

Programme for the Use of Real World Data in Health Technology Assessment

Using real-world data for health technology assessment

Series of methodological developments
of the Spanish Network of Agencies
for Assessing National Health System
Technologies and Performance

REPORTS, STUDIES AND RESEARCH



Programme for the Use of Real World Data in Health Technology Assessment

Using real-world data for health technology assessment

Series of methodological developments
of the Spanish Network of Agencies
for Assessing National Health System
Technologies and Performance

REPORTS, STUDIES AND RESEARCH



Programme for the Use of Real world Data in Health Technology Assessment.

Using real-world data for health technology assessment / Celia Muñoz Fernández, Guillermo Pérez García, Lucía Prieto Remón, Carlos Tellería Orriols, Hugo Hernández Alemán, Sandra García Armesto - Madrid: Ministerio de Sanidad. Zaragoza: Instituto Aragonés de Ciencias de la Salud (IACS) - 120 p.; 24 cm (Collection: Reports, studies and research. Series: Series of methodological developments of the Spanish Network of Agencies for Assessing National Health System Technologies and Performance. IACS)

NIPO: 133-23-125-8

ISBN: 978-84-09-69227-9

DOI: https://doi.org/10.46994/ets_42

1. Methodological framework. 2. Health Technology Assessment. 3. Real World Data. 4. Spanish Network of Agencies for Assessing National Health System Technologies and Performance

I. Aragón. Instituto Aragonés de Ciencias de la Salud (IACS). II. España. Ministerio de Sanidad.

Edition: 2025

Edits: Ministerio de Sanidad
Instituto Aragonés de Ciencias de la Salud

NIPO: 133-23-125-8

ISBN: 978-84-09-69227-9

DOI: https://doi.org/10.46994/ets_42

Formatting: Gambón, S.A. Zaragoza

This document has been produced by the Instituto Aragonés de Ciencias de la Salud (IACS), within the framework of funding from the Ministry of Health, Consumer Affairs and Social Welfare aimed at developing the activities included in the Annual Work Plan of the Spanish Network of Agencies for Assessing National Health System Technologies and Performance, approved by the Consejo Interterritorial del SNS of the Spanish National Health System (SNS) on 26 May 2021.

This document may be reproduced in part or in full for non-commercial use, provided that the source is acknowledged.

To cite this report:

Muñoz-Fernández C, Pérez-García G, Prieto-Remón L, Tellería-Orriols C, Hernández-Alemán H, García-Armesto S. Using real-world data for health technology assessment. Ministerio de Sanidad. Instituto Aragonés de Ciencias de la Salud; 2025. Series of methodological developments of the Spanish Network of Agencies for Assessing National Health System Technologies and Performance: IACS.



MINISTERIO
DE SANIDAD



RED ESPAÑOLA DE AGENCIAS DE EVALUACIÓN
de Tecnologías y Prestaciones del Sistema Nacional de Salud



INSTITUTO ARAGONÉS
DE CIENCIAS DE LA SALUD

Index

List of figures	9
List of tables	11
List of boxes	13
Authors and collaborators	15
Declaration of conflict of interest	17
Abbreviations	19
1. Scope and objectives	21
2. Real-world data for preadoption HTA	23
2.1. Workflow	24
3. The decision model	27
3.1. Key aspects of modelling and their application to the use of RWD	28
3.2. Standard Models	28
3.3. Patient-level simulation models and further developments for decision modelling	30
3.3.1. When is simulation the appropriate modelling technique?	31
3.4. Making probabilistic models	32
3.5. Decision model development	34
4. Data request	37
4.1. Gathering background information	37
4.1.1. Systematic review protocol	38
4.1.2. Data extraction	42
4.2. Working with data experts	44
4.2.1. Defining cohort criteria	45
4.2.2. Defining research questions	48
4.2.3. Constructing a data model specification	51
4.3. Implementing the data extraction process	54
5. Data description and analysis	59
5.1. Exploratory data analysis	59
5.1.1. Analysing variation	62
5.1.2. Analysing covariation	71
5.2. Addressing data availability and quality	79

6. Modelling	83
6.1. Building and executing the model	83
6.1.1. Calculate patient-specific transition probabilities	83
6.1.2. Incorporate patient-specific resource use	87
6.1.3. Estimating patient-specific outcomes	88
6.1.4. Additional parameters in our decision model	89
6.1.5 Model simulation	90
6.2. Assumptions	91
6.3. Model validation	92
6.4. Obtaining results	93
6.5. Dealing with uncertainty	94
6.5.1 Making our decision model probabilistic	95
7. Limitations	99
8. Conclusions and further work	101
Annexes	103
Annex I. Bibliography	103
Annex II. ICD-SCD use case indicators	109
Annex III. ICD-SCD use case data model specification description	113
Annex IV. RWD section in HTA protocol for direct oral anticoagulants (DOAC) quantification	114
Annex V. Exploratory data analysis tools	119

List of figures

Figure 1. Prioritisation-appraisal cycle for health technologies in Spain	23
Figure 2. Workflow for the use of RWD in preadoption HTA	25
Figure 3. Schematic representation of a decision tree model ²⁴	29
Figure 4. Schematic representation of a Markov model ²⁵	30

