at the top level of the hierarchy, and these represent the main branches of the hierarchy. The latest SNOMED CT release includes 19 Top Level Concepts (see Figure 2.6)

SNOMED CT content is described using three types of components: Concepts, Descriptions and Relationships. Figure 2.7 provides an illustrative example.

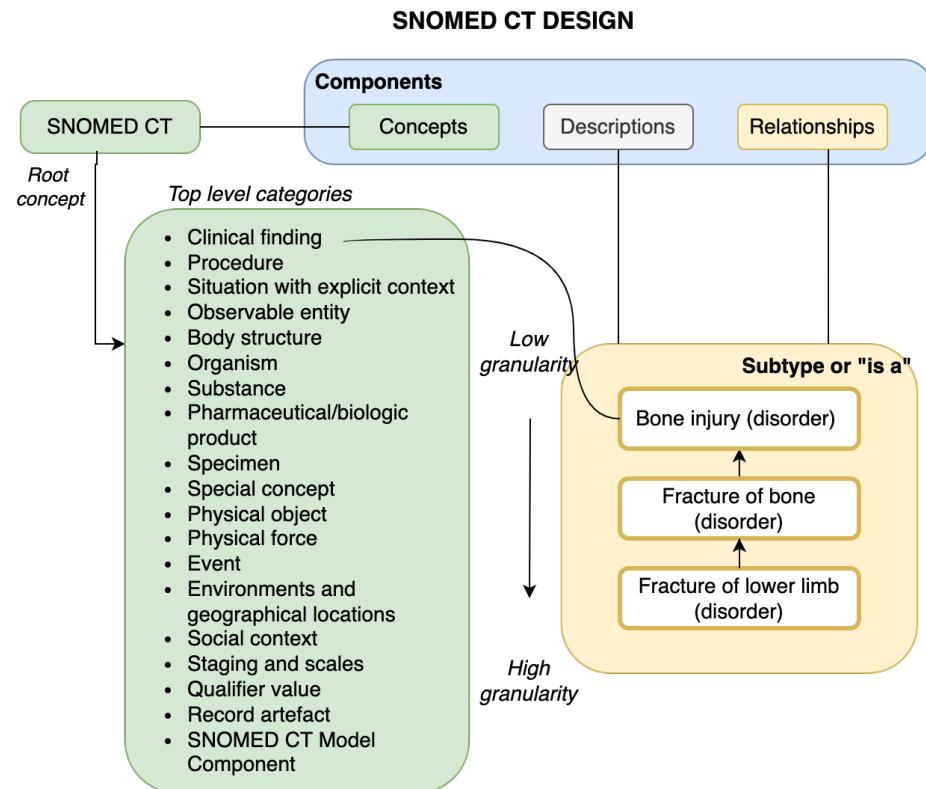


Figure 2.6: SNOMED CT structure and top-level concepts

- **Concepts:** represents a unique clinical term with a unique numeric concept identifier (conceptId). For example, a ‘Fracture of lower limb (disorder)’ is a concept with a unique identifier ‘46866001’; see example in Figure 2.7.
- **Descriptions:** represents the link or association between a human-readable phrase (term) and a particular SNOMED CT concept. A concept can be associated with several descriptions, often synonyms. Descriptions also link terms in other languages.
- **Relationships:** represents the link between two concepts. There are two types of relationships: the subtype or ‘is a’ relationship and the ‘attribute’ relationship. We will give a more detailed explanation later.

In order for a machine to understand its structure, SNOMED CT uses two types of relationships. The most used is the **‘subtype’** or also called ‘is a’ relationship. It is a directional relationship that indicates that a source concept is a part or a subtype of another concept. All concepts have at least one ‘is a’ relationship; the only exception is the root concept ‘SNOMED CT concept’, which is the top general concept. The specificity of the tree increases with the depth of the hierarchies, but because a concept can be linked to two or more other concepts, it can be called a ‘polyhierarchy’ [SNOMED International 2022c].

The **attribute** relationship is used to associate characteristics that distinguish a concept

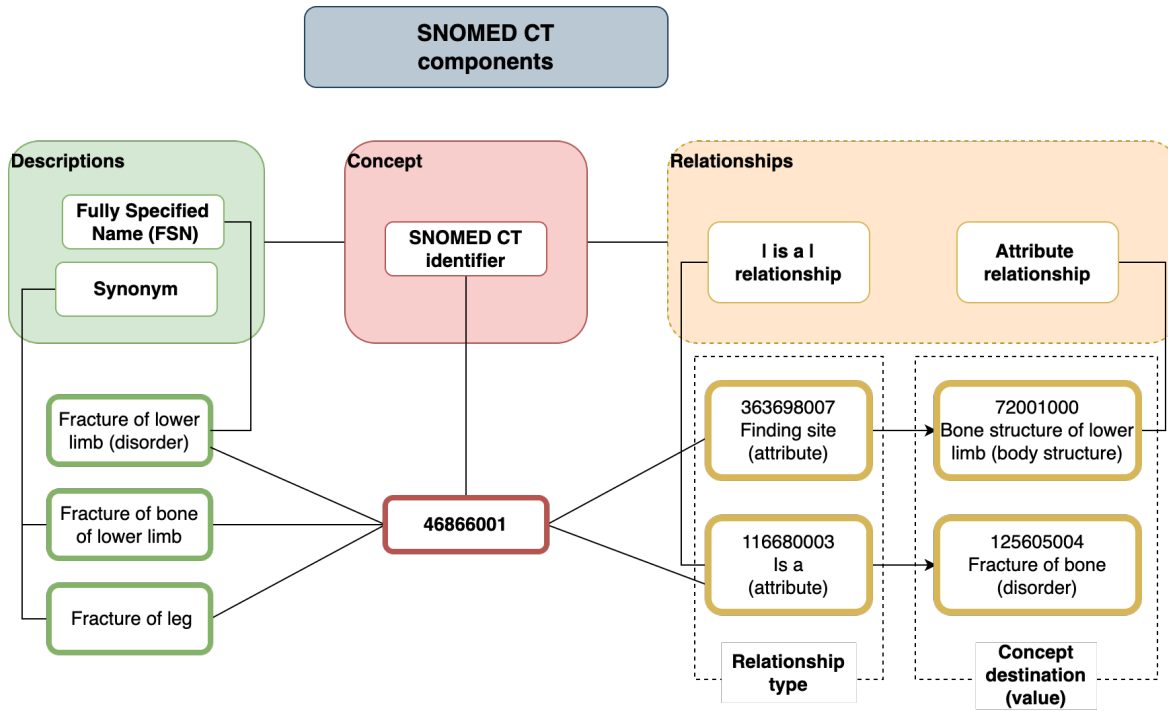


Figure 2.7: SNOMED CT components explained

from others. This relationship has two elements: a ‘relationship type’ and its ‘value’. For example, following the example in Figure 2.7 and the definition stated by SNOMED CT, we find that a ‘Fracture of lower limb (disorder)’ has as a type of attribute ‘Finding site’, and its value is ‘Bone structure of lower limb’.

SNOMED CT also provides the Expression Constraint Language (ECL), which supports the construction of computer-readable expressions for searching a specific set of concepts using SNOMED CT concepts’ structure and attributes. It is important to highlight that ECL does not support querying over EHR content (for instance, retrieving a specific resource from a database); but it is possible to build ECL queries using the SNOMED CT expressions and query the SNOMED CT hierarchy.

The Expression Constraint Language (ECL) is not a formal computer query language but resembles it as is possible to build a set of rules (queries) that can be executed over SNOMED CT hierarchy [SNOMED International 2022b] and return a concept or set of concepts that meet the specified rules. ECL queries could contain one or more focus concepts joined by either a conjunction, disjunction or exclusion. ECL also supports the refinement specification, meaning that a query can include specific constraints related to relationships, attributes and values. Figure 2.8 is a simplification of the ECL logical model; in this example the query is designed to search all subtypes of the given concept ‘*Disorder of pregnancy*’ or ‘*Ectopic pregnancy*’ (the focus concept part of the query) and filtering only the concepts that have the attribute ‘*Interprets*’ and value ‘*Pregnancy observable*’ (refinement part of the query). For

instance, an application of ECL queries is related to the exploration of SNOMED CT content in order to search and retrieve expressions to be displayed in software applications and make it available to final users inputting EHR details.

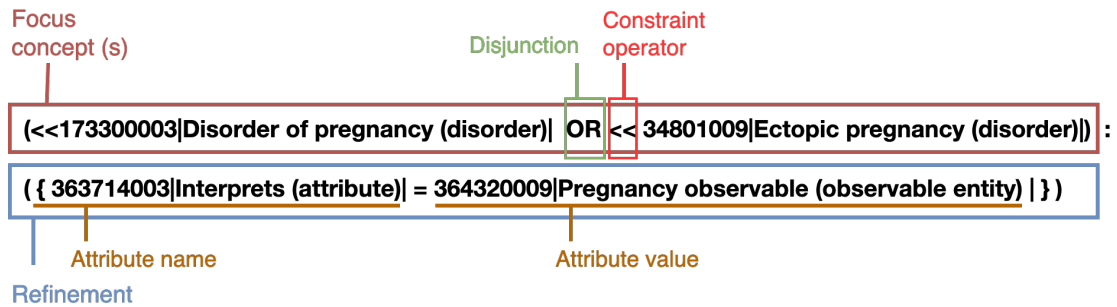


Figure 2.8: SNOMED CT Expression Constraint Language example

SNOMED CT has plenty of documentation regarding ECL specification, including a thorough description of the syntax, the description of the logical model used to search the hierarchy structure, and plenty of examples [SNOMED International 2022b]. Additionally, the Australian e-Health Research Centre of the CSIRO²⁴ has made available a web-based terminology browser ‘Shrimp’²⁵ [Metke-Jimenez et al. 2018] which can be used to explore SNOMED CT taxonomy and build assisted ECL queries. SNOMED CT Browser²⁶ also provides a web service to explore its taxonomy; however, the tools to build ECL queries are still under development.

SNOMED CT was released in RF2²⁷ format for years; more recently, with the advancement of Semantic Web Technologies, SNOMED CT provides support for OWL²⁸ format. The OWL reference sets are used to distribute ontology reference information and axioms representing the formal logical definitions of SNOMED CT concepts. The prefix name ‘sct:’ is for SNOMED CT concept identifiers, and the namespace URI is <http://snomed.info/id/> [SNOMED International 2022a]. For instance, the URI for a SNOMED CT concept such as ‘Fracture of lower limb (disorder)’ can be one of the following:

- <http://snomed.info/id/46866001>
- `sct:46866001`

²⁴<https://aehrc.csiro.au/about/>

²⁵<https://ontoserver.csiro.au/shrimp/ecl/>

²⁶SNOMED CT Browser: <https://browser.ihtsdotools.org/>

²⁷Release Format 2 (RF2) is the current release format developed by the International Health Terminology Standards Development Organisation (IHTSDO) to support the creation of reference sets.

²⁸Web Ontology Language (OWL): https://en.wikipedia.org/wiki/Web_Ontology_Language

Finally, SNOMED CT can be used as a consistent vocabulary for recording patients' clinical information. Appropriate software and information technologies support the use of this controlled terminology, in this way clinicians record patients' clinical events and stored them as electronic health records.

Synthetic electronic health record dataset

A large and growing body of literature has investigated the area of synthetic healthcare data generation [Walonoski et al. 2018; Imtiaz et al. 2021; Bhanot et al. 2022] and its objective is the generation of data that fits a real dataset while excluding any actual information. Domains such as healthcare, biometrics and energy consumption benefit from accessing useful synthetic data while ensuring reasonable privacy protections.

As described in Section 2.1 access to electronic health records (EHR) data can improve the provision of emergency services [Zorab, Robinson, and Endacott 2015; Patterson et al. 2019; Morales Tirado, Daga, and Motta 2021], medication dispensing [VanLangen and Wellman 2018] among other areas. However, access to it is often restricted by data protection laws and regulations [Centers for Disease Control and Prevention 1996; UK Government 2007a; European Parliament 2016], limiting research activities such as reproducibility and reuse of data.

Several privacy-preserving techniques have been used in an effort to overcome privacy concerns. For example, obfuscation, masking, and cryptography, to name a few.

- Data obfuscation techniques use data sets to lower individual item data accuracy. It is a process that hides original data by modifying its content in a systematic, controlled and statistical way [Bakken et al. 2004].
- Data masking methods hide sensitive data, usually to test information with software and database developments. The objective is to identify sensitive data and alter it using different methods.
- Data cryptography uses different mathematical techniques to transform information into a not understandable piece of information, interpretable only by those who can restore the data to its original form using a key.

However, there is still a risk of data loss, breaches or information leaks, particularly when large amounts of data are shared with a third party. For instance, third-party organisations could potentially combine data collected by them with the information provided by healthcare organisations and identify people, for example, the use of health data by Google

in partnership with the University of Chicago Medicine. Both organisations teamed up on a project that used Machine Learning to improve the prediction of healthcare. The university shared vast amounts of health data as part of the project. Although it was de-identified (a HIPAA²⁹ law requirement HHS 1996), it still contained dates and timestamps. Concerns about how Google can potentially use this information in combination with people's visited venues and locations led to both facing a lawsuit in 2019 [The New York Times 2019]. This is a clear example of the difficulties when accessing and using healthcare in the research domain.

In this context, synthetic healthcare data generation is a viable alternative to real healthcare data. According to Georges-Filteau and Cirillo (2020), methods to generate synthetic data can be classified as either theory-driven [Walonoski et al. 2018] (theoretical, mechanistic or iconic) or data-driven (empirical or interpolatory) modelling [Imtiaz et al. 2021]. On the one hand, theory-driven approaches generate data based on models of clinical workflow and disease progression; these models rely heavily on knowledge collected from domain experts. On the other hand, data-driven approaches infer data representation from a sample distribution [Georges-Filteau and Cirillo 2020].

At the time of writing, several data-driven approaches are gaining momentum, particularly the ones using Generative Adversarial Networks (GAN), a specific type of Deep Learning model [Hernandez et al. 2022]. Georges-Filteau and Cirillo (2020) reviewed literature in the area and collected a list of open-source tools that use empirical methods for synthetic healthcare data generation, for instance, MedBGAN³⁰ [Baowaly et al. 2019].

However, early synthetic data generation tools were proprietary primarily (e.g., Patient-Gen³¹), documentation was scarce, the number of EHR generated was minimal, and data quality assessment was limited to statistical comparison, which in turn was preventing repeatability and data reusability [Walonoski et al. 2018]. In contrast, Synthea was made available as open-source software and supported by well-structured documentation. As Synthea is a theory-driven solution, it relies heavily on disease evolution modelling made by knowledge experts, so there is a risk of simplification and assumptions during the modelling task. Although not a limitation, Synthea generates data based on US statistics by default; specific format and data statistics are required if there is a need to create records for a given location or different country.

²⁹HIPAA: Health Insurance Portability and Accountability Act of 1996

³⁰MedBGAN: <https://github.com/baowaly/SynthEHR>

³¹Michigan Health Information Network Shared Services (MiHIN) <https://mihin.org/services/patientgen/>

In our research, we use Synthea³² [Walonoski et al. 2018] to model the medical history of synthetic patients. . Synthea uses as data input publicly available information and health statistics; it then generates complete health records (a patient’s lifetime record) instead of focusing on a specific health disease. Each patient’s health record is generated independently and simulates the health registers from birth to death through modular representations of various diseases. The synthetic electronic health records generated are deep and extensive, including demographic data, appointments, patient conditions, procedures, care plans, medication, allergies, and observations. Table 2.1 lists all the information generated by Synthea.

The synthetic dataset of health records uses SNOMED CT, standard terminology for clinical content in electronic health records. As healthcare records are increasingly digitised, FHIR (Fast Healthcare Interoperability Resources) [HL7 2019] is the standard specification adopted to represent the health records dataset.

Table 2.1: List of data generated by Synthea

Information generated by Synthea software	Description
Demographic data of the patient	Name, surname, place of birth, passport number, ethnicity, race, gender, and complete address.
Appointments/encounters in the patient record	Encounters: Each encounter has a code and description of the encounter A description of all the conditions and the diagnoses found during an Encounter.
Details of the patient condition/s	The conditions are recorded over time with a start and finish date Ongoing conditions will not have a finish date
Procedures	Details about any action taken related to a condition or any surgery are called procedures
Care plans	Care plans detail. The support needed, type of care and the reason it has been given
Medication	Medication details, the medication and the reason it is prescribed
Allergies	Description of the temporary and ongoing allergies
Observations	Observations or tests taken and the results
Details about the providers	Hospitals and doctors details Details about the insurance companies

³²Synthea by Mitre. <https://github.com/synthetichealth/synthea/wiki/Basic-Setup-and-Running>

Chapter 3

Requirements' analysis

In this chapter, we expand on the scenario that motivated our research on the use of health records to support the information needs of emergency services (Section 3.1). To complete our vision of how organisations manage personal data when planning for a fire emergency, in Section 3.2 we analyse a real scenario and identify actors involved in the handling of such data (e.g., data managers and providers). The analysis results are the basis for formalising the requirements that our proposed intelligent system should consider to support the use of health records by emergency services. The analysis is also a pivotal step to situating our solution in the context of a Smart City ecosystem (Section 3.3) and provides the basis for answering our fourth research question about developing an intelligent system capable of identifying people requiring assistance during an emergency.

3.1 Expanded Motivating scenario

In Chapter 1 we described the importance of using electronic health records in emergency settings in different scenarios. In this section, we elaborate on the first scenario, that of a fire emergency, capitalising on the literature on Smart City applications [R. Srinivasan, Mohan, and P. Srinivasan 2016; Palmieri et al. 2016]. This work was dedicated to analysing how Smart City systems manage large data sources, particularly EHR containing personal data, and understanding how they deal with privacy issues when this information is exchanged and used to attend emergencies.

We consider a fire event in a large organisation. The building has in place an Access Control System (ACS), which registers employees' access to the premises using identity cards. All employees use their access cards to enter the building, and visitors must register as they enter or leave the premises. As stated in the organisation's procedures, all employees

should inform the Health and Safety Department (HSD) if they have a long-term condition or a temporary disability. Following this notification, the HSD must ensure that each employee has an emergency evacuation plan (PEEP) tailored to their needs. The HSD follows governmental guidelines and internal regulations to record the evacuation plan.

A fire starts on the fourth floor of a building, it is spreading quickly, and emergency services are alerted. The HSD may be able to identify the number of people currently on-premises and their identities by accessing the ACS. Furthermore, the HSD may be able to identify people with special needs by retrieving the PEEP records. However, a number of issues reduce the effectiveness of this approach. In the absence of digital infrastructure, PEEP files may be impossible to retrieve efficiently. But also, in the case of a database, the completeness and accuracy of the data are questionable. Compiling the PEEP requires the sharing of health information that could be considered very sensitive by employees. Many people may not want to share this type of information with the line manager or the colleagues appointed as fire wardens. For example, anxiety or other mental health conditions can be typically hard to disclose. In addition, the information included in the PEEP may be outdated. Crucially, visitors may not be included in the records. Having precise information about vulnerable people could help emergency services react promptly and take the right decisions when planning resources.

In this context, accessing the health records of the National Health provider by a Smart City system constitutes a substantial opportunity to retrieve up-to-date information and accurately recognise people requiring support. However, obtaining such an amount of fine-grained and specialised data could be overwhelming for firefighters and fire wardens because:

- Healthcare data is highly specialised and may be difficult to interpret by the personnel involved in supporting the evacuation (e.g., firefighters).
- A large amount of data makes it difficult to find relevant information.
- Exchange of sensitive information might put citizens' privacy at risk.

Therefore, finding a solution that can access healthcare data, identifies the relevant information, and processes it to deliver only meaningful data while preserving citizens' privacy is imperative. In principle, an Intelligent System could act as a mediator between the healthcare data provider and the emergency services to balance the trade-off between utility and sensitivity.

This review helped us to have a general view of the overall data management process, the information that could support emergency services activities and the different issues that arise during the exchange of personal information. We selected a paradigmatic scenario of how Smart City systems operate in case of emergencies [R. Srinivasan, Mohan, and P. Srinivasan 2016]. The purpose was to state clearly the issues that Smart City systems and applications are facing in the context of emergency events. By the end of this Chapter, we aim to establish the evaluation scenario of our proposed solution, identify the actors and illustrate the environment in which our solution could be most useful.

3.2 Survey on data management practices for emergency - The OU case

We took as an example a large organisation to learn how privacy and security issues are handled during emergencies. The organisation considered was The Open University, as we had unrestricted access to the information elaborated by the Health and Safety Department. We chose to focus on The Open University fire procedures because it provides a significant amount of well-structured documentation (see Appendix D.1) about emergencies. Furthermore, the Health and Safety regulations of The Open University comply with Statutory Instruments and British Standards. Therefore, it constitutes a good, paradigmatic, and concrete case study that helps us understand how organisations collect and use personal data when planning for emergency events. Consequently, we expect to extract comprehensive information that can be generalised to other emergencies.

The standard procedure indicates that everybody should familiarise themselves with the organisation's fire safety guidance. The guidance includes the definition of roles (specifically the Health and Safety Department, Fire Warden, Appointed Helper, and Employee) and their responsibilities. Also, all employees should understand how to perform a standard evacuation plan, which basically states how to follow the evacuation signs to reach the nearest exit and gather at the meeting point.

Besides, if an employee is not capable of reaching the nearest exit, then a Personal Emergency Evacuation Plan (PEEP) must be elaborated. The PEEP constitutes an excellent case in which personal information is stored and used to handle emergencies. As stated in the organisation's procedures, all employees must notify their line manager in case they have a permanent or temporary disability. If they do not report any disability, there is no need to carry out additional actions. If a disability is reported, then both parties should discuss and

identify if they need to elaborate a plan. The PEEP form (see Appendix D.2) follows the formats established by the organisation, and it is intended to give a detailed explanation of the steps/tasks (e.g. reaching an evacuation area, opening a door, using of evacuation lift, etc.) that people should follow to leave the premises in case of evacuation.

Generally, a person designated by the HSD interviews the employee and evaluates his/her capacity to perform the plan. Typically, factors to take into account and negotiate a suitable evacuation plan are (a) type of disability, (b) the employee's capacity to perform a plan, and (c) the means of escape available in the building. Once identified the type of assistance required, the following action is to register a tailored Personal Emergency Evacuation Plan (PEEP). For the elaboration of a plan, the UK governmental guidelines [UK Government 2008] provide a comprehensive list of disabilities and recommended options for escape as well as essential guidance for assessing and arranging the appropriate means of evacuation for the employee.

The PEEP form collects employee information (name, telephone number, email address, staff number, unit/department, description of disability or disabilities, any special aids), as well as detailed information on their expected actions during the evacuation. The escape plan must be tailored to the employee's needs. When a person requires personal assistance in evacuating safely, appropriate support should be provided, and all details must be contained within the PEEP. Appropriate support refers to any specialist equipment like flashing lights, evacuation lifts or chairs, or braille evacuation route signs, to mention a few. If an employee is using specialist equipment, it has to be described in the PEEP.

For those with mobility problems, a simple buddy system may suffice where a nominated colleague will assist the disabled person in the event of an evacuation; such appointed helpers must not be Fire Wardens. The details (name, location, and telephone number) of the Appointed Helpers and Fire Wardens must be stated in the PEEP, together with any evacuation equipment they have to use. Any issue with the form or its content should be directly discussed with the fire wardens or the Health and Safety Department.

Once the PEEP form is completed, it must be communicated to the relevant parties: fire wardens, all appointed helpers, and the relevant staff at the Health and Safety Department. The PEEP should be updated regularly, and it is stored by the Health and Safety Department. A generic PEEP should be devised and used where there are visitors or casual users of the building who may be present infrequently or on only one occasion. In general, staff data will be kept for six years after their affiliation ends, although some health and safety and occupational health data will be kept for 40 years as these may have a long-term liability

(according to the Staff, workers, and applicants - Privacy notice). Figure 3.1 summarises the flow of information according to the Health and Safety regulations at The Open University in the case of a fire event.

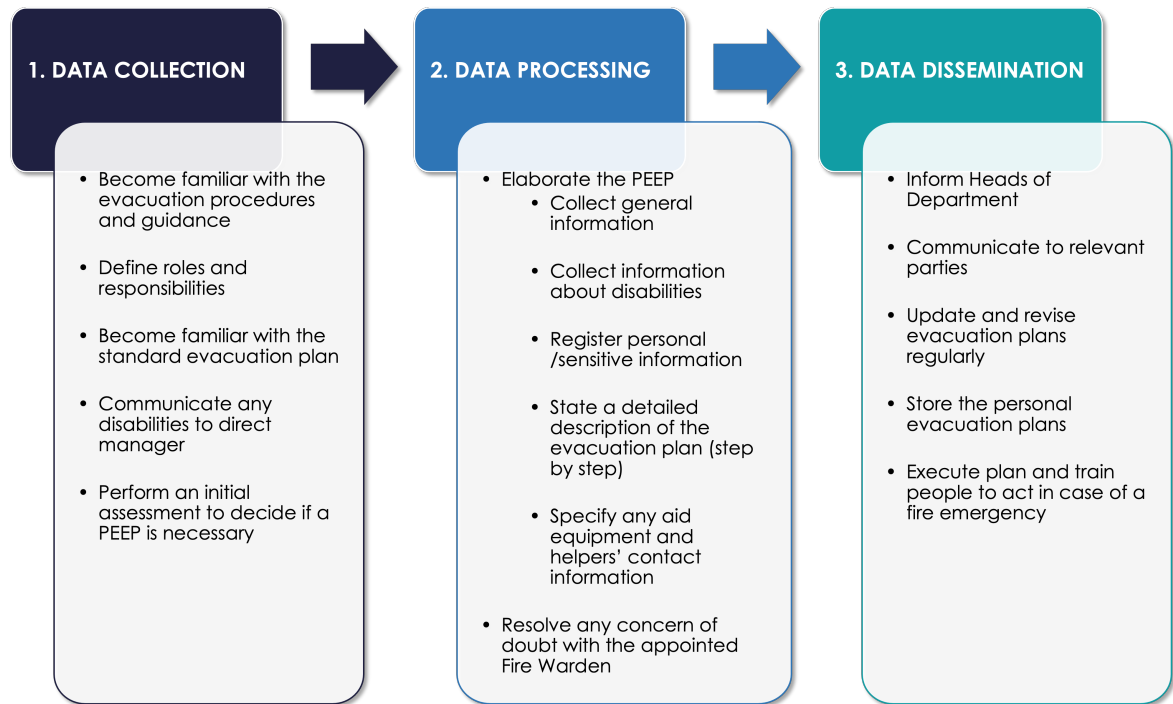


Figure 3.1: Summary of Health and Safety data flow for fire planning

In what follows, we summarise the key issues identified during our analysis of personal data management for emergency planning.

- The PEEP form registers personal and sensitive data, such as employee health-related information. For instance, it records the description of disabilities.
- Besides registering information about an employee, the PEEP form also collects personal information about any appointed helper or helpers.
- Personal and sensitive information is collected in paper forms.
- Once the PEEP form is complete, it must be communicated to the relevant parties: fire wardens, all appointed helpers, and Health and Safety Department, and must also be made available to whoever is involved in attending the emergency.
- The PEEP should be updated regularly, and it is stored by the Health and Safety Department. This happens independently from an actual incident in which the information may be used.

- Health information is considered sensitive data, and often employees with a 'hidden disability' (for example, depression, anxiety, epilepsy, asthma, heart problems, or other mental conditions) could choose not to disclose their status as the PEEP is stored and communicated to a number of people.
- In case of a fire event, the information contained in the PEEP form represents valuable information for emergency responders, although it might not be easily accessible.

In the next section, we give a summary of the main findings of our scenario-based analysis. In particular, we list all the actors involved in the data management process, we detail the requirements an intelligent system should address in order to support the emergency teams during a fire emergency, and we also illustrate its role in the context of a broader Smart City environment.

3.3 Requirements elicitation

In the previous sections, we carried out a scenario-based analysis of data management during fire emergencies. Organisations aim to identify people requiring assistance during an emergency. Therefore, they collect and use health-related data that indicate if a person has a disability or any issue that prevents them from performing an evacuation during a fire event. The organisations' main data requirement is to have access to health-related information that allows them to carry out this assessment. Moreover, emergency services require swift access to this assessment since information about vulnerable people could facilitate planning and rescue operations. However, gaining access to this specific data is not a straightforward process. There are recurrent issues that arise when managing health-related information.

Although medical information enables the identification of people in a vulnerable situation, processing and interpreting detailed and large data sources could be a time-consuming task, and crucially helpful information might be overlooked. Moreover, swift access to physical records may be challenging during a fire event.

In addition, as data is collected a priori, there is a risk of having inaccurate information when an emergency occurs. For example, in large organisations, HSD may take some time to record and elaborate a plan for an employee with a recent fracture of the ankle.

Finally but not least important, health information is sensitive data; therefore, its collection and exchange might put citizens' privacy at risk. It is possible that employees might prefer not to disclose medical data knowing that it could be stored and communicated in var-

ious instances. For instance, often, employees with a ‘hidden disability’¹ (e.g., depression, anxiety, epilepsy, asthma, heart problems, or other mental conditions) might feel uncomfortable about sharing this information.

We summarise the key requirements an intelligent system should meet in order to address the aforementioned issues:

- Requirement 1 - Provide access to up-to-date information about current² health issues.
- Requirement 2 - Support swift identification of useful data.
- Requirement 3 - Minimise the disclosure of personal information.

In order to meet these general requirements, we envision a novel method for handling personal data in Smart Cities environments. We aim to design an Intelligent System (IS) capable of (a) identifying useful data for assisting emergency services and able to (b) minimising the exchange of sensitive information appropriately. The Smart City environment should facilitate the data exchange between the different entities in a fire emergency scenario.

In what follows, we expand on these requirements by detailing specific tasks and identifying knowledge components that should be part of an IS that aims to address the main requirements. As a core component of our approach, we propose to make use of Semantic Web technologies. Crucially, by describing the different components of the IS, we will start outlining answers to the research questions formulated in Chapter 1.

Requirement 1 - Provide access to up-to-date information about current disabilities and health issues. On the one hand, organisations use information about disabilities to identify people that require special assistance or to make arrangements to perform an evacuation plan. On the other hand, emergency services benefit from knowing in advance about people in vulnerable situations. Clearly, medical information is a key element in both cases. Access to health records can potentially reveal the latest health issues and picture the ongoing health status of a person, facilitating organisations’ data management during a fire emergency.

Electronic health records (EHR) have become more accessible thanks to the adoption of information technology facilitating data exchange and processing of medical information. EHR are the ultimate source of the latest medical records. In order to provide up-to-date information, we envision an intelligent system capable of:

¹Disabilities: <https://www.gov.uk/definition-of-disability-under-equality-act-2010>

²We use the term ‘**current**’ to refer to a health condition affecting the citizen during the emergency.

- (a) Accessing electronic health records. In the Smart City context, the acquisition of medical information is viable thanks to the iteration with the national health provider. Although our contributions do not cover technical aspects of data exchange between organisations, for evaluation purposes, we assume this access is granted (see Chapter 1, Assumptions).
- (b) Processing electronic health records by taking advantage of established standards for the exchange and representation of medical information, particularly the use of FHIR³ and SNOMED CT⁴.

Requirement 2 - Support swift identification of useful data. Having access to electronic health records is not enough to identify vulnerable people; fire wardens and emergency services still have to analyse historical records and promptly detect ongoing issues. Ideally, an intelligent system should be able to automatically analyse if a medical event is ongoing and how it evolves over time. A formal representation of 'health evolution' is needed to guide our system and support the evaluation of current health issues. Therefore, the proposed intelligent system should incorporate:

- (a) A knowledge component, specifically an ontology that formally represents 'health evolution' and supports the reasoning on ongoing health events.
- (b) A Knowledge Graph (KG) that stores information regarding health evolution.
- (c) A Health Event Evolution Reasoner component that contains the rules for reasoning on health evolution and evaluates if a health condition is ongoing at a certain point in time. It interacts with the previous two components to achieve its objectives.

Requirement 3 - Minimise the disclosure of personal information. As described in the proposed scenarios, health-related data is considered sensitive information. Clearly, emergency services collect and use this information with the objective of saving lives; however, its management is not exempt from potential, sometimes unintentional, data disclosure. Organisations should comply with data regulations, which often impede collaboration and exchange of valuable information with third-party organisations. In a Smart City environment, we envision an interaction among three main actors:

³FHIR: Fast Healthcare Interoperability Resources

⁴SNOMED CT: the Systematised Nomenclature of Medicine Clinical Terms

- Data subject: the employees/the person involved in the fire event. The individual whom the particular data is about [European Parliament 2016].
- Data controller: in this case, the health provider and the university both decide on the rationale and the need to collect and use the data [European Parliament 2016].
- Data processor: in this case, the emergency service body that is given data by a third party (for instance, the health provider or the university sends information to firefighters). [European Parliament 2016].

Data controllers shoulder the highest level of compliance responsibility and can be held liable for non-compliance according to the GDPR legislation [Information Technology Industry Council 2018]. Therefore, we envision a solution that facilitates data controllers' decisions to exchange health information. Ideally, the intelligent system can act as a service activated by a data manager, which processes the relevant information and produces an output addressing the requirements of the emergency services. The organisation where the fire event develops should share the identification of the people involved in the event by retrieving information from its Access Control System (ACS).

Sharing extensive and detailed data, such as electronic health records, clearly is not desirable for the patients nor for the health providers. Therefore, an intelligent system interacting in a Smart City ecosystem should be able to act as a mediator between healthcare providers and emergency services and should also be able to manage the trade-off between data *sensitivity* and their *value* for the emergency services.

Our aim is to design an intelligent system that meets the listed requirements; in this way, we achieve our goal of automatically analysing people's health records and, therefore, providing emergency services with additional information about people requiring assistance during an emergency.

In the following chapter, we describe in detail how we developed each knowledge component and demonstrate its functionality through extensive experiments.

Chapter 4

Knowledge representation of health condition evolution

In Chapter 2, we described the limitations associated with using time validity representation to detect vulnerable people in an emergency. A more sophisticated description of how health conditions develop over time can enhance the system’s precision, detect ongoing conditions and provide means to calculate the recovery duration.

In this chapter, we propose a methodology for designing a model to represent condition evolution, allowing the inference of ongoing health issues. We introduce the notion of *Health Condition Evolution Statement (HES)* to solve the problem of representing and reasoning over a person’s health status at a given point in time. First, we examine how health condition evolution is expressed in natural language and use this analysis as the basis for devising the features of the conceptual model. With this information, we design a model for representing the evolution of health conditions, which comprises a set of Health Condition Evolution Statements (HES).

4.1 Methodology

Ontologies are largely used in applications in the Semantic Web field and are widely adopted in Knowledge Engineering in various domains. Uschold and Gruninger (1996) proposed the first guidelines for building ontologies based on their experience developing complex projects in the business domain. Furthermore, methodological frameworks like [López 1999; Presutti et al. 2009] also cover aspects of the ontology development process, life cycle, methods and techniques. For instance, the NeOn Methodology [Suarez-Figueroa, A. Gomez-Perez, and Fernandez-Lopez 2012] proposes a more flexible approach with particular atten-

tion to ‘ontology engineering by reuse’ and collaborative development. Our methodology follows these established frameworks and adapts the different steps from each method according to our system and knowledge requirements for health condition evolution representation. For instance, we follow [Gruninger and Fox 1995; Uschold and Gruninger 1996] to identify the ontology’s purpose and motivating scenario and to capture knowledge requirements expressed as Competency Questions (CQs). Moreover, we follow best practices of reusing established ontologies to minimise the addition of new ontology terms [Suarez-Figueroa, A. Gomez-Perez, and Fernandez-Lopez 2012].