List of tables

Table 1. Definition of general data requirements for the ICD-SCD use case	39
Table 2. Example cohort characterisation questions and indicators	50
Table 3. Example health outcome questions and indicators	50
Table 4. Example treatment pathways questions and indicators	51
Table 5. Example use of resources questions and indicators	51
Table 6. List of entities and variables in the ICD-SCD use case	53
Table 7. Common problems and solutions during data extraction	55
Table 8. Objectives and questions in exploratory data analysis	59
Table 9. Data quality dimensions according to the EMA Data Quality Framework ⁵⁰	79
Table 10. How to address data reliability	80
Table 11. How to deal with missing or incomplete data	80
Table 12. Statistical techniques for populating models with RWD	87

List of boxes

Box 1. ICD-SCD use case presentation	38
Box 2. Definition of specific data requirements for the ICD-SCD use case	43
Box 3. Cohort entry events and inclusion/exclusion criteria in the ICD-SCD use case	47
Box 4. A problem of unavailable variables in the ICD-SCD use case	56
Box 5. A problem of unstructured records in the ICD-SCD use case	57
Box 6. Summarising categorical data	63
Box 7. Summarising numerical data	65
Box 8. Visualising categorical variables	67
Box 9. Numerical data visualisation	69
Box 10. Summarising two categorical variables	72
Box 11. Summarising numerical vs categorical variables	73
Box 12. Summarising two numerical variables	74
Box 13. Visualising two categorical variables	76
Box 14. Visualising numerical vs. categorical variables	77
Box 15. Visualising two numerical variables	78
Box 16. Baseline patient characteristics	84
Box 17. Transforming a Markov model into an individual patient simulation	86
Box 18. Simulating resource utilisation	88
Box 19. Simulating patient-specific outcomes	89
Box 20. Key distributions commonly used in stochastic modelling	96

Authors and collaborators

Authorship

Celia Muñoz Fernández. Instituto Aragonés de Ciencias de la Salud (IACS).

Guillermo Pérez García. Instituto Aragonés de Ciencias de la Salud (IACS).

Lucía Prieto Remón. Instituto Aragonés de Ciencias de la Salud (IACS).

Carlos Tellería Orriols. Instituto Aragonés de Ciencias de la Salud (IACS).

Hugo Hernández Alemán. Instituto Aragonés de Ciencias de la Salud (IACS).

Sandra García Armesto. Instituto Aragonés de Ciencias de la Salud (IACS).

Collaborators (in alphabetical order)

Blanca Novella Arribas. Unidad de Evaluación de Tecnologías Sanitarias de Madrid (UETS-Madrid).

Carmen Guirado Fuentes. Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC). Servicio de Evaluación del Servicio Canario de la Salud (SESCS).

Cristina Valcárcel Nazco. Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC). Servicio de Evaluación del Servicio Canario de la Salud (SESCS).

Eva Reviriego Rodrigo. Fundación Vasca de Innovación e Investigación Sanitaria (BIOEF). Osasun Teknologien Ebaluazioko Zerbitzua (Osteba).

Francisco José Rodríguez Salvanés. Unidad de Evaluación de Tecnologías Sanitarias de Madrid (UETS-Madrid)

Janet Puñal Riobóo. Unidad de Asesoramiento Científico-técnico (Avalia-t). Agencia Gallega para la Gestión del Conocimiento en Salud (ACIS).

Jessica Ruiz Baena. Agència de Qualitat i Avaluació Sanitàries de Catalunya (AQuAS)

Lidia García Pérez. Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC). Servicio de Evaluación del Servicio Canario de la Salud (SESCS).

Montserrat Carmona Rodríguez. Instituto de Salud Carlos III (ISCIII).

Rosa María Vivanco Hidalgo. Agència de Qualitat i Avaluació Sanitàries de Catalunya (AQuAS).

Soledad Isern de Val. Instituto Aragonés de Ciencias de la Salud (IACS).

Silvia Moler Zapata. Instituto Aragonés de Ciencias de la Salud (IACS).

External reviewers (in alphabetical order)

Beatriz González López-Valcárcel. Universidad de Las Palmas de Gran Canaria (ULPGC).

Daniel Prieto Alhambra. University of Oxford and Erasmus MC.

Jaime Pinilla Domínguez. Universidad de Las Palmas de Gran Canaria (ULPGC).

The external reviewers of the paper do not necessarily endorse each and every one of the final considerations and conclusions, which are the sole responsibility of the authors.

Declaration of conflict of interest

The authors declare they have no conflicts of interest that could compromise the primary interest and objectives of this report or influence their professional judgement in this regard.

Abbreviations

APTT	Partial Thromboplastin Time
ARB	Angiotensin II Receptor Blocker
ARNI	Angiotensin Receptor Neprilysin Inhibitor
ATC	Anatomical, Therapeutic, Chemical classification system
CDM	Common Data Model
CKD	Chronic Kidney Disease
CLD	Chronic Liver Disease
CRF	Case Report Forms
CRT	Cardiac Resynchronisation Therapy
CV	Coefficient of Variation
DMS	Data Model Specification
DOAC	Direct Oral Anticoagulants
dTT	Dilute Thrombin Test
EDA	Exploratory Data Analysis
EHDS	European Health Data Space
EIT	Pulmonary Electrical Impedance Tomography
ENDS	National Health Data Space
GRADE	Grading of Recommendations, Assessment, Development, and Evaluation
HIFU	High-Intensity Focused Ultrasound
HR	Hazard Ratio
HT	Health Technology
HTA	Health Technology Assessment
ICD	Implantable Cardioverter Defibrillator
ICER	Incremental Cost-Effectiveness Ratio

IQR	Interquartile Range
LVEF	Left Ventricular Dysfunction
NA	Not Applicable
NYHA	New York Heart Association
OR	Odds Ratio
P25	25th percentile
P50	50th percentile
P75	75th percentile
PICO	Population, Intervention, Comparison, Outcomes
PREM	Patient-reported Experience Measures
PROM	Patient-reported Outcome Measures
PSA	Probabilistic Sensitivity Analysis
PT	Prothrombin Time
Q1	Quartile 1
Q2	Quartile 2
Q3	Quartile 3
QALY	Quality Adjusted Life Year
RedETS	Spanish Network of Agencies for Assessing National Health System Technologies and Performance
RR	Risk Ratio
RWD	Real World Data
RWE	Real World Evidence
SCA	Sudden Cardiac Arrest
SCD	Sudden Cardiac Death
SD	Standard Deviation
SNOMED	Systematized Nomenclature of Medicine
SNS	Spanish National Health System
VOI	Value of Information

1. Scope and objectives

The Spanish Network of Agencies for Assessing National Health System Technologies and Performance (RedETS) is the organisation responsible for conducting health technology assessment (HTA) to support decision-making processes about non-pharmacological technologies in Spain. This includes the assessment of medical devices and other non-pharmacological interventions, diagnostic tests, screening programmes, and emergent technologies.

This manual focuses on uses of real-world data (RWD) for HTA of health technologies in **pre-implementation phase** (often referred to as “preadoption” in our context) and will be further updated to consider assessments in the postadoption phase, along with the appraisal of real-world evidence (RWE) provided by stakeholders or found during the usual assessment process for HTA.

The relevant information is organised into two levels:

- **Methodological:** methods and tools for the specification of questions and data processing.
- **Procedural:** step-by-step actions to be carried out to collect the data and use it for answering questions in the report, including relevant information to be considered in each step.

To develop this methodological framework, the working group initially held several meetings to define the **primary uses** of RWD for RedETS in the short term¹. In this context, “use” refers to functionalities that could enhance RedETS’ ability to inform decision-making throughout the lifecycle of health technologies. Two primary uses for the preadoption phase were identified: 1) a better adjustment of the assessments to the context of the Spanish population; 2) a live assessment from the early stages of the introduction of a technology until its eventual widespread adoption, enabling the fine-tuning of organisational strategies.

Next, a **manual review** was conducted on national and international initiatives that provide guidance on the use of RWD in HTA. Forty-one initiatives were identified, from which 20 documents were reviewed²⁻²¹ and used as a reference to build a first approach to the methodology. This approach was subsequently discussed within the working group.

It was recognized that the type of technology to assess warrants a specific evaluation framework and sets the array of topical assessment

questions to be tackled, leading to a considerable casuistry. However, common pathways for utilising RWD in HTA were also identified, be they interventional, diagnostic, prognostic, or other. The working group agreed to outline and explain the key overarching steps and provide general guidelines for working with RWD, developing as illustration a **use case** for an interventional technology.

The **BIGAN**²² platform was chosen as the source of RWD for constructing this use case. BIGAN is the Big Data project of the Department of Health of the Government of Aragon, created to improve healthcare using data routinely collected within Aragon's public healthcare system. BIGAN gathers all data collected in the health system on a data lake integrated into a technological platform for processing and curating. It also offers advanced analytical tools for authorised healthcare professionals, healthcare managers, researchers and educators to conduct their queries and analyses.

The methods and procedures described in this document draw upon the concepts and recommendations derived from the manual review of handbook documents, as well as the authors' own experience in implementing the use case.

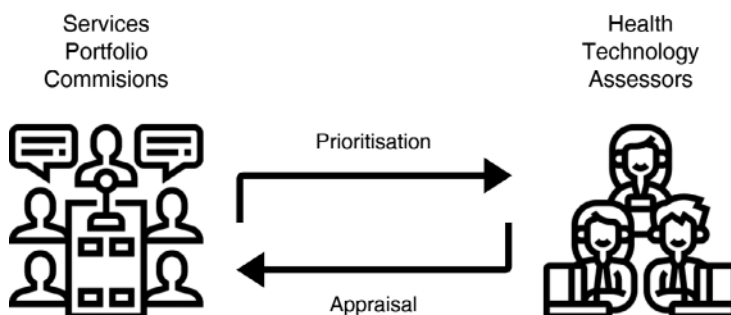
2. Real-world data for preadoption HTA

Preadoption assessments are carried out for technologies that have not yet been implemented in the health system; therefore, they are still under **consideration for inclusion in the benefits or services basket**. This is the most frequent type of assessment conducted by the Spanish Network of Agencies for Health Technology and Services Assessment of the National Health System (RedETS).