In what follows, we describe the methodology used to develop the Health Condition Evolution Ontology (HECON), which is our proposed model for representing the evolution of health events over time and facilitating a system’s reasoning on health records. We follow ontology engineering good practices and methodologies [Uschold and Gruninger 1996; Presutti et al. 2009; Suarez-Figueroa, A. Gomez-Perez, and Fernandez-Lopez 2012], and devise the following design process:

1. **Abstracting the scenario.** Identify the scope and purpose of the ontology, and identify the data sources that describe health condition evolution.
2. **Identify knowledge requirements and formulate Competency Questions (CQs).** Identify the knowledge reasoning requirements of a system that reasons on health records to assess a person’s recent health issues, and formulate them as CQs.
3. **Ontology design and construction.** Describe the concepts and relations that constitute the ontology and use this vocabulary to express the ontology in a formal language.
4. **Ontology evaluation.** Evaluate the ontology considering the consistency of the model and the fulfilment of the CQs.

The estimation of health condition evolution is a complex task; typically, the convalescence process varies from person to person, making it difficult to establish a set time frame; instead, doctors and medical professionals give an approximate duration.

With this representation, we aim to:

- Build a formal representation of health condition evolution that abstracts the estimation of recovery or evolution of a condition over time. So that, it is possible to evaluate if a condition is ongoing automatically (for instance, when an emergency occurs).

- Design a model that captures a measurable estimation of the time required to recover from a particular health condition, providing a system with the knowledge to evaluate the convalescence time.
- Represent cases where conditions are chronic or deteriorate over time.
- By following a Linked Data approach, we aim to provide a structured database of conditions' evolution over time, which is easy to access and fits our system requirements.
- Represent health conditions according to well-known clinical terminology taxonomies such as SNOMED CT and allowing direct linkage to FHIR events.

In the following sections, we report on the process that led to the Health Condition Evolution Ontology (HECON), a sophisticated description of how health conditions develop over time. We start by providing examples of how health condition evolution is expressed in natural language; these examples were collected from publicly available sources such as NHS England and MAYO Clinic. In this thesis, we use the term **health condition** or simply **condition** interchangeably and adopt the FHIR definition to refer to any '*...clinical condition, problem, diagnosis, or other events, situation, issue, or clinical concept that has risen to a level of concern*'¹.

4.2 Abstracting the scenario

In Section 2.2 we discussed the challenges of extracting relevant information from health records specifically to support emergency responders. The main goal of our ontology is to provide an intelligent system with enough knowledge to allow the detection of ongoing health issues using health records as the data source. We used the scenario reported in Chapter 3 as a guide for identifying key ontology concepts and relationships that define 'health evolution' terms.

In our scenario, we described a fire emergency in a large organisation. In the context of a Smart City, an intelligent system can leverage the organisation's Access Control System (ACS) to determine the people (employees and visitors) in the building at the moment of the emergency. The ACS provides the list of people on the premises; therefore, the health-care provider can retrieve their electronic health records (see Section 1.2, Assumption 1). The public health provider uses SNOMED CT and FHIR as standards for accessing and

¹FHIR 'condition' definition: <http://hl7.org/fhir/condition.html> (Appendix A)

exchanging EHRs (see Section 1.2, Assumption 2). Therefore, the **intelligent system can use people’s healthcare data to identify up-to-date health issues and provide emergency services with information that could indicate a person is in a vulnerable situation due to a recent health issue**. The main question that our system should answer using the ontology is:

What are a person’s ongoing health issues at a certain point in time?

Once we identified the ontology’s scope and purpose, we focused on identifying data sources about health condition evolution. Such resources are necessary for two reasons: (a) to develop the knowledge model about health conditions’ evolution; and (b) to populate a database of health conditions’ evolution (Chapter 5). The sources should comply with certain requirements to be considered for our analysis:

- Available: the source should be of easy access, with no restrictions.
- Extensive: the source should contain a considerable number of conditions’ descriptions.
- Case-oriented: the information provided should contain descriptions of health recovery, duration or references that allow the estimation of the length of convalescence time.
- Authoritative: the information should be generated by a reliable source; this could be a private organisation or a public governmental authority that performs constant and regular maintenance (review and update) of the data on health conditions.

Here we relied on two health organisations: NHS England and MAYO Clinic. NHS England is the largest health website in the UK, and it provides straightforward access to content about symptoms, conditions, and treatments. The MAYO Clinic is a non-profit organisation oriented to clinical practice, education, and research, providing comprehensive and easy access to condition descriptions. NHS England website displays information on 972 health conditions and MAYO Clinic, 1186 health conditions. Both include sections that describe the ‘recovery’ and ‘treatment’ where we can find condition evolution information.

4.3 Knowledge reasoning – principles

Next, attention was given to examining the selected sources and understanding how condition evolution is expressed in natural language. The objective was to identify the elements

and features that comprise condition evolution and use them as a guide for modelling the ontology.

4.3.1 Thematic Analysis

This section describes the task of abstracting the features or elements that can be used to represent and build a formal model for condition evolution. Therefore, we select a random list of conditions (about 1% from each data source: 10 conditions from NHS England and 15 from MAYO Clinic); and manually examine the content on the web pages looking for condition evolution information:

1. Read thoroughly the text, looking for references and complete descriptions on how conditions evolve, the time it takes to recover or if they are chronic or permanent.
2. Extract the identified text.

Table 4.1 displays a few examples of text found in the dataset, which are used as an initial guide for the construction of the model. From this sample, we learned that often the evolution process (indicating recovery or deterioration) is expressed in one sentence, and these sentences contain expressions such as *fully recover*, *last between*, *is a progressive condition/disease*, *lifelong condition*, *no specific cure*, among others. For instance, the ‘Bronchitis’ web page contains sentences such as ‘*In most cases, acute bronchitis clears up by itself within a few weeks without the need for treatment.*’. We also found that often conditions descriptions may have more than one health recovery description or none. Continuing with the ‘Bronchitis’ example, we found a second sentence: ‘*If symptoms last for at least 3 months, it’s known as chronic bronchitis.*’.

Next, to define the elements that define condition evolution and all values that represent them, we consider it essential to expand this process and gather more examples from the data sources. Therefore, we decided to analyse a larger subset of sentences.

To build this subset of sentences, we manually:

- (a) Compile a list of text snippets indicating condition recovery. We review the text that is part of the data sources and manually identifies text snippets to build a list of recovery expressions. Table 4.2 displays examples of expressions found in the text (see Appendix F.1 for a complete list).
- (b) Use this list of snippets and cosine similarity measure to find complete sentences within the descriptions.

Table 4.1: Examples - text describing health evolution.

Condition (*name collected from web sources)	Sentence (* text collected from web sources)	Source
Fracture of ankle	<i>'A broken ankle usually takes 6 to 8 weeks to heal, but it can take longer.'</i>	NHS
Addisson's disease	<i>'Addison's disease symptoms usually develop slowly, often over several months.'</i>	MAYO
Jakob Creutzfeldt disease	<i>'There's currently no cure for CJD, so treatment aims to relieve symptoms and make the affected person feel as comfortable as possible.'</i>	NHS
Blood tests	<i>Only a small amount of blood is taken during the test so you shouldn't feel any significant after-effects.'</i>	NHS

Table 4.2: Examples of text snippets describing health evolution

Text snippet used for cosine similarity matching	Original sentence extracted from data source
<i>rest home day two</i>	<i>rest at home for a day or two</i>
<i>2 weeks</i>	<i>for 2 weeks</i>
<i>usually progresses slowly</i>	<i>usually progresses very slowly</i>
<i>no cure</i>	<i>There's currently no cure for</i>
<i>symptoms last least 3 months known chronic lifelong condition</i>	<i>If symptoms last for at least 3 months, it's known as chronic is a lifelong condition</i>

We use this subset of sentences to analyse its structure, and as a result, we abstracted three dimensions or components that are present in health evolution descriptions: type of condition, pace, and time range. For each dimension, we extract annotations that define all the possible values they could take. We do this by grouping expressions with similar meanings. Table 4.3 lists all the three dimensions found in the sentences, all the possible annotations or values each dimension could take, a definition, and a compilation of *expressions* as found in the text (the expressions are not limited to the ones found in the table).

For instance, the dimension 'type' of condition indicates if a health condition improves, declines or becomes chronic over time. Then, sentences often provide details about the speed of the recovery or 'pace'. In this case, sentences contain different expressions such as 'rapidly, gradually or long time', among others; thus, we classify them as: fast, moderate and slow. Finally, 'time range' provides a measurable account of how long it lasts (expressed in days, weeks, months, and years). This last dimension has two elements, a minimum convalescence time frame when people either recover good health or deterioration starts and a maximum convalescence time frame when deterioration starts and before it becomes

permanent. From our analysis, we identified six general time range categories.

Table 4.3: Summary of expressions used abstract health condition evolution features

Dimension	Annotation	Annotation definition	Expressions found in sentences
TYPE	IMPROVEMENT	Indicates recovery of good health condition. Always have a minimum and maximum recovery time	<i>'improve in', 'no more than X minutes', 'last between/within/around X', 'take around', 'less than', 'fully recover', 'temporary'</i>
	DECLINE	Indicates that gradually becomes worse. A developing condition that leads to death or causes disability.	<i>'deteriorates', 'develop slowly/rapidly/gradually', 'is a progressive condition/disease', 'gets gradually worse over time'.</i>
	PERMANENT	Long-lasting and never goes away or not regaining full pre-condition status.	<i>'lifelong condition', 'no specific cure', 'cannot be cured', 'it is a long-term condition/complication'</i>
	UNAFFECTED	Describe administrative procedures not affecting health	
PACE	FAST	That progresses quickly.	<i>'rapidly', 'in less than a few/X days', 'is a straightforward process', 'very quick'.</i>
	MODERATE	That progresses at a moderate pace.	<i>'develop gradually', 'several weeks', 'within a few months'</i>
	SLOW	That progresses over a long period.	<i>'develop slowly', 'several months to years', 'over years/-many years/several years', 'often over several months', 'long time to recover (from few months to a year)', 'progresses slowly'</i>
TIME RANGE	FROM	A minimum period of convalescence or deterioration starts also called a lower bound (LB).	<i>Expressed in hours, days, months or years, several/few years/months/weeks. E.g., 'for 4 to 6 week', 'within 2 weeks', 'it lasts up to 3 weeks'</i>
	TO	A maximum period of convalescence (to recover good health) or deterioration (before it becomes a chronic/permanent condition), also called upper bound (UB).	<i>Expressed in hours, days, months or years, several/few years/months/weeks.</i>

With this abstraction, a sentence or a piece of text can be represented in a machine-readable way as a combination of these features. We name this representation **Health Condition Evolution Statement (HES)**. Figure 4.1 illustrates the structure of our model.

However, not all combinations of features are meaningful; as a result, there are some constraints we considered when building a HES (see Figure 4.1):

- **Constraint one:** Unaffected and Permanent do not combine with 'Pace' and 'Time range'. For example, the annotations Unaffected and Permanent express either the absence of a condition evolution expression or a condition evolution expression that, although relevant, does not change over time.
- **Constraint two:** combination of pace and time range features should be coherent. For

example, if the time range expression is ‘from 2 months to 6 months’, it cannot be ‘fast’.

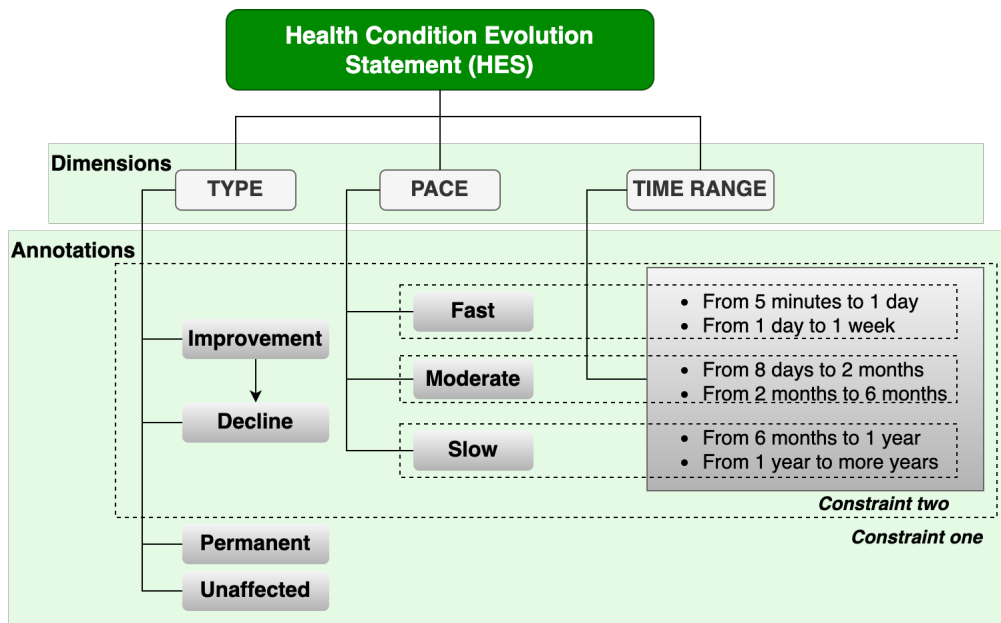


Figure 4.1: Health Condition Evolution Statement (HES) representation

Finally, we illustrate the application of the HES model by taking a few sentences and annotating them manually. We assign an annotation for each dimension of the HES using the definitions in Table 4.3. For instance, we start with the dimension ‘type’ and evaluate if the sentence indicates improvement, decline, permanent or unaffected. We repeat the exercise for pace and time range. Table 4.4 contains a list of sentences and their reformulation as HES statements. For example, the sentence ‘*A broken ankle usually takes 6 to 8 weeks to heal, but it can take longer.*’ indicates that after some time, a person may recover (improve); the estimated recovery time is six to eight weeks which falls into the ‘from 8 days to 2 months’ category.

4.3.2 Competency Questions

In this section, we extract the knowledge requirements from the perspective of a system that handles the scenario proposed in Chapter 3. Building around the described scenario and the central question proposed in Section 4.2, we expressed the knowledge requirements as a set of Competency Questions (CQs) that our proposed ontology should answer. The core CQs were expressed as follows:

- **CQ1:** What is the health evolution information of a given SNOMED concept?
- **CQ2:** How does a given condition evolve over time?

Table 4.4: Examples - sentences and Health Evolution Statement representation.

Condition Name	Sentence	Health Condition Evolution Statement (HES)		
		Type	Pace	Time
Fracture of ankle	<i>'A broken ankle usually takes 6 to 8 weeks to heal, but it can take longer.'</i>	Improvement	Moderately	from 8 days to 2 months
Addison's disease	<i>'Addison's disease symptoms usually develop slowly, often over several months.'</i>	Decline	Slowly	from 6 months to 1 year
Jakob Creutzfeldt disease	<i>'There's currently no cure for CJD, so treatment aims to relieve symptoms and make the affected person feel as comfortable as possible.'</i>	Permanent		
Blood tests	<i>'Only a small amount of blood is taken during the test so you shouldn't feel any significant after-effects.'</i>	Unaffected		

- **CQ3:** What is the pace at which a given SNOMED concept evolves?
- **CQ4:** What are the expected minimum and maximum recovery times for a given SNOMED concept?
- **CQ5:** If an emergency happens after the expected maximum recovery time, is a condition still ongoing?
- **CQ6:** If an emergency happens before the expected minimum recovery time, is a condition still ongoing?
- **CQ7:** If an emergency happens between the expected minimum and maximum recovery time, is a condition still ongoing?

The data used to build the database of health evolution is retrieved from different sources, identified during the abstraction of the scenario. These include two authoritative and reliable data sources: NHS England² and MAYO Clinic³. We expect the population of the ontology to be a multi-faceted effort, combining different methods and techniques, this includes the participation of domain experts and the use of other data sources. In any event, assessing the accuracy and validity of the data is of paramount importance; therefore, we consider it essential to represent this knowledge by including the associated source and explaining the context of how information was generated, and if available, including measures that support the quality assessment of the annotations. We expressed these provenance requirements as follows:

²NHS England: <https://www.nhs.uk/conditions/>

³MAYO Clinic: <https://www.mayoclinic.org/diseases-conditions>

- **CQ8:** What activity generates specific health evolution information of a given SNOMED CT concept (e.g., a user study, an automatic knowledge extraction)?
- **CQ9:** What is the source of the health evolution information (e.g., authoritative sources, domain experts)?
- **CQ10:** What/Who is the organisation/person providing this information?
- **CQ11:** Is there additional information indicating the information's quality?

In addition, we consider it relevant to describe the design constraints derived from the scope of our system and best ontology engineering practices:

- Reuse established ontologies when possible, minimising the addition of new ontology terms. In our case, to represent the time duration, we utilise the Time Ontology [Hobbs and Feng 2006] and PROV Ontology [Lebo et al. 2013] for provenance information.
- Reuse standards in the health domain. We link the condition evolution information to SNOMED CT healthcare terminology. We also use the FHIR standard to facilitate the management of health-related data.

4.4 Building HECON - Health Condition Evolution Ontology

In this section, we describe the process for designing and building the model representing the evolution of health events over time. As described in previous sections, the process of recovering from a health situation is not limited to a fixed convalescence time. Our ontology aims to support the identification of ongoing health events at a given point in time by representing conditions' evolution information.

We linked the health evolution data to the SNOMED CT taxonomy and extended it by aggregating information on the recovery time of conditions. We motivated the design rationale of the ontology by considering a fire emergency scenario as our reference application, which provides both a source for requirements and a validation set.

In what follows, we describe the core concepts of the **Health Condition Evolution Ontology (HECON)** and the integration of SNOMED CT clinical terms. Next, we explain how we incorporate Time [Hobbs and Feng 2006] and PROV-O [Lebo et al. 2013] Ontologies and their role in our model.

4.4.1 HECON core concepts

In this section, we defined a number of ontological concepts in order to meet the knowledge requirements captured as CQs. Figure 4.2 displays an overview of the core components of HECON. The HECON terms are under the following namespace:

`http://kmi.open.ac.uk/conrad/hecon`

First, we defined the core concept of HECON Ontology, the `HealthEvolutionStatement` (HES) class. As described previously, health condition evolution information is available in natural language. Its main elements are the types of health evolution, the velocity at which it changes, and its duration. The `HealthEvolutionStatement` class is an abstraction of these components, and indicates the recovery process [Morales Tirado, Daga, and Motta 2021; Morales Tirado, Daga, and Motta 2022b]. We identified four main types of health evolution:

- **Improvement** - the evolution is favourable and indicates recovery of good health.
- **Decline** - the evolution is adverse and gradually worsens after some time.
- **Permanent** - indicates a persistent condition.
- **Unaffected** - describes a health event or SNOMED CT concepts with no effect on the state of health, for example, blood tests or administrative procedures.

We define the class `Progress` to represent the type of conditions that evolve over time. Constraints described in section Thematic Analysis 4.3.1 indicate that there are two types of conditions that evolve over time: improvement and decline, therefore, we create two classes `Improvement` and `Decline`. For instance, a ‘Fracture of ankle’ gets better after some time; therefore, its evolution type is ‘Improvement’.

The velocity at which a health condition evolves is represented by class `Pace` and could take values such as `fast`, `moderate` and `slow` described by property `hasPace`.

The estimated duration is defined by properties `hasMinDuration` and `hasMaxDuration`. These two boundaries (min and max duration) express time extent, and could take any numerical value in *minutes*, *hours*, *days*, *weeks*, *months* or *years*; we represent these elements with class `time:Duration` and properties `time:numericDuration`, `time:unitType` from the Time Ontology [Hobbs and Feng 2006].

Health events (for example, Bronchitis or Fracture) are represented in health records using a standard terminology system; in our case, SNOMED CT taxonomy. Therefore, each

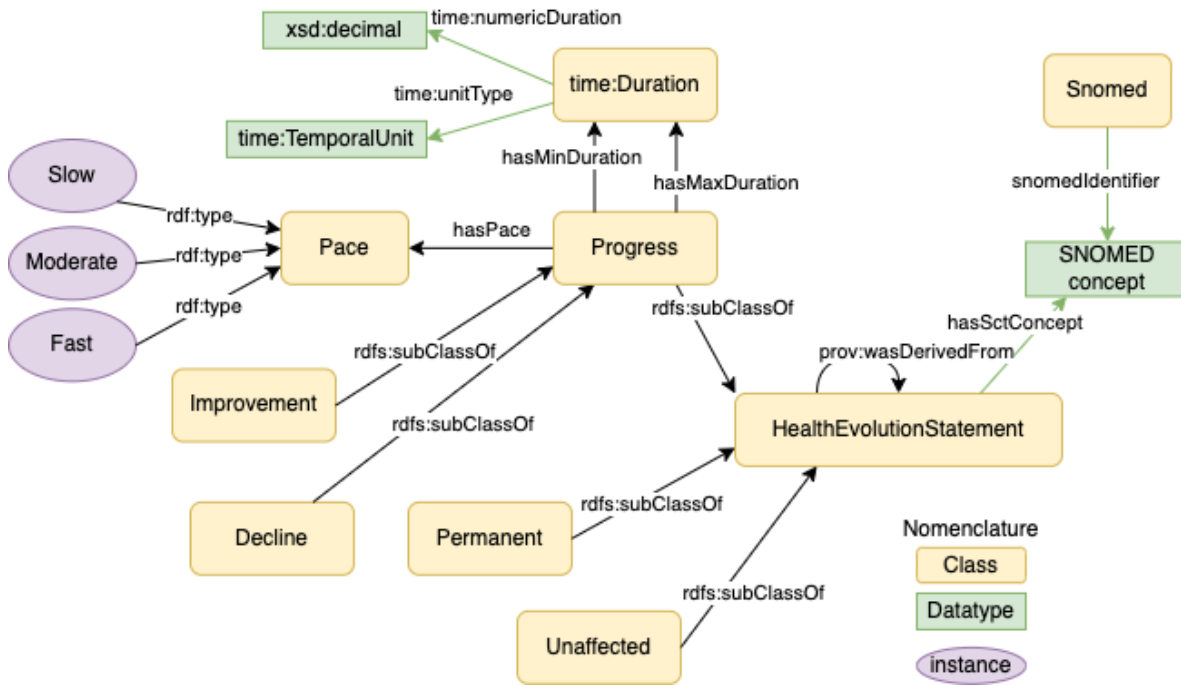


Figure 4.2: HECON Ontology core concepts.

expression of health evolution is linked to its correspondent SNOMED CT identifier using property `hasSctConcept`. We reuse the entities of SNOMED CT Ontology; for example, a health event such as ‘Fracture of ankle’ is represented as follows: `http://snomed.info/id/16114001` or `sct:16114001`⁴.

A system using HECON can predict whether a health condition is still ongoing by calculating the time that has passed between the date it was recorded (in a health record) and its estimated recovery date (the minimum and maximum duration). For example, a ‘Fracture of ankle’ (see Table 4.4) that occurred recently and has not reached its minimum duration (e.g., six weeks) is more likely to impact a person’s health condition than if the same health event happened two months ago (max. duration, e.g., eight weeks).

The type of health evolution also influences the impact on a patient’s state of health; a condition that improves is different from one that is permanent; we cover the implementation of the reasoning on health evolution in Chapter 6.

4.4.2 Provenance representation

Part of the requirements state the necessity to provide context on how the Health Evolution Statement (HES) was generated, the data sources used, and the organisations or persons supporting this information. Therefore, we used PROV-O basic classes and properties to repre-

⁴`sct` is the prefix name associated to SNOMED CT concept identifiers and the namespace URI, defined by SNOMED CT.

sent the origin of each HES. Figure 4.3 illustrates the abstraction of provenance information used for HECON.

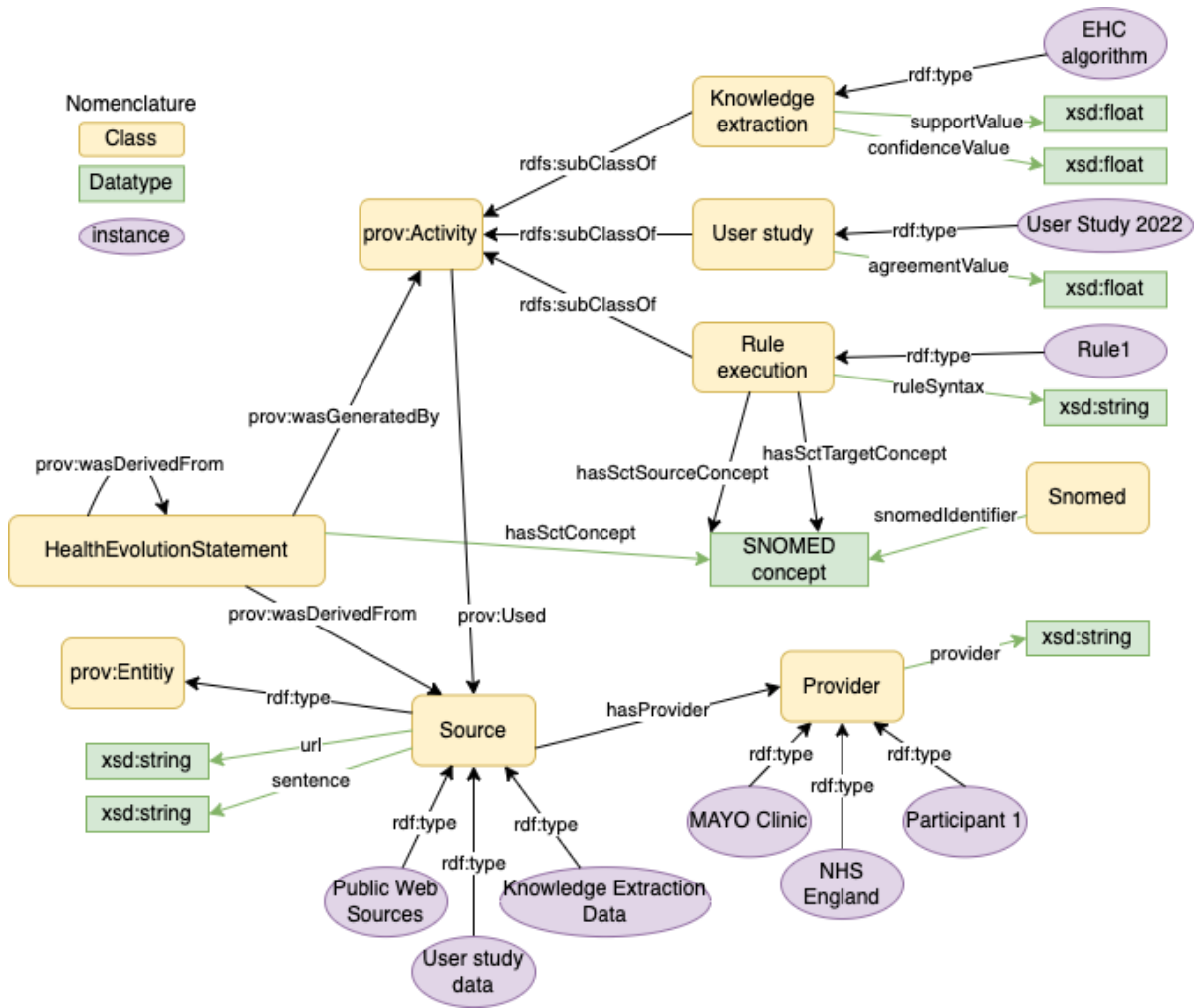


Figure 4.3: HECON Ontology provenance representation.

A SNOMED CT concept can be linked to one or many Health Evolution Statements. Each HES is a subclass of `prov:Entity` that was generated (`prov:wasGeneratedBy`) using one or many approaches or activities - `prov:Activity`. Initially, we identified one activity that could generate `HealthConditionEvolution` entities: a knowledge extraction process. However, new HES statements may be generated by domain experts or by using SNOMED CT taxonomy to derive HES that may be shared by similar SNOMED CT concepts [Morales Tirado, Daga, and Motta 2022b; Morales Tirado, Daga, and Motta 2022c]. In what follows, we describe these three activities:

Knowledge extraction. Health evolution statements can be built using knowledge acquisition techniques for extracting information from unstructured data sources. This activity could generate one or more recommended HES; therefore, it is important to indicate each statement's most frequent combination of annotations; thus, we used confidence and support

metrics. Every knowledge extraction process (each time the activity is performed) is an entity of `KnowledgeExtraction` class, which in turn is a subclass of `prov:Activity` and has properties `support` and `confidence`. For instance, the statement ‘Improvement moderate from 8 days to 2 months’ was generated by the ‘KnowledgeExtraction’ activity ‘EHC algorithm’, and its support value is ‘0.0016’.

Rule execution. Health evolution statements can be generated as a result of a knowledge completion process (as described in detail in [Morales Tirado, Daga, and Motta 2022c]). The objective is to create a set of rules (based on the relationships and attributes in SNOMED CT taxonomy) that indicate a ‘target’ SNOMED CT concept can inherit the same HES as another SNOMED CT ‘source’ concept that already has a HES. The relation between `RuleExecution` and a SNOMED CT concept is expressed as `hasSctTargetConcept` and `hasSctSourceConcept` respectively. The rule(s) used by this activity are represented by the data property `ruleSyntax`. For instance, the HES ‘Improvement Slow from 2 months to 6 months’ linked to ‘Bacterial sinusitis’ (`sct:703470001`) was generated by the activity ‘Rule 1’ and entity of ‘`RuleExecution`’ class. It has as source concept ‘Sinusitis’ (`sct:36971009`) and target concept `Bacterial sinusitis` (`sct:703470001`).

User study. In this case, the HES is the result of a manual evaluation performed by domain experts (e.g., doctors, nurses, emergency responders) who either create new HES or evaluate the accuracy of the recommended HES statements. This activity also provides data on the agreement between compiled responses in terms of an `agreementValue`.

An activity, for example, Knowledge Extraction, can use one or multiple sources, represented by class `Source` which in turn is an entity of (`prov:Entity`). Sources are linked to a provider, represented by class `Provider` and `hasProvider`; this could be an organisation or a person, for example, NHS England or a domain expert. Some sources optionally include details of public URL or exact text that generated the HES, expressed with properties `url` and `sentence`. Following the previous example of ‘Fracture of ankle’, we can express that it was derived from ‘Public Web Sources’, and its provider is ‘MAYO Clinic’. It is possible to represent all the possible Sources used by a certain activity with the property `prov:Used`.

4.5 Ontology evaluation

In this section, we describe how HECON can be used to meet the knowledge requirements of an intelligent system which automatically detects ongoing health issues when an emergency occurs.

We used Protégé⁵ and the Hermit OWL Reasoner⁶ to check the formal consistency of the ontology. We based the consistency check on a random data sample selected from the dataset built in the reasoning requirements section 4.3.

In order to evaluate the expressivity of the ontology and the completeness (fitness for use) of the related KG for our task, we encode the Competency Questions (CQs) listed in Section 4.3.2 into SPARQL queries. We used Blazegraph database⁷ to store the triples and evaluate the CQs by executing SPARQL queries on the data. We use the same dataset of patients' synthetic health records used in [Morales Tirado, Daga, and Motta 2021] and compare the outcome of the queries against the expected results. The dataset is described using FHIR and SNOMED CT standards, emulating UK national information standards requirements, and therefore we make sure our approach is compliant with real implementations. In what follows, we group closely related CQs to give a description of the evaluation process and the results.

Table 4.5: Competency Questions - requirements on health evolution information

Competency Questions	Requirement
CQ1 What is the health evolution information of a given SNOMED concept?	Health condition evolution information
CQ2 How does a given condition evolve over time?	
CQ3 What is the pace at which a given SNOMED concept evolves?	
CQ4 What are the expected minimum and maximum recovery times for a given SNOMED concept?	

Competency Questions one to four (CQ1 to CQ4). These competency questions inquire about the duration, pace and type of change characterising the evolution of a condition (see Table 4.5).

As described previously the Health Evolution Statement is a compilation of three pieces of information: type of event (represented by: Improvement, Decline, Permanent or Unaffected), pace (Slow, Moderate or Fast) and time range. Listing 4.1 lists the

⁵Open-source software for ontology edition: <https://protegewiki.stanford.edu/wiki/ProtegeDesktopUserDocs>

⁶Given an OWL file, Hermit can determine whether or not the ontology is consistent, identify subsumption relationships between classes, and much more: <http://www.hermit-reasoner.com/>

⁷Blazegraph database: <https://blazegraph.com/>

triples that reflect the representation of a health condition evolution (e.g., Fracture of ankle) using HECON.

Listing 4.1: HES: type, pace and time range. E.g., Fracture of ankle

```

1 :ab270431d30ca755300021fdb7af5f5e
2     rdf:type                hecon:Improvement , hecon:HealthEvolutionStatement ;
3     hecon:hasMaxDuration    <http://kmi.open.ac.uk/conrad/kg/hecon/2/MONTH> ;
4     hecon:hasMinDuration    <http://kmi.open.ac.uk/conrad/kg/hecon/8/DAY> ;
5     hecon:hasPace           hecon:MODERATE ;
6     hecon:hasSctConcept     sct:16114001 ;
7     prov:wasDerivedFrom     <http://kmi.open.ac.uk/conrad/kg/hecon/NHS/7
c205d8cb4e8941a83a164c9dacf4a88> ;
8     prov:wasGeneratedBy     <http://kmi.open.ac.uk/conrad/kg/hecon/activity/
extraction/ab270431d30ca755300021fdb7af5f5e> .

```

By retrieving information about the type of event (CQ2), the pace (CQ3) and the expected recovery time (CQ4), it is also possible to satisfy CQ1 at the same time. We built a query that retrieves all the three dimensions that constitute the health condition evolution statement. Listing 4.2 displays the query that an intelligent system can use to answer the first four questions. This query takes as input the SNOMED CT identifier of a given condition. For instance, if an entry in the EHR indicates a ‘Fracture of ankle’ event (`sct:16114001`), an intelligent system can use the proposed query and retrieve such information. The result is a list of HES linked to the associated SNOMED CT concept, as shown in Table 4.6. An intelligent system can use these results to assess the validity of the given condition.

Listing 4.2: Competency questions one to four SPARQL.

```

1 prefix sct: <http://snomed.info/id/>
2 prefix prov: <http://www.w3.org/ns/prov#>
3 prefix hecon: <http://kmi.open.ac.uk/conrad/ontology/hecon#>
4
5 SELECT DISTINCT ?snmdIdentifier ?typeEvent ?pace ?maxDuration ?minDuration
6 FROM <http://conrad.kmi.open.ac.uk/hecon/kg/HES>
7 WHERE {
8     VALUES (?snmdIdentifier) { (sct:16114001) }# fracture of ankle
9     ?s rdf:type ?typeEvent ;
10        hecon:hasSctConcept ?snmdIdentifier ;
11        prov:wasGeneratedBy ?anActivity_IRI .
12     ?anActivity_IRI rdf:type hecon:KnowledgeExtraction .
13     FILTER ( ?typeEvent != hecon:HealthEvolutionStatement ) .
14     OPTIONAL { ?anActivity_IRI hecon:confidenceValue ?confidenceVal } .
15     OPTIONAL { ?s hecon:hasPace ?pace ;
16                hecon:hasMaxDuration ?maxDuration ;
17                hecon:hasMinDuration ?minDuration ; }
18 } ORDER BY ?snmdIdentifier DESC ( ?confidenceVal )

```

Table 4.6: Results execution SPARQL

Data	Value
snmdIdentifier	sct:16114001
typeEvent	hecon:Improvement
pace	hecon:MODERATE
maxDuration	<http://kmi.open.ac.uk/conrad/kg/hecon/2/MONTH>
minDuration	<http://kmi.open.ac.uk/conrad/kg/hecon/8/DAY>

Table 4.7: Competency Questions - time range information

Competency Questions	Requirement
CQ5 If an emergency happens after the expected maximum recovery time, is a condition still ongoing?	Time range
CQ6 If an emergency happens before the expected minimum recovery time, is a condition still ongoing?	detailed
CQ7 If an emergency happens between the expected minimum and maximum recovery time, is a condition still ongoing?	information

Competency Questions five to seven (CQ5 to CQ7). This group of questions (see Table 4.7) request information on the time range that allows an intelligent system to estimate and infer if the condition is ongoing at the time of the emergency. Listing 4.3 displays the triples that reflect the representation of the time range using HECON; we continue with the ‘Fracture of ankle’ example.