During this phase of the technology life-cycle, the value of utilising RWD lies in providing information to describe the **target population within our context**, data about the **existing comparator(s) available in our context and their performance**—that is, which technologies are currently in use for the same problem and what outcomes they deliver— and data to describe and predict the **actual use of health resources**. This information may be compared with data obtained from a systematic literature review on the new technology.

Moreover, RWD might be very helpful for **health authorities and regulators** in setting priorities for HTA at both national and regional levels. Rapid queries can offer insights into the relative frequency and severity of the clinical conditions involved, the size of the target population (patients likely to benefit from the technology), the number of potential users of a technology (providers or care settings liable to adopt it), or the potential direct cost variations of implementing the technology.

Figure 1. Prioritisation-appraisal cycle for health technologies in Spain



Additionally, the preadoption assessments conducted by RedETS using RWD have the potential to guide future research by identifying current information gaps that require further investigation regarding health technologies. By publishing the explicit methods of analysis (i.e. source code) there is also an opportunity to enhance the external review process of HTA, allow the reproducibility of analyses, increase public participation, and promote results dissemination. This can also set a model for the industry to present studies with RWD, highlighting which data and quality concerns are essential for the decision process.

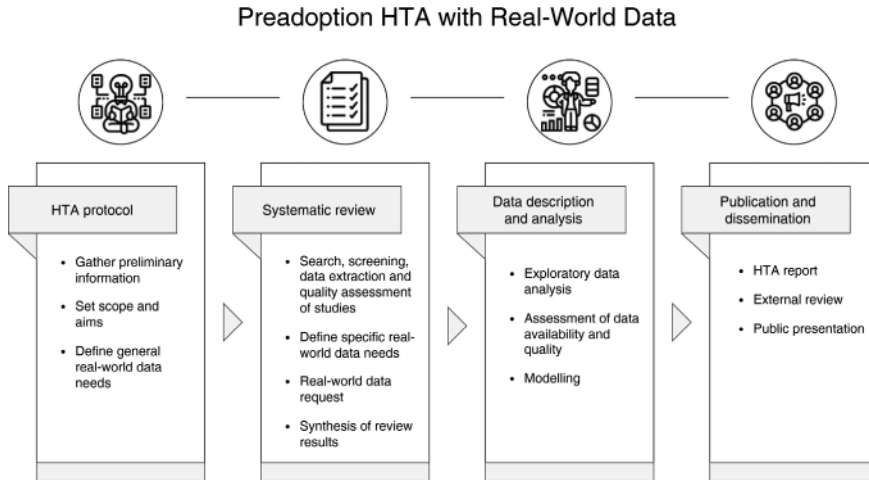
Furthermore, by identifying the populations that are most likely to benefit from a particular technology, policymakers can proactively develop strategies to guide and prioritise its introduction, before widespread adoption takes place. Manufacturers and business partners can also find this information useful in adjusting their marketing plans to align with the identified target population's needs and preferences.

2.1. Workflow

Based on our own experience, adding RWD to the HTA process does follow the same usual steps as when it is only based on a systematic literature review. RWD complements this process by adding information about what is happening in routine clinical practice in our context. Adequate data from a technology not adopted in the health system will often not be obtained from RWD. Therefore, a systematic review of the evidence in experimental settings will still be needed most of the times. As we will see, this will provide us with information on the new technology and help us plan the data request and analysis.

Thus, the phases for enabling work with RWD would be incorporated from the beginning of the usual HTA workflow. Already during the protocol phase, we should start defining possible data requirements. These requirements would be further specified during the systematic review stage and then translated into a data request that would take place before the evidence synthesis. Once the necessary data are gathered, a phase of description and analysis of the real-world information should be added to the evidence results to produce a final report. In the analysis phase, emphasis is placed on constructing a decision model, as the results of the analysis of primary data are not enough themselves in the case of preadoption HTA. Therefore, incorporating RWD into the process will not eliminate or replace anything that was already being done, but will add additional information of a supplementary nature.

Figure 2. Workflow for the use of RWD in preadoption HTA



3. The decision model

Models can be defined as abstract and simplified representations of perceived realities or theories. They use mathematical language to construct schematic representations of the underlying processes in a complex situation. This may help to explain a particular problem and study the effect of different parameters, variables and relationships in the system, predicting their evolution and thereby facilitating decision-making.

The use of RWD elevates modelling and makes it a key element that should be considered early in the reporting process and iterated upon as more background information about the technology and the characteristics of the target population becomes available. It enables adaptive modelling, in which models are continuously updated and refined as new RWD becomes available. This ensures that models remain relevant and improve their performance over time.

Expanding the use of modelling in health technology assessment offers a powerful tool to combine information from different sources employed in our assessment. RWD will only provide us with information about the current state of clinical practice, i.e. we will be able to inform the comparison branch. Against this usual practice scenario, we will need to consider a hypothetical scenario, based on the literature reviewed, where the new technology is incorporated. These models can capture the relevant factors and dynamics associated with the topic of interest. They enable exploration of potential outcomes under different scenarios, assessment of intervention impacts, and informed decision-making across various levels, from clinical practice to health policy.

In this context, RWD provides an opportunity to employ modelling techniques throughout the evaluation, not exclusively focused on deriving efficiency indicators. Modelling with RWD provides value when:

- *Complementing clinical studies and trials:* Models assist in restructuring and synthesising information, particularly when it originates from different sources, to provide decision-makers with useful insights. RWD can contain valuable information regarding the effectiveness, safety, and resource use of health technologies in standard practice that may not be captured in traditional clinical trials. Models help integrate and analyse diverse data sources to offer a comprehensive evaluation of current practice performance.

- *Overcoming follow-up time limitations:* Models help extend findings from routinely collected data and clinical studies to assess long-term health outcomes and the impact of a health technology over time. RWD can help us to make these projections more realistic, and can be adjusted as more information is learned about current practice. Further guidance on how and when adjustments are appropriate will be provided in the update of the document addressing the postadoption phase.

3.1. Key aspects of modelling and their application to the use of RWD

Traditionally, in health technology assessment (HTA), decision models have aimed to inform decision-making for the entire population, considering factors such as efficacy, safety, efficiency, and the overall perspective of patients, while accounting for the inherent uncertainty surrounding these values.

In some cases, alternatives to traditional modelling approaches would need to be implemented, towards an approach focused on capturing individual patient and healthcare characteristics in order to develop simulations.

The use of RWD can significantly enhance the estimates provided to decision-makers. Since RWD originates from the same population to which the evaluated technologies will be applied, the availability of context-specific data allows us to explore much more of the uncertainty surrounding the estimates as applied to the target population.

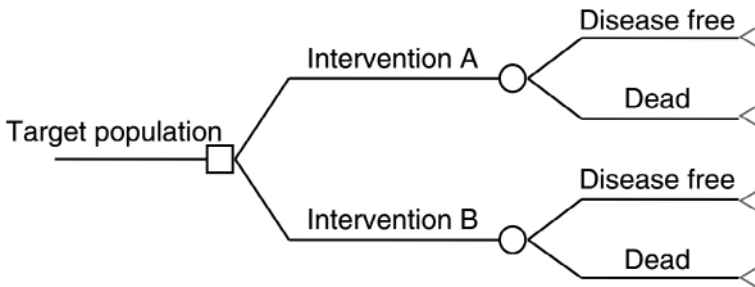
3.2. Standard Models

The main objective of a decision model is to describe the outcomes of different options in a manner that is suitable for the specific problem²³. One critical aspect to consider is whether the model ought to summarise the experience of a typical patient from a group with the same characteristics or if it should explicitly account for individual patients and variations between them.

Models generally known as “cohort models” are the most commonly used in HTA because of their simplicity and mathematical robustness. Such models can be a first step towards the use of RWD in HTA assessments. We can broadly classify standard models according to their structure (mutually exclusive alternatives):

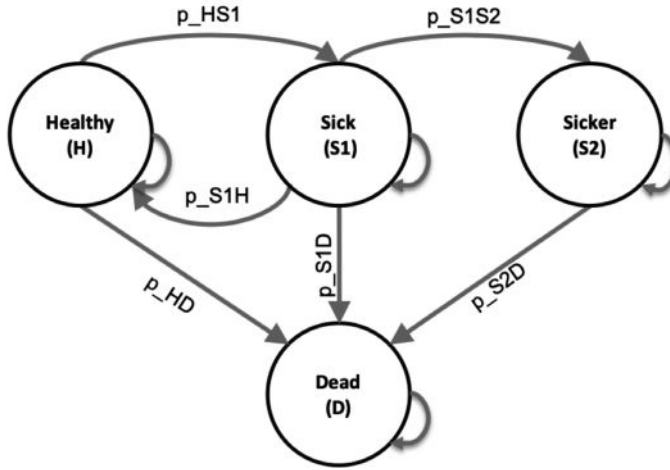
- **Decision tree models:** These models quantitatively and systematically represent a clinical decision-making situation, explicitly incorporating probabilities of event occurrence and their consequences (Figure 3). The key features are:
 - Decision points: Represented as square decision nodes, these points indicate alternative options.
 - Chance points: Represented as circular nodes, these points represent situations where multiple alternative events for a patient are possible. The actual event that will occur is uncertain.
 - Pathways: These are mutually exclusive sequences of events that form routes through the tree.
 - Probabilities: Likelihoods of specific events occurring at chance nodes. Subsequent probabilities are conditional, meaning they depend on whether an earlier event has or has not occurred. By multiplying probabilities along pathways, the overall pathway probability can be estimated.

Figure 3. Schematic representation of a decision tree model²⁴



- **Markov models:** These mathematical models are used in decision analysis to simulate the progression of patients through different health states over time. Markov models can be either time-varying (Markov processes) or constant (Markov chains), and they transform risks associated with different health states into transition probabilities (Figure 4). Their main components include:
 - Health states representing different disease stages or phases in the treatment procedure (Healthy, sick, sicker, dead).
 - Transition probabilities (pHS1, pHD, pS1S2, etc.) expressing the likelihood of moving from one state to another over a given time period (cycles).