Listing 4.3: Fracture of ankle - time range information.

```

1 <http://kmi.open.ac.uk/conrad/kg/hecon/2/MONTH>
2     rdf:type                time:Duration ;
3     time:numericDuration    "2"^^xsd:float ;
4     time:unitType           time:unitMonth .
5
6 <http://kmi.open.ac.uk/conrad/kg/hecon/8/DAY>
7     rdf:type                time:Duration ;
8     time:numericDuration    "8"^^xsd:float ;
9     time:unitType           time:unitDay .

```

We encoded a SPARQL query that retrieves a condition’s maximum and minimum duration (see Listing 4.4). The query takes as input the date the fire started and the patient’s unique identifier. The query retrieves all the records that indicate a condition and calculates the number of days that have passed between the start of the health event and the fire event, `?days`. In this example, we calculate the number of days; however, it should be noted that minimum and maximum duration could be expressed in a `time:TemporalUnit` other than days. Therefore, values should be converted accordingly. For this example, we set the date the fire started as ‘2019-10-17’ and calculate the time in `?days`.

To answer CQ5, the intelligent system can use the number of `?days` and the `?max`

Listing 4.4: Fracture of ankle health condition evolution information time range.

```

1 PREFIX hecon: <http://kmi.open.ac.uk/conrad/ontology/hecon#>
2 PREFIX fhir: <http://hl7.org/fhir/>
3 PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
4 PREFIX prov: <http://www.w3.org/ns/prov#>
5 PREFIX time: <http://www.w3.org/2006/time#>
6
7 SELECT ?patientId ?snmdIdentifier ?reasonDescription ?typeEvent ?pace ?startDate
8 ?fireEvent ?days ?minDurationValue ?minDurationUnit ?maxDurationValue ?maxDurationUnit
9 FROM <http://conrad.kmi.open.ac.uk/hecon/kg/EHR>
10 WHERE {
11   VALUES (?patientId)
12   { ( <http://kmi.open.ac.uk/emergency/hr/618e8418-9eb9-42f2-9905-617649e99c13> ) }
13   {?id fhir:Patient.identifier ?patientId ;
14     fhir:Condition.period ?blankNode ;
15     fhir:Condition.codeableConcept ?reasonBlank .
16     ?blankNode fhir:Period.start ?startDate ;
17     fhir:Period.end ?endDate .
18     ?reasonBlank skos:inScheme ?snmdIdentifier ;
19     fhir:CodeableConcept.text ?reasonDescription .
20   }.{
21     GRAPH <http://conrad.kmi.open.ac.uk/hecon/kg/HES>{
22       ?s rdf:type ?typeEvent ;
23       hecon:hasSctConcept ?snmdIdentifier ;
24       prov:wasGeneratedBy ?anActivity_IRI .
25       ?anActivity_IRI rdf:type hecon:KnowledgeExtraction .
26       FILTER ( ?typeEvent != hecon:HealthEvolutionStatement) .
27       OPTIONAL { ?anActivity_IRI hecon:confidenceValue ?confidenceVal } .
28       OPTIONAL {?s hecon:hasPace ?pace ;
29         hecon:hasMaxDuration ?maxDuration_IRI ;
30         hecon:hasMinDuration ?minDuration_IRI .
31         ?maxDuration_IRI time:numericDuration ?maxDurationValue ;
32         time:unitType ?maxDurationUnit .
33         ?minDuration_IRI time:numericDuration ?minDurationValue ;
34         time:unitType ?minDurationUnit . }
35     }
36   } BIND ( "2019-10-17"^^<http://www.w3.org/2001/XMLSchema#dateTime> AS ?fireEvent ) .
37   BIND ((?fireEvent)-(?startDate) AS ?days).
38 }ORDER BY DESC ( ?confidenceVal )

```

DurationValue (converted to the appropriate time unit) to estimate if the condition is still ongoing. For example, if the number of days that have passed is greater than the maximum duration time and the type of condition is ‘improvement’, then the health condition is not relevant, and the person has already recovered good health. Figure 4.4 displays the values retrieved using, as an example, the EHR of a person that recently suffered ‘Sinusitis’. To answer CQ6, we follow the same rationale, and this time the system should compare the days that have passed against the minimum duration time. For CQ7, the comparison this time

should consider both the maximum and minimum duration (the upper and lower bound). The information retrieved using the query allows the intelligent system to infer if the health event is still ongoing. The reasoning rationale and the implications of the type of HES will be explained in detail in Chapter 6. We performed the test using reference dates that simulate an emergency happening before the minimum duration time, in between the minimum and maximum duration, and after the maximum duration; the queries and results are available in the HECON repository ⁸.

snmIdentifier	reasonDescription	typeEvent	pace	confidence Val	startDate	fireEvent	days	min Duration Value	min Duration Unit	max Duration Value	max Duration Unit
sct:36971009	Sinusitis (disorder)	hecon:Improvement	hecon:SLOW	0.00826446	16/10/2019	17/10/2019	0.958333333	2	time:unitMonth	6	time:unitMonth
sct:65363002	Otitis media	hecon:Improvement	hecon:MODERATE	0.00257069	16/02/1981	17/10/2019	14121.95833	2	time:unitMonth	6	time:unitMonth
sct:65363002	Otitis media	hecon:Improvement	hecon:MODERATE	0.00161812	16/02/1981	17/10/2019	14121.95833	8	time:unitDay	2	time:unitMonth
sct:36971009	Sinusitis (disorder)	hecon:Improvement	hecon:MODERATE	0.00161812	16/10/2019	17/10/2019	0.958333333	8	time:unitDay	2	time:unitMonth
sct:36971009	Sinusitis (disorder)	hecon:Permanent		0.00115207	16/10/2019	17/10/2019	0.958333333				
sct:65363002	Otitis media	hecon:Improvement	hecon:FAST	0.00103627	16/02/1981	17/10/2019	14121.95833	5	time:unitMinute	1	time:unitDay

Figure 4.4: Query outcome for CQs5-7.

Table 4.8: Competency Questions - provenance information

Competency Questions		Requirement
CQ8	What activity generates specific health evolution information of a given SNOMED CT concept (e.g., a user study, an automatic knowledge extraction)?	Provenance information
CQ9	What is the source of the health evolution information (e.g., authoritative sources, domain experts)?	
CQ10	What/Who is the organisation/person providing this information?	
CQ11	Is there additional information indicating the information's quality?	

Competency Questions eight to eleven (CQ8 to CQ11). The last set of questions (see Table 4.8) are related to provenance information. As described previously, a health evolution statement could be generated using different data sources and executing various activities. Listing 4.5 displays the triples that reflect provenance information.

We encoded a query that an intelligent system can use to retrieve the Activity (CQ8), the Source (CQ9) and the Provider (CQ10) that support the generation of a HES for a given SNOMED CT concept (see Listing 4.6). The query results are displayed in Table 4.9.

In summary, these results show that the HECON Ontology allows us to represent and access knowledge about the evolution of conditions recorded in health records. We selected a random sample of health records and queried their HES using HECON and the KG. Retrieved data showed that our ontology meets requirements successfully.

⁸HECON Ontology repository: <https://github.com/albamoralest/HECON-Ontology>

Listing 4.5: Fracture of ankle health HES provenance information.

```

1 <http://kmi.open.ac.uk/conrad/kg/hecon/NHS/7c205d8cb4e8941a83a164c9dacf4a88>
2     rdf:type          hecon:Source ;
3     rdfs:label         "Public Web Sources" ;
4     hecon:hasProvider  ont:NHS ;
5     hecon:sentence     "A broken ankle usually takes 6 to 8 weeks to heal, but it
6     can take longer." ;
7     hecon:url          "https://www.nhs.uk/conditions/broken-ankle/" .
8 <http://kmi.open.ac.uk/conrad/kg/hecon/activity/extraction/
9     ab270431d30ca755300021fdb7af5f5e>
10    rdf:type          prov:Activity , hecon:KnowledgeExtraction ;
11    rdfs:label         "EHC Algorithm - 16114001" ;
12    hecon:confidenceValue "0.0016181229773462"^^xsd:float ;
13    hecon:supportValue   "0.0001932740626207"^^xsd:float ;
14    prov:used           <http://kmi.open.ac.uk/conrad/kg/hecon/NHS/7
15    c205d8cb4e8941a83a164c9dacf4a88> .

```

Listing 4.6: Query provenance information.

```

1 PREFIX sct: <http://snomed.info/id/>
2 PREFIX prov: <http://www.w3.org/ns/prov#>
3 PREFIX hecon: <http://kmi.open.ac.uk/conrad/ontology/hecon#>
4
5 SELECT DISTINCT ?activity ?sourceName ?provider ?confidenceVal ?supportVal
6 WHERE {
7     VALUES (?snmdIdentifier) { (sct:36971009) } # sinusitis
8     ?s rdf:type ?typeEvent ;
9     hecon:hasSctConcept ?snmdIdentifier ;
10    prov:wasGeneratedBy ?anActivity_IRI ;
11    prov:wasDerivedFrom ?aSource_IRI .
12    ?anActivity_IRI rdf:type ?activity .
13    ?aSource_IRI rdf:type hecon:Source ;
14        rdfs:label ?sourceName ;
15        hecon:hasProvider ?provider .
16    FILTER ( ?typeEvent != hecon:HealthEvolutionStatement
17        && ?activity != prov:Activity ) .
18    OPTIONAL { ?anActivity_IRI hecon:confidenceValue ?confidenceVal ;
19        hecon:supportValue ?supportVal .}
20    OPTIONAL {?anActivity_IRI hecon:hasSourceSnmdConcept ?sourceConcept ;
21        hecon:hasTargetSnmdConcept ?targetConcept ;
22        rdfs:label ?ruleName .}
23    OPTIONAL {?s hecon:hasPace ?pace ;
24        hecon:hasMaxDuration ?maxDuration ;
25        hecon:hasMinDuration ?minDuration ; }
26 } ORDER BY ?snmdIdentifier DESC ( ?confidenceVal )

```

Table 4.9: Results - provenance information

Data	Value
activity	hecon:KnowledgeExtraction
sourceName	Public Web Sources
provider	<http://kmi.open.ac.uk/conrad/kg/hecon/NHS>
confidenceVal	0.001618123
supportVal	0.000193274
typeEvent	hecon:Improvement

4.6 Discussion

In this chapter, we presented the Health Condition Evolution Ontology (HECON), a formal model for representing and reasoning health events' evolution over time. We followed best knowledge engineering practices to collect knowledge requirements of an intelligent system that automatically estimates a person's current state of health in the context of a fire emergency. Although the examples in this chapter relate to a fire emergency, the ontology is generic and can support other types of emergencies, leaving the assessment of how the condition has to be handled to the emergency support system relying on it [Morales Tirado, Daga, and Motta 2021]. Furthermore, the design of the ontology includes basic classes and properties to represent the provenance of each health condition evolution statement (HES) and classes to assess their accuracy and validity.

As part of the design constraints we minimised the addition of new ontology terms reusing the Time Ontology [Hobbs and Feng 2006] and PROV Ontology [Lebo et al. 2013]. Similarly, we reuse standards in the health domain by linking HES to SNOMED CT ontology and using the FHIR resources description.

Handling privacy requirements is not part of the ontology; however, in the following chapters, we will explore a reasoning system that allows pruning all the events that are not relevant to the emergency by exploiting this model [Morales Tirado, Daga, and Motta 2021].

Chapter 5

Knowledge acquisition and knowledge graph construction of health condition evolution

In the previous chapter, we presented the first component required for an intelligent system to inspect health records and identify people needing special assistance by reasoning on the evolution of medical events. We developed HECON, the Health Condition Evolution Ontology, a model for representing and reasoning on condition evolution. The second element is knowledge about health condition evolution, specifically structured data that will support the annotation of conditions according to HECON. Unfortunately, there is a lack of structured resources regarding health condition evolution and recovery time. However, unstructured information available on the web could help domain experts build this database.

In this chapter, we address this gap by designing a Knowledge Acquisition (KA) approach focused on extracting Health Evolution Statements (HES) from natural language sources, symbolic reasoning, and domain experts' knowledge.

In the next section, we give an overview of the methodology proposed to build a structured database of health condition evolution information. Section 5.2 describes the collection of unstructured data and corpus preparation process. In Section 5.3 we describe the extraction of knowledge components from text. We use knowledge classification techniques such as Machine Learning to classify sentences according to HES features. Section 5.4 focuses on the process of expanding the HES annotations. We take advantage of SNOMED CT semantic features to propagate health evolution statements to a more significant number of SNOMED CT concepts. Section 5.5 describes the capture of domain experts' knowledge by including a Human-in-the-loop module. Section 5.6 presents the resulting Knowledge Graph of HES.

Finally, Section 5.7 summarises the contributions of the knowledge acquisition process.

5.1 Approach overview

As reported in previous chapters, we identified a lack of resources regarding condition evolution. In contrast, we identified unstructured data sources from which we can acquire such information. To address this gap, we rely on knowledge acquisition techniques such as machine learning to capture knowledge from natural language. We use a supervised learning approach, specifically a multiclass classification. We focus on building a gold standard dataset which serves as our training data and uses the HES definition to establish our categorical class labels. We also perform a knowledge completion task and take advantage of SNOMED CT features to expand the annotations of SNOMED CT concepts. Finally, we include a Human-in-the-loop step with the primary goal of building a curated and high-quality database. Figure 5.1 provides a graphical summary of the four-step approach. Our contributions are the following:

- A methodology implements a Knowledge Acquisition pipeline for building a Knowledge Graph of Health Evolution.
- A training dataset and the machine learning models trained to classify text according to HES.
- A set of propagation rules built using SNOMED CT taxonomy features that escalate the annotation of SNOMED CT concepts.
- An approach to capture domain experts' knowledge and a tool that instantiates the HITL module.
- The first database of health evolution information published as a Knowledge Graph.

In what follows, we summarise the steps of our proposed approach.

1. Corpus preparation. The first step of the pipeline is dedicated to identifying data sources that describe health evolution and preparing the corpus to be used for the classification task. The sources should comply with characteristics such as: being an authoritative source, publicly available, extensive and including a description of health evolution. The aim is to collect text describing diseases, procedures, and conditions (e.g., asthma, appendicitis, bronchitis) and link them to the corresponding concept in SNOMED CT taxonomy. The result is a corpus of health conditions and their descriptions.

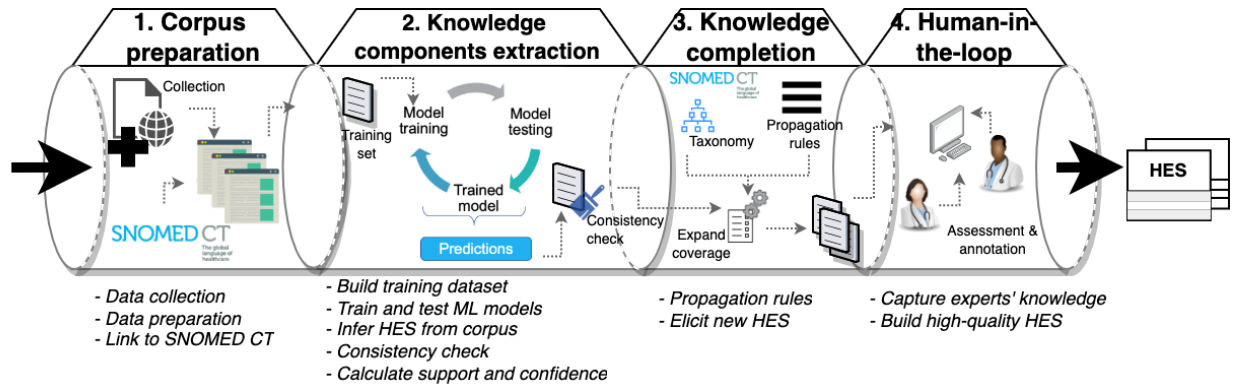


Figure 5.1: Knowledge Acquisition pipeline

2. Knowledge components extraction. The output from the previous phase is a large corpus, and only a few sentences provide information on the evolution of health conditions. Therefore, the next task is to identify the sentences that contain relevant information and then classify them according to the HES components defined by HECON Ontology. We rely on Machine Learning techniques and develop a pipeline that includes the training and testing of a set of models [Kotsiantis, Zaharakis, and Pintelas 2007] for each feature of the HES statement. Next, the best-performing models are used to predict a HES for each sentence in the corpus. Since a condition can have one or more recommended HES, the next task is to clean inconsistent and redundant repeated HES. Lastly, we apply an algorithm that uses support and confidence as metrics to rank the most frequent combination of annotations. The output of this step is a collection of SNOMED CT concepts linked to one or more recommended HES.

3. Knowledge completion. The recommended annotations generated in the previous step have limited coverage of SNOMED CT concepts; therefore, we exploit the semantic structure of SNOMED CT taxonomy to find similar concepts that could share the same HES. Specifically, we use the SNOMED CT concepts' features to identify patterns and derive propagation rules. The rules expand the coverage of the HES to other concepts in SNOMED CT and make it possible to elicit a large dataset of SNOMED CT concept annotations.

4. Human-in-the-loop. Until now, the proposed methodology generated one or more recommended HES for each condition. Selecting the more accurate HES requires additional knowledge. Therefore, in this step, the objective is to build a high-quality database of health condition information; therefore, we rely on domain experts to capture accurate knowledge on health evolution. Domain experts contribute in two ways: (a) by validating and curating the recommended annotations and (b) by creating new ones. The final output is a curated

database of health evolution statements.

In the following sections, we describe each step of the proposed approach in detail.

5.2 Corpus preparation

The first step is dedicated to performing two tasks (a) identifying data sources describing health condition evolution and (b) collecting the information. As detailed in the previous chapter, data sources should come from (a) an authoritative organisation and (b) publicly available. Also, sources should be (c) extensive and (d) contain descriptions of condition evolution. We identified two health websites that comply with these requirements: NHS England¹ and MAYO Clinic². NHS England is the largest health website in the UK, providing straightforward access to content about symptoms, conditions, and treatments. The MAYO Clinic is a non-profit organisation; its website provides comprehensive and easy access to condition descriptions. NHS England website displays information on 979 health conditions and MAYO Clinic, 1,186. Between the two websites, we compile 1,774 different conditions in total.

We collected the HTML files (2,165 web pages in total) that contain descriptive guides to most common health conditions, including symptoms and treatments. Notably, both sites include sections that describe the 'recovery' and 'treatment' where we found condition evolution information. Then we cleaned the text by removing HTML tags, line breaks, special characters and empty spaces. We also cleaned web page sections that contained unrelated information such as tables of content, other related topics, update and review dates and references.

As described in Section 4.3.1 - Thematic Analysis, we manually reviewed web page descriptions (1% of conditions), looking for condition evolution information. From this analysis, we compiled a set of textual descriptions of condition evolution. Once we reviewed the content, we found that condition evolution is usually described in one sentence. For instance, the evolution of 'Broken ankle' is described as '*A broken ankle usually takes 6 to 8 weeks to heal, but it can take longer.*'. Other conditions such as 'Cataract surgery' has more than one description: '*It can take 2 to 6 weeks to fully recover from cataract surgery.*' and '*These side effects usually improve within a few days, but it can take 4 to 6 weeks to recover fully.*'. Therefore, in order to find complete information on condition evolution, we decided to organise the corpus in sentences. The dataset contains 208,838 sentences in total, grouped

¹NHS England website <https://www.nhs.uk/conditions/>

²MAYO Clininc website <https://www.mayoclinic.org/diseases-conditions>

by health conditions. Table 5.1 presents a summary of the total conditions and documents that are part of the dataset.

Table 5.1: Summary of data collected from web sources

Source	# of health conditions	Total web pages	Total sentences	Mean value of sentences per condition
NHS England	979	1,523	71,973	73.51
MAYO Clinic	1,186	2,336	136,865	119.76

Typically, EHR use SNOMED CT as a standard to describe clinical conditions [NHS Digital services 2022b]. Since health conditions collected from web sources are not linked to SNOMED CT, the next step is to align the conditions’ names (as they were collected from web sources) to SNOMED CT identifiers. In this way, we will be able to link health conditions and their health condition evolution statement to the events recorded in EHR. To perform this alignment, we use Levenshtein distance [Levenshtein et al. 1966] to search for similar strings, the objective is to find SNOMED CT terms that match with conditions’ names. SNOMED CT concepts could have one or more term descriptions; therefore, we perform an automatic evaluation of SNOMED CT identifier candidates. For instance, we compare ‘Sore throat’ (the name of the condition as it was collected from the web page) against all SNOMED terms and alternative labels. Taking as example the list of conditions displayed in Table 5.2, we designed an algorithm that automatically obtain the Levenshtein distance between ‘Sore throat’ and all SNOMED CT concept term in column ‘B’ and C. Once we find a Levenshtein value of zero or one, the algorithm stops and store results. Table 5.2 displays the names of conditions (column A), the terms (columns B and C), the Levenshtein value (columns A+B, A+C), and the corresponding SNOMED CT identifier. Next, we run a manual review of all the matching results. The final corpus is a collection of sentences grouped by health conditions (1,774 different conditions in total), where each health condition is linked to its corresponding SNOMED CT identifier.

5.3 Knowledge components identification

Here, the focus is on extracting Health Evolution Statements recommendations from natural language (see Figure 5.1, step 2). Since the corpus is extensive (see Table 5.1), we implemented a supervised text classification approach to automatically annotate sentences according to the health condition evolution statement features described by HECON ontol-

Table 5.2: Finding matching SNOMED CT concepts using Levenshtein distance

Condition Name *as collected from web pages (A)	SNOMED CT Concept term (B)	Lvns distance (A+B)	SNOMED CT Alternative label (C)	Lvns distance (A+C)	SNOMED CT URI
Abdominal aortic aneurysm	Abdominal aortic aneurysm (disorder)	0	AAA - Abdominal aortic aneurysm		http://snomed.info/id/233985008
Abdominal aortic aneurysm screening	Abdominal aortic aneurysm screening (procedure)	0			http://snomed.info/id/698356002
Abortion	Termination of pregnancy (procedure)		Abortion	0	http://snomed.info/id/386639001
Sore throat	Pain in throat (finding)		Sore throat	0	http://snomed.info/id/162397003
Lumps	Mass (morphologic abnormality)		Lump	1	http://snomed.info/id/4147007
Lumps	Mass of body structure (finding)		Lump	1	http://snomed.info/id/300848003

ogy. In what follows, we describe in detail the workflow for performing the classification task at hand.

5.3.1 Building a gold standard dataset of HES

In order to perform the classification task, we need a dataset to feed the machine learning models. Therefore, we built a gold standard dataset using sentences that refer to condition evolution.

In the previous chapter (refer Chapter 4, section 4.2), we built a dataset of sentences containing health condition evolution. This dataset is used to abstract the dimensions and characteristics of the health condition evolution statement (HES) and the design of the HECON ontology.

Here we use the same dataset and manually annotate each sentence in the dataset with its corresponding HES components: type of condition (improvement, decline, permanent), pace (fast, moderate, slow) and duration (maximum and minimum duration). Since health condition descriptions (collected from websites) are extensive, on average, one health condition has 98 sentences, and only a few of them describe health condition evolution; we also added negative annotations to the training set. We emulate this scenario by adding sentences without this information and annotating them as ‘NONE’. The output is a manually curated gold standard of 1,987 sentences and their corresponding HES, built by the lead investigator of this research. Table 5.3 summarises the total number of sentences grouped by HES.

Table 5.3: Number of sentences per HES in the training dataset

Health Condition Evolution Statement (HES)			Total
Type	Pace	Duration	
NONE			1437
PERMANENT			141
IMPROVEMENT	MODERATE	8 DAYS TO 2 MONTHS	106
IMPROVEMENT	FAST	5 MINUTES TO 1 DAY	74
DECLINE	SLOW	1 YEAR TO MORE YEARS	56
IMPROVEMENT	MODERATE	2 MONTHS TO 6 MONTHS	53
IMPROVEMENT	FAST	1 DAYS TO 1 WEEKS	37
IMPROVEMENT	SLOW	1 YEAR TO MORE YEARS	37
IMPROVEMENT	SLOW	6 MONTHS TO 1 YEAR	30
DECLINE	FAST	1 DAY TO 1 WEEK	6
DECLINE	SLOW	6 MONTHS TO 1 YEAR	4
DECLINE	MODERATE	8 DAYS TO 2 MONTHS	4
DECLINE	MODERATE	2 MONTHS TO 6 MONTHS	2
TOTAL sentences			1987

5.3.2 Training and testing Machine Learning algorithms for the classification task

In this task, the focus is on implementing, training and testing different Machine Learning (ML) algorithms [Kotsiantis, Zaharakis, and Pintelas 2007]. The objective is to classify sentences according to the dimensions used in HECON Ontology [Morales Tirado, Daga, and Motta 2022b].

Before training the models, we prepare the gold standard dataset that will feed our selected models:

- We randomly divide the dataset into training and test datasets, both with the same proportion of class labels as the gold standard. We used an 80/20 and 70/30 proportion to divide the dataset; according to the size of the training set. Also, we ensure that the data is shuffled before splitting.
- We then tokenise, count and normalise the data. In this case, our data is parsed to lowercase before tokenizing; we use words as tokens ($n=1$) and TF-IDF³ method for feature extraction.

³Term frequency–inverse document frequency (TF-IDF) <https://en.wikipedia.org/wiki/Tf%E2%80%93idf>

First of all, we perform a preliminary binary classification (C0). We train a boolean classifier to discriminate sentences that describe health condition evolution from those that do not contain such description; thus, we use two labels, ‘NO’ for the sentences annotated as ‘NONE’ and ‘YES’ for all the other annotations (see Figure 5.2). For instance, text such as ‘*Landing awkwardly from a jump*’ is classified as NO (it does not describe health evolution) and ‘*There’s currently no cure for chronic obstructive pulmonary disease (COPD)*’ is classified as ‘YES’, as it represents health evolution. In this way, we aim to identify sentences containing condition evolution information and increase the number of true positives. Table 5.4 (column C0) displays the accuracy obtained using different ML algorithms.

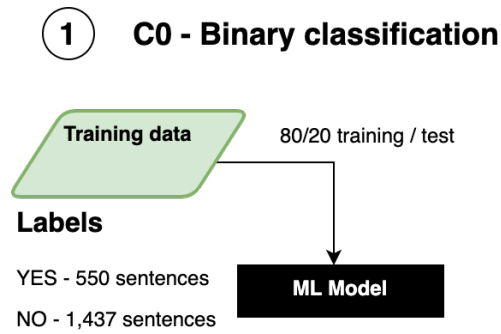


Figure 5.2: Binary classification training

Next, we consider the definition of the Health Evolution Statement features: type of health condition, pace and duration (each one of them could take different values) and train different models using the three dimensions of the HES, each in a separate task. Figure 5.3 illustrates the overall classification workflow.

First, we trained the models to classify sentences using ‘type’ dimension classes: improvement, decline, permanent; thus, the feeding dataset is formed by all the sentences that have an annotation different than ‘NONE’, a total of 550 sentences (see Figure 5.3, step 2). Second, we train the ML models according to the ‘pace’ dimension: slow, moderate, and fast. In total, 409 sentences are part of the dataset since we filter the sentences annotated as ‘permanent’ (as described in the previous chapter, permanent has no pace nor time duration values) (see Figure 5.3, step 3). Third, similar to the other dimensions, we feed the models with sentences annotated according to time duration using six different classes (see Figure 5.3, step 4).

Finally, Table 5.4 shows the list of the ML algorithms we trained and the accuracy of each model grouped by HES features. Highlighted in bold are the models with the best performance. Table 5.5 summarises the final hyper-parameters configuration for each algorithm

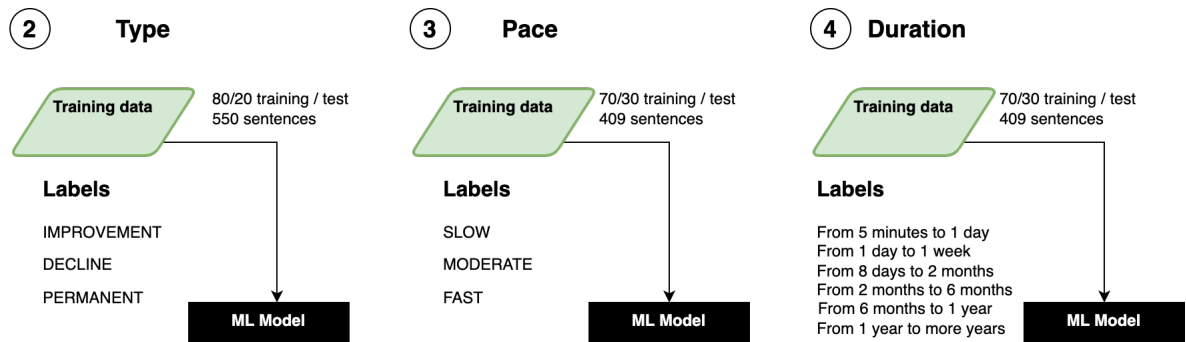


Figure 5.3: ML Training process

and additionally can be found in the repository ⁴.

Table 5.4: ML training results: Accuracy per algorithm & HES features

ML Algorithms	C0	Type	Pace	Duration
<i>Logistic Regression</i>	0.8907	0.9727	0.8148	0.8114
<i>Decision Tree</i>	0.8337	0.9272	0.8934	0.8606
Linear SVC	0.8789	0.9545	0.8271	0.8360
MLP Classifier	0.8337	0.9545	0.7530	0.6803
Naïve Bayes	0.4181	-	-	-
Multinomial NB	0.8136	0.8636	0.6790	0.5737
Random Forest Classifier	-	0.8818	0.7901	0.8442

Table 5.5: Best performing ML algorithms hyper-parameters configuration

ML Algorithms	C0	Type	Pace	Duration
Logistic regression	n_jobs=1	n_jobs=1		
	C=1000	C=1000		
	solver='liblinear'	solver='newton-cg'		
	multi_class='ovr'	multi_class='multinomial'	-	-
	max_iter=1000	max_iter=1000		
	class_weight='balanced'	class_weight='balanced'		
	random_state=12	random_state=12		
Decision Tree	-	-	min_impurity_decrease=0.2	min_impurity_decrease=0.2

5.3.3 Application of the machine learning approach.

Once we are satisfied with the results obtained in the previous task, we use the best-performing models for each feature of the HES (see Table 5.4) and make predictions on the entire corpus, Figure 5.4 illustrates the overall prediction process.

⁴Knowledge Graph of Health Condition Evolution: <https://github.com/albamoralest/Health-Condition-Evolution-database>

First, we run predictions using the best ‘C0’ classification model (see Figure 5.4, step 1). A total of 5,174 out of 208,838 sentences were classified as providing information about condition evolution. Table 5.6, column ‘Sentence’, shows examples of sentences identified as positives.

Second, we take this reduced dataset and run an independent classification process for each dimension of the HES. The first dimension is the type of health condition: improvement, decline or permanent (see Figure 5.4, step 2). For instance, in Table 5.6 column ‘HES’, the sentence are classified as an improvement, decline and permanent. As described in HECON, only values ‘improvement’ and ‘decline’ have *progress* dimension. Thus, only 4,306 sentences, annotated as improvement or decline, were selected to complete the following two classification tasks (Pace and Duration, see Figure 5.4, steps 3 and 4). The example in Table 5.6 illustrates these cases. The sentence in the first row has a value for pace (‘moderate’) and duration (‘8 days to 2 months’), unlike the sentence in the fourth row that is classified as ‘permanent’. The output is a dataset of 5,174 sentences annotated according to the different features of the HES model.

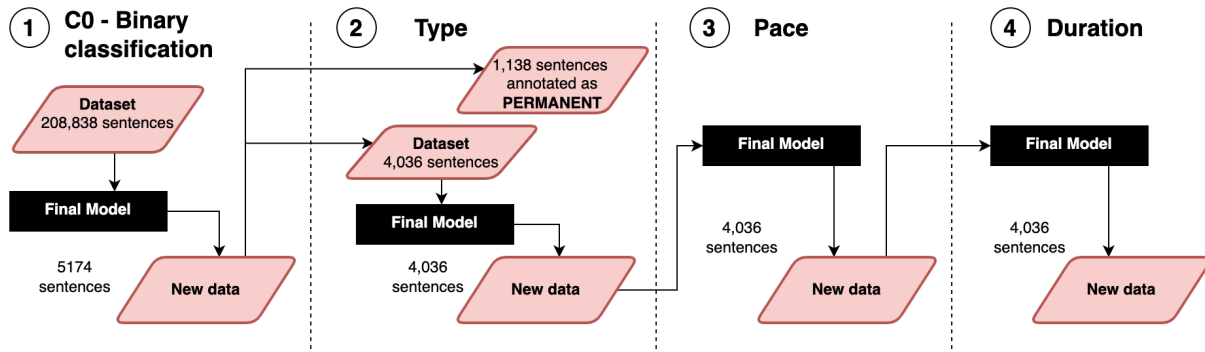


Figure 5.4: Prediction process

5.3.4 Consistency check

In this task, our objectives were: to (a) clean any inconsistencies or repeated HES that may arise from the classification and (b) produce metrics that allow the selection of the best HES (the HES statement is built from the combination of *type* + *pace* + *duration*) among the recommended annotations.

First, as some sentences were annotated with the same HES, we deleted repeated combinations of ‘*condition* + *sentence* + *HES*’, leaving us with a total of 3,635 sentences. Next, we proceed to verify that the combination of features forms a coherent HES. For example, inconsistent combinations may have a pace annotation such as ‘fast’ while duration indi-

cates a long recovery 'from 6 months to 1 year'. We rely on the pace and duration features to remove incoherent HES combinations.

Second, the classification task generates one or more recommended HES; therefore, we provide metrics to select the best statement. We use an association rule learning method⁵ to identify how likely it is for a combination of HES features to represent a health condition and rely on support and confidence scores. In the first instance, we define the *Health Condition Evolution Statement - HES* as the **union** of three sets: let T be the type of health condition, P the pace at which the condition evolves and D the duration. The HES is then represented as the union of these three elements (see Table 4.3 for a list of annotations): $HES (T \cup P \cup D)$.

In order to identify the most frequent combination of HES sets for each condition, we define the specific rule: we look at how frequently the combination of a given 'health condition' and $T \cup P \cup D$ appear in the dataset. For instance, the frequency of the combination for a given health condition '*Abdominal aortic aneurysm*' and a HES statement (e.g., '*improvement + moderate + from 8 days to 2 months*') appear in the dataset, which can be stated as $(HES \cap condition)$. In Equation 5.1, $Support(HES \cap condition)$ is calculated using the number of predictions that contain a combination of type, pace and duration for a given condition divided by the total number of predictions.

$$Support(HES \cap condition) = \frac{\text{number of predictions containing T, P and D, and condition}}{\text{total number of predictions}} \quad (5.1)$$

Next, we calculate how often the combination of *health condition and HES is valid*, and we use the confidence metric $Confidence(HES \cap condition)$. In Equation 5.2, confidence is calculated using the $Support(HES \cap condition)$ and the $Support(HES)$ value; in Equation 5.3 we also define how support for the combination of HES $(T \cup P \cup D \cup)$ it is calculated.

$$Confidence(HES \cap condition) = \frac{Support(HES \cap condition)}{Support(HES)} \quad (5.2)$$

$$Support(HES) = \frac{\text{number of predictions containing T, P and D}}{\text{total number of predictions}} \quad (5.3)$$

Table 5.6 shows a list of recommended HES for 'Chronic obstructive lung disease' ranked by confidence. The dataset resulting from this phase has a total of 1,324 SNOMED CT concepts annotated with recommended HES.