Figure 4. Schematic representation of a Markov model²⁵



RWD analysis allows estimating model parameters in a way that is more tailored to local conditions and more accurate than through a literature review, enhancing the information available to decision-makers to make choices about the adoption and implementation of new technologies in clinical practice.

3.3. Patient-level simulation models and further developments for decision modelling

A patient-level simulation takes into account the unique characteristics of each individual patient within a group, providing estimates of outcomes for that particular group or cohort of patients. This approach offers several advantages over traditional models, which estimate outcomes for the entire patient group without considering individual patient characteristics. While widely used models may incorporate some patient variability based on predefined characteristics, they are not considered patient-level simulation models unless outcomes are evaluated at the patient level.

One specific type of patient-level simulation is **discrete event simulation**, which tracks individual patients and their events over time, accounting for the inherent randomness in patient characteristics and clinical events. This method allows for greater flexibility in modelling individual patient pathways and capturing the variability in costs and outcomes that may be overlooked by usual models.

In addition to patient-level simulation models, other modelling approaches, such as **survival analysis** or **dynamic models** used for simulating infectious diseases²⁶, can also be employed to address specific aspects of interest in decision making, depending on the research question and available data. We acknowledge the potential of utilising RWD to develop these complex models for tackling specific decision problems. These approaches will be thoroughly analysed and further developed in future updates of this handbook.

3.3.1. When is simulation the appropriate modelling technique?

Here are some criteria that can help identify situations where developing a simulation model is a preferable strategy over using standard modelling techniques:

- **Patient-specific response:** Patient-level simulation is beneficial when the outcomes of the model depend on individual patient characteristics. Simulating patients individually allows for capturing the complex interactions between patient characteristics and intervention effects.
- **Event-based patient progression:** If the patient pathways in the model are influenced by factors such as time since the last event or the history of previous events, a patient-level simulation is appropriate. It enables the modelling of individual patient pathways based on their specific clinical history.
- **Overcoming limitations of discrete time intervals:** Traditional models often use discrete time intervals, which may limit the ability to capture dynamic changes within short time frames. Patient-level simulation allows for more flexible time modelling, allowing events to occur at any time. This flexibility is valuable in situations that require precise temporal resolution.
- **Developing a flexible model for future analyses:** Patient-level simulations offer the advantage of being flexible and adaptable for future analyses. They can be updated and refined to incorporate new data, interventions, or patient characteristics as they become available. This adaptability makes patient-level simulation a valuable long-term investment.
- **Modelling systems with interactions:** Patient-level simulation is particularly suitable for modelling systems where interactions

between individuals, resources, or other people play a significant role. It allows for capturing complex dependencies and feedback loops within the system, providing a more realistic representation of real-world interactions. For example, in the case of a simulation model for infectious disease transmission, the interactions between individuals influence the spread of the disease, and the model needs to account for these interactions and their impact on transmission dynamics.

- Probabilistic sensitivity analysis for decision uncertainty: If there is a need to assess decision uncertainty using probabilistic sensitivity analysis, patient-level simulation is often preferred. It enables the incorporation of parameter uncertainty and variation across individual patients, allowing for the generation of probabilistic results and more robust decision-making.

Downsides of patient-level simulation models are the increased complexity, data requirements and computational costs. Analysts should assess the trade-off between the complexity and utility of these models, as well as the feasibility of conducting a simulation model given the expertise and resources available in their specific context. By considering these criteria, researchers can determine when patient-level simulation is a preferable strategy for their specific modelling needs.

3.4. Making probabilistic models

The choice of model type depends on the research question, available data, and specific evaluation requirements. Models can be classified into two categories based on their nature of uncertainty:

- Deterministic Models: In deterministic models, the variables of interest (e.g., treatment effects, survival probabilities, individuals in each health state) are directly inputted and computed using algebraic formulas without employing simulation techniques. Deterministic models assume that these variables are free of uncertainty.
- Stochastic Models: Stochastic models are probabilistic approaches that incorporate uncertainty into the calculations. These models use randomization techniques to simulate the probabilities of events that could occur due to chance. Stochastic variation can occur in two ways:

- At the decision node or transition probability, allowing random variation in the model trajectory for an individual patient, irrespective of their characteristics at the start of the model.
- At the parameter level, where parameter values are sampled from a distribution reflecting the uncertainty in their population mean. This uncertainty is propagated through the model to determine the resulting uncertainty in the expected model outcomes.

At RedETS, the models we have developed to inform decision-making are generally stochastic in nature, except in cases where parameter sparsity limits further investigation of uncertainty. However, our assessments are typically documented using parameters obtained from clinical studies or economic evaluations identified through systematic reviews, which may come from health systems other than our own. In some instances, access to microdata from the national health information systems has been possible, but the parameters obtained are point estimates, and their dispersion measurements are often unavailable.

The use of RWD can facilitate the transition from stochastic models with limitations (or even deterministic models) to more contextualised stochastic models. RWD, which reflects the variability and complexity of clinical practice and patient populations, can provide more accurate estimates of parameters while accounting for inherent uncertainty. This enables a better capture of the probabilistic nature of healthcare interventions and allows for consideration of a broader range of potential scenarios, leading to more robust and informed decision-making. The main advantages of the data-driven probabilistic approach based on RWD are as follows:

- Improved accuracy: RWD provides more accurate estimates of probabilities and outcomes as it is based on actual patient outcomes and reflects real-world variability and complexity.
- Enhanced generalizability: RWD captures the diversity of patient populations, settings, and practice patterns, leading to more generalizable results compared to models relying on simplified assumptions or literature-based stochastic models. By better reflecting the real-world context, these stochastic models may yield more externally valid and relevant findings for decision-makers.
- Increased transparency: RWD enables transparent and reproducible results when best practices are applied, such as protocol

registration and code sharing. This allows for better understanding and scrutiny of modelling assumptions and inputs.

- **Enhanced decision-making:** Robust stochastic models, driven by RWD, facilitate the consideration of a wider range of potential scenarios. This provides decision-makers with a more comprehensive understanding of uncertainties and risks associated with different interventions, aiding in informed decision-making.

Despite the advantages, some limitations of probabilistic models should be noted, such as the overreliance on model inputs. This highlights the need to understand the limitations of the data and assumptions underlying the models when informing decision-making.

3.5. Decision model development

Modelling makes it possible to present in an explicit and simplified way the possible courses of action of two or more technologies that we need to compare for different aspects (effectiveness, safety, efficiency, patient perspective, organisational implications, legal, etc.).

The development of a decision model for health technology assessment consists of several stages that involve key choices regarding the nature of the evaluation:

- **Specifying the decision problem:** The first step is to clearly identify the question to be addressed in the analysis. This includes defining the population and subpopulations, which typically consist of patients exposed to standard clinical practice, but may also include non-patients in the case of screening and primary prevention technologies. Specific details about individual characteristics, locations, and settings where comparator options are delivered need to be considered. Additionally, the specific comparator options must be detailed, indicating if they are main interventions or companion technologies or sequences of treatments with particular starting and stopping rules. These elements should be pre-specified and registered in an evaluation protocol before proceeding to data extraction and analysis.
- **Prioritising the patient cohort:** Models are abstractions of reality, so choices must be made regarding which comparator options and outcomes will be explicitly represented in the model. Careful consideration and prioritisation of the patient cohort to be

included in the formal model are crucial. This requires a deep understanding of the decision problem and exploration of characteristics, potential outcomes, and associated uncertainties. For example, when assessing specific coagulation tests, the inclusion of patients on occasional anticoagulation therapy due to surgery should be decided.

- **Determining the Structure of Consequences:** Choices must be made regarding the structure of possible consequences for the options being evaluated within the context of the decision problem and model boundaries. The structure of the model is influenced by the characteristics of the interventions being assessed (e.g., diagnostic accuracy, treatment efficacy), the natural history of the condition, and the impact of the options on that process. There are no general rules for the appropriate model structure, but certain features should be considered:
 - The type of disease (acute or chronic) and the occurrence of health-related events over time, as well as the time horizon.
 - The risks associated with these events, whether they change over time or not.
 - The duration and time-limited effects of interventions' effectiveness.
 - Assumptions about patients' future health profile after treatment cessation.
 - The probability of health-related events, which may depend on past patient experiences

Choosing the appropriate model structure often requires an iterative process of consulting RWD. Conducting a literature review can provide initial insights into the type of patients, settings, and positioning of the new technology. This information helps refine the model structure by adding nodes, creating transition phases, or simplifying based on the actual characteristics and treatment pathways of patients in the real world.

Depending on the natural evolution of the clinical condition, the type of technology, and the available data, various models can be deployed.

4. Data request

As previously reported, the use of RWD is certainly valuable for informing the HTA process. However, gaining access to this data can be a complex and sensitive procedure, as it often involves navigating issues of privacy, confidentiality, and adherence to legal regulations and ethical concerns.

Access to RWD typically follows similar pathways as in health research projects. Researchers are usually required to submit data requests to **data holders**, which may include healthcare organisations, government agencies, or other entities that collect, curate and store data. These requests must provide detailed information about the study and the specific data elements needed, as well as assurances that the data will be handled in accordance with legal and ethical guidelines.

In the same way, HTA analysts need to collaborate with data holders to **determine which data will be required** for the assessment. Access to data may be granted based on regulation and the specific decision-making needs.

In the near future, the availability of the **Spanish Healthcare Data Space (ENDS)**, the acronym in Spanish) and **European Health Data Space (EHDS)** may streamline the data request process. Regardless, some of the described steps will still be necessary, such as gathering background information, working with data experts, and optimising the data extraction process. The following sections will describe them in more detail.

4.1. Gathering background information

RWD is not arranged in a way that immediately allows us to fill our information gaps. Data lakes aggregate multiple information sources that were not originally designed for research purposes, and only a small portion of this vast dataset may be of interest to the question at hand.