⁵Association rule learning: https://en.wikipedia.org/wiki/Association_rule_learning

Table 5.6: HES best confidence value

SNOMED Concept	SNOMED Identifier	HES	Conf.	Sentence	Source
Abdominal aortic aneurysm (disorder)	233985008	IMPROVEMENT MOD- ERATE FROM 8 DAYS TO 2 MONTHS	0.0032	Full recovery is likely to take a month or more.	MAYO
Abdominal aortic aneurysm screening (procedure)	698356002	IMPROVEMENT FAST FROM 5 MINUTES TO 1 DAY	0.0020	Screening for AAA involves a quick and painless ultrasound scan of your tummy.	NHS
Chronic obstructive lung disease (disorder)	13645005	DECLINE SLOW FROM 1 YEAR TO MORE YEARS	0.0036	Although COPD is a progressive disease that gets worse over time, COPD is treatable.	MAYO
Chronic obstructive lung disease (disorder)	13645005	PERMANENT	0.0034	There's currently no cure for chronic obstructive pulmonary disease (COPD), but treatment can help slow the progression of the condition and control the symptoms.	NHS
Cyst of ovary (disorder)	79883001	IMPROVEMENT SLOW FROM 2 MONTHS TO 6 MONTHS	0.0082	In most cases, ovarian cysts disappear in a few months without the need for treatment.	NHS
Cyst of ovary (disorder)	79883001	IMPROVEMENT SLOW FROM 2 MONTHS TO 6 MONTHS	0.0082	The majority disappears without treatment within a few months.	MAYO

5.4 Knowledge completion

The HES generated in the previous phase has limited coverage of SNOMED CT. A total of 1,324 SNOMED CT concepts which represents less than 1% of the total of SNOMED CT concepts (SNOMED CT taxonomy includes 353,567 terms). The main hypothesis that guides the knowledge completion process is that SNOMED CT concepts with similar features may also share the same health condition evolution (HES). For instance, the condition ‘Sinusitis’ (sct : 36971009) might share the same HES with ‘Acute sinusitis’ (sct : 15805002) as both concepts have the same features: ‘Finding site’ → Nasal sinus structure and ‘Associated morphology’ → Acute inflammation.

Therefore, in this step, the task is to define heuristics and encode them as rules. We take advantage of SNOMED CT taxonomy and analyse the relationships and attributes of a given concept with the aim of finding similar concepts that could share the same HES. The objective is to identify patterns and create general propagation rules that guide an automatic HES expansion from source SNOMED CT concepts (with HES) to target SNOMED CT concepts (without HES)⁶, as illustrated in Figure 5.1.

⁶The term ‘source concept’ is used to refer to a SNOMED CT concept that already has a HES annotation, and ‘target concept’ to refer to a SNOMED CT concept that has no HES assigned.

The logic model of SNOMED CT taxonomy includes components that represent two types of relationships [SNOMED International 2017] (a detailed review can be found in Chapter 2, Section 2.4.2):

- **Subtype relationship.** This is the most used relationship and is known as “*is a*” relationship or hierarchical relationship because they form the hierarchies in SNOMED CT. This means that the clinical detail of a concept increases with the depth of the hierarchies. For example, ‘Elbow fracture’ \rightarrow *is a* \rightarrow ‘Fracture of upper limb’.
- **Attribute relationship.** This relationship contributes to the definition of the source concept by associating it with defining characteristics. The characteristics are called attributes and are specified by (a) the *relationship type* and (b) the *value* provided by the destination of the relationship. For example, ‘Diabetes mellitus’ attribute \rightarrow is ‘Finding site (attribute)’ and its value is \rightarrow ‘Structure of endocrine system (body structure)’.

Using these SNOMED CT taxonomy definitions, we follow a set of steps to find patterns and derive generalised rules of propagation; in what follows, we enumerate the steps taken:

1. Select a source concept manually and analyse its features: subtype relationships such as a number of parents and descendants, and attribute relationships.
2. Analyse if the features of the source concept are shared by other target concepts. For example, sharing the same parents and similar attribute relationships or sharing only parents.
3. If identical or similar relationships (subtype or attributes) are found, then build a general query using SNOMED CT Expression Constraint Language (ECL) [SNOMED International 2022b] and retrieve all concepts sharing the identified relationships (see Section 2.4 SNOMED CT for details of ECL).
4. Select a random number of concepts retrieved after executing the query and manually verify that the results share the same HES.
5. If the target concepts correctly share the same HES, then formalise the query by building a general query.

In what follows, we describe the rules created using the SNOMED CT features. Table 5.7 presents an overview of the rules and corresponding ECL query examples.

Table 5.7: Propagation rules details.

Rules	General rule description (* source concept inherit HES to)	Example queries (ECL syntax)
Rule 1	All descendants of an administrative related concept procedure	<<120646007 Antibody screen (procedure)
Rule 2	All immediate descendants of a source concept	<! 23406007 Fracture of upper limb
Rule 3	All target concepts that share two or more attributes similar to source concept	(*):([1..1]363698007 Finding site (attribute) =<<955009 Bronchial structure (body structure) ,116676008=4532008 AND ...
Rule 4	All target concepts with one attribute and are direct descendant of the given source concept	(102482005 Growing pains (finding) OR <!102482005 Growing pains (finding)):([1..1]363698007 Finding site (attribute) =<<66019005 Limb structure (body structure))
Rule 5	All target concepts with same source parents OR source is parent AND similar attributes	((<<301098007 Heart valve finding (finding) AND <<56265001 Heart disease (disorder)) OR <<368009 Heart valve disorder (disorder)):[1..1]363698007 Finding site (attribute) = <<17401000 Cardiac valve structure (body structure)
Rule 6	All target concepts with two or more similar parents	(<!111273006 Acute respiratory disease (disorder) AND <!32398004 Bronchitis (disorder) AND <!128482007 Acute inflammatory disease (disorder))

- **Rule 1.** The hypothesis is that concepts describing administrative procedures do not affect people's health. Therefore, we identify SNOMED CT concepts that indicate administrative procedures and build queries that retrieve descendants of such concepts. For example, 'Antibody screen' is the source concept annotated as 'UNAFFECTED', and the retrieved target concepts share the same annotation such as 'Indirect platelet antibody screening (procedure)'.
- **Rule 2.** The hypothesis is that target concepts with a direct 'is a' relationship inherit the source's HES. For example, 'Elbow fracture' *is a* 'Fracture of upper limb' and therefore inherits the source's HES.
- **Rule 3.** In this case, the target concept with two or more attributes similar to the source concept inherits the HES. For example, 'Acute bronchitis (disorder)' shares its HES with concepts with similar attributes (e.g. concepts with Finding site, Associated morphology and Clinical course attributes).
- **Rule 4.** For the cases where the source concept has only one attribute, target concepts should additionally be a direct descendant of the source concept. In this way, we avoid expanding to broad terms that are not related to the source concept. For instance, 'Growing pains' has one attribute, 'Finding site'. Therefore, we add this constraint to retrieve closer related target concepts.

- **Rule 5.** In this case, a target concept with the same parents and attributes as the source concept inherits the HES statement. For instance, ‘Inflammation of aortic valve (disorder)’ inherits the ‘Abnormality of aortic valve (disorder)’ statement.
- **Rule 6.** Here, a target concept with two or more similar parents as the source concept inherits the HES. This rule does not take into account concepts with one parent only because the retrieved concepts are too broad.

This set of rules supports two tasks: (1) identify the source concepts that could inherit their HES annotations to other target concepts. For instance, different rules apply to certain source concepts that meet requirements in terms of the number of attributes, the number of parents or the type of branch. For example, source concepts with only one attribute will only be eligible to execute Rule 4. The second task (2) is the automatic construction and execution of ECL queries. We automatically build the ECL queries based on the requirements and restrictions described above and summarised in Table 5.7. For each source concept, we execute the ECL queries and obtain a list of candidate target concepts that could inherit the HES annotations. To execute the ECL queries, we use the SNOMED CT API server for terminology querying, Snowstorm⁷.

By using these rules, we manage to annotate 96,253 different concepts in total, 27% of SNOMED CT taxonomy. Until now, we have used semi-automatic processes to generate HES recommendations; however, we understand that knowledge components generation and knowledge completion process are prone to unintentional errors or generalisations; therefore, we recognise the necessity of validating the extraction process and results generated by experts. The following section describes the process envisioned to ensure the construction of a high-quality database of health condition evolution.

5.5 Human-in-the-loop

In order to scale up the construction of the health evolution KG and build high-quality data, it is imperative to include domain experts in the loop; therefore, the last step of the pipeline focuses on capturing human knowledge (see Figure 5.1).

This knowledge can be captured in three ways, by providing experts with (a) a list of recommended HES for each condition or (b) a list of recommended target concepts that can

⁷Snowstorm provides the terminology server API for the SNOMED International Browser, including the International Edition and around fourteen national Extensions - <https://github.com/IHTSDO/snowstorm>

share the same HES as the source concept; thus, they can assess the most accurate option swiftly. Also, experts can (c) build a new HES according to their best judgement.

We provide experts with a tool that reflects the options described above. The first interface displays the name of a condition and the list of candidate statements obtained in the knowledge components extraction (see Section 5.3). Experts' task is to select the 'Correct' HES according to their best judgement. The second interface uses the responses generated in the previous interface and the output from the knowledge completion step (See, Section 5.4). The tool displays a source concept, the HES that was selected as 'Correct' in the first interface and the recommended conditions that could share the given HES. Experts' task is to assess and select as 'Correct' all the target concepts that effectively share the same HES as the source concept. When there is no recommendation available, experts can use a third interface and input a new HES using the different elements of the statement (type, pace, duration).

In this way, we secure that the HES generated is evaluated by domain experts, ultimately meeting our goal of building a quality database of health evolution.

The final Knowledge Acquisition pipeline's output is a curated Health Evolution Statement (HES) database linked to its corresponding SNOMED CT concept. The KA pipeline is reproducible, and all the resources are available in the repository⁸.

5.6 Knowledge Graph

In order to make the newly created database available in a structured and machine-readable format, we built a Knowledge Graph following the HECON Ontology presented in the previous chapter and stores the data generated as result of executing the aforementioned approach. The Health Condition Evolution Ontology is a formal representation of the evolution of health events over time. Each HES in the curated database is linked to a SNOMED CT concept identifier; likewise, each SNOMED CT concept could be linked to one or more HES. The KG also stores data that represents the relationships between the data sources (MAYO Clinic and NHS England) and the process used to generate the annotation (knowledge component extraction, knowledge completion or HITL). This structured information is a fundamental component in order to reason on the evolution of health conditions over time and the identification of ongoing health issues from EHR.

The KG was built using SPARQL Anything [Daga, Asprino, et al. 2021], a system that

⁸Repository: <https://github.com/albamoralest/Health-Condition-Evolution-database>

allows generating RDF data from different types of formats with plain SPARQL CONSTRUCT queries⁹. Finally, we store RDF data in TTL data format and Blazegraph¹⁰, Listing 5.1 exemplifies a HES triple part of KG.

Listing 5.1: KG data - triple example

```

1 ont:ef5bf5170cc99fe754329997c93f269e
2     rdf:type          hecon:Improvement , hecon:HealthEvolutionStatement ;
3     hecon:hasMaxDuration <http://kmi.open.ac.uk/conrad/kg/hecon/2/MONTH> ;
4     hecon:hasMinDuration <http://kmi.open.ac.uk/conrad/kg/hecon/8/DAY> ;
5     hecon:hasPace       hecon:MODERATE ;
6     hecon:hasSctConcept sct:233985008 ;
7     prov:wasDerivedFrom <http://kmi.open.ac.uk/conrad/kg/hecon/MAY0/
8     e81b2e02f7a379011a000e6fdaba97a2> ;
9     prov:wasGeneratedBy <http://kmi.open.ac.uk/conrad/kg/hecon/activity/
10    extraction/ef5bf5170cc99fe754329997c93f269e> .

```

5.6.1 Data statistics

The resulting Knowledge Graph contains 12,062,340 triples. From these triples, 96,253 represent unique SNOMED CT concepts and 355,551 HES. A SNOMED CT concept is linked to one or more HESs using the corresponding SNOMED CT identifier. Figure 5.5 displays an example of the SNOMED concept ‘Abdominal aortic aneurysm (disorder)’ as part of the KG.

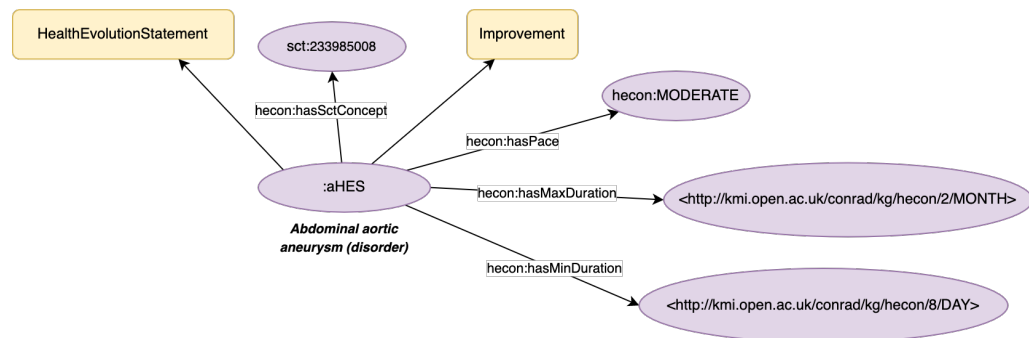


Figure 5.5: HES KG representation

Table 5.8 presents a summary of the number of entities by each OWL class. For instance, from the total number of HES in the KG, 64% are of type improvement (230,953 in total), 13% decline (44,332 in total), 21% permanent (73,823 in total) and 2% unaffected (6,443 in total).

⁹All the queries and results are available in the HECON repository

¹⁰Blazegraph <https://blazegraph.com/>

Table 5.8: Core. A number of instances per class.

	HES	Improvement	Decline	Permanent	Unaffected	time:Duration
Total Entities	355,551	230,953	44,332	73,823	6,443	8

In terms of provenance information, the activities that generate the HESs are the *Health Evolution Extraction Algorithm* (a Knowledge Extraction activity) and the *Knowledge Completion Process* (a Rule Execution activity) and the participation of domain experts (a User Study activity). We described each activity in detail in Chapter 4, section 4.4.2. A total of 1,264 SNOMED CT concepts have at least one HES generated by the Health Evolution Extraction Algorithm activity, whereas 96,253 SNOMED CT concepts have at least one HES generated by the Knowledge Completion Process activity (Rule Execution). Because the Knowledge components identification step (represented by the class Knowledge Extraction) generates one or more HES recommendations, we have a total of 2,805 entities in the KG. The same applies to the Knowledge completion task, represented by the RuleExecution class; two or more rules could expand a HES to the same SNOMED CT concept; therefore, the KG contains a total of 355,518 entities. Table 5.9 presents the summary statistics for each class. Additionally, two organisations are represented: NHS England and MAYO Clinic.

Table 5.9: Provenance. Number of instances per class.

	Activity	Source	Knwl. Ext	Rule Exc.	User Study
Total Entities	358,323	2,577	2,805	355,518	16

The number of SNOMED CT concepts represented in our KG is approximately 27% of the total of concepts in the SNOMED CT taxonomy (352,567 concepts, January 31, 2020 publication).

5.7 Conclusions

In this chapter, we developed an original knowledge acquisition pipeline to build a database of health evolution information. The approach included the implementation of knowledge components such as text classification and completion. It also includes a Human-in-the-loop step to capture knowledge from domain experts.

The main goal of this work was to build a curated and high-quality database of health condition evolution. Our work has demonstrated that extracting health evolution statements

(HES) from natural language is a feasible task. Furthermore, exploiting SNOMED CT features accelerates the production of recommendations, hence the coverage of SNOMED CT.

A key strength of this research was the inclusion of the Human-in-the-loop module. The inclusion of domain experts as part of the methodology accelerates the construction of the KG. More importantly, it ensures the capture of their valuable and accurate knowledge. With these results, we fill a gap in the literature and provide structured data on health evolution.

Chapter 6

Reasoning with Health Condition

Evolution: the CONRAD system

In our research, we followed a knowledge engineering approach that supports the development of core knowledge components required to recognise ongoing health issues from information contained in electronic health records. These components are:

- a model that formally represents the features that compose the evolution of health conditions;
- a database with information about health conditions' evolution and their estimated convalescence time.

The identification and implementation of these components answer two of our research questions. We demonstrated that it is possible to collect and extract data on how health conditions progress over time (RQ2) and that by using our proposed model, this information can be formally represented in a machine-readable format (RQ1).

In this chapter, we concentrate on the core task of the identification of ongoing health issues derivable from electronic health records to benefit emergency services in the Smart City context. Our proposed approach integrates the knowledge components listed above and intends to answer the research questions RQ3 and RQ4. For this purpose, here we focus on three main tasks:

- Design a knowledge component that uses HECON and the Knowledge Graph of health condition evolution to *reason* automatically on health records and extracts information about current or progressing health issues.

- Design a method and the intelligent system software architecture based on the requirements elicited in Chapter 3 that integrates the knowledge components developed into a unified end-to-end solution.
- Develop a prototype of an intelligent system that demonstrates the proposed architecture.

The Intelligent System aims to be a prototypical instantiation of our approach, which we use to test our method in a paradigmatic environment of Smart Cities for reasoning automatically on EHR and recognising current health issues.

In the next section, we describe the context of a Smart City where we envision implementing our proposed solution. Section 6.2 reports on the characteristics of the electronic health records dataset used for instantiating the intelligent system. Section 6.3 moves on to the design of the reasoning module in charge of identifying ongoing health issues. Section 6.4 summarises the design of the intelligent system architecture. Finally, Section 6.5 is dedicated to describing the Health Condition Radar - CONRAD, an instantiation of the proposed design based on the knowledge requirements extracted from the analysis of the expanded scenarios. The design illustrates the data flow between healthcare providers and emergency services and assists in identifying people in vulnerable situations during fire emergencies.

6.1 Context: the Smart City scenario

Electronic health records (EHR) have been adopted as a vital source of information for emergency support systems [Hussain et al. 2015; Cook et al. 2018] as they contain information about people's medical events (for instance, recent procedures or conditions), revealing ongoing health issues and identifying people in a vulnerable situation. For instance, a record indicating a fracture in one of the lower extremities, if recent, may suggest that a person requires special assistance to evacuate during a fire emergency. However, identifying relevant information could become a challenging task for emergency responders. Processing extensive, specific and sensitive healthcare data could be time-consuming; furthermore, critical health issues could be overlooked and challenging to interpret [Morales Tirado, Daga, and Motta 2020; Morales Tirado, Daga, and Motta 2021].

The adoption of the Smart City paradigm and the implementation of smart solutions provides a platform to access and analyse large data sources effectively; in an emergency

scenario, this translates into providing up-to-date information that can support emergency services operations.

To solve the problem of identifying people requiring assistance during a fire emergency, we propose a solution that benefits from the iteration of the different elements of a Smart City environment. Our solution uses Semantic Web technologies to develop the knowledge components required to reason on health conditions' evolution over time and to process electronic health records. In broad terms, we envision an intelligent system able to process electronic health records of people involved in a fire emergency in order to distinguish ongoing health issues that potentially require assistance to evacuate.

In our scenario, a fire starts on the fourth floor of a building, emergency services are alerted, emergency evacuation plans are executed. In a typical scenario, firefighters collect information about people requiring assistance from witnesses or fire wardens; often, this information is scarce.

In the Smart City environment, the interaction between smart systems leverages the collection of such information. For instance, an organisation's Emergency Alert System (EAS) can automatically gather information about people on premises by querying the building's Access Control Systems (ACS) records. The EAS will immediately notify firefighters about the number of people involved and general information about the incident. At the same time, the EAS pass on employees' information to the healthcare service provider (HS); thus, this organisation can identify patients and retrieve their EHR.

Then, the healthcare service provider should assess and decide what EHR information to exchange with emergency services. To this purpose, our proposed intelligent system CONRAD - Health Condition Radar [Morales Tirado, Daga, and Motta 2022a], acts as a service that assists a data engineer in the task of discriminating the information to be shared with emergency services. CONRAD analyses the EHR of each person and identifies health issues that are valid or ongoing at the moment of the emergency. The person may require special assistance if at least one condition is tagged as valid. Furthermore, the system processes the data and, ideally, classifies the relevant health events according to UK governmental guidelines on types of disabilities [UK Government 2007b; UK Government 2008] which ultimately reduces the amount of sensitive information analysed and exchanged. The system delivers the findings to the data engineer or sends them directly to the firefighters attending the emergency. Figure 6.1 illustrates the flow of information and the iteration among the described elements.

The final output is a list of people with ongoing health issues that potentially require

assistance to evacuate during a fire emergency; this information supports emergency services in the planning and execution of rescue operations.

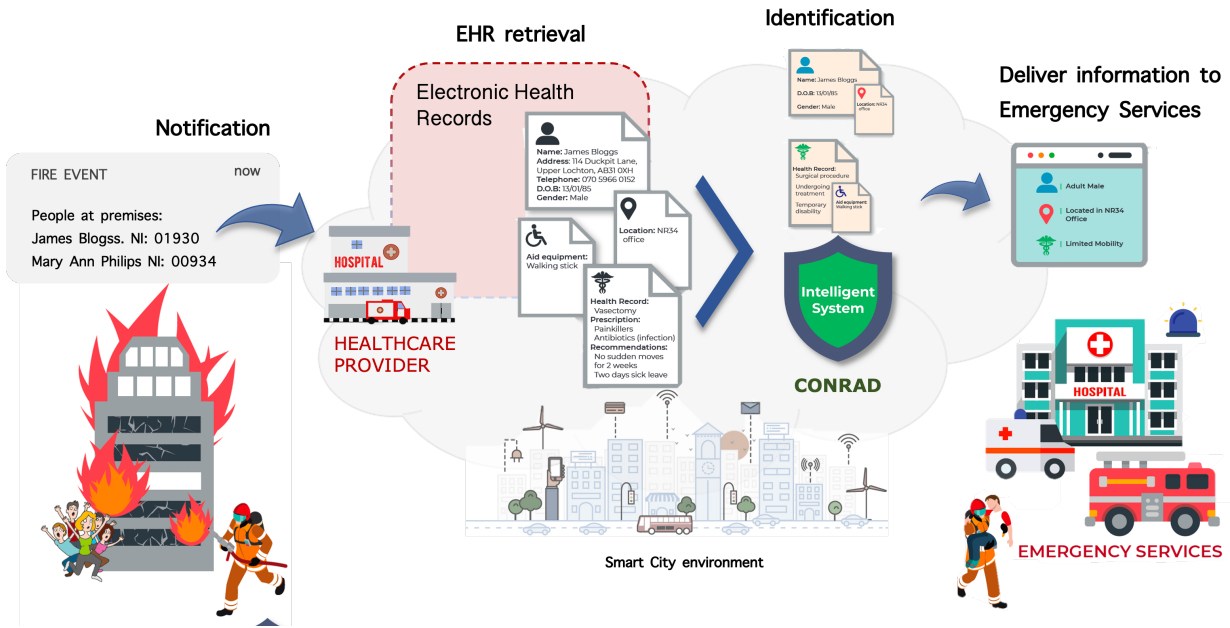


Figure 6.1: CONRAD - data flow and implementation setting

In our demonstration, we use synthetic EHRs generated using Synthea software [Walonoski et al. 2018]. The EHR is encoded employing established standards, such as the Fast Healthcare Interoperability Resources (FHIR) [HL7 2019], for the exchange of EHR, and the Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) [SNOMED International 2022c] for standard concept descriptions. The following section is dedicated to describing the electronic health records dataset and the elements used by CONRAD to infer health issues that may affect the patient during an emergency.

6.2 Dataset of health records

Health-related data is considered highly sensitive information. Therefore, a good research practice is to make use of synthetic data [Walonoski et al. 2018; Hernandez et al. 2022]. Hence, in our research, we use synthetic healthcare data to prevent any disclosure of private information. Synthea [Walonoski et al. 2018] is open-source software that generates synthetic electronic health records. Synthea generates a comprehensive medical history of synthetic patients that covers their lifespan. Synthetic electronic health records are extensive as they include demographic data and medical observations such as appointments, patient conditions, procedures, care plans, medication, allergies, and observations. We decided to represent the data using FHIR (Fast Healthcare Interoperability Resources) standard spec-

ification for exchanging healthcare information electronically. Specifically, we adopted a Linked Data approach and represented the synthetic dataset using the FHIR ontology¹.

In what follows, we focus on analysing the electronic health records generated using Synthea software and report on the attributes used to perform the evaluation of health conditions.

6.2.1 FHRI representation and annotation

We generated a sample of 10,000 patients; we intended to create a diverse dataset of synthetic patients that included as many diseases as possible. The patients' ages range from 20 to 80 years old, as we try to simulate the ages of employees of a large organisation. From the generated dataset (in CSV format), we identify 120 attributes grouped into 12 different files (see Appendix E.1 for details of attributes), each one representing an FHIR resource type. In total, the synthetic dataset is represented by 178,598,778 triples and was stored using Blazegraph². Figure 6.2 illustrates the representation of the EHR for our synthetic dataset.

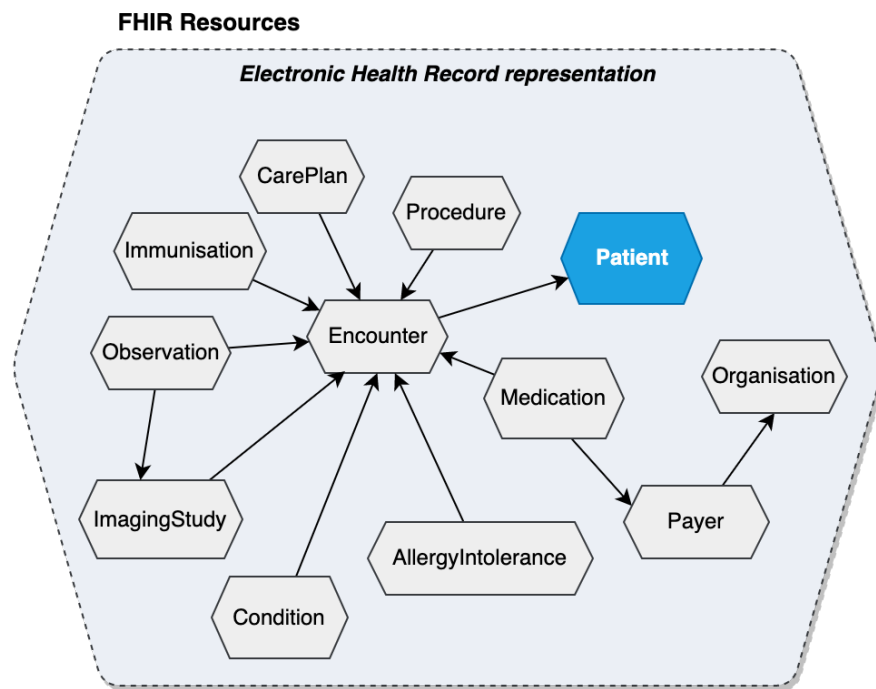


Figure 6.2: Synthetic EHR representation using FHIR resources

We inspect the content of our health record dataset and its data schema to reason over the data's features that can help answer our main research question. Crucially, we want to distinguish useful data features; particularly, we observe that data points³ describing health

¹We used Apache Jena API: https://jena.apache.org/tutorials/rdf_api.html#ch-Jena-RDF-Packages

²Blazegraph: <https://blazegraph.com/>

³The term *data point* refers to an event or entry in electronic health records.

conditions that have a temporal validity. In what follows, we summarise some of the findings:

- The health records have a **timestamp** that makes it easy to identify recent encounters, conditions or procedures which could be of interest.
- Data from encounters, conditions, procedures, medications, and allergies often record only the start date and no end date, which could represent an ongoing issue/situation that may be relevant.
- The descriptions of any medical event are encoded using SNOMED CT taxonomy.
- Healthcare data is considered sensitive personal data according to the GDPR [European Parliament 2016]; therefore, it is imperative to balance the trade-off between utility and sensitivity.

There are infinite considerations that could be made about potential utility in the context of an emergency. Here we focus specifically on identifying if a condition is still ongoing when a fire emergency occurs; therefore, we concentrate on the use of attributes such as the start date (typically represented as `fhir:Period.start` or `fhir:*.recorded`⁴) and the description of the medical event, e.g., condition, allergy, procedure, care plan, encounter (represented by `fhir:*.code`⁵).

The following section moves on to describe in greater detail how we use these attributes and the knowledge components developed previously to derive valid health records at a certain point in time.

6.3 Health Condition Evolution Reasoning on health records validity

This section explains the implementation of the last component required to identify ongoing health conditions from EHR, the Knowledge Reasoning (KR) component. The KR uses the health condition evolution statements (HES) to predict whether a specific condition holds at the time of an emergency. In what follows, we first describe the logical rules built using the HES dimensions and later illustrate their use by taking examples from the dataset.

⁴We use * to indicate that the entity name changes according to the type of resource.

⁵We use * to indicate that the name of the entity changes according to the type of resource. E.g., `fhir:Allergy.code` or `fhir:Procedure.code`

Typically, a condition recorded in a health record is relevant if it is still ongoing. A record is relevant in any of these three cases: recovery time has not passed yet, the condition is chronic, or the situation deteriorates in time.

We use the HES feature ‘type’ with two purposes: a) to identify the *relevant* data points and b) to guide the reasoning on a time range. In what follows, a description of the logic involved for each annotation (defined by HECON ontology) is given:

1. **UNAFFECTED**: Identifies health conditions that do not affect a person’s health. Therefore, if the type’s annotation is ‘Unaffected’, then the system marks this health record as *‘not relevant’*.
2. **PERMANENT**: Represents a chronic condition; therefore, if a condition has an annotation ‘Permanent’, then the situation is *always valid*.
3. **IMPROVEMENT**: Describes a health condition that lasts for a certain amount of time. The HES represents the convalescence time using the annotations FROM and TO. Typically, in the best-case scenario, the convalescence period could last a minimum time (FROM) or lower bound (LB). In the worst-case scenario, a maximum time (TO) or upper bound (UB). (see Figure 6.3a)
4. **DECLINE**: This annotation describes a condition that deteriorates over time. For instance, a person that has good health but shows certain symptoms. At some point, the person is diagnosed (a record is created: RD). This person receives care, but due to the nature of the condition, the person’s health worsens (FROM/LB). After some time it becomes chronic or permanent (TO/UB) (see Figure 6.3b).

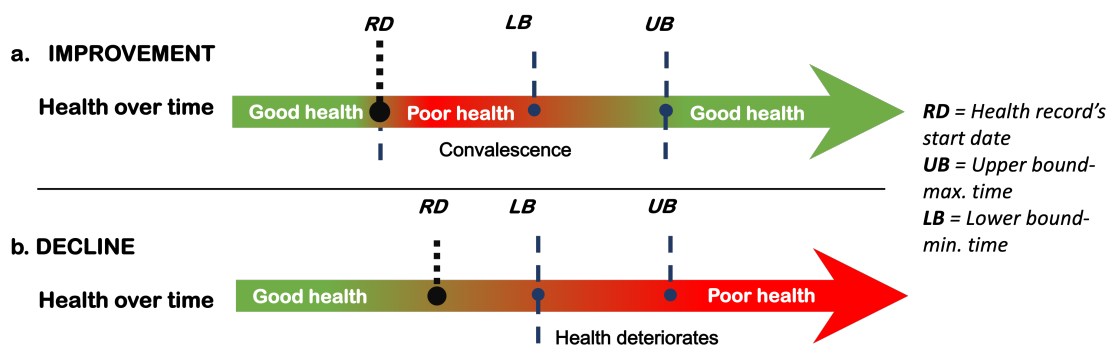


Figure 6.3: Reasoning on HES: Direction interpretation.

After evaluating the type of condition, we move to reason on ‘time range’ and how a system should interpret it in combination with the feature direction. The time range dimension has two elements: a lower bound (LB or FROM) value and an upper bound (UB or TO)

value. The validity of an event registered in the health record depends on whether the emergency happened before, in between or after these two boundaries. These cases are explained under three headings, which are:

1. **The emergency happens after the UB (TO).** On the one hand, if type annotation is IMPROVEMENT, then the convalescence period has ended (see Figure 6.4a). Thus the health record is *not relevant*. On the other hand, if the direction is DECLINE, then the condition has become permanent; hence the health record indicates a current health issue (see Figure 6.4b), and therefore, it is *relevant*.
2. **The emergency happens in between LB (FROM) and UB (TO).** The health record is *always relevant*.
3. **The emergency happens before the LB (FROM).** In the case of IMPROVEMENT annotation, the condition is valid as the person is still suffering from the given condition (see Figure 6.4c). In the case of DECLINE, it is possible that the person may be showing symptoms related to a condition, but the deterioration process starts after the LB (see Figure 6.4d).

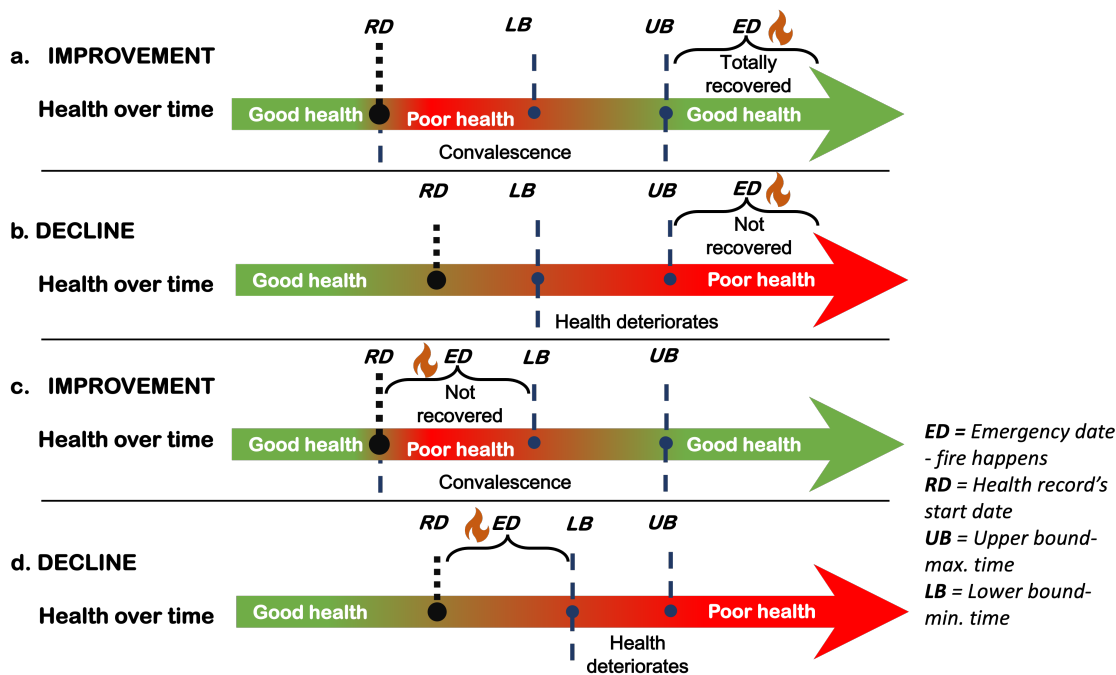


Figure 6.4: Reasoning on HES: Time range interpretation.

As seen previously, a single event registered on the EHR has two key attributes: the start date or the date it was recorded (RD) and the description of the medical event that is a link to a SNOMED CT identifier. Table 6.1 provides examples of events found in the EHR of three

different people. In order to assess the validity of these records, the system uses these two attributes, the date the fire started and the rules described above.

Let's examine how a given data point is evaluated using the KR component. First, the system retrieves the HES using the SNOMED CT identifier stored in the EHR and proceeds to analyse the statement feature: 'type of condition'.

1. If the type of condition annotation is 'Unaffected', then the record is marked as 'not relevant', and no further process is required.
2. On the contrary, if it is 'Permanent', it is always valid; in Table 6.1 the second row belongs to a wheelchair user's health record. The condition 'Fracture of the vertebral column' identifies a person with mobility issues, therefore, requiring assistance.
3. If the type of condition is 'Improvement' or 'Decline', then the system evaluates the time duration and the number of days that have passed between the health condition start date and the emergency date. The system uses these dates to estimate if the fire occurred before LB, between LB and UB, or after UB.

For example, suppose the condition 'Sprain of the ankle' (Table 6.1, row three) improves around eight days (LB) to two months (UB). If the fire emergency occurs at some point before the LB (eight days), then the system estimates the condition is valid (see Figure 6.4c); therefore, the person might need assistance. On the contrary, if the fire happens after UB, for example, three months after this event was recorded, the system estimates that the condition is not valid, therefore, does not require assistance.

Table 6.1: Reasoning on EHR - examples

Condition	Date	Quartile	Severity score	HES annotation
Chronic obstructive bronchitis (disorder)	05/10/2019	LB	1	IMPROVEMENT MODERATE FROM 2 MONTHS TO 6 MONTHS
Fracture of the vertebral column with spinal cord injury	08/04/2008	LB	1	PERMANENT
Sprain of ankle	03/10/2019	Q2	3	IMPROVEMENT MODERATE FROM 8 DAYS TO 2 MONTHS

6.3.1 Severity score

The HES representation provides the opportunity to assess if a person is recently recovering from a health condition and, therefore, likely to require assistance. For instance, if a condition improves with time, the closer the emergency date is to the minimum recovery time, the higher the probability of requiring assistance (see Figure 6.5a). On the contrary, if the condition declines, the closer the emergency date is to the minimum time for the condition to start deteriorating, the lower the probability of requiring assistance (see Figure 6.5b). If a condition is PERMANENT, the condition is ongoing; therefore, severity is always the highest.

In order to calculate the *severity score*, we establish a six-level system. Level one represents the period before LB, levels two to five are the quartiles between LB and UB, and finally, level 6 is the period after UB. The system calculates the severity score according to the proximity to the ‘Poor health’ zone and assigns the severity score accordingly. Figure 6.5 summarises the use of quartiles for severity score calculation.

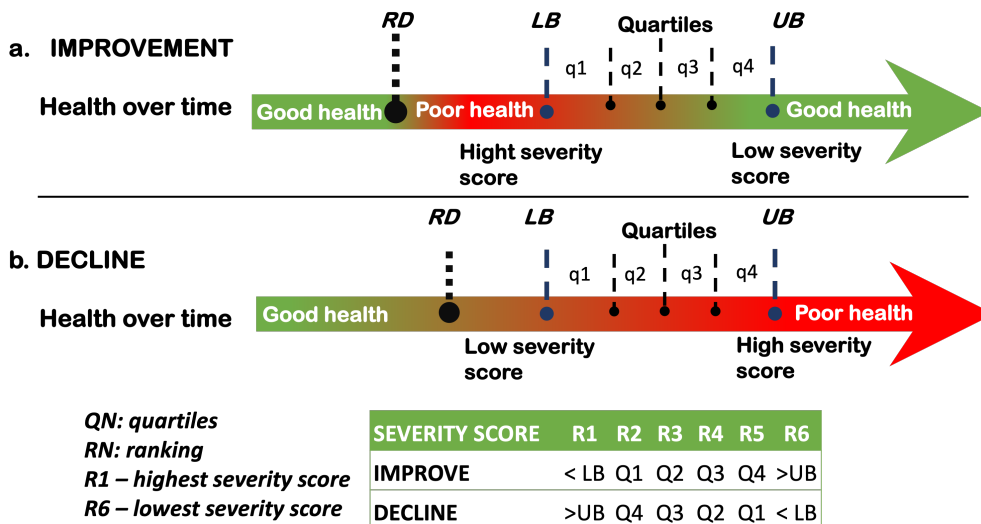


Figure 6.5: Reasoning on HES: Severity score.

Following the examples from the previous section, Table 6.1 shows the severity score for the wheelchair user (row three) and the person with lung disease (row one). In both cases, the severity score is one, the highest score. The ‘vertebral fracture’ indicates a PERMANENT condition; therefore, it is always relevant and has high severity. Although ‘chronic bronchitis’ improves after a certain period, in the example, the condition started recently (considering a fire emergency 25 days after it was first recorded), closer to LB. We can assume that the person might have difficulties breathing or walking; therefore, the system assigns a high score.

6.3.2 Supporting interpretation

In this section we describe the initial work developed to support the interpretation of the EHR identified as valuable to support emergency services. Health data typically require a level of knowledge in order to interpret health conditions and their implications. Furthermore, first responders execute their activities under time and pressure constraints. For these reasons, we explore an approach to facilitate the analysis of the EHR identified as useful during the previous steps. The final objective is to provide information that could ideally describe if a person's EHR are related to a type of disability, in this way supporting firefighters during a fire evacuation.

To support the interpretation task, we use the categories that represent disabilities (listed in Table 6.2) according to UK guidelines to identify people with disabilities during fire emergencies [UK Government 2007b; UK Government 2008]. In order to bridge the gap between the categories in the classification and the description of the health records, we use a common-sense knowledge base: ConceptNet⁶.

Table 6.2: Types of disabilities and correspondent Key Concept.

Category Description [UK Government 2007b]	Key Concept
Electric wheelchair and wheelchair user	<i>wheelchair_user</i>
Mobility impaired person	<i>movement_disorder</i>
Asthma and breathing issues	<i>respiratory_disease</i>
Visually impaired person	<i>visual_impairment</i>
Dyslexic and orientation disorders	<i>disorientation</i>
Learning difficulty and autism	<i>learning_difficulty</i>
Mental Health problems	<i>mental_health_problem</i>
Dexterity problems	<i>indexterity</i>
Hearing impaired person	<i>hearing_impaired</i>

First, for each category in the list of disabilities, we manually find a key concept in ConceptNet that represents it (see Table 6.2, column Key Concept). Second, to match the health record data points with the most related type of disability, we use the ConceptNet API

⁶ConceptNet: <https://conceptnet.io/>

Table 6.3: Ranked top 3 reasons for assistance

rank	category	score
1	Asthma and breathing issues	0.368
	<i>/relatedness?node1=/c/en/respiratory_disease&node2=/c/en/injury_of_tendon_of_the_rotator_cuff_of_shoulder</i>	0.126
	<i>/relatedness?node1=/c/en/respiratory_disease&node2=/c/en/pulmonary_emphysema</i>	0.610
2	Electric Wheelchair and wheelchair user	0.201
	<i>/relatedness?node1=/c/en/wheelchair_user&node2=/c/en/injury_of_tendon_of_the_rotator_cuff_of_shoulder</i>	0.371
	<i>/relatedness?node1=/c/en/wheelchair_user&node2=/c/en/pulmonary_emphysema</i>	0.031
3	Mobility impaired person	0.198
	<i>/relatedness?node1=/c/en/movement_disorder&node2=/c/en/injury_of_tendon_of_the_rotator_cuff_of_shoulder</i>	0.103
	<i>/relatedness?node1=/c/en/movement_disorder&node2=/c/en/pulmonary_emphysema</i>	0.293

[Speer, Chin, and Havasi 2017]. ConceptNet provides an API that allows us the comparison of two terms, and it returns a ‘relatedness value’ indicating how connected the two terms are; the higher the value, the more related each pair of terms is. Hence, we query the API to obtain the relatedness value between each valid data point and each key concept. For example, in Table 6.3, we obtain the relatedness value for two data points: injury of tendon and pulmonary emphysema against the first three categories of disabilities.

After comparing all valid data points against the types of disabilities, our system calculates the average score, allowing us to deliver a ranked list of the most related types of disabilities associated with the time-valid condition extracted from the health record. In this case, the average is considered a suitable measure as it allows us to obtain a representative value of the most likely type of disability connected to the list of all relevant conditions, meaning that relevant conditions’ scores are evaluated together and not in an isolated manner. For instance, in Table 6.3, ‘*pulmonary emphysema*’ is clearly related to an *Asthma and breathings issues* type of disability, and in combination with an ‘*injury of tendon*’, this person might also have some sort of mobility issue and be related to the use of an electric wheelchair.

6.4 Intelligent System architecture

This section presents the system architecture proposed to derive ongoing health issues from electronic health records. The architecture illustrates the flow of information that receives as input electronic health records and delivers valuable data to emergency responders; it is composed of four components, as shown in Figure 6.6.

The *first component* is the Health Evolution Ontology (HECON). The ontology is a formal representation of entities and relationships that characterise the definition of health progress or evolution over time. The ontology also supports the link to the SNOMED CT taxonomy (a clinical terminology scheme representing medical terms) and facilitates the reasoning to detect ongoing health issues. The *second component* is the Knowledge Graph (KG), which is a structured compilation of information about health evolution information.

The *third component* is the Health Evolution Reasoner module (HER). This module contains the rules for reasoning on health evolution and evaluates if a health condition (represented by a SNOMED CT identifier) is ongoing at a certain point in time. It performs this task using the HECON Ontology and the KG of Health Evolution Statements (HES). The module receives as input the electronic health records; then, it processes each data point (a record in the EHR) applying the reasoning rules and returns as output the data points that indicate an ongoing health issue.

The *fourth component* is the Data fitting module (DF). The module performs two tasks, (a) evaluates the severity of the health condition according to the time that has passed since it was first reported and the date of the fire; and (b) includes information about the type of disability that might be related to the health issue. The final output is a reduced list of patients' data points, each representing a current health condition, indicating a potential impediment to performing an evacuation plan.

6.5 CONRAD - Health Condition Radar

CONRAD is the prototype system that demonstrates the proposed architecture. This section reports on the scenario in which CONRAD leverages the identification of people in a vulnerable situation during a fire emergency evacuation.

The prototype system uses synthetic EHRs generated with Synthea software [Walonoski et al. 2018] and described in section 6.2. The EHR is encoded by employing (FHIR) [HL7 2019] and SNOMED CT [SNOMED International 2017] for standard concept descriptions.

In case of a fire emergency, CONRAD would be able to access up-to-date electronic

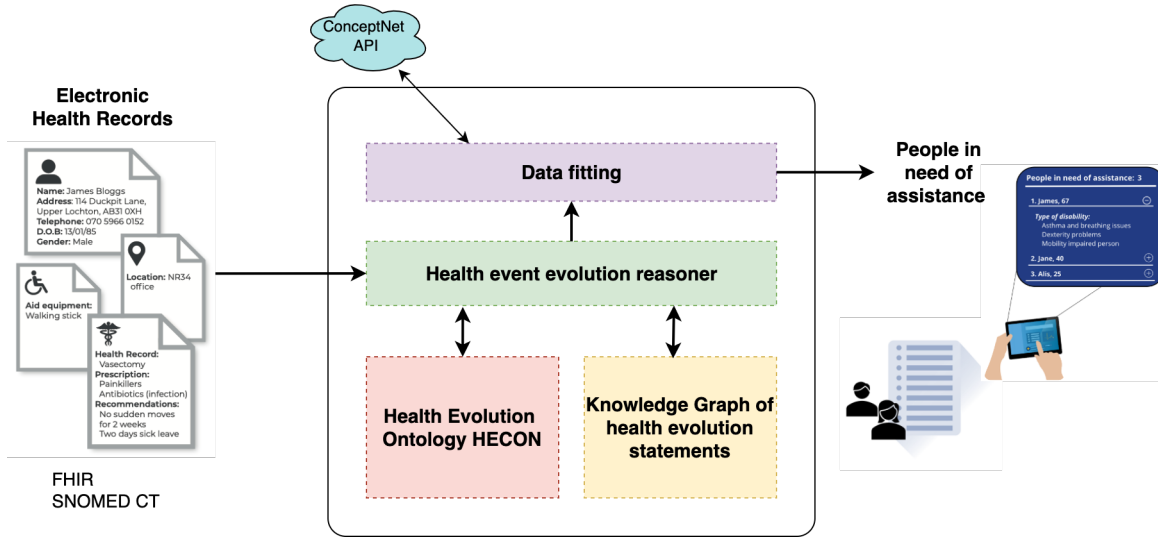


Figure 6.6: CONRAD system architecture.

health records. We relied upon the **Assumption 1** (Section 1.2) describing a Smart City environment in which an organisation’s Emergency Alert System (EAS) notifies immediately about the fire to emergency services. This notification includes the delivery of data to the Healthcare Service (HS) provider (acting as the data controller, the organisation that collects and decides on the management of the data) about people present at the premises when the fire happened. Because the HR acts as the data controller, it analyses the EHR in order to identify people requiring assistance and consequently, it is in charge of notifying firefighters.

Once CONRAD has access to EHR, it starts analysing each data point or entry in the EHR. The Health Evolution Reasoner (HER) module is responsible for this task. The reasoner uses the condition description, in this case, linked to a SNOMED CT identifier to retrieve the health evolution description (a Health Evolution Statement - HES); it uses SPARQL queries and the Knowledge Graph of Health Evolution Statements (HES). For instance, if a person’s EHR contains an entry such as ‘Fracture of ankle’ (`sct:16114001`), the reasoner module uses the SNOMED CT identifier to retrieve the HES; in this case, ‘*Improvement Moderate from 6 weeks to two months*’.

The HES is composed of three dimensions: type of condition, pace and time duration. In the first instance, the reasoner analyses the type of condition dimension (see Figure 6.3). In our example, as the type is ‘Improvement’, the system initially considers this data point as relevant and proceeds to review the ‘*time duration*’ dimension.

For the time duration evaluation, the system calculates if the fire happened before the minimum recovery time (FROM or LB), after the maximum recovery time (UB), or between LB and UB. Continuing with our example, if the fire happens eight weeks after the health record date (RD), then the condition is likely to be still ongoing as it falls between six weeks

(LB) and two months (UB). In this example, the system estimates that this data point is *relevant*.

The intelligent system will repeat this process for each patient's health record. The module's output is a reduced number of health records estimated to be ongoing or in progression at the moment of the fire. Only patients that report at least one relevant data point will be considered as requiring assistance.

The reasoner's output is then passed to the Data Fitting module (DF). This module is in charge of providing a quantifiable measure of how current is the health condition and information on the type of disability related to it. In our example, the progression falls in Q1, meaning a severity score of two; see Figure 6.5. Finally, the type of disability related to the given data point is *Mobility impaired person* once calculated using ConceptNet.

The system's final output is a list of people requiring assistance, their ongoing condition(s), and an estimation of the severity and related information on the type of disability. This information is delivered to the Health Service provider, who then decides how to share the information with emergency services.

6.6 Conclusions

In this chapter, we presented the proposed architecture of an intelligent system capable of reasoning about health records to identify relevant information and support firefighters during a fire emergency. We developed the knowledge component responsible for automatically estimating valid health records. The knowledge component contains the formal reasoning rules for analysing health records, which answers RQ3.

We described the software architecture design that uses knowledge components such as HECON ontology, the Knowledge Graph of HES and the knowledge reasoner to support this task. The proposed design integrates all knowledge components into an intelligent system capable of identifying people requiring assistance during an emergency, which answers RQ4.

We developed an intelligent system, the Health Condition Radar - CONRAD, an intelligent system software prototype that implements the proposed architecture to process Electronic Health Records (EHR) and performs the automatic identification of people in a vulnerable situation during an emergency. CONRAD was developed following the three general requirements detailed in Chapter 3.

Chapter 7

Evaluation

In the previous chapter, we explained how emergency services could be supported in the task of evaluating if people require special assistance during a fire emergency. We designed an intelligent system architecture that integrates knowledge acquisition, representation and reasoning components to facilitate the automatic identification of ongoing issues from electronic health records data; furthermore, we instantiated our approach in a paradigmatic environment of Smart Cities.

In Chapter 6, we focused on the technical feasibility of the approach built on the integration of these knowledge components developed in previous chapters. These include the HECON Ontology (Chapter 4), the creation of a high-quality Knowledge Graph of health condition evolution (Chapter 5), the rules of reasoning on EHR and the instantiation of an end-to-end solution (Chapter 3 and Chapter 6) that supports the proposed approach. So far, we have relied on two hypotheses: that **(a)** it is possible to design an intelligent system that uses electronic health records to analyse, extract and deliver relevant information to emergency services and that **(b)** this analysis is facilitated by the evaluation of health conditions' progression over time.

However, each of these hypotheses requires investigation on its own. Therefore, we performed two evaluations. First, we executed a user study to assess the accuracy of the recommendations generated as a result of the KA approach. We relied on quantitative data analysis methods and invited domain experts to evaluate the HES recommendations. We expect to gain insights into the feasibility of building a curated knowledge graph of HES with domain experts' contributions. We measure the practicality of the task in terms of the effort required to expand the annotations and coverage of the SNOMED CT taxonomy.

Second, we performed a series of experiments to measure the performance of CONRAD in terms of its ability to identify correctly people requiring assistance. We developed alter-

native hypotheses considering the use of time range, one of the dimensions of the health condition evolution statements. To validate the results of the experiments, we built a gold standard dataset based on the synthetic electronic health records and then compared it with CONRAD results; insights into the expected results were acquired as well as observations about the accuracy and recall.

We dedicate the next section to describing the study methodology (Section 7.1.1), the objectives and the results obtained (Sections 7.1.2, 7.1.3, 7.1.4). Section 7.2 reports on the overall evaluation of CONRAD, the gold standard dataset built for this purpose (Section 7.2.1) and a more detailed explanation of the different experiments and hypotheses used to assess the accuracy of the approach (Section 7.2.3) and the results of the experiments (Section 7.2.4).

7.1 User study - Towards a KG of health condition evolution

In Chapter 5, we presented an original Knowledge Acquisition (KA) approach to build a curated database of health evolution information. We collected descriptions of health condition evolution from authoritative public sources. The approach took advantage of natural language processing techniques and performed a supervised text classification task relying on machine learning models to annotate health condition evolution according to the HECON Ontology. We included a post-processing reasoning component to clean inconsistencies that may be generated during the classification process. Because these recommendations were generated automatically, we decided to produce metrics that support the evaluation of HES. The second key component of the KA approach comprises means to expand the coverage of health condition evolution statements; specifically, we use the HES recommendations and exploit SNOMED CT taxonomy features to devise heuristic rules of HES propagation. Lastly, in order to accelerate the annotation process and ensure a high-quality HES generation, we designed an approach to include domain experts' contributions. The overall approach facilitated the extraction of knowledge about health condition evolution.

Building this database and making it available as a Knowledge Graph of health condition evolution is essential for the overall objective of identifying ongoing issues in electronic health records. CONRAD's reasoning module (Chapter 6) draws on the Knowledge Graph data about the evolution of medical events as a critical component of the process to estimate if a condition is in progress at the time of an emergency. For instance, to recognise how a

given condition evolves and its convalescence time.

The information compiled in the Knowledge Graph can directly affect the accuracy of an intelligent system trying to identify ongoing medical events at a certain point in time [Morales Tirado, Daga, and Motta 2021]. For instance, if a person's health records hold information about a 'Fracture of vertebral column (disorder)' (a 'Permanent' health issue), but it is recorded as lasting a few days, this person might not be identified as requiring assistance when a fire occurs.

Therefore, it is imperative to validate the correctness of the recommendations generated as a result of the KA process. In what follows, we detail the objectives, methodology and user study results.

7.1.1 User study methodology

The user study's main objective is to validate the overall approach implemented to extract health condition evolution from natural language and the viability of annotating health conditions by incorporating a Human-in-the-loop (HITL) step. Consequently, we sought to assess this in three ways:

- (a) by evaluating the proportion of relevant HES recommendations generated. Specifically, the HES resulting from the knowledge components identification step, in a quantitative analysis of the precision of the recommendations and the feasibility of the task;
- (b) by quantifying how much of the recommended HES, generated by the knowledge completion step, is useful to support the participants' task of expanding the coverage of SNOMED CT taxonomy. We produce a quantitative analysis of the number of correct HES and the number of achievable annotations;
- (c) by inspecting how sustainable and feasible the proposed methodology is to populate a database of health evolution that includes humans in the loop; in this case, we perform an analysis on the effort required to populate SNOMED CT taxonomy based on the performing results of the previous activities.

What follows is an account of the methodology employed in the study, including the study's design, a description of the participants' recruitment process, and the data collection and analysis.

Design. The user study was designed to simulate the process in which domain experts should decide how accurately a health condition evolution statement (HES) represents a medical condition's progression. We instantiated the tool proposed in the Human-in-the-loop step (Chapter 5, section 5.5) and provided participants with a set of recommendations (HES statements) to accelerate the annotation tasks. The *first task* was dedicated to evaluating the HES generated by the knowledge components identification step (Chapter 5, section 5.3), and the *second task*, to evaluate the HES produced by the knowledge completion step (Chapter 5, section 5.4). It is worth stating that in this Chapter, we use the terms: health condition, condition, SNOMED CT concept and concept interchangeably to refer to an item under evaluation.

The session started with an introductory phase, where participants were given a short presentation about the context of the user study, followed by an explanation of the features used to build a health condition evolution statement. Then, a description of the tasks they were going to perform, exemplified by a guided example on how to annotate the health conditions.

The *first task* focused on evaluating the recommendations produced by the knowledge components identification step and was structured as follows:

1. **Understand the context and annotation task.** Participants were asked to familiarise themselves with a scenario¹ in which they have to assess if a condition is still ongoing or not. The tool shows the name of the health condition and a list of recommended HES (the number of recommendations varies for each item, in our study, we had a maximum of eight HES).
2. **Indicate their level of familiarity.** Before assessing the validity of the HES, participants must indicate how familiar they are with the given health condition in the form of a Likert-scale choice: (1) *Unfamiliar*, (2) *Partially familiar* or (3) *Familiar*.
3. **Indicate what annotations are accurate.** Domain experts were asked to indicate whether each HES listed is relevant, in the form of Likert-scale questions: (5) *Incorrect*, (4) *Partially incorrect*, (3) *Neither correct nor incorrect*, (2) *Partially incorrect* and (1) *Correct*.
4. **Input their best guess.** If suitable, participants can provide a correct health condition evolution statement. In this case, they should use the features of the HES to create a

¹Specifically, we described a fire scenario in which they should look at health records and decide if the condition is relevant (in progress).

new one for the given condition. This step is optional.

Figure 7.1 presents the web interface developed for this purpose and reflects the structure of the user study described above.

1. Understand context and annotation task →

Scenario
A fire emergency occurs in a University campus. Among the persons in the building is Anne. Her health record includes the condition displayed on the right. Our Intelligent system needs to decide whether Anne requires special assistance during the emergency. Therefore, it needs to decide whether the condition found in the health record may still be ongoing at the time of the emergency event. The system is required to make a timely assessment based on the information in the health record only. Luckily, the system can use a dataset of Condition Evolution Statements.

Developmental co-ordination disorder (dyspraxia) in children
Snomed CT Concept Dyspraxia (finding) (6950007)
Source(s):
NHS [Overview](#)
NHS [Treatment](#)

2. Express familiarity →

How familiar are you with this concept?
☒ Unfamiliar ☐ Partially familiar ☐ Familiar

3. Indicate accuracy →

Considering the scenario above, how do you agree with the following Condition Evolution Statements?

Condition Evolution Statement	Incorrect <input type="radio"/>	Partially incorrect <input type="radio"/>	Neither correct nor incorrect <input type="radio"/>	Partially correct <input type="radio"/>	Correct <input type="radio"/>
PERMANENT	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>

4. Provide (input) the correct or most accurate statement (optional) →

Can you help us improve this Condition Evolution Statement?

Developmental co-ordination disorder (dyspraxia) in children

Direction: Pace: From: MINUTES To: MINUTES ☒ No, I can't think of a better CES

Next

Figure 7.1: User study: interface task one.

The *second task* concentrates on evaluating the recommendations produced by the knowledge completion step and was structured as follows:

- 1. Understand the annotation task.** Participants were asked to familiarise themselves with the given health condition (a source SNOMED CT concept) and the HES and features (relationships provided by SNOMED CT taxonomy). This information was provided as an aid for participants to expedite the estimation of a HES recommendation as the annotation for another health condition, a target SNOMED CT concept.
- 2. Indicate what annotations are accurate.** Domain experts were asked to indicate whether each of the target SNOMED CT concepts listed share the same HES as the source, in the form of Likert-scale questions: (5) *Incorrect*, (4) *Partially incorrect*, (3) *Neither correct nor incorrect*, (2) *Partially incorrect* and (1) *Correct*.

Figure 7.2 presents the web interface developed to reflect the structure of the user study described above.

Recruiting participants and sampling data. We invited medical students, interns, nurses, general practitioners, paramedics and first responders who are knowledgeable on how health events (medical procedures, health conditions, diseases) evolve. We advertised the study

1. Understand the annotation task

2. Assess the accuracy of the statement

Endoscopy of stomach (procedure) (386831001)

Condition Evolution Statement(s) (CES):
IMPROVE MODERATELY FROM 5 MINUTES TO 2 WEEK

Attribute(s):
Method (attribute) : Inspection - action (qualifier value)
Procedure site - Direct (attribute) : Stomach structure (body structure)
Using device (attribute) : Endoscope, device (physical object)

Direct Parent(s):
Abdomen endoscopy (procedure)
Gastrointestinal tract endoscopy (procedure)
Procedure on stomach (procedure)

Considering the health condition "Endoscopy of stomach (procedure)" and its Condition Evolution Statement(s):
How do you agree with the following list of health conditions *sharing* the same Condition Evolution Statement (CES) as Endoscopy of stomach (procedure) ?

Condition Description	Incorrect	Partially Incorrect	I do not know	Partially Correct	Correct
Group 3: This condition shares similar attribute(s) as the main health condition					
Esophagogastroduodenoscopy with endoscopic ultrasound of upper gastrointestinal tract (procedure)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Operative esophagogastrosctoscopy (procedure)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Operative esophagogastroduodenoscopy (procedure)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Group 5: This condition shares same parent(s) and/or similar attribute(s) as the main health condition					
Endoscopy of stomach (procedure)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Operative esophagogastroduodenoscopy (procedure)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Esophagogastrosctoscopy through stoma (procedure)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Next

Figure 7.2: User study: interface task two.

among local general practitioners (GPs²), hospitals, health care service providers and medical universities. Seven people agreed to participate all with a background in healthcare. Our sample included participants with different levels of expertise, from medical students to trainee doctors, as shown in Table 7.1. We consider the sample of participants an excellent result and a good sample, considering the recruitment process was open for five months, and experts were under pressure at work due to the COVID emergency and the rise in infections.

Table 7.1: Participants by level of expertise

Total	Expertise	Current role	Specialisation
1	Research	Project Officer	Palliative Care for Cancer Patients
1	Doctor	Trainee doctor	Psychiatry
2	Nurse	Nurse	Respiratory care
		Nurse practitioner	Minor illnesses in a GP surgery
3	Student	Intern	Gynecology
		3rd year undergraduate student	Medicine
		3rd year PhD student	Clinical medicine research

²General practitioners (GPs) treat all common medical conditions and refer patients to hospitals and other medical services for urgent and specialist treatment. <https://www.healthcareers.nhs.uk/explore-roles/doctors/roles-doctors/general-practitioner-gp/general-practitioner>

For the *first task*, we used a randomly selected dataset of SNOMED CT concepts resulting from the knowledge components identification output. For each concept, we presented the recommended HES and asked participants to assess the accuracy of each statement. Participants were given thirty minutes to annotate as many health conditions as possible.

For the *second task*, the tool displays the information of a source SNOMED CT concept and the list of potential target concepts that might share or inherit the source's HES statement. These potential target concepts were generated using six different rules of propagation (refer to Chapter 5, section 5.4). The list of target SNOMED CT concepts was created taking into consideration that:

- (a) Not all propagation rules apply to every concept; therefore, the list of target concepts varies for each source SNOMED CT concept.
- (b) Since the execution of the propagation rules produces hundreds and, in some cases, thousands of new statements, we considered that asking participants to annotate these number recommendations manually was not feasible. Therefore, we simplified the participants' task and randomly selected five target concepts from each rule where it applies.

We built the dataset of source SNOMED CT concepts from the results of the previous task and following two constraints: only the concepts that were (a) marked as *Familiar or Partially familiar* and (b) their HES annotation was evaluated as *Correct*. If participants created new HES, these were also included.

Participants had to indicate whether a target SNOMED CT concept shares the same HES as the source concept, as illustrated in Figure 7.2.

Data collection. During the experiment, we gathered three types of data: the decisions that were taken by the participants in terms of the relevance of the annotations generated by (a) the knowledge components identification and (b) the knowledge completion processes, and (c) newly created HES based on participants' expertise. All sessions were run remotely³. We attended the study acting as supervisors, providing support to participants in case they needed clarification on the annotation process or the use of the web tool. We made sure to avoid any intervention on how to answer the questions and what annotations were most suited.

³Following The Open University COVID-19 regulations

Data analysis. We divided the data analysis into three parts. First, we performed an agreement analysis to quantitatively measure the relevance of the recommendations generated by the knowledge extraction process. We expect that the data collected produced one of the following results: (a) participants find HES effectively represents health evolution and, therefore, HES can be extracted from natural language; (b) a low agreement among participants and, therefore, the text classification task should be revised; and (c) there is no agreement; therefore, the current approach is not suitable for the extraction of health evolution information from text. We present the results in Section 7.1.2.

Second, we assess the knowledge completion task’s effectiveness by quantifying the propagation rules’ reliability to expand the HES coverage to other ontologically similar SNOMED CT concepts. We present this analysis in Section 7.1.3. We expect that data supports one of the following hypotheses:

- (a) participants find that most of the HES can be inherited from other conditions. Therefore, using SNOMED CT features is a useful method to populate SNOMED CT or,
- (b) participants find a few HES can be inherited to other conditions. Therefore, a thorough revision of the propagation rules is required to assess the approach’s effectiveness and the use of SNOMED CT features.

Third and final, we analyse the effort required to obtain a curated database of HES that expands SNOMED CT taxonomy. We expect that results reflect the feasibility of the task by (a) providing domain experts with HES recommendations generated can accelerate the construction of a curated knowledge graph of HES. Alternatively, (b) even using recommended HES, the participants find it difficult to perform the tasks; therefore, the methodology is not sustainable. The results are presented in Section 7.1.4.

7.1.2 Precision analysis

In this section, we present the analysis of the system’s precision using the data collected in the *first task*. Additionally, we analyse the feasibility of the task by measuring the number of annotations participants were able to assess during the given time frame. The analysis is based on the data collected in the *first annotation task*.

Participants had to indicate their level of familiarity with a given concept and whether the HES is correct or not using the five-category Likert scale. Also, they could input a new HES according to their best judgment. Table 7.2 displays the total number of concepts annotated per participant and the new HES generated.

Table 7.2: First task: number of annotated SNOMED concepts per participant

	Part 1	
	SNOMED concepts annotated	New HES generated
P1	60	25
P2	47	0
P3	31	6
P4	67	0
P5	30	3
P6	53	6
P7	26	6
Avg	44.86	7

The data shows that participants were able to use the recommendations and the instantiation of the human-in-the-loop (HITL) tool to annotate an average of 45 conditions in 30 minutes. Also, they generated, on average, seven new HES.

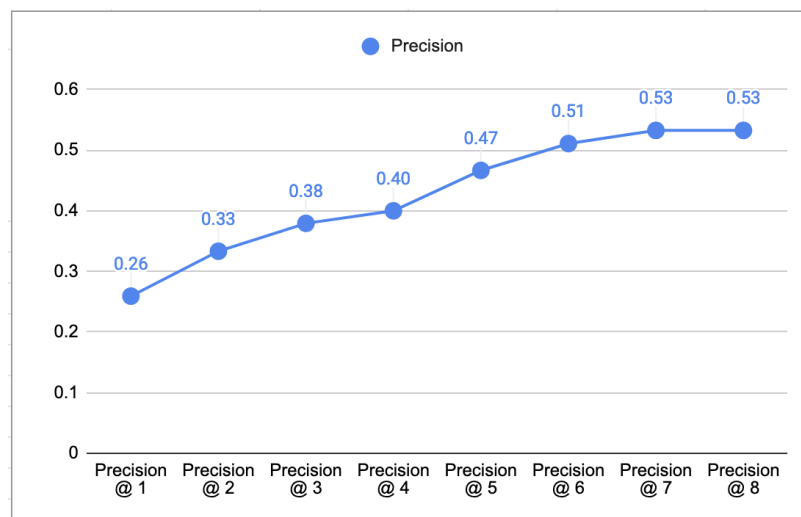


Figure 7.3: Precision @ k

Regarding the precision evaluation, it is essential to remember that the classification process generates one or more recommended HES per condition⁴; therefore, to measure

⁴Participant P4 annotated the more significant number of conditions, which included conditions with eight

Table 7.3: First task: Precision@k per participant

Precision @ k	1	2	3	4	5	6	7	8
P1	0.20	0.28	0.31	0.36	0.38	0.39	0.39	-
P2	0.37	0.49	0.57	0.62	0.65	0.65	0.66	-
P3	0.31	0.41	0.49	0.53	0.57	0.57	-	-
P4	0.31	0.42	0.46	0.50	0.53	0.55	0.56	0.56
P5	0.26	0.30	0.37	0.42	0.47	0.49	-	-
P6	0.28	0.34	0.39	0.42	0.42	-	-	-
P7	0.23	0.30	0.38	0.38	0.38	0.38	-	-
Median	0.28	0.34	0.39	0.42	0.47	0.52	0.56	0.56

the number of relevant HES, we calculate Precision@k. It can be seen from the data in Table 7.3 that the system was able to *provide relevant recommendations in more than half of the cases*, with a median Precision@8 of 0.56 (see also Figure 7.3). These results show that the extraction of HES is an achievable task.

In the final part, we evaluate the inter-rater reliability. As described previously, the participants annotated the same dataset and performed the task independently. Therefore, we considered it appropriate to use two measurements: Krippendorff’s alpha⁵ [Krippendorff 2011] and Fleiss’ kappa⁶.

First, we used Krippendorff’s alpha coefficient, which applies to incomplete data and any number of raters, sample sizes and categories. To calculate this coefficient, we used all the data resulting from the annotation task, including the conditions that were annotated only by one participant and obtained a measure of 0.4685, indicating low agreement.

Second, we assess the reliability of agreement using Fleiss’ kappa measure, which can be used to calculate the agreement between three or more raters and when elements are annotated using categorical ratings. In our case, we use a Likert scale to assess the precision of the HES recommendation. Importantly, Fleiss’ kappa requires that each element (a HES recommendation) is evaluated as many times as the number of raters. The agreement was

HES.

⁵Krippendorff’s alpha: https://en.wikipedia.org/wiki/Krippendorff%27s_alpha

⁶Fleiss’ kappa: https://en.wikipedia.org/wiki/Fleiss%27_kappa

Table 7.4: First task: Total annotations by familiarity

	Familiar	Partially familiar	Unfamiliar
P1	27	10	23
P2	9	19	19
P3	17	4	10
P4	33	22	12
P5	13	6	11
P6	14	17	22
P7	10	13	3
<i>Total</i>	<i>123</i>	<i>91</i>	<i>100</i>
<i>Standard Deviation</i>	<i>9.05</i>	<i>6.73</i>	<i>7.30</i>
<i>Standard Deviation</i>		<i>12</i>	<i>7</i>
<i>(*Familiar and Partially familiar as one group)</i>			
<i>Proportion</i>	<i>0.39</i>	<i>0.29</i>	<i>0.32</i>
<i>Proportion</i>		<i>0.7</i>	<i>0.3</i>
<i>(*Familiar and Partially familiar as one group)</i>			

calculated using the conditions annotated by all participants, a total of 26 conditions. We obtained a *Fair agreement* of (0.2335). This value reflects that there is shared expertise and fair consistency in judgement.

These results are also supported by data in Table 7.4, which indicates that participants were somehow familiar with 7 out of 10 concepts.

7.1.3 Evaluating the knowledge completion task

The purpose of the *second task* was to evaluate to what extent the rules of propagation, created using the ontological organisation of SNOMED CT taxonomy, are useful to extend SNOMED CT coverage. We asked participants to indicate whether a target SNOMED CT concept shares the same HES as the source concept. Similar to the *first task*, participants should answer using a five-category Likert scale. The source concept sample is constituted from the SNOMED CT concepts annotated as “Correct” in the first part of the study and the

concepts for which the participants provided a new HES.

As shown in Table 7.5, each participant was able to review an average of 33 source concepts and annotate 147 target concepts. In comparison with results in the previous task, where participants annotated an average of 45 concepts, what stands out is that the exploitation of SNOMED CT taxonomy not only produces more recommendations but once participants are given a list of potential HES statements, they were able to annotate *three times more (312%) recommended HES*.

Table 7.5: Second task: Number of annotated SNOMED concepts per participant

	Part 2	
	Source concepts	Target concepts annotated
P1	18	70
P2	32	143
P3	29	126
P4	62	284
P5	27	117
P6	37	162
P7	29	126
Avg	33	147

Further analysis shows that participants reviewed 1,028 recommendations in total (see Table 7.6), which means that each participant reviewed an average of 147 target concepts. The top three rules that contributed the more significant number of recommendations were rules five, three and four (see Section 5.4 for a detailed explanation of the rules). We noticed that rule two did not display any target concepts; after checking the dataset, we concluded that rule two did not apply to any of the source concepts.

Turning now to recommendations' relevance, as shown in Table 7.7, in total 501 recommended target concepts were annotated as 'Correct', as a collective effort and on average, 72 per participant. Figure 7.4 compares the results of the total annotations and the total correct HES.

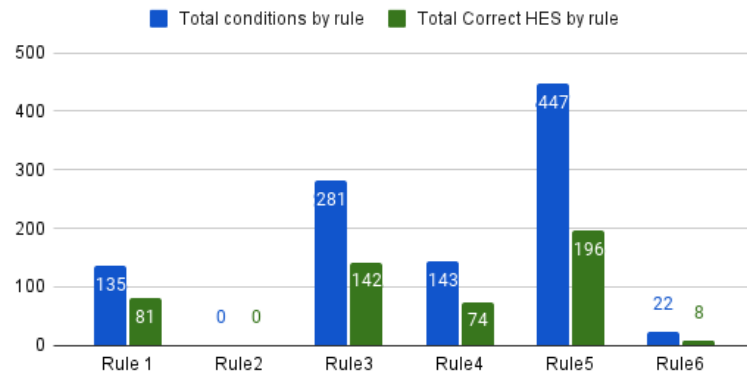


Figure 7.4: Comparison total number of annotations and correct answers

These results demonstrate that the recommendations generated by the knowledge completion method are *useful in half of the cases* to swiftly populate the part of SNOMED that was not originally covered by the web sources.

Table 7.6: Second task: Total annotations grouped by rule

	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Total
P1	18	0	17	10	25	0	70
P2	16	0	39	20	66	2	143
P3	21	0	33	15	53	4	126
P4	23	0	91	37	131	2	284
P5	21	0	24	21	47	4	117
P6	20	0	43	24	70	5	162
P7	16	0	34	16	55	5	126
Total	135	0	281	143	447	22	1028
Avg	19	0	40	20	64	3	147

Finally, we analyse the results by type of rule. It can be seen from the data in Table 7.7, that the three top rules that contributed with the most significant number of ‘Correct’ annotations were rules five, three and one. However, we observe that in proportion rules one, four and three demonstrated to be more efficient in finding similar concepts and producing ‘Correct’ recommendations.

Table 7.7: Second task: Total Correct annotations per participant

	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Total
P1	13	0	4	4	6	0	27
P2	13	0	7	13	19	0	52
P3	9	0	19	11	28	3	70
P4	15	0	61	20	73	0	169
P5	19	0	16	14	32	4	85
P6	11	0	25	8	20	0	64
P7	1	0	10	4	18	1	34
<i>Total</i>	<i>81</i>	<i>0</i>	<i>142</i>	<i>74</i>	<i>196</i>	<i>8</i>	<i>501</i>
<i>Proportion</i>	0.60	-	0.51	0.52	<i>0.44</i>	<i>0.36</i>	<i>0.49</i>
<i>Avg</i>	<i>12</i>	<i>0</i>	<i>20</i>	<i>11</i>	<i>28</i>	<i>1</i>	<i>72</i>

7.1.4 Sustainability

In order to assess the sustainability and feasibility of adopting the proposed methodology to build a curated knowledge graph of HES, we examine the results obtained in tasks one and two. We gave an account of the effort (expressed in ‘person-month’) required to populate the KG. We take as a reference the last edition of SNOMED CT, which included 353,567 concepts (published on January 31, 2020).

On the one hand, in task one, one participant annotated an average of 25 correct concepts in 30 minutes; therefore, we can derive that one participant can annotate 50 correct concepts per hour and 350 a day (7 working hours). Suppose we only use recommendations generated by the knowledge component extraction step. In that case, it will take approximately four years and a half (55.20 person-months) to curate and populate the SNOMED CT taxonomy. On the other hand, in task two, one participant annotated an average of 144 correct annotations per hour (1,008 a day); it will take approximately a year and a half (19.16 person-months) to complete the task.

We calculated only one person’s effort, yet experts could perform the task simultaneously. For instance, with seven experts (emulating our user study) and the effort required in task one, it will take approximately eight months (7.88 person-months) to curate the database. Likewise, the task is reduced to approximately three months (2.73 person-months) consider-

ing the effort in tasks two and seven participants. From these results, we can conclude that the approach is sustainable and attainable.

7.1.5 Discussion

The results show that it is possible to extract health condition evolution knowledge from natural language; more importantly, this knowledge can be used to support domain experts in building a high-quality knowledge graph of HES.

Task one results demonstrate that participants could assess and decide whether a certain HES was appropriate for a given condition. Participants annotated at least 26 concepts in 30 minutes; considering the short time participants were given to familiarise themselves with the health condition evolution statement (HES) model, it is safe to say that none of them had problems interpreting the recommendations while judging the relevance of the statements.

The results, as shown in Table 7.4, indicate that although we selected the concepts for evaluation randomly, participants were familiar with the given concepts in seven out of ten cases (with and standard deviation of 12). We hypothesise that selecting a dataset based on participants' specific domain of expertise could improve the agreement results.

Another important finding is that the rules created to expand SNOMED CT coverage proved useful and pertinent. Participants could annotate as correct three times the number of items they reviewed in the first task. This result is significant as, on average, half of the recommended HES were correct. Further work may consider working on these results for rules refinement.

We consider the inclusion of the Human-in-the-loop step a key strength of the knowledge acquisition approach, and the results support the hypothesis that the task is feasible. More importantly, it ensures the capture of experts' valuable and accurate knowledge.

In summary, the results demonstrate that extracting health evolution statements (HES) from natural language is possible and exploiting SNOMED CT features accelerates the production of recommendations, hence the coverage of SNOMED CT. Furthermore, the results confirm that the recommendations facilitate participants' task and ensure the contribution of domain experts.

7.2 Intelligent System Evaluation

As it was pointed out in the introduction to this section, so far, we have relied on the hypothesis that it is possible to design an intelligent system that automatically analyses electronic health records to identify people's ongoing health issues and deliver this information to emergency services. Having designed CONRAD and developed the prototype system in a paradigmatic Smart City environment, our final objective is to assess its performance when performing its task of detecting people requiring assistance.

In this section, we present the experimental evaluation of CONRAD, which is based on the proposed scenarios in Chapter 3 and uses a dataset of synthetic electronic health records (detailed in Chapter 6, section 6.2) for analysis, detection of ongoing health issues and identification of people in a vulnerable situation. *The objective is to compare the performance of our system against a typical assessment a person or a team would do when a fire emergency occurs in a large organisation.*

As mentioned in the previous chapter, we generated an initial dataset of 10,000 synthetic patients. By creating this large dataset, we aimed to have a diverse collection of records, which would include as many heterogeneous clinical histories as possible. However, for the experiments, we focus on a randomly selected set of 1,012 patients' electronic health records, whose age ranges from 20 to 80 years old, as we try to simulate the wide spectrum of ages associated with the employees of a large organisation, such as The Open University.

In the next section, we detail the development of a well-curated Gold Standard Dataset that is the point of reference to measure the validity of our results. In order to make the best use of the HES representation, we designed three experiments, each using the time range feature in different ways; this is detailed in Section 7.2.3. Finally, in Section 7.2.4 we present the results of the experiments and discuss the findings.

7.2.1 Gold standard dataset

To evaluate the results of our experiments, we developed a Gold Standard dataset (GSD) based on a collection of annotated patients' health records answering the following questions:

- Q1: Who needs special assistance in case of a fire evacuation?
- Q2: What type of assistance the person needs?

The GSD was developed by two people independently, referred to as the reviewers. It is worth mentioning that the reviewers are members of a large organisation (The Open University), and their competence is comparable to that of a fire warden.

To support the reviewers in building the GSD, we developed a web interface that for each sample displays: a) the question to be answered, b) the patient's details (name, last name, age), and c) a section with the patient's health record (description, reason, type of record, start date, end date). To answer Q1 about a person requiring assistance, we present the reviewer with the options 'Yes' or 'No'. The reviewer should read the health records and detect any condition that could reveal a person's impediment to evacuating the building.

For Q2, in addition, to the data displayed in Q1, we list the type of disabilities [UK Government 2008] and ask the reviewers to choose at least one item that better describes the patient's impediment or health issue. It is essential to mention that we used the same list of disabilities as our system. Additionally, Q2's sample is composed only of the samples annotated as 'Yes' in Q1. The GSD was initially built by two reviewers using the following process:

- Annotate the GSD individually.
- Identify discrepancies by reviewing the differences between their answers.
- Discuss each difference, explanations and evidence for answering 'Yes' or 'No', including external sources, such as the NHS website.
- Take a motivated decision: evaluate the evidence and reach an agreement.
- Annotate the reasons for the agreement: write down any comments and explanations to ensure consistency across decisions.

The resulting GSD is an account of how a person typically involved in supporting a fire evacuation may interpret the content of health records, having sufficient time and resources.

7.2.2 CONRAD overview

In Chapter 6 we described CONRAD's proposed software architecture in detail. Here we will give a brief overview of the input, process and output expected from the execution of the intelligent system to evaluate our experiments. The intelligent system takes as input the EHR. Then it processes the records following our approach to finally deliver the number of people requiring assistance and the type of help required.

Input: the primary data input are the electronic health record (EHR) of the people involved in the emergency.

Process: For each person in the building, CONRAD identifies the valid health condition; this implies an evaluation of each record according to its temporal validity. Each data point that represents a health condition has a health condition evolution (HES) that indicates the type of disease and its recovery time. A health condition is relevant if the medical event is ongoing at the time of the emergency. If a person has at least one relevant health condition, it means that this person requires assistance. On the contrary, if the person has no relevant records, then no assistance is required. Our system identifies all the people that require assistance and the health records that support this result. Next, the system evaluates the type of disability, which defines the reason for the assistance and the severity score. The system uses the types of disabilities taken from the UK government guidelines [UK Government 2008].

Output: As exemplified in Figure 7.5, the system returns a list of people requiring help. Additionally, the system provides information about the best matching type of disability according to the relevant medical conditions.

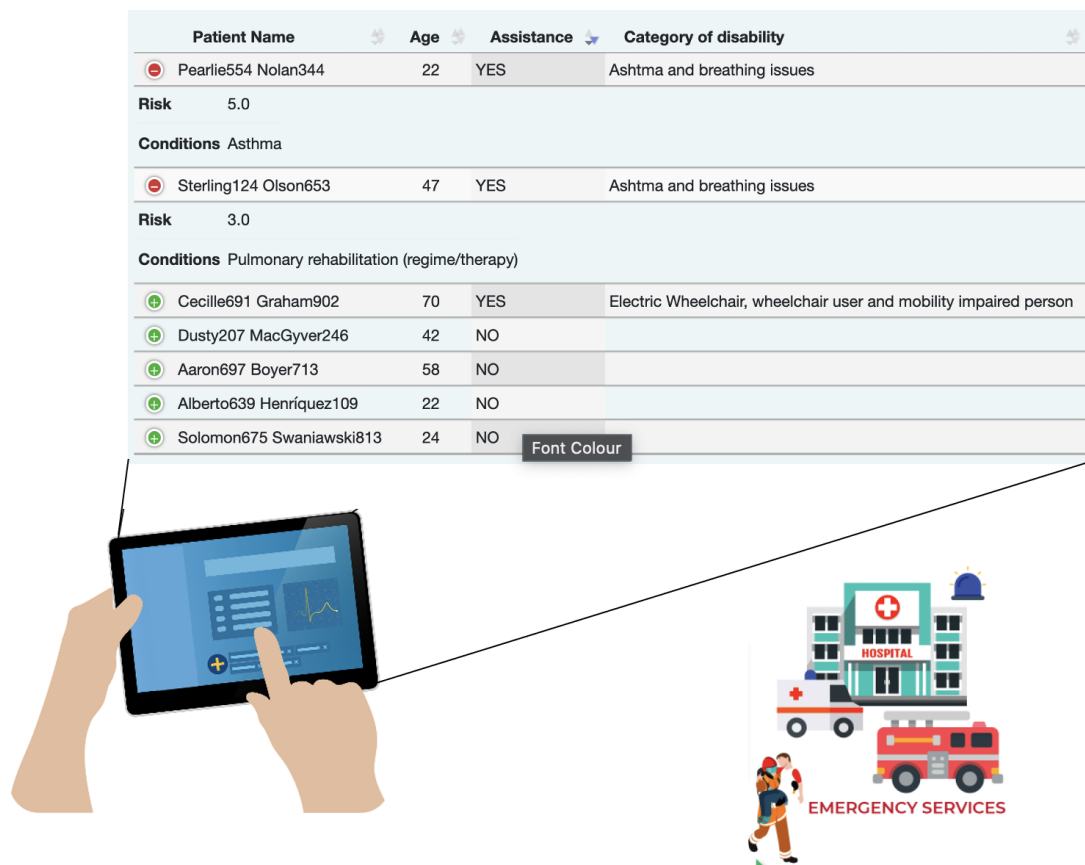


Figure 7.5: CONRAD's output - People in need of assistance and type of vulnerability

7.2.3 Experiments design

The current HES model establishes a detailed and precise description of condition evolution, and it is defined by three dimensions: type of condition, pace and duration. The time range dimension estimates the recovery time or convalescence period, which is not a fixed number; instead, it uses a minimum and maximum time range. Therefore, we designed three experiments to best use this representation. Each experiment uses the time range components in different ways. In what follows, we describe each experiment and the relevant hypotheses.

Experiment one - Pessimistic approach or upper bound (UB). For this experiment, the assumption is that a condition develops in the most prolonged period. For example, suppose a condition improves within two to eight days (UB). In this case, the system assumes the disease takes the longest time to recover (eight days). Therefore, if an emergency happens in this period (including the eighth day), the condition is valid, and the person might need assistance.

Experiment two - Optimistic approach or lower bound (LB). For this experiment, the hypothesis is that a condition develops in the shortest period. Following the previous example, if a condition is expected to improve in two (LB) to eight days (UB), the system assumes that the condition takes the shortest time to recover (two days). Therefore, if an emergency happens before the LB, then the condition is valid. If the emergency occurs on the third day or after, the system identifies the condition as not relevant, meaning that the person is assumed to have already recovered their good health, thus not requiring help.

Experiment three - Average or median approach (Av). This experiment assumes that a health condition develops in a period considered the median between LB and UB. Continuing with the previous example, the median time to recovery is three days for a condition that improves between two to eight days. In this case, the system estimates the convalescence period for the given condition as five days. If the emergency happens before that day (including the fifth day), they might still require help, meaning the condition is relevant.

7.2.4 Results and discussion

We compare the decisions of our system against the gold standard developed previously. The objective is to identify who requires special assistance during an evacuation. To measure the performance of our system, we use the following metrics:

- Accuracy, to evaluate our system as a boolean classifier and measure its ability to distinguish whether an individual needs assistance or not.
- Precision, to measure the percentage of people identified as vulnerable that were correctly classified.
- Recall, to measure the percentage of actual people in need of assistance that were correctly classified. Recall is a particularly relevant measure for our system as we want to minimise the risk of missing a person in need.
- F-Measure, to measure the system's performance considering both precision and recall.

As seen in Table 7.8, all three experiments reported accuracy of 0.90 or higher; therefore, our system correctly identifies 90% to 92% of the people that either need or do not need assistance.

Table 7.8: Experiments results

Experiments	Accuracy	Precision	Recall	F-score
Pessimistic approach or Upper bound (UB)	0.92	0.73	0.81	0.77
Optimistic approach or lower bound (LB)	0.90	0.79	0.51	0.62
Average or median approach (Av)	0.92	0.80	0.72	0.76

The main aim of our system is to maximise the possibility of identifying people requiring assistance; thus, in our study, we consider Recall as the most important indicator. Hence, while there is very little difference in overall accuracy and f-measure between the *Average or median approach (Av)* and the *Pessimistic approach (UB)*, we consider the latter as the most appropriate method, as it maximises recall without affecting precision too much. In particular, through this approach, we succeeded in identifying 81% of the people who need assistance.

Although the results show good accuracy and recall, it is still possible to improve further the quality of the recommendations by considering the severity of a condition, in the context of an emergency situation. To this purpose, our system also associates a *severity score* to each condition. In this way, it further leverages the predictions of UB, which in comparison to LB and AV, has lower precision but better F-score.

Table 7.9 shows the severity score for a wheelchair user and a person with lung disease. In both cases, the severity score is 1, the highest score. The ‘vertebral fracture’ indicates a PERMANENT condition. At the same time, ‘chronic bronchitis’ improves over a certain period. However, as it started recently (considering a fire emergency 25 days later), the system assumes the person has some difficulties as the recovery period has not finished.

Table 7.9: Examples severity score

Condition	Date	Quartile	Severity score	HES annotation
Chronic obstructive bronchitis (disorder)	05/10/2019	LB	1	IMPROVE ERATELY 2 MONTHS TO 6 MONTHS
Fracture of the verte- bral column with spinal cord injury	08/04/2008	LB	1	PERMANENT

Turning to analysing the type of disability, we focus on why people need assistance. We compare the results of our system against the gold standard dataset, question two. In order to evaluate the ability of our system to provide a precise ranking of the most relevant categories of disabilities, we use Precision at K. The results obtained are summarised in Table 7.10. The results show that overall, in experiment one, the Pessimistic approach attains a high precision when identifying the first three more likely reasons concerning a disability. It is also important to mention that in evaluating the type of disability, we asked participants to select at least one type of disability and the one they consider most important. Having information about the most relevant type of disability complements the results obtained as part of the severity score assessment. For instance, delivering a score measure could help to identify health conditions that are recent, such as bronchitis affecting breathing, and possibly indicate that some kind of support is required to evacuate a building. Our approach succeeds in providing data on the degree of impact or level of importance of the ongoing health condition.

Table 7.10: Precision @ 3 categories of disability

Precision @ 3	1	2	3
Pessimistic approach or Upper bound (UB)	0.50	0.59	0.87
Optimistic approach or lower bound (LB)	0.44	0.67	0.70
Average or median approach (Av)	0.45	0.52	0.77

7.3 Chapter conclusions

This chapter presented the work done to evaluate:

- (a) A novel approach to extracting health condition evolution information from natural language;
- (b) The overall performance of our intelligent system, CONRAD, with respect to its ability to identify people requiring assistance.

The results of this study support our hypothesis that extracting health condition evolution information from natural language is feasible. Furthermore, knowledge classification techniques and knowledge completion tasks based on ontological features can assist in the creation of the knowledge graph of HES.

Our results also demonstrate that it is possible to include domain experts in the process of building an authoritative database of health condition evolution statements. Furthermore, these results confirm that recommendations facilitate participants' task of assessing the accuracy of the statements, therefore accelerating its production and improving the coverage of SNOMED CT.

Even, more importantly, the results regarding the performance of CONRAD demonstrate that our system effectively identifies people that require assistance 92% of the time. Furthermore, from the analysis, the best-performing approach is the *Pessimistic approach or upper bound (UB)*. The results demonstrate that the system correctly identifies a person requiring assistance in 73% of the cases (precision). Importantly, in the fire emergency use case, the recall indicates that the system correctly identifies the people able to evacuate and, therefore, does not require assistance 81% of the time. In our case, we value recall as it is the measure that indicates that the system minimises the incorrect classification of people actually requiring assistance (reducing false negatives from the perspective of a classifier system). Overall, our experiments achieved a high accuracy with respect to the gold standard built to test our

system.

The insights gained from the user study and the experiments provide empirical confirmation that the knowledge components developed during this research effectively support the identification of people requiring assistance when a fire occurs, which was the key objective and research question of this work.

Chapter 8

Discussion and Conclusions

This research aimed to address the problem of extracting timely and valuable information from Electronic Health Records (EHR). The initial work was dedicated to the definition and elicitation of requirements for developing an intelligent system that addresses the problem of recognising ongoing health issues using EHR (Chapter 3). The following chapters (Chapters 4- 6) reported on the work conducted to develop the knowledge components¹ required to achieve this goal.

First we set out to understand how the evolution of health conditions was described in natural language and medical sources (Chapter 4) and built a formal representation of health conditions development, the Health Condition Evolution Ontology - HECON. Next, we gathered data about health condition evolution, that is, the convalescence time of health conditions. We then built a Knowledge Graph of health condition evolution information, the *second knowledge component*. This process was supported by a Knowledge Acquisition approach, which classifies text and generates (semi-automatically) Health Condition Evolution Statements (HES). The approach involved domain experts, who curated the first Knowledge Graph of condition evolution (Chapter 5 and Section 7.1). The *third knowledge component* supports the reasoning rules that support the recognition of the health conditions that must be taken into account when an emergency occurs (Chapter 6). These three components provided the basis for implementing the intelligent system for emergency support (Chapter 6). Finally, experiments were conducted to evaluate the validity of the proposed approach as well as the system's performance (Section 7.2).

The following section analyses to what extent this work has addressed the research questions and discusses whether the initial hypotheses have been verified. Section 8.2 reviews and discusses the limitations of our proposed approach along with future work that could

¹Knowledge components were introduced in Chapter 1, Section 1.4

expand the research presented in this thesis. This chapter closes with Section 8.3 dedicated to the conclusions and final remarks on the research work.

8.1 Discussion

The contribution of this research focuses on developing the knowledge components that support the analysis of EHR to identify health events that are in progress automatically. Therefore, we concentrated on **(a)** providing relevant information to decision-makers, **(b)** facilitating their task of assessing if a person requires assistance and **(c)** minimising the amount of sensitive information they exchange. In particular, the first hypothesis was that, *a model for representing the evolution of health events over time can support the detection of ongoing health issues in electronic health records (H1)*. Therefore, the first research question focuses on:

Research question 1: How to formally represent health evolution to support the identification of current/ongoing health issues?

We reviewed in detail how public and authoritative health sources described the progression of medical events. This analysis served to derive the characteristics that define health condition evolution. We identified three key elements, including the type of health condition (improvement, decline, permanent and none), the pace at which they develop (slow, moderate and fast) and the duration (a minimum and a maximum convalescence time). These three elements constitute the Health Condition Evolution Statement (HES).

We focused on creating a model that indicates condition progression independently of other medical events occurring at the same time. In this way, we ascertained to provide an intelligent system with enough knowledge to recognise all the relevant medical records and leave decision-makers, in this case, firefighters, the task of assessing if a person has an impediment or a disability that puts them in a vulnerable situation. Moreover, this is the first model that represents condition evolution from the perspective of emergency responders, who need to familiarise themselves with the emergency by gathering information about the people involved and their immediate disabilities.

The first knowledge component is the model that abstracts the definition of health condition evolution and was formalised as an ontology, the Health Condition Evolution Ontology - HECON. Considering the ontology evaluation performed and detailed in Section 4.5 was satisfactory in terms of **(1)** consistency and **(2)** expressivity. Moreover, the results proved

that HECON addressed the knowledge requirements described in Chapter 3.3 and later formalised as Competency Questions in Section 4.3.2, we can reasonably conclude that HECON is a knowledge component that answers RQ1 sufficiently.

However, the representation of health condition evolution is not enough for an intelligent system to identify ongoing health conditions; the system also requires specific information about how each condition evolves over time to assess whether it is relevant. This brings us to the next research question:

Research question 2. How to build a database of health condition evolution?

The hypothesis was that *it is possible to build a structured database of health evolution information using open, public and authoritative data sources (H2)*. We identified a lack of structured resources regarding health condition evolution. Instead, unstructured data was available in text format. We collected this data from two websites, NHS England and MAYO Clinic. However, only a few sentences in the corpus described health condition progression. Consequently, we decided to rely on knowledge extraction techniques (such as Natural Language Processing and Machine Learning) and perform a supervised classification task to annotate the sentences in the corpus. We classified the text according to the three main components of a health condition evolution statement (HES): type of condition (improvement, decline, permanent, unaffected), pace (slow, moderate, fast) and duration (minimum and maximum duration). The overall process generated one or more HES recommendations for a given health condition, represented by a SNOMED CT concept.

Although in this initial step, we managed to annotate 1,324 SNOMED CT concepts with at least one health condition statement recommendation, the electronic health records are rich and can refer to any clinical event. Therefore, we decided to expand the coverage by exploiting SNOMED CT ontological features. We analysed SNOMED CT relationships (particularly 'subtype' and 'attribute') and identified patterns that guided the automatic expansion of SNOMED CT source concepts with HES to other concepts without HES. Based on these patterns, we created six propagation rules encoded using Expression Constraint Language (ECL) [SNOMED International 2022b] and expanded the coverage of SNOMED CT. However, we understood that the automatic generation of recommended HES could contain errors due to the generalisation of the rules or a HES wrongly recommended in the first step. Therefore, it was imperative to include domain experts in the loop to capture errors generated in previous steps and crucially accelerate the annotation of HES while ensuring the development of a high-quality database. We developed a tool that captures knowledge from domain experts by asking them to evaluate the recommendations generated in the previous two steps

and build a new HES according to their best judgement.

In order to validate the overall approach, we carried out a user study involving domain experts. The study's main goal was to measure the precision of the recommended HES, the validity of the rules of propagation, and the sustainability of the approach. The user study results (Section 7.1) indicate that overall, the proposed methodology to populate a database of health condition evolution is sustainable. The HES recommendations generated using the knowledge acquisition techniques support the domain experts' task of swiftly annotating the database. Although participants had a short time to familiarise themselves with the tool and the terminology used to represent health condition evolution, they expressed an understanding of the task and the information they were evaluating. An important finding was that the SNOMED CT taxonomy features and the generation of propagation rules derived three times more HES recommendations than the text classification process, which accelerated the generation of recommended HES annotations.

Another key aspect of the user study was the number of experts that agreed to be part of the survey (seven in total). In an ideal scenario, we expected to recruit a higher number of participants in a shorter space of time. However, external factors influenced the recruitment process; at the time of the user study, experts were under pressure due to the latest wave of COVID-19² infections.

Finally, to make the newly created database available in a structured and machine-readable format, we built a Knowledge Graph following the HECON Ontology model; this constitutes the second knowledge component and answers the research question mentioned above.

For an intelligent system to estimate whether a health condition stored in an electronic health record holds at a certain point, it is necessary to rely on knowledge that guides this assessment. This leads to the next research question:

Research question 3. How to automatically reason on EHR to identify ongoing health issues?

This research question refers to the third knowledge component and comprises the reasoning rules to assess if a specific condition is ongoing. Therefore, the hypothesis was that *it is possible to formalise the rules for estimating health condition evolution automatically (H3)*.

The rules were developed to simulate the rapid evaluation that first responders and decision-makers should perform when attending an emergency. Typically, a health condition is ongoing

²COVID-19 pandemic: <https://en.wikipedia.org/wiki/COVID-19>

ing if the health event is developing (has not passed yet), the condition is chronic, or it deteriorates over time. We designed a set of reasoning rules that use the ‘type of condition’ and the ‘duration’ dimensions of the health condition evolution statement (HES) to evaluate if the medical event is relevant when the emergency occurs. The rules are thoroughly explained in Section 6.3. The reasoning component was evaluated as part of the final experiments to identify people requiring special assistance due to an ongoing health issue in Section 7.2.4. Results demonstrate that the reasoning component supports identifying ongoing health conditions and correctly identifies people needing assistance 92% of the time. Therefore, we can conclude that it is possible to reason on EHR automatically, which answers RQ3.

So far, we have identified the knowledge components necessary to estimate if a medical event is current/ongoing. We also built these components separately, which validated the first three hypotheses. However, we did not know if the integration of these components were enough to reason on EHR and fulfil the primary requirements (Chapter 3) of (a) providing relevant information, (b) facilitating the identification of a person requiring assistance and (c) minimising the amount of sensitive information exchanged with emergency services. Therefore, our last hypothesis was that *by relying on the developed components, it is possible to design an intelligent system that uses health records data to deliver relevant information to emergency services (H4)*. This brings us to the last research question:

Research question 4. *How to leverage the developed knowledge components to build an intelligent system capable of identifying people requiring assistance during an emergency?*

To answer this question, we examined how a large organisation manages healthcare data, particularly how medical information is handled during a fire event. We took the example of The Open University as we had access to documentation, and we considered it an exemplification of a large organisation.

We learnt that organisations are bound by law to ensure measures are taken to protect people, especially people in vulnerable situations, for instance, people with disabilities, pregnant women, children, elderly [UK Government 2008]. As a result, organisations collect health data to elaborate evacuation plans and put in place measures to evacuate people with special needs. This information is also needed by emergency services at the fire scene to evaluate the situation and plan rescue operations. As reported in the introductory chapter, this data (collected by organisations) hardly reaches emergency services, and when it is available, decision-makers should use it with caution as there is a risk that it is outdated. Furthermore, managing personal information, crucially sensitive data, raises concerns regarding employ-

ees' privacy, preventing the direct exchange of such data between health service providers (who hold the most up-to-date medical information) and emergency services.

On the one hand, health service providers hold crucial medical information about people's disabilities or health issues. On the other hand, emergency services require up-to-date and relevant information suitable to carry on their activities and help them identify people requiring help. Both organisations should observe data regulations in order to ensure the correct handling of sensitive information, such as healthcare data. Precisely, in this scenario, we propose the implementation of *CONRAD - Health Condition Radar*, an intelligent system that acts as a mediator and a facilitator when balancing the trade-off between the exchange of relevant data and the protection of citizens' privacy.

Therefore, we designed a software architecture that integrates the knowledge components (developed previously) to analyse electronic health records and detect if a medical event is ongoing at the time of the fire event. By obtaining the list of ongoing health issues, it is possible to identify if people have an impediment that puts them in a vulnerable situation, which supports the last hypothesis **H4**. Crucially, minimising the amount of data to be exchanged between organisations, since only relevant information is listed. In what follows, we summarise the design of CONRAD and recap how the different knowledge components assist in identifying people requiring assistance during an emergency.

Before building a system's software architecture, we should analyse the type of data the system should process. In this case, a dataset of electronic health records. Since data disclosure is one of the main concerns, we decided to explore other options and use a dataset of synthetic electronic health records as data input. The dataset was generated using Synthea [Walonoski et al. 2018], open-source software that simulates patients' medical history. EHR generated by the Synthea software proved to be extensive as they covered a great variety of health issues and simulated the progression of different conditions over a patient's lifetime.

In a Smart City ecosystem, CONRAD could access the EHR of people present at the emergency scene (see Section 1.2, Assumptions 1 and 2). The system's reasoner component (Section 6.3) is the key module of CONRAD as it analyses the EHR; it takes the description of the medical event recorded in the EHR to retrieve the health evolution description (a Health Evolution Statement - HES). This process is supported by HECON Ontology and the Knowledge Graph of health conditions. Then the system estimates if the health event is relevant at the time of the emergency, this means if it happened recently, the person is still affected or has already recovered from a medical condition.

In addition, if we consider that emergency responders often have different levels of ex-

expertise and training, it could be helpful to provide additional information about the types of disabilities related to the current health issues and a numeral representation of how recently the condition was recorded. Therefore, we integrated a ‘data fitting’ component as part of CONRAD’s data analysis that uses ConceptNet (a common-sense knowledge base) to find the most related type of disability given a particular health event.

To evaluate the performance of CONRAD for detecting ongoing medical events and identifying people requiring assistance, we used a synthetic dataset of electronic health records representing 1,012 patients’ medical histories. Additionally, we developed a Gold Standard dataset as the point of reference to measure the validity of the results. The experiment results reported 90% to 92% accuracy, which demonstrated that overall the system is able to distinguish whether a person requires or not assistance; we can safely say that the system has a high performance which answers RQ4. Further analysis reveals that among all the patients identified as requiring assistance, the system was able to correctly identify 81% of them (Experiment one - pessimistic approach, Section 7.2.3).

In conclusion, we successfully demonstrated that it is possible to provide emergency services with additional information about people requiring assistance during an emergency by automatically analysing people’s health records, through the application of a knowledge engineering approach to the problem of representing and reasoning with health condition evolution. Furthermore, the application of such an approach and the implementation of the different knowledge components provide means to identify relevant information (in this case, a minimal part of EHR data). Therefore, facilitating organisations’ tasks in terms of making decisions on the amount and value of information that can be exchanged with third-party organisations.

8.2 Limitations of the study and Future work

This research has demonstrated that it is possible to reason on electronic health records to identify ongoing health issues and derive if a person requires special assistance in the case of an emergency. However, during our work, we also identified some limitations of the proposed solution. This section will discuss the limitations and sketch initial ideas for future work. We structure the rest of the section by grouping issues according to the knowledge components they most relate to.

Knowledge representation of health condition evolution. The HECON Ontology was built based on how health condition evolution is expressed in natural language. We take as

input descriptions published by two authoritative health organisations: NHS England and MAYO Clinic.

Although the natural language description can be considered correct, it leaves out the case when condition evolution may depend on other factors cooperating with the condition. First, some conditions may evolve differently depending on the patient's age. Further work is needed to evaluate the factors influencing health condition evolution. We speculate that the same could apply to gender and that not all conditions may need this specification.

Nevertheless, the current design of HECON already represents a SNOMED CT concept with several HES annotations, each supported by properties that express the reliability of the annotation. Therefore, further work is needed to examine and develop the concepts required to reflect on the characteristics mentioned above (age, gender) as part of the estimation of health condition evolution. Although it could add complexity to the model, we believe it will ultimately benefit the identification of vulnerable people as it will provide a more accurate definition of convalescence time.

Knowledge acquisition and knowledge graph construction. Regarding the user study, we recruited experts with medical background and asked them to annotate the same randomly selected list of SNOMED CT. We did not consider their areas of expertise and assumed their general knowledge was sufficient to annotate SNOMED CT concepts. Although results show that participants were familiar with the given concepts in 7 out of 10 cases, the data collected allowed the analysis of raters' agreement based on a reduced number of conditions (23, in total, the number of conditions that all participants annotated). In the future, it will be important to explore the potential of presenting participants with a dataset tailored according to their domain of expertise, considering specific branches of SNOMED CT taxonomy. Ideally, curating the dataset, considering participants' backgrounds, will accelerate the annotation task and effectively capture humans' knowledge.

The user study's main objective was to validate the overall approach to extracting health condition evolution from natural language. We concentrated on evaluating the relevance and accuracy of the recommendations as a final result of the approach. The results indicate that the recommendations generated as part of knowledge completion can considerably accelerate the annotation of SNOMED CT. Therefore, we consider it important to explore the refinement of the propagation rules proposed in (Section 5.4) and develop new rules that could expedite the propagation to all SNOMED CT taxonomy.

Reasoning with Health Condition Evolution and the CONRAD system. In this work, we considered the use of synthetic electronic health records generated by the state-of-the-art software available at the time this research started. The decision to use synthetic data for this research was based mainly on the lack of real data available for research purposes; access to that data is still very restricted due to data sensitivity and ethical issues [Aviñó, Ruffini, and Gavalda 2018]. Therefore, we decided to overcome this limitation by using a synthetic dataset. We relied on Synthea, an open-source software that generates patients' historical medical data. Synthea was one of the first synthetic healthcare data generators and proved reliable in modelling healthcare data [Chen et al. 2019]. Although the dataset analysed was synthetic, the medical records were an excellent simulation of a real scenario. Data such as the type of medical event (represented by the SNOMED CT concept) and the date of the event are always part of electronic health records.

The analysis of the quality of the synthetic EHR dataset was not part of this research. However, we understand that the output is only as good as the data received as input. Therefore, future work should contemplate a set of experiments processing real electronic health records in a controlled scenario or use case. Crucially, the performance of CONRAD should be tested under realistic considerations.

Other important aspects that can be studied further are the resources required to analyse large data sources such as EHR. In our research, the time and computational resources required to process a patient's electronic health records were not part of the data collected during the experiments; however, an analysis is required in order to verify that valuable information could reach emergency services on time. Considerations about processing resources will become even more relevant in coming years [Allam and Jones 2020; Shah and Khan 2020]. Indeed, a report by Stanford Medicine [Stanford Medicine 2017] estimated that by the end of 2020, the health sector will generate globally over 2314 exabytes (1 exabyte = 1 billion gigabytes) of data. This increase, especially in the medical field, is attributed to the adoption of information technologies in every area of healthcare services and that its collection is no longer limited to hospital records.

A natural progression of this work is to analyse and expand on the utility of providing context data about the type of disability and the severity of the medical event. As part of the CONRAD system design, we included a module dedicated to adding information that could support less experimented emergency responders examining health records. Further research is needed to investigate the use of common-sense knowledge bases such as ConceptNet in order to support the interpretation of health records for timely emergency response in the

Smart City.

8.3 Conclusions

The work presented in this thesis is the result of four years of research during which several challenges and issues were overcome. In an attempt to validate the research hypothesis that it is possible to provide emergency services with information about people requiring assistance by automatically analysing people's health records, we explored how to capture knowledge about health condition evolution and the reasoning behind the process of estimating if the condition is on hold when an emergency occurred. We developed four knowledge components with the aim of solving the lack of resources regarding the representation of health evolution and estimation of it, particularly needed in an emergency scenario.

The CONRAD system is the instantiation of the approach proposed to solve the problem of identifying ongoing health issues by analysing electronic health records (EHR). The results of the experiments that tested CONRAD performance confirm that EHR can be used to reason on health condition evolution to identify people requiring assistance in emergencies.

Bibliography

- Abu-Elkheir, Mervat, Hossam S. Hassanein, and Sharief M.A. Oteafy (Sept. 2016). ‘Enhancing emergency response systems through leveraging crowdsensing and heterogeneous data’. en. In: *2016 Int Wireless Communications and Mobile Computing Conference (IWCMC)*. IEEE, pp. 188–193. URL: <http://ieeexplore.ieee.org/document/7577055/>.
- Abu-Salih, Bilal et al. (2022). ‘Healthcare Knowledge Graph Construction: State-of-the-art, open issues, and opportunities’. en. In: p. 29.
- Alfattni, Ghada, Niels Peek, and Goran Nenadic (Aug. 2020). ‘Extraction of temporal relations from clinical free text: A systematic review of current approaches’. In: *Journal of Biomedical Informatics* 108, p. 103488.
- Allam, Zaheer and David S. Jones (Feb. 2020). ‘On the Coronavirus (COVID-19) Outbreak and the Smart City Network: Universal Data Sharing Standards Coupled with Artificial Intelligence (AI) to Benefit Urban Health Monitoring and Management’. en. In: *Healthcare* 8.1, p. 46.
- Allen, Adrienne S. and Thomas D. Sequist (Nov. 2012). ‘Pharmacy Dispensing of Electronically Discontinued Medications’. en. In: *Annals of Internal Medicine* 157.10, p. 700.
- Aviñó, Laura, Matteo Ruffini, and Ricard Gavalda (July 2018). ‘Generating Synthetic but Plausible Healthcare Record Datasets’. en. In: arXiv:1807.01514. arXiv:1807.01514 [cs, stat]. URL: <http://arxiv.org/abs/1807.01514>.
- Ayaz, Muhammad et al. (July 2021). ‘The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities’. In: *JMIR Medical Informatics* 9.7, e21929.
- Bakıcı, Tuba, Esteve Almirall, and Jonathan Wareham (June 2013). ‘A Smart City Initiative: the Case of Barcelona’. en. In: *Journal of the Knowledge Economy* 4.2, pp. 135–148.
- Bakken, D.E. et al. (Nov. 2004). ‘Data obfuscation: anonymity and desensitization of usable data sets’. en. In: *IEEE Security and Privacy Magazine* 2.6, pp. 34–41.

- Baowaly, Mrinal Kanti et al. (2019). ‘Synthesizing electronic health records using improved generative adversarial networks’. en. In: *Journal of the American Medical Informatics Association* 26.3, p. 14.
- Bartoli, G. et al. (Mar. 2015). ‘A novel emergency management platform for smart public safety: A Novel Emergency Management Platform’. In: *International Journal of Communication Systems* 28.5, pp. 928–943.
- Ben-Assuli, Ofir and Moshe Leshno (Sept. 2016). ‘Assessing electronic health record systems in emergency departments: Using a decision analytic Bayesian model’. en. In: *Health Informatics Journal* 22.3, pp. 712–729.
- Benson, Tim and Grahame Grieve (2021). *Principles of Health Interoperability: FHIR, HL7 and SNOMED CT*. en. Health Information Technology Standards. Cham: Springer International Publishing. URL: <https://link.springer.com/10.1007/978-3-030-56883-2>.
- Bergman, Michael K. (July 2019). *A Common Sense View of Knowledge Graphs*. Adaptive Information, Adaptive Innovation, Adaptive Infrastructure Blog. URL: <http://www.mkbergman.com/2244/a-common-sense-view-of-knowledge-graphs/>.
- Berners-Lee, Tim and Mark Fischetti (1999). *Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor*. 1st ed. San Francisco: Harper-SanFrancisco.
- Bertino, Elisa (June 2016). ‘Data Security and Privacy: Concepts, Approaches, and Research Directions’. en. In: *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*. Atlanta, GA: IEEE, pp. 400–407. URL: <http://ieeexplore.ieee.org/document/7552042/>.
- Bhanot, Karan et al. (Mar. 2022). ‘Downstream Fairness Caveats with Synthetic Healthcare Data’. en. In: arXiv:2203.04462. arXiv:2203.04462 [cs]. URL: <http://arxiv.org/abs/2203.04462>.
- Bodenreider, Oliver, Ronald Cornet, and Daniel Vreeman (2018). ‘Recent Developments in Clinical Terminologies — SNOMED CT, LOINC, and RxNorm’. In: *Yearbook of Medical Informatics*.
- Boukerche, Azzedine and Rodolfo W. L. Coutinho (June 2018). ‘Smart Disaster Detection and Response System for Smart Cities’. en. In: *2018 IEEE Symposium on Computers and Communications (ISCC)*. Natal: IEEE, pp. 01102–01107. URL: <https://ieeexplore.ieee.org/document/8538356/>.
- Bowerman, B et al. (2000). ‘The vision of a smart city’. In: *2nd International Life Extension Technology Workshop*. Paris, p. 2000.

- Britain, Great, Department for Communities, and Local Government (2006). *Fire safety risk assessment: educational premises*. English. London: Department for Communities and Local Government.
- Caird, S, L Hudson, and G. Kortuem (2016). *A Tale of Evaluation and Reporting in UK Smart Cities*, p. 51.
- Cassandras, Christos G. (June 2016). ‘Smart Cities as Cyber-Physical Social Systems’. In: *Engineering* 2.2, pp. 156–158.
- Centers for Disease Control and Prevention (1996). *Health insurance portability and accountability act of 1996 (HIPAA)*. URL: <https://www.cdc.gov/%20phlp/publications/topic/hipaa.html>.
- Chang, Eunsuk and Javed Mostafa (Aug. 2021). ‘The use of SNOMED CT, 2013-2020: a literature review’. en. In: *Journal of the American Medical Informatics Association* 28.9, pp. 2017–2026.
- Chaudhri, Vinay, Naren Chittar, and Michael Genesereth (May 2021). *An Introduction to Knowledge Graphs*. URL: <https://ai.stanford.edu/blog/introduction-to-knowledge-graphs/>.
- Chehade, Samer, Nada Matta, Jean-Baptiste Pothin, et al. (Oct. 2018). ‘Data Interpretation Support in Rescue Operations: Application for French Firefighters’. en. In: *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–6.
- Chehade, Samer, Nada Matta, Jean-Baptiste Pothin, et al. (Sept. 2020). ‘Handling effective communication to support awareness in rescue operations’. en. In: *Journal of Contingencies and Crisis Management* 28.3, pp. 307–323.
- Chen, Junqiao et al. (Dec. 2019). ‘The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures’. en. In: *BMC Medical Informatics and Decision Making* 19.1, p. 44.
- Clark, Sophie Jane et al. (Aug. 2019). ‘Using deterministic record linkage to link ambulance and emergency department data: is it possible without patient identifiers?: A case study from the UK’. en. In: *International Journal of Population Data Science* 4.1. URL: <https://ijpds.org/article/view/1104>.
- Cook, Diane J. et al. (Apr. 2018). ‘Using Smart City Technology to Make Healthcare Smarter’. In: *Proceedings of the IEEE* 106.4, pp. 708–722.
- Cunningham, Hamish (2005). ‘Information extraction, automatic’. In: *Encyclopedia of language and linguistics*, 3.8, p. 10.

- Curzon, James, Abdulaziz Almeahmadi, and Khalil El-Khatib (Apr. 2019). 'A survey of privacy enhancing technologies for smart cities'. en. In: *Pervasive and Mobile Computing* 55, pp. 76–95.
- Daga, Enrico, Luigi Asprino, et al. (Aug. 2021). 'Facade-X: An Opinionated Approach to SPARQL Anything'. In: *Volume 53: Further with Knowledge Graphs*. Vol. 53. IOS Press, pp. 58–73. URL: <http://oro.open.ac.uk/78973/>.
- Daga, Enrico, Mathieu d'Aquin, et al. (2015). 'Propagation of Policies in Rich Data Flows'. en. In: *Proceedings of the Knowledge Capture Conference - K-CAP 2015*. Palisades, NY, USA: ACM Press, pp. 1–8. URL: <http://dl.acm.org/citation.cfm?doid=2815833.2815839>.
- Al-Dahash, Hajer, Menaha Thayaparan, and Udayangani Kulatunga (2016). 'Understanding the terminologies: Disaster, crisis and emergency'. In: *Proceedings of the 32nd Annual ARCOM Conference, ARCOM 2016 (pp. 1191-1200)*. London South Bank University.
- Dalianis, Hercules (2018). *Clinical Text Mining*. en. Cham: Springer International Publishing. URL: <http://link.springer.com/10.1007/978-3-319-78503-5>.
- Daoud, Mariam et al. (N.D.). 'Integrating Semantic Medical Entity Relations for Disease Prediction Using SNOMED-CT Terminology'. en. In: p. 13.
- Dodig-Crnkovic, Gordana (2002). 'Scientific Methods in Computer Science'. In: *Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden*, p. 7.
- Earley, Seth (Jan. 2016). 'Really, Really Big Data: NASA at the Forefront of Analytics'. In: *IT Professional* 18.1, pp. 58–61.
- Ehrlinger, Lisa and Wolfram Wöß (2016). 'Towards a definition of knowledge graphs.' In: *SEMANTiCS (Posters, Demos, SuCCESS)* 48.1-4, p. 2.
- European Parliament (2016). *General Data Protection Regulation (GDPR)*. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&rid=1#d1e1374-1-1>.
- Fensel, Dieter et al. (2020). *Knowledge Graphs: Methodology, Tools and Selected Use Cases*. en. Cham: Springer International Publishing. URL: <http://link.springer.com/10.1007/978-3-030-37439-6>.
- Galton, Antony and Michael Worboys (2011). 'An Ontology of Information for Emergency Management'. In: *Proc. of ISCRAM*, p. 10.
- Garcia, Laura et al. (Sept. 2018). 'System for Detection of Emergency Situations in Smart City Environments Employing Smartphones'. en. In: *2018 International Conference on*

- Advances in Computing, Communications and Informatics (ICACCI)*. Bangalore: IEEE, pp. 266–272. URL: <https://ieeexplore.ieee.org/document/8554654/>.
- Georges-Filteau, Jeremy and Elisa Cirillo (Nov. 2020). *Synthetic Observational Health Data with GANs: from slow adoption to a boom in medical research and ultimately digital twins?*
- Gharaibeh, Ammar et al. (2017). ‘Smart Cities: A Survey on Data Management, Security, and Enabling Technologies’. In: *IEEE Communications Surveys & Tutorials* 19.4, pp. 2456–2501.
- Gomez-Perez, Jose Manuel et al. (2017). ‘Enterprise Knowledge Graph: An Introduction’. en. In: *Exploiting Linked Data and Knowledge Graphs in Large Organisations*. Ed. by Jeff Z. Pan et al. Springer International Publishing, pp. 1–14. URL: http://link.springer.com/10.1007/978-3-319-45654-6_1.
- Greater London Authority (2018). *Smarter London Together*. Government. URL: https://www.london.gov.uk/sites/default/files/smarter_london_together_v1.66_-_published.pdf.
- Grishman, Ralph (Sept. 2015). ‘Information Extraction’. In: *IEEE Intelligent Systems* 30.5, pp. 8–15.
- Gruninger, Michael and Mark S Fox (1995). ‘Methodology for the Design and Evaluation of Ontologies’. In: *Workshop on Basic Ontological Issues in Knowledge Sharing*, p. 10.
- Hashem, Ibrahim Abaker Targio et al. (Oct. 2016). ‘The role of big data in smart city’. en. In: *International Journal of Information Management* 36.5, pp. 748–758.
- Healy, Michael, Achilles Kameas, and Roberto Poli (2010). *Theory and applications of ontology*. eng. Dordrecht London: Springer.
- Hernandez, Mikel et al. (Apr. 2022). ‘Synthetic Data Generation for Tabular Health Records: A Systematic Review’. en. In: *Neurocomputing*.
- HHS (1996). *Health Insurance Portability and Accountability Act of 1996 (HIPAA)*.
- HIMSS (2022). *Healthcare Information and Management Systems Society*. URL: <https://www.himss.org/resources/interoperability-healthcare>.
- HL7 (2019). *HL7FHIR*. URL: <https://www.hl7.org/fhir/overview.html>.
- Hobbs, Jerry R. and Pan Feng (Sept. 2006). *Time ontology in OWL*. URL: <https://www.w3.org/TR/owl-time/>.
- Hogan, Aidan (2020). ‘Resource Description Framework’. In: *The Web of Data*. Cham: Springer International Publishing, pp. 59–109. ISBN: 978-3-030-51580-5. DOI: [10.1007/978-3-030-51580-5_3](https://doi.org/10.1007/978-3-030-51580-5_3). URL: https://doi.org/10.1007/978-3-030-51580-5_3.

- Hogan, Aidan et al. (2021). *Knowledge Graphs*. en. Cham: Springer International Publishing. URL: <https://link.springer.com/10.1007/978-3-031-01918-0>.
- Hommersom, Arjen and Peter J. F. Lucas (2015). ‘An Introduction to Knowledge Representation and Reasoning in Healthcare’. In: *Foundations of Biomedical Knowledge Representation*. Ed. by Arjen Hommersom and Peter J.F. Lucas. Vol. 9521. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 9–32. URL: http://link.springer.com/10.1007/978-3-319-28007-3_2.
- Huang, Zhisheng et al. (2017). ‘Constructing Knowledge Graphs of Depression’. In: *Health Information Science*. Ed. by Siuly Siuly et al. Vol. 10594. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 149–161. URL: http://link.springer.com/10.1007/978-3-319-69182-4_16.
- Hussain, Aamir et al. (Dec. 2015). ‘Health and emergency-care platform for the elderly and disabled people in the Smart City’. In: *Journal of Systems and Software* 110.7.
- Imtiaz, Sana et al. (July 2021). ‘Synthetic and Private Smart Health Care Data Generation using GANs’. In: *2021 International Conference on Computer Communications and Networks (ICCCN)*. Athens, Greece: IEEE, pp. 1–7. URL: <https://ieeexplore.ieee.org/document/9522203/>.
- Information Commissioner’s Office (2016). *Guide to the UK General Data Protection Regulation (UK GDPR)*. URL: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/>.
- Information Technology Industry Council (2018). *Information Technology Industry Council, Cloud Healthcare Pledge*. en. Cloud Healthcare Pledge. URL: <https://www.itic.org/public-policy/CloudHealthcarePledge.pdf>.
- Institute of Medicine (Mar. 2012). ‘Crisis Standards of Care: A Systems Framework for Catastrophic Disaster Response.’ In: *Crisis Standards of Care: A Systems Framework for Catastrophic Disaster Response*. Vol. 3. Washington (DC): National Academies Press (US). URL: <https://www.ncbi.nlm.nih.gov/books/NBK201058/>.
- Jensen, Peter B., Lars J. Jensen, and Søren Brunak (June 2012). ‘Mining electronic health records: towards better research applications and clinical care’. en. In: *Nature Reviews Genetics* 13.6, pp. 395–405.
- Ji, Shaoxiong et al. (Feb. 2022). ‘A Survey on Knowledge Graphs: Representation, Acquisition, and Applications’. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.2, pp. 494–514.

- Johannesson, Paul and Erik Perjons (2021). *An Introduction to Design Science*. Springer International Publishing. URL: <https://link.springer.com/10.1007/978-3-030-78132-3>.
- Juric, Damir et al. (Apr. 2020). ‘A System for Medical Information Extraction and Verification from Unstructured Text’. en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.08, pp. 13314–13319.
- Kadhim, Ammar Ismael (June 2019). ‘Survey on supervised machine learning techniques for automatic text classification’. en. In: *Artificial Intelligence Review* 52.1, pp. 273–292.
- Keogh, B (2013). *Transforming Urgent and Emergency Care Services in England. Urgent and Emergency Care Review*. Leeds. URL: <https://www.nhs.uk/NHSEngland/keogh-review/Documents/UECR.Ph1Report.FV.pdf>.
- Kim, Jung-Hoon and Joo-Young Kim (Mar. 2022). ‘How Should the Structure of Smart Cities Change to Predict and Overcome a Pandemic?’ en. In: *Sustainability* 14.5, p. 2981.
- Kotsiantis, SB, I Zaharakis, and P Pintelas (2007). ‘Supervised machine learning: A review of classification techniques.’ In: *Emerging artificial intelligence applications in computer engineering*. IOS Press,US (1 Oct. 2007), pp. 3–24.
- Kowsari et al. (Apr. 2019). ‘Text Classification Algorithms: A Survey’. en. In: *Information* 10.4, p. 150.
- Krippendorff, Klaus (2011). ‘Computing Krippendorff’s alpha-reliability’. In.
- Lakemeyer, Gerhard and Bernhard Nebel (1994). ‘Foundations of knowledge representation and reasoning’. en. In: p. 12.
- Lamprinakos, Georgios et al. (2014). ‘Using FHIR to develop a healthcare mobile application’. In: *Proceedings of the 4th ICWMCH*.
- Lebo, Timothy et al. (2013). *PROV-O: The PROV ontology*.
- Lehne, Moritz, Sandra Luijten, and Sylvia Thun (Sept. 2019). ‘The Use of FHIR in Digital Health - A Review of the Scientific Literature’. en. In: *German Medical Data Sciences*, p. 7.
- Leroux, Hugo, Alejandro Metke-Jimenez, and Michael J. Lawley (2017). ‘Towards achieving semantic interoperability of clinical study data with FHIR’. In: *Journal of Biomedical Semantics*.
- Levenshtein, Vladimir I et al. (1966). ‘Binary codes capable of correcting deletions, insertions, and reversals’. In: *Soviet physics doklady*. Vol. 10. 8. Soviet Union, pp. 707–710.
- Lewis, Bernard T. and Richard P. Payant (2003). *The facility manager’s emergency preparedness handbook*. New York: AMACOM.

- Li, Fang et al. (July 2020). ‘Time event ontology (TEO): to support semantic representation and reasoning of complex temporal relations of clinical events’. In: *Journal of the American Medical Informatics Association* 27.7. DOI: [10.1093/jamia/ocaa058](https://doi.org/10.1093/jamia/ocaa058).
- Lombardi, Patrizia et al. (June 2012). ‘Modelling the smart city performance’. en. In: *Innovation: The European Journal of Social Science Research* 25.2, pp. 137–149.
- López, Fernández (1999). ‘Overview Of Methodologies For Building Ontologies’. In: p. 13.
- Loukil, Faiza et al. (2017). ‘Privacy-Aware in the IoT Applications: A Systematic Literature Review’. en. In: *On the Move to Meaningful Internet Systems. OTM 2017 Conferences*. Ed. by Hervé Panetto et al. Vol. 10573. Cham: Springer International Publishing, pp. 552–569. URL: http://link.springer.com/10.1007/978-3-319-69462-7_35.
- Lung, Claudiu, Attila Buchman, and Sebastian Sabou (Oct. 2018). ‘Smart City Emergency Situations Management System Based on Sensors Network’. en. In: *2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging (SIITME)*. Iasi: IEEE, pp. 288–291. URL: <https://ieeexplore.ieee.org/document/8599257/>.
- Majumder, Sumit et al. (Oct. 2017). ‘Smart Homes for Elderly Healthcare—Recent Advances and Research Challenges’. In: *Sensors* 17.11, p. 2496.
- McClelland, S (2013). *A Strategic Review of Welsh Ambulance Services*. Cardiff. URL: <https://www.ambulance.wales.nhs.uk/assets/documents/f06e69f9-3921-4946-a55a-aad53637c282635179619910478381.pdf>.
- McGraw, Deven and Kenneth D. Mandl (Dec. 2021). ‘Privacy protections to encourage use of health-relevant digital data in a learning health system’. en. In: *npj Digital Medicine* 4.1, p. 2.
- McNeill, Fiona et al. (May 2019). ‘Communication in Emergency Management through Data Integration and Trust: an introduction to the CEM-DIT system’. In: *Proc. of ISCRAM*, p. 12.
- Mekruksavanich, Sakorn (Aug. 2016). ‘Medical expert system based ontology for diabetes disease diagnosis’. In: *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. Beijing, China: IEEE, pp. 383–389. URL: <http://ieeexplore.ieee.org/document/7883091/>.
- Mello, Blanda Helena de et al. (Jan. 2022). ‘Semantic interoperability in health records standards: a systematic literature review’. en. In: *Health and Technology*. URL: <https://link.springer.com/10.1007/s12553-022-00639-w>.
- Metke-Jimenez, Alejandro et al. (Dec. 2018). ‘Ontoserver: a syndicated terminology server’. en. In: *Journal of Biomedical Semantics* 9.1, p. 24.

- Mhadhbi, Linda and Jalel Akaichi (2017). ‘DS-Ontology: A Disease-Symptom Ontology for General Diagnosis Enhancement’. en. In: *Proceedings of the 2017 International Conference on Information System and Data Mining - ICISDM '17*. Charleston, SC, USA: ACM Press, pp. 99–102. URL: <http://dl.acm.org/citation.cfm?doid=3077584.3077586>.
- Möller, Manuel and Sonntag, Daniel and Ernst, Patrick (2013). ‘Modeling the International Classification of Diseases (ICD-10) in OWL’. In: *Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Ed. by Ana Fred et al. Vol. 272. Communications in Computer and Information Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 226–240. URL: http://link.springer.com/10.1007/978-3-642-29764-9_16.
- Morales Tirado, Alba Catalina, Enrico Daga, and Enrico Motta (Sept. 2020). ‘Effective Use of Personal Health Records to Support Emergency Services’. In: *Knowledge Engineering and Knowledge Management*. Springer International Publishing, pp. 54–70. DOI: [10.1007/978-3-030-61244-3_4](https://doi.org/10.1007/978-3-030-61244-3_4).
- (Dec. 2021). ‘Reasoning on Health Condition Evolution for Enhanced Detection of Vulnerable People in Emergency Settings’. en. In: *Proceedings of the 11th KCAP Conference*. ACM, pp. 9–16. URL: <https://dl.acm.org/doi/10.1145/3460210.3493551>.
- (May 2022a). ‘CONRAD - Health Condition Radar: an Intelligent System for Emergency Support’. In: *5th Workshop on Semantic Web solutions for large-scale biomedical data analytics*.
- (May 2022b). ‘HECON Health: Condition Evolution Ontology’. In: *5th Workshop on Semantic Web solutions for large-scale biomedical data analytics*.
- (Sept. 2022c). ‘Towards a Knowledge Graph of Health Evolution’. In: *Knowledge Engineering and Knowledge Management. EKAW 2022*.
- Moreira, João, Marten van Sinderen, and Luís Ferreira Pires (2019). ‘SEMIoTICS: Semantic Model-Driven Development for IoT Interoperability of Emergency Services’. en. In: *Proc. 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, p. 13.
- Newell, Allen (1982). ‘The knowledge level’. In: *Artificial intelligence* 18.1, pp. 87–127.
- NHS Digital services (2022a). *FHIR UK Core*. URL: <https://digital.nhs.uk/services/fhir-uk-core>.
- (2022b). *SNOMED CT is a clinical vocabulary readable by computers*. URL: <https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct>.

- NHS England (July 2019). *Planning to Safely Reduce Avoidable Conveyance: Ambulance Improvement Programme*. URL: <https://aace.org.uk/wp-content/uploads/2019/08/safetly-reduce-avoidable-conveyance-v2.0.pdf>.
- Nunavath, Vimala, Andreas Prinz, and Tina Comes (2016). ‘Identifying First Responders Information Needs: Supporting Search and Rescue Operations for Fire Emergency Response’. In: *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*.
- Olex, Amy L. and Bridget T. McInnes (June 2021). ‘Review of Temporal Reasoning in the Clinical Domain for Timeline Extraction: Where we are and where we need to be’. en. In: *Journal of Biomedical Informatics* 118, p. 103784.
- Oyelade, Olaide Nathaniel et al. (2021). ‘A semantic web rule and ontologies based architecture for diagnosing breast cancer using select and test algorithm’. en. In: *Computer Methods and Programs in Biomedicine Update* 1, p. 100034.
- Palmieri, Francesco et al. (Nov. 2016). ‘A cloud-based architecture for emergency management and first responders localization in smart city environments’. en. In: *Computers & Electrical Engineering* 56, pp. 810–830.
- Patterson, Rebecca et al. (Dec. 2019). ‘Paramedic information needs in end-of-life care: a qualitative interview study exploring access to a shared electronic record as a potential solution’. en. In: *BMC Palliative Care* 18.1, p. 108.
- Pellicer, Soledad et al. (July 2013). ‘A Global Perspective of Smart Cities: A Survey’. In: *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. IEEE, pp. 439–444. URL: <http://ieeexplore.ieee.org/document/6603712/>.
- Phillips, Brenda D, David M Neal, and Gary Webb (2016). *Introduction to emergency management*. CRC Press.
- Porter, Alison et al. (Feb. 2020). ‘Electronic health records in ambulances: the ERA multiple-methods study’. en. In: *Health Services and Delivery Research* 8.10, pp. 1–140.
- Pramanik, Md Ileas et al. (Nov. 2017). ‘Smart health: Big data enabled health paradigm within smart cities’. en. In: *Expert Systems with Applications* 87, pp. 370–383.
- Prasanna, Kankanamge (2010). ‘Information Systems for Supporting Fire Emergency Response’. en. PhD thesis. URL: <https://hdl.handle.net/2134/7648>.
- Presutti, Valentina et al. (2009). ‘eXtreme Design with Content Ontology Design Patterns’. In: *Proceedings Workshop on Ontology Patterns*.

- Qu, Jia (July 2022). ‘A Review on the Application of Knowledge Graph Technology in the Medical Field’. en. In: *Scientific Programming* 2022. Ed. by Jianping Gou, pp. 1–12.
- Rahman, Sk. Md. Mizanur et al. (Sept. 2016). ‘Privacy preserving secure data exchange in mobile P2P cloud healthcare environment’. en. In: *Peer-to-Peer Networking and Applications* 9.5, pp. 894–909.
- Rajput, Ahmed Raza et al. (2019). ‘EACMS: Emergency Access Control Management System for Personal Health Record Based on Blockchain’. en. In: *IEEE Access* 7, pp. 84304–84317.
- Rego, Albert et al. (Nov. 2018). ‘Software Defined Network-based control system for an efficient traffic management for emergency situations in smart cities’. In: *Future Generation Computer Systems* 88, pp. 243–253.
- Riaño, David et al. (June 2012). ‘An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients’. en. In: *Journal of Biomedical Informatics* 45.3, pp. 429–446.
- Rocha, Pacheco et al. (Aug. 2019). ‘Smart Cities and Healthcare: A Systematic Review’. en. In: *Technologies* 7.3, p. 58.
- Romanou, Anna (Feb. 2018). ‘The necessity of the implementation of Privacy by Design in sectors where data protection concerns arise’. en. In: *Computer Law & Security Review* 34.1, pp. 99–110.
- Rosenfield, Daniel, Gregory Harvey, and Karim Jessa (Jan. 2019). ‘Implementing electronic medical records in Canadian emergency departments’. en. In: *CJEM* 21.1, pp. 15–17.
- Rotmensch, Maya et al. (Dec. 2017). ‘Learning a Health Knowledge Graph from Electronic Medical Records’. en. In: *Scientific Reports* 7.1, p. 5994.
- Saripalle, Rishi, Christopher Runyan, and Mitchell Russell (June 2019). ‘Using HL7 FHIR to achieve interoperability in patient health record’. en. In: *Journal of Biomedical Informatics* 94, p. 103188.
- Schiaffonati, Viola and Mario Verdicchio (Sept. 2014). ‘Computing and Experiments: A Methodological View on the Debate on the Scientific Nature of Computing’. In: *Philosophy & Technology* 27.3, pp. 359–376.
- Semenov, Ilia et al. (Dec. 2019). ‘Experience in Developing an FHIR Medical Data Management Platform to Provide Clinical Decision Support’. en. In: *International Journal of Environmental Research and Public Health* 17.1, p. 73.
- Shah, Shahid Munir and Rizwan Ahmed Khan (2020). ‘Secondary Use of Electronic Health Record: Opportunities and Challenges’. en. In: *IEEE Access* 8, pp. 136947–136965.

- Shi, Longxiang et al. (2017). ‘Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services’. In: *BioMed Research International* 2017, pp. 1–12.
- Shibuya, Yuya and Hideyuki Tanaka (2019). ‘Detecting Disaster Recovery Activities via Social Media Communication Topics’. en. In: *Proc. 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, p. 13.
- Smith, Holly (Aug. 2017). *Pre-hospital emergency department data (PHED)*. Health. URL: <https://www.nuffieldtrust.org.uk/project/pre-hospital-emergency-department-data-phed>.
- SNOMED International (July 2017). *SNOMED CT Starter Guide*. URL: <http://snomed.org/sg>.
- (Jan. 2022a). URL: <https://confluence.ihtsdotools.org/display/DOCOWL/SNOMED+CT+OWL+Guide>.
- (2022b). *Expression Constraint Language - Specification and Guide*. en. URL: <http://snomed.org/doc>.
- (2022c). *SNOMED CT International*. URL: <http://www.snomed.org/snomed-ct/five-step-briefing>.
- (2022d). *SNOMED CT Starter Guide*. en. URL: https://www.snomed.org/SNOMED/media/SNOMED/documents/doc_StarterGuide_Current-en-US_INT_20140731.pdf.
- Soyiri, Ireneous N. and Daniel D. Reidpath (Jan. 2013). ‘An overview of health forecasting’. en. In: *Environmental Health and Preventive Medicine* 18.1, pp. 1–9.
- Speer, Robyn, Joshua Chin, and Catherine Havasi (2017). ‘ConceptNet 5.5: An Open Multilingual Graph of General Knowledge’. In: *AAAI 31*, p. 8.
- Srinivasan, Arunkumar et al. (2006). ‘Semantic Web Representation of LOINC: an Ontological Perspective’. en. In: p. 1.
- Srinivasan, Ramya, Apurva Mohan, and Priyanka Srinivasan (Apr. 2016). ‘Privacy conscious architecture for improving emergency response in smart cities’. en. In: *2016 Smart City Security and Privacy Workshop (SCSP-W)*. Vienna, Austria: IEEE, pp. 1–5. URL: <http://ieeexplore.ieee.org/document/7509559/>.
- Stanford Medicine (2017). *Harnessing the Power of Data in Health*. Stanford, CA, USA. URL: <https://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicine%20HealthTrendsWhitePaper2017.pdf>.
- Suarez-Figueroa, Mari Carmen, Asuncion Gomez-Perez, and Mariano Fernandez-Lopez (2012). ‘The NeOn Methodology for Ontology Engineering’. In: *Ontology Engineering in a Networked World*, pp. 9–34.

- Tayefi, Maryam et al. (Nov. 2021). ‘Challenges and opportunities beyond structured data in analysis of electronic health records’. en. In: *WIREs Computational Statistics* 13.6. URL: <https://onlinelibrary.wiley.com/doi/10.1002/wics.1549>.
- The New York Times (June 2019). *Google and the University of Chicago Are Sued Over Data Sharing*. URL: <https://www.nytimes.com/2019/06/26/technology/google-university-chicago-data-sharing-lawsuit.html>.
- Thummavet, P. and S. Vasupongayya (Sept. 2013). ‘A novel personal health record system for handling emergency situations’. en. In: *2013 International Computer Science and Engineering Conference (ICSEC)*. IEEE, pp. 266–271. URL: <http://ieeexplore.ieee.org/document/6694791/>.
- Turoff, Murray and Michael Chumer (2004). ‘The Design of a Dynamic Emergency Response Management Information System (DERMIS)’. en. In: p. 36.
- UK Government (Sept. 2006). *Addressing lessons from the emergency response to the 7 July 2005 London bombings*. English, p. 15. URL: <https://www.statewatch.org/news/2006/sep/uk-ho-7-july-2005-report.pdf>.
- (2007a). *Data protection and sharing guidance for emergency planners and responders*. URL: <https://www.gov.uk/government/publications/data-protection-and-sharing-guidance-for-emergency-planners-and-responders>.
- (2007b). *Fire safety risk assessment: means of escape for disabled people*. URL: <https://www.gov.uk/government/publications/fire-safety-risk-assessment-means-of-escape-for-disabled-people>.
- (Feb. 2008). *Identifying People Who Are Vulnerable in a Crisis. Guidance for Emergency Planners and Responders*. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/61228/vulnerable_guidance.pdf.
- (2013a). *Emergency response and recovery guidance*. Cabinet Office. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/253488/Emergency_Response_and_Recovery_5th_edition_October_2013.pdf.
- (2013b). *Preparation and planning for emergencies: responsibilities of responder agencies and others*. Government. URL: <https://www.gov.uk/guidance/preparation-and-planning-for-emergencies-responsibilities-of-responder-agencies-and-others>.
- United Nations (2022). *World Urbanization Prospects 2018*. en. URL: <https://population.un.org/wup/Publications/>.
- Uschold, Mike and Michael Gruninger (June 1996). ‘Ontologies: principles, methods and applications’. In: *The Knowledge Engineering Review* 02.

- VanLangen, Kali and Greg Wellman (May 2018). 'Trends in electronic health record usage among US colleges of pharmacy'. en. In: *Currents in Pharmacy Teaching and Learning* 10.5, pp. 566–570.
- Verma, Rupali (Jan. 2022). 'Smart City Healthcare Cyber Physical System: Characteristics, Technologies and Challenges'. en. In: *Wireless Personal Communications* 122.2, pp. 1413–1433.
- Walonoski, Jason et al. (Mar. 2018). 'Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record'. In: *Journal of the American Medical Informatics Association* 25.3. DOI: [0.1093/jamia/ocx079](https://doi.org/10.1093/jamia/ocx079).
- Webb, William and Chai Keong Toh (July 2020). 'The Smart City and Covid-19'. en. In: *IET Smart Cities* 2.2, pp. 56–57.
- Wiljer, David et al. (Oct. 2008). 'Patient Accessible Electronic Health Records: Exploring Recommendations for Successful Implementation Strategies'. In: *Journal of Medical Internet Research* 10.4, e34.
- Wilson, Jennifer and Arthur Oyola-Yemaiel (Oct. 2001). 'The evolution of emergency management and the advancement towards a profession in the United States and Florida'. en. In: *Safety Science* 39.1–2, pp. 117–131.
- World Health Organization (2022). *Classification of Diseases (ICD)*. URL: <https://www.who.int/standards/classifications/classification-of-diseases>.
- World Wide Web Consortium (W3C) (2008). *SPARQL Query Language for RDF*. URL: <https://www.w3.org/TR/rdf-sparql-query/>.
- (2012). *OWL 2 Web Ontology Language*. URL: <https://www.w3.org/TR/2012/REC-owl2-quick-reference-20121211/>.
- (2014a). *RDF 1.1 N-Triples*. URL: <https://www.w3.org/TR/n-triples/>.
- (2014b). *RDF 1.1 Turtle*. URL: <https://www.w3.org/TR/turtle/>.
- Xiang, Dingyi and Wei Cai (Oct. 2021). 'Privacy Protection and Secondary Use of Health Data: Strategies and Methods'. en. In: *BioMed Research International* 2021. Ed. by Lei Zhang, pp. 1–11.
- Xu, Boyi et al. (May 2014). 'Ubiquitous Data Accessing Method in IoT-Based Information System for Emergency Medical Services'. In: *IEEE Transactions on Industrial Informatics* 10.2, pp. 1578–1586.

- Yu, Wenbin et al. (Mar. 2017). ‘Privacy-preserving design for emergency response scheduling system in medical social networks’. en. In: *Peer-to-Peer Networking and Applications* 10.2, pp. 340–356.
- Zahid, Arnob et al. (May 2021). ‘A systematic review of emerging information technologies for sustainable data-centric health-care’. en. In: *International Journal of Medical Informatics* 149, p. 104420.
- Zhang, Joe et al. (Mar. 2020). ‘Interoperability in NHS hospitals must be improved: the Care Quality Commission should be a key actor in this process’. en. In: *Journal of the Royal Society of Medicine* 113.3, pp. 101–104.
- Zhou, Li and George Hripcsak (Apr. 2007). ‘Temporal reasoning with medical data—A review with emphasis on medical natural language processing’. en. In: *Journal of Biomedical Informatics* 40.2, pp. 183–202.
- Zorab, Ollie, Maria Robinson, and Ruth Endacott (Dec. 2015). ‘Are prehospital treatment or conveyance decisions affected by an ambulance crew’s ability to access a patient’s health information?’ en. In: *BMC Emergency Medicine* 15.1, p. 26.
- Zygiaris, Sotiris (June 2013). ‘Smart City Reference Model: Assisting Planners to Conceptualize the Building of Smart City Innovation Ecosystems’. en. In: *Journal of the Knowledge Economy* 4.2, pp. 217–231.

Appendix A

Appendix - Glossary

This section is dedicated to providing a clear definition and clarification of technical and specific terms used in this thesis. Each term definition also references the sources used to support the stated definition.

Table A.1: Glossary

Term	Definition	Source / Reference
Cabinet Office	Department of the United Kingdom Government responsible for supporting the Prime Minister and Cabinet	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Category 1 responder	A person or body is likely to be at the core of the response to most emergencies. As such, they are subject to the full range of civil protection duties in the Act.	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]

Table A.1 continued from previous page

Category 2 responder	A person or body. These are cooperating responders who are less likely to be involved in the heart of multi-agency planning work but will be heavily involved in preparing for incidents affecting their sectors. The Act requires them to cooperate and share information with other Category 1 and 2 responders	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
CES	See Health Evolution Statement (HES). CES or Condition Evolution Statement is deprecated, and it was the first version of the Condition Evolution model and the name of the database of health evolution statements	Reference found in [Morales Tirado, Daga, and Motta 2021]
Civil Contingencies Act (2004)	The Act of 2004 established a single framework for Civil Protection in the United Kingdom. Part 1 of the Act establishes a clear set of roles and responsibilities for Local Responders; Part 2 of the Act establishes emergency powers. <i>Note: in the UK civil protection context, the CCA may often be referred to as 'The Act'.</i>	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Condition	A clinical condition, problem, diagnosis, or other events, situation, issue, or clinical concept has risen to a level of concern. This resource is used to record detailed information about a condition, problem, diagnosis, or other events, situations, issues, or clinical concept that has risen to a level of concern.	FHIR terminology, taken from: http://hl7.org/fhir/condition.html

Table A.1 continued from previous page

Crisis	1. General definition: an inherently abnormal, unstable and complex situation that represents a threat to strategic objectives, reputation or existence of an organisation.	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Crisis	2. Specific definition - emergency of magnitude and/or severity requiring the activation of central government response	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Current health issues	We use the term ‘current’ to refer to a health condition affecting the citizen during an emergency.	Reference described in [Morales Tirado, Daga, and Motta 2021]
Decline	A Progress type represents medical events that evolve adversely and gradually worsen over time.	Reference found in [Morales Tirado, Daga, and Motta 2022b]
Disaster	Emergency (usually but not exclusively of natural causes) causing or threatening to cause, severe and widespread disruption to community life through death, injury, and/or damage to property and/or the environment	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Emergency	An event or situation which threatens serious damage to human welfare in a place in the UK, the environment of a place in the UK, or the security of the UK or of a place in the UK. Note: to constitute an emergency, this event or situation must require the implementation of special arrangements by one or more Category 1 responder.	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]

Table A.1 continued from previous page

Emergency management	A multi-agency approach to emergency management entails six key activities –anticipation, assessment, prevention, preparation, response and recovery	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Emergency medical services (EMS)	Prehospital care is provided by emergency medical services, who are the initial healthcare providers at the scene of a disaster.	[Institute of Medicine 2012]
Emergency services	Also, emergency service. Generic term for police, fire and rescue, and health agencies; may also include HM Coastguard and other responders.	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Health condition	See Condition	
Health Condition Evolution Ontology (HECON)	The formal model represents the evolution of health events over time.	[Morales Tirado, Daga, and Motta 2022b]
Health Evolution Statement (HES)	Health Condition Evolution Statement is an abstraction of the components that describe the health recovery process. Initially called CES but then changed to generalise the fact that we are not only talking about conditions but also procedures, etc.	[Morales Tirado, Daga, and Motta 2022b]

Table A.1 continued from previous page

Impact	The scale of the consequences of a hazard, threat or emergency is expressed in terms of a reduction in human welfare, damage to the environment and loss of security	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Improvement	A Progress type represents medical events that evolve favourably, indicating recovery of good health.	[Morales Tirado, Daga, and Motta 2022b]
Incident	Event or situation that requires a response from the emergency services or other responders. <i>Note: emergency (or major incident) refers to a specific type of incident requiring special deployment by one or more category 1 responder</i>	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Pace	The velocity at which a medical event evolves. It could take values such as FAST, MODERATE and SLOW	[Morales Tirado, Daga, and Motta 2022b]
Permanent	A HES type that indicates a persistent medical event.	[Morales Tirado, Daga, and Motta 2022b]
Prehospital care	Prehospital care is an essential part of the continuum of emergency health care that is frequently initiated by a 911 call to a dispatch centre.	[Institute of Medicine 2012]
Progress	A HES type represents medical events that evolve over time.	[Morales Tirado, Daga, and Motta 2022b]
Responder	Organisations are required to plan and prepare a response to an emergency. See Category 1 responder; Category 2 responder.	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]

Table A.1 continued from previous page

Statutory	Prescribed in legislation	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]
Threat	Intent and capacity to cause loss of life or create adverse consequences to human welfare (including property and the supply of essential services and commodities), the environment or security.	Emergency Response and Recovery guidelines by the UK HM Government [UK Government 2013a]

Appendix B

Appendix - Summary software

In this section, we list the contributions of our work, a brief description, and the repositories where software, datasets and pilots can be found.

Table B.1: Contributions and repository description

Name	DOI	Chapter
Health Condition Evolution Ontology - HECON The ontology for representing and reasoning about the evolution of health events over time.	https://doi.org/10.5281/zenodo.7121514	Ch. 4
Knowledge Acquisition pipeline A pipeline that allows the <i>semi-automatic</i> extraction of information about health evolution and includes humans in the loop to curate the generated data.	https://doi.org/10.5281/zenodo.7121535	Ch. 5
Knowledge graph Defines an abstraction of the available data about health evolution. The KG includes information that extends the descriptions of the clinical concepts provided by SNOMED CT.	https://doi.org/10.5281/zenodo.7121532	Ch. 5

Table B.1 continued from previous page

<p>CONRAD prototype and experiments results</p> <p>Health Condition Radar is the intelligent system that implements the proposed methodology. It uses as data input a sample of randomly selected synthetic health records. The final output is a list of people with ongoing health issues that potentially require assistance to evacuate during a fire emergency.</p>	<p>https://doi.org/10.5281/zenodo.7122115</p>	<p>Ch. 6</p>
<p>User study results and human-in-the-loop tool</p> <p>This repository provides all the data collected as part of the knowledge acquisition pipeline and the efforts to capture human knowledge. The tool used for this task is also available in the repository.</p>	<p>https://doi.org/10.5281/zenodo.7121569</p>	<p>Ch. 7</p>

Appendix C

Appendix - Emergency response

This section is dedicated to providing a concise description of Emergency management literature, the type of emergencies and the emergency management cycle.

C.1 Emergency management

In this section, we provide a further review of the literature regarding Emergency management and emergency response

According to its nature, emergencies can be grouped into three different categories [Lewis and Payant 2003]:

- Natural: emergencies are the result of weather or environmental conditions. Examples include earthquakes, fires, storms, hurricanes, tornadoes, tidal waves, floods, and droughts.
- Technological: for example, hazardous material incidents, telecommunications failures, and electrical power outages.
- Human origin: such as crime (assault, theft, robbery, civil disturbances, and vandalism), bomb threats, terrorist attacks, medical crises, cyberterrorism, and environmental accidents.

Additionally, emergency events are classified, taking into account their impact on life, property damage, health or environment; the aim is to allocate the appropriate resources according to the priority and affectation of such situations. In this context, events can be classified as everyday life emergencies and crises and disasters; the crisis and disaster terms share common features, and they can be used interchangeably [Al-Dahash, Thayaparan, and Kulatunga 2016].

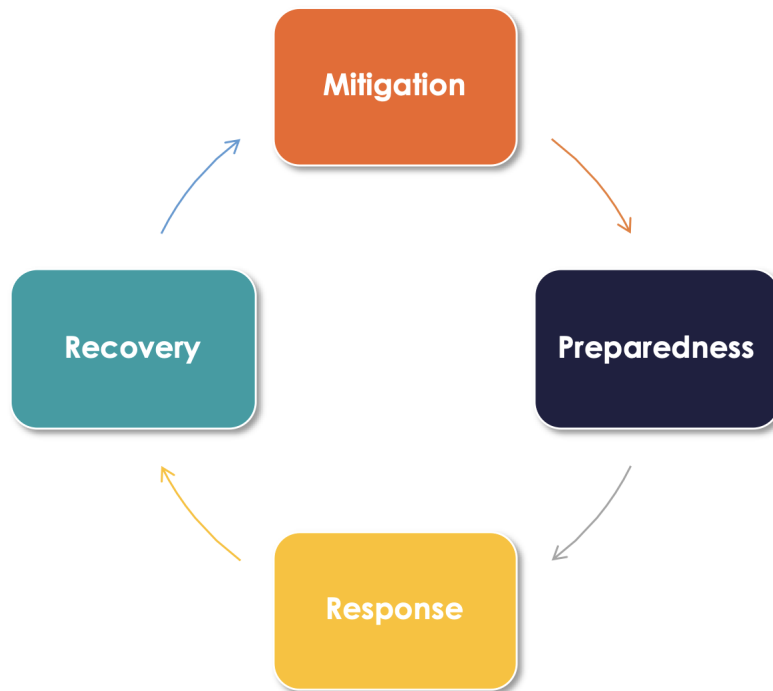


Figure C.1: Emergency management cycle

Emergency Managers need information that could help them act promptly and make use of appropriate resources [Phillips, Neal, and G. Webb 2016]; also, smart solutions to manage emergencies involve complex tasks related to the following:

- Mitigation/Monitoring and forecasting – includes the activities that help to determine and eliminate or reduce beforehand the risks of the occurrence of emergencies or disasters. Smart solutions should help in adopting specific monitoring activities and analysis of results to prevent natural or human-made emergencies and disasters.
- Preparedness/Planning – includes solutions that help to prepare action plans to be adopted in case of disaster. It consists of the activities of developing the response plan and training first responders to save lives and reduce emergencies and disaster damage. Also, it should take into consideration the identification of resources and the agreements on the procedures they are in charge of.
- Emergency responding – management and coordination of the operations of the first responders, which follow an emergency or disaster to reduce the probability of other damages while minimising recovery operations.
- Recovering – handling short- or long-term post-emergency activities to coordinate, design and verify the restoration works for a rapid return to normal living conditions.

Appendix D

Appendix - Requirements' analysis

This section gathers information that supports the Requirements' analysis process presented in Chapter 3. Here we provide a list of the documentation generated by The Open University regarding fire preparation and response, essential to analyse the current handling of data during emergencies (Section D.1). In Section D.2 we provide the example of the Personal Emergency Evacuation Plan - PEEP used by the university to collect data regarding the special needs of people with disabilities. Finally, in Section D.3 we present the poster 'Towards privacy-aware intelligent systems for emergency response', which summarises the first steps of our research.

D.1 Fire related documentation

This section lists the documentation generated by The Open University Health and Safety Department. We list the main topics of the documentation and do not provide access to them as they are considered only for internal use of the organisation.

Health & Safety policies:

- Statement of Intent
 - Organisational Detail
 - General Arrangements
 - Health and Safety Privacy Notice
- Unit Health and Safety Policy Template:
 - Unit Health & Safety Policy Template
- Additional Contact Lists:

- Unit Safety Coordinator and Department Safety Advisors List
- Union Health and Safety Representatives

Health & Safety FRE – Fire Documentation:

- Introduction
- Roles and responsibilities
- Fire risk assessment and Fire prevention
- Fire safety information and training
- Emergency management and evacuation
- Testing and inspection of fire alarm systems and emergency lighting
- Fire classification, fire fighting and drills
- Persons with disabilities and special requirements
- Legislation/HSE information
- Fire Action Notice
- Fire Warden Monthly Inspection Form
- Fire Warden List - Walton Hall
- Fire Warden List - Regional and National Offices
- Fire Assembly Points and Fire Evacuation Lifts
- Educational Establishments, including Universities
- Means of Escape for Disabled Persons
- Fire Warden Leaflet

PEEP - Personal Emergency Evacuation Plan:

- Introduction
- Roles and responsibilities
- Emergency Evacuation

- Types of Disability
- Evacuation Aids and Other Considerations
- Legislation/HSE information
- Personal Emergency Evacuation Plan

D.2 Personal Emergency Evacuation Plan - PEEP

In what follows, we give an example of the Personal Emergency Evacuation Plan used by The Open University to perform a risk assessment for disabled people working at the OU (see Figure D.1).


Personal Emergency Evacuation Plan (PEEP)				 The Open University	
<p>An individual PEEP and risk assessment must be completed for all disabled employees and research degree students who would require assistance in evacuating a building. This applies to anyone who is permanently or temporarily disabled. The plan must be shared with everyone who is named in it and it must be practised regularly to ensure it works. A copy should be sent to The Open University Health and Safety Department. This form should be used in conjunction with the information provided on PEEP on the intranet.</p>					
Personal Details					
Name		Tel Number		Email address	
Staff Number		Unit/Dept		Date	
Section 1 Brief description of disability					
Section 2 Any special aids used by individual (wheelchair, crutches, walking stick, pager, etc)					
Section 3 Agreed Evacuation Plan					
1					
2					
3					
4					
5					
6					
Section 4 Line Manager					
Name		Location		Tel No	

Figure D.1: Personal Emergency Evacuation Plan - PEEP form

D.3 Poster

Towards privacy-aware intelligent systems for emergency response. This poster was presented during the International Semantic Web Research Summer School in Italy in 2019. It was awarded the best poster from a 1st year PhD Student, Figure D.2.

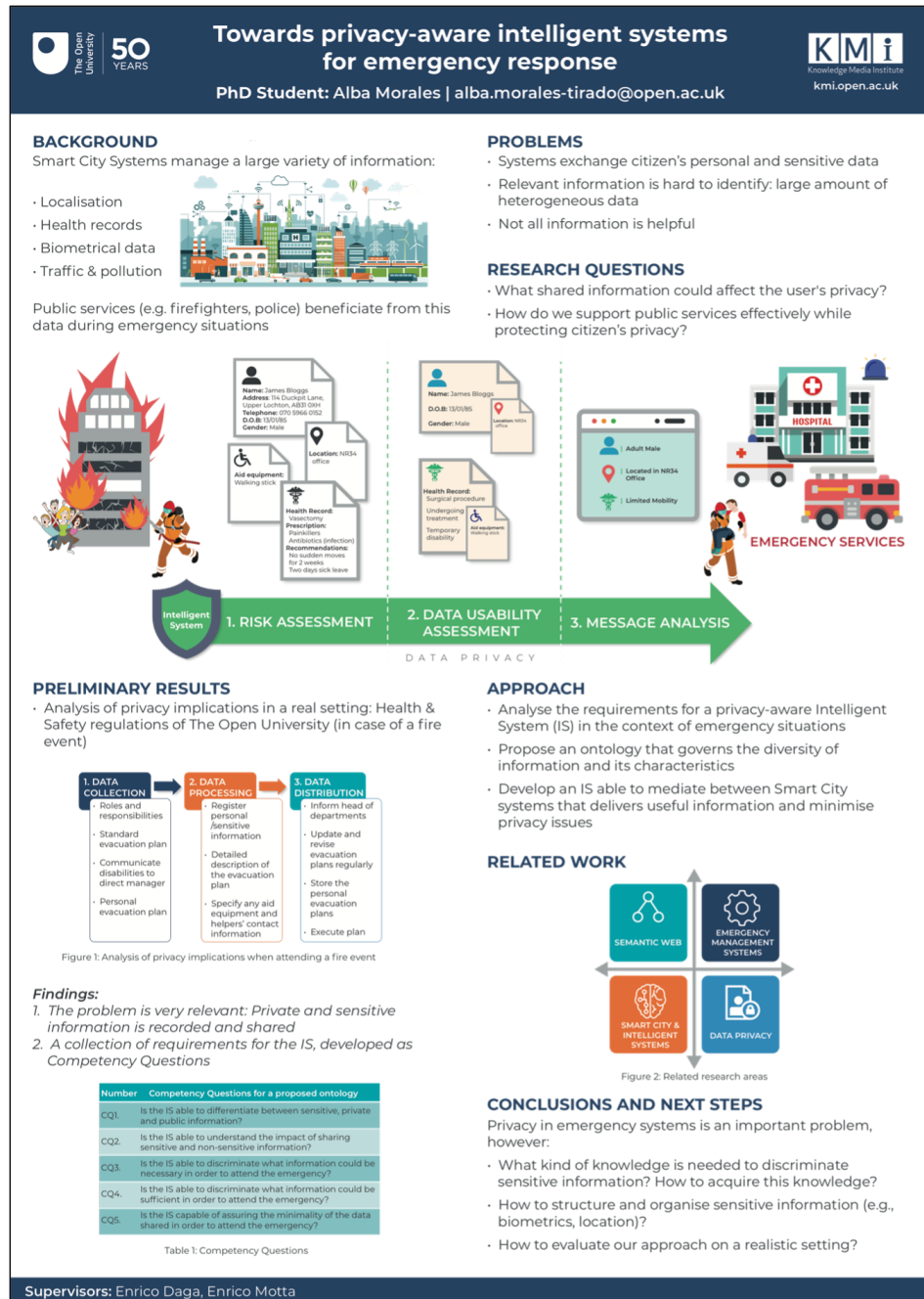


Figure D.2: Poster: Towards privacy-aware intelligent systems for emergency response

Appendix E

Appendix - Intelligent System Design

In this section, we list all the attributes that describe the synthetic dataset of electronic health records. In Chapter 6, Section 6.2, we described the dataset used to experiment on the evaluation of health conditions. This dataset comprises extensive electronic health records of patients; we used Synthea software for this task. The data is generated in CSV format and grouped into twelve different files; each file groups a number of attributes that describe an FHIR resource type.

In detail, the twelve FHIR resources used are: Patient (see Figure E.1), Medication (see Figure E.1), Encounter (see Figure E.2), Condition (see Figure E.2), Procedure (see Figure E.2), Allergy (see Figure E.2), Careplan (see Figure E.3), Imaging study (see Figure E.3), Organisation (see Figure E.3), Immunisation (see Figure E.3), Observation (see Figure E.4) and Payer (see Figure E.4).

E.1 FHIR specification

PATIENT		MEDICATION	
Column Name	Property	Column Name	Property
Id	fhir:Patient.identifier	Start	fhir:MedicationRequest.dispenseRequest.validityPeriod
BirthDate	fhir:Patient.birthDate	Stop	fhir:MedicationRequest.dispenseRequest.validityPeriod
DeathDate	fhir:Patient.deceasedDate	Patient	fhir:MedicationAdministration.subject
SSN	fhir:Identifier.value	Payer	fhir:Coverage.payer
Drivers	fhir:Identifier.value	Encounter	fhir:MedicationAdministration.encounter
Passport	fhir:Identifier.value	Code	fhir:Medication.code
Prefix	fhir:HumanName.prefix	Description	fhir:Medication.code
First	fhir:HumanName.given	Base_Cost	fhir:MedicationKnowledge.cost.cost
Last	fhir:HumanName.family	Payer_Coverage	fhir:Coverage.costToBeneficiary.valueMoney
Suffix	fhir:HumanName.suffix	Dispenses	fhir:MedicationAdministration.dosage.rateQuantity
Maiden	fhir:HumanName.family	TotalCost	fhir:Invoice.totalGross
Marital	fhir:Patient.maritalStatus	ReasonCode	fhir:MedicationAdministration.reasonCode
Race	fhir:Extension.value	ReasonDescription	fhir:MedicationAdministration.reasonCode
Ethnicity	fhir:Extension.value		
Gender	fhir:Patient.gender		
BirthPlace	fhir:Extension.uri		
Address	fhir:Address.line		
City	fhir:Address.city		
State	fhir:Address.state		
County	fhir:Address.district		
Zip	fhir:Address.postalCode		
Lat	fhir:Extension.uri		
Lon	fhir:Extension.uri		
Healthcare_Expenses	fhir:Coverage.costToBeneficiary.valueMoney		
Healthcare_Coverage	fhir:Coverage.costToBeneficiary.valueMoney		

Figure E.1: FHIR mapping. Part 1/4

ENCOUNTER		CONDITION	
Column Name	Property	Column Name	Property
Id	fhir:Encounter.identifier	Start	fhir:Period.start
Start	fhir:Period.start	Stop	fhir:Period.end
Stop	fhir:Period.end	Patient	fhir:Patient.identifier
Patient	fhir:Patient.identifier	Encounter	fhir:Condition.encounter
Provider	fhir:Encounter.serviceProvider	Code	fhir:Condition.code
Payer	fhir:Organization.identifier	Description	fhir:Condition.code
EncounterClass	fhir:Encounter.class		
Code	fhir:Encounter.type		
Description	fhir:Encounter.type		
Base_Encounter_Cost	fhir:Coverage.costToBeneficiary.valueMoney		
Total_Claim_Cost	fhir:Claim.total		
ReasonCode	fhir:Encounter.reasonCode		
ReasonDescription	fhir:Encounter.reasonReference		
PROCEDURE		ALLERGY	
Column Name	Property	Column Name	Property
Date	fhir:Procedure.recorded	Start	fhir:AllergyIntolerance.recordedDate
Patient	fhir:Patient.identifier	Stop	fhir:AllergyIntolerance.onsetDateTime
Encounter	fhir:Procedure.encounter	Patient	fhir:AllergyIntolerance.patient
Code	fhir:Procedure.code	Encounter	fhir:AllergyIntolerance.encounter
Description	fhir:Procedure.code	Code	fhir:AllergyIntolerance.code
ReasonCode	fhir:Encounter.reasonCode	Description	fhir:AllergyIntolerance.code
ReasonDescription	fhir:Encounter.reasonReference		

Figure E.2: FHIR mapping. Part 2/4

CAREPLAN		IMAGING_STUDY	
Column Name	Property	Column Name	Property
Id	fhir:CarePlan.identifier	Id	fhir:ImagingStudy.identifier
Start	fhir:CarePlan.created	Date	fhir:ImagingStudy.started
Stop	fhir:Period	Patient	fhir:ImagingStudy.subject
Patient	fhir:Patient.identifier	Encounter	fhir:ImagingStudy.identifier
Encounter	fhir:CarePlan.identifier	Body Site Code	fhir:ImagingStudy.series.bodySite
Code	fhir:CarePlan.activity.detail.code	Body Site Description	fhir:ImagingStudy.series.description
Description	fhir:CarePlan.description	Modality Code	fhir:ImagingStudy.modality
ReasonCode	fhir:CarePlan.activity.detail.reasonCode	Modality Description	fhir:ImagingStudy.modality
ReasonDescription	fhir:CarePlan.activity.detail.description	SOP Code	fhir:ImagingStudy.series.instance.sopClass
		SOP Description	fhir:ImagingStudy.series.instance.title
ORGANISATION		IMMUNISATION	
Column Name	Property	Column Name	Property
Id	fhir:Organization.identifier	Date	fhir:Immunization.occurrenceDateTime
Name	fhir:Organization.name	Patient	fhir:Immunization.patient
Address	fhir:Organization.contact	Encounter	fhir:Immunization.identifier
City	fhir:Organization.contact	Code	fhir:Immunization.reasonCode
State	fhir:Organization.contact	Description	fhir:Immunization.reasonCode
Zip	fhir:Organization.contact	Cost	fhir:Immunization.value
Lat	fhir:Location.position.latitude		
Lon	fhir:Location.position.longitude		
Phone	fhir:Organization.telecom		

Figure E.3: FHIR mapping. Part 3/4

OBSERVATION		PAYER	
Column Name	Property	Column Name	Property
Date	fhir:Observation.effectiveDateTime	Id	fhir:InsurancePlan.identifier
Patient	fhir:Observation.subject	Name	fhir:InsurancePlan.name
Encounter	fhir:Observation.identifier	Address	fhir:InsurancePlan.contact.address
Code	fhir:ObservationDefinition.code	City	fhir:InsurancePlan.contact.address
Description	fhir:ObservationDefinition.code	State_Headquartered	fhir:InsurancePlan.contact.address
Value	fhir:Observation.component.value	Zip	fhir:InsurancePlan.contact.address
Units	fhir:Observation.valueQuantity	Phone	fhir:InsurancePlan.contact.telecom
Type	fhir:Observation.valueQuantity	Amount_Covered	fhir:InsurancePlan.coverage.benefit.limit.value

Figure E.4: FHIR mapping. Part 4/4

Appendix F

Appendix - Building HECON Ontology - resources

F.1 Text snippets

In this section, we list a comprehensive set of expressions indicating condition recovery. The text snippets listed were found in the data source and used as the base to cluster the different types of condition evolution types.

Table F.1: List of text snippets used to build a subset of sentences.

Text snippet used for cosine similarity matching	Original sentence extracted from data sources
rest home day two	rest at home for a day or two
2 weeks	for 2 weeks
usually lasting minutes hours	usually lasting for a few minutes to hours
2 3 days	After 2 to 3 days
hospital 3 4 days	be in hospital for 3 or 4 days
stay hospital 5 days open surgery	stay in hospital for up to 5 days (open surgery)
start effect within day 2	start to have an effect within a day or 2
usually takes 10 15 minutes	usually takes about 10 to 15 minutes.

usually takes 15 45 minutes	usually takes between 15 and 45 minutes
takes around 30 45 minutes	takes around 30 to 45 minutes
takes 15 45 minutes	takes between 15 and 45 minutes
last around 30 90 minutes	last around 30 to 90 minutes
week every 2 weeks	once a week or once every 2 weeks
6 months	after 6 months
usually lasting six 12 months	usually lasting six to 12 months
least 3 6 months fully recover	at least 3 to 6 months to fully recover
least 4 6 months	at least 4 to 6 months
take 3 6 months fully heal	can take between 3 and 6 months to fully heal
take months even years fully heal	can take months or even years to fully heal
6 12 months	be on them for 6 to 12 months.
within 7 days job involves sitting desk physical may need stay work 2 weeks	within 7 days if your job involves sitting at a desk, but if it's more physical, you may need to stay off work for up to 2 weeks.
week	in about a week
within week 2	within a week or 2.
within 2 3 weeks	within 2 to 3 weeks
two four weeks	for two to four weeks
take 4 6 weeks fully recover	it can take 4 to 6 weeks to fully recover
usually return normal activity four six weeks	You usually can return to normal activity four to six weeks

usually takes 6 8 weeks heal	usually takes 6 to 8 weeks to heal
around 4 8 weeks	around 4 to 8 weeks
treatment may last year longer	treatment may last a year or longer if you
longterm mental illness	have a long-term mental illness
take 4 6 weeks fully recover	It can take 4 to 6 weeks to fully recover
may able return work 2 weeks	may be able to return to work in 2 weeks
often develops decades	often develops over decades
usually progresses slowly	usually progresses very slowly
gets slowly worse time	gets slowly worse over time
develop gradually many years	develop gradually over many years
cause gradually worsening problems	can cause gradually worsening problems
get gradually worse several years	get gradually worse over several years
usually takes several years develop	it usually takes several years to develop
eventually causes permanent progressive damage	eventually causes permanent, progressive damage
usually takes several years	it usually takes several years
develops slowly several years	develops slowly over several years
progresses slowly	progresses so slowly
sometimes day cases	sometimes day cases
4 6 weeks open surgery	for 4 to 6 weeks after open surgery
usually takes couple weeks make full recovery	It usually takes a couple of weeks to make a full recovery
around 14 days	around 14 days
within 5 days	within 5 days

you'll take 6 months	You'll have to take them for about 6 months
usually takes 10 minutes	usually takes no more than 10 minutes
overnight recover	overnight to recover
24 hours	for 24 hours
day	same day
able go home day day	You should be able to go home on the day of, or the day after
often need continued week	These often need to be continued for up to a week
persist year	can persist for more than a year
around 2 hours	around 2 hours
around 2 hours	for around 2 hours
full recovery within 2 weeks	full recovery within 2 weeks
2 weeks fully recover	up to 2 weeks to fully recover
feel completely better 2 weeks	you should feel completely better after about 2 weeks
lasts 3 weeks	It lasts up to 3 weeks
need stay hospital around 6 hours	need to stay in hospital for around 6 hours
full recovery within 12 weeks	full recovery within 12 weeks
usually takes 10 minutes although whole consultation may take 30 minutes	usually takes about 10 minutes, although the whole consultation may take about 30 minutes.
less half hour	less than half an hour
usually takes 30 minutes	usually takes up to 30 minutes

often get better within day	often get better on their own within a day
necessary stay hospital overnight afterwards	it's not necessary to stay in hospital overnight afterwards
likely gone within days	likely be gone within a few days.
last minutes	only last a few minutes
usually takes minutes	It usually only takes a few minutes
discomfort may last days weeks	discomfort may last from a few days to a few weeks.
last longer 6 months	last no longer than 6 months
often improve time	often improve over time
spread months	spread over a few months
may improve child gets older	may improve as a child gets older
improve significantly even clear completely	it can improve significantly, or even clear completely
long frustrating process	can be a long and frustrating process.
last around 4 years last period although	last around 4 years after your last period, although
women experience much longer	some women experience them for much longer
last months	last for months
longterm problems	long-term problems
last months years	can last for months to years.
last months years	can last for months or years.
get better without need treatment	get better without the need for treatment
dont need treatment	don't need treatment
get better without treatment	get better without any treatment

get better without need medical treatment	get better without the need for medical treatment.
might require treatment	They might not require treatment
dont cause problems need treatment	don't cause problems or need treatment
however eye floaters dont require treatment	However, most eye floaters don't require treatment
lead perfectly normal life	can lead a perfectly normal life
clears without causing problems	clears up by itself without causing problems
dont require treatment	don't require treatment.
dont need treatment	don't need treatment.
doesnt require treatment	It doesn't require treatment
get better without treatment	get better without treatment
dont require surgery treatment	don't require surgery or other treatment
harmless dont require treatment	are harmless and don't require any treatment
typically dont reduce fertility require treatment	They typically don't reduce fertility or require treatment
might require treatment	might not require treatment
dont symptoms may need treatment	If you don't have symptoms, you may not need treatment
wont need treatment	won't need treatment
may get better without treatment	may get better without any treatment
may get better without treatment	may get better without treatment.
get better without treatment	get better without treatment.

may get better without surgical treatment	may get better without surgical treatment
true age spots dont need treatment	True age spots don't need treatment
dont require specific treatment	don't require specific treatment
need treatment	need treatment
blushing common problem embarrassing	Blushing is a common problem that can be
affect day day life	embarrassing and affect your day to day life
probably wont need extensive treatment	probably won't need extensive treatment
reduce risk getting	can reduce your risk of getting
theres specific treatment growing pains	There's no specific treatment for growing pains
hair loss isnt usually anything worried	Hair loss isn't usually anything to be worried about
persists without causing problems	persists without causing problems
theres cure	There's no cure
longterm condition cannot cured	is a long-term condition and cannot be cured
lifelong condition sometimes cause serious disability	It's a lifelong condition that can sometimes cause serious disability
symptoms last least 3 months known chronic	If symptoms last for at least 3 months, it's known as chronic
theres cure	There's no cure for
theres currently specific treatment cure	there's currently no specific treatment or cure.
theres currently cure	There's currently no cure for
longterm condition	is a long-term condition

longterm condition	a long-term condition
lifelong condition	is lifelong condition
long difficult frustrating process	It can be a long, difficult and frustrating process
perfectly healthy able lead normal life	perfectly healthy and able to lead a normal life
disorder cant cured	the disorder can't be cured.
rest life	for the rest of your life
people regain full preinjury	few people regain full pre-injury
youll atrial fibrillation permanently	You'll have atrial fibrillation permanently
lifelong condition	is a lifelong condition
theres currently cure	There's currently no cure
theres cure	there's no cure
cant cured	can't be cured
last long time	can last a long time
may get worse years remain steady	It may get worse for a few years but then remain steady
usually lifelong condition	is usually a lifelong condition
rest life	for the rest of your life.
usually takes several years	usually takes several years
people get allergic rhinitis months time	Some people only get allergic rhinitis for a few months at a time
long lasting	are long lasting