Thus, in the same way we start by designing a search strategy when working with bibliographic databases, some kind of logic is warranted to narrow down what we are looking for. On the one hand, we need to draw the fragment of all the information available that is most pertinent to our assessment. On the other hand, we need to keep in mind that the metrics of the results coming from these data should be similar to those published by clinical trials; that is, if we want to guarantee its comparability.

The first step is to **estimate our data requirements**, which should be kept to the minimum necessary. Then, we must outline these data needs, as specifically as possible, leveraging the information gathered through our usual systematic review process. Although these steps may appear complex initially, they will ultimately save us considerable time and help facilitate our data request and analysis with minimal setbacks. The decision problem serves as the starting point for defining these data requirements.

Box 1. ICD-SCD use case presentation

We formulated a case for leveraging RWD in the assessment of implantable cardiac defibrillators (ICDs) for the prevention of sudden cardiac death (SCD). The Agency for Health Quality and Assessment of Catalonia (AQuAS) had already assessed this technology using conventional HTA methods (systematic review of literature on randomised controlled trials –RCTs– and traditional modelling techniques)²⁷. We specifically selected this technology because we judged that the outcomes and target population groups could be easily translated into computer language, and comprehensive data would be available for nearly all relevant variables.

We aimed to test procedures and methods to describe the real-world scenario of ICDs between 2011-2018, a period corresponding to the bibliographic searches conducted for the original assessment. Building upon the background information derived from AQuAS’s systematic review, we defined our data requirements and assessment questions. The aim was to characterise the specific target patient populations within the Spanish context and generate insights into the effectiveness and safety under the technologies applied in routine clinical practice during that time frame.

Target Population	Patients with indications for primary and secondary prevention of sudden cardiac death (SCD).
Intervention	Implantable cardioverter defibrillator (ICD).
Comparison	Optimal pharmacological treatment, cardiac resynchronisation therapy (CRT), or cardiac pacemaker.
Outcomes	Sudden cardiac death, death by any cause, adverse events, use of health resources, costs.
Study Period	From 2011-12-01 to 2018-05-31.

Refining the data requirements is an iterative process, but we can actually identify two key moments to enhance it: 1. during the development of the systematic review protocol and 2. After the data extraction.

4.1.1. Systematic review protocol

When planning an HTA protocol, we usually gather general insights about the technology proposed for assessment through an exploratory bibliographic search, expert consultation, examination of manufacturer’s documentation and review of technical specifications of the technology. During this phase,

the following elements are established: the target population for the technology and potential subgroups, current alternatives or comparators, and the primary and secondary outcomes.

This is usually the starting point for establishing inclusion criteria of studies in a systematic review. However, we may also start thinking about how this information is registered within real-life health systems, and thus define the **primary clinical codes** that will serve as filters for data extraction. Additionally, we can start identifying which clinical variables and resource use parameters would be interesting to measure.

By taking into account both the literature-based information and the RWD aspects, the HTA protocol can be more comprehensive and tailored to capture relevant information for the assessment. This early consideration of clinical codes, variables, and resource use parameters sets the foundation for a more robust and efficient analysis of the technology’s impact and effectiveness within the targeted population.



Use tools such as eCIEMaps²⁸ (Spanish Health Ministry) or ATC/DDD Index²⁹ (WHO) to look for preliminary diagnostic and treatment codes.

Below, we present our initial thought process focused solely on the PICO question (Population, Intervention, Comparison, Outcomes) for the Implantable Cardioverter Defibrillator (ICD) for the prevention of sudden cardiac death (SCD), i.e. ICD-SCD use case.

Table 1. Definition of general data requirements for the ICD-SCD use case

Target population	<p>Structuring the definition:</p> <p>Patients with indication for <u>primary prevention</u> of sudden cardiac death (SCD):</p> <ul style="list-style-type: none"> • Patients with high risk for SCD, due to left ventricular dysfunction (LVEF \leq 35%) of ischemic and nonischemic origin, including diagnosis of: <ul style="list-style-type: none"> • Acute myocardial infarction • Chronic heart failure • Dilated cardiomyopathy • Patients with indication for <u>secondary prevention</u> of SCD: <ul style="list-style-type: none"> • History of aborted SCD • History of sustained ventricular arrhythmia
--------------------------	---

Target population	Constructing data requirements:
	<p>We need data from patients in contact with the health system starting from two main points (entry events):</p> <ul style="list-style-type: none"> • Patients who <u>did not suffer a sudden cardiac arrest (SCA)</u>, but have already been diagnosed with a myocardial infarction, chronic heart failure, and dilated cardiomyopathy. <p>They have been possibly admitted to a hospital, but also registered in primary care, their care episodes being coded with ICD-10 diagnostic codes I21 (acute myocardial infarction), I50 (heart failure), and I420 (dilated cardiomyopathy).</p> <p>Not all of them are of interest, only those with LVEF $\leq 35\%$. This is most likely measured in a hospital, through echocardiography.</p> <ul style="list-style-type: none"> • Patients who <u>suffered a SCA</u>. They may have been assisted for other health problems after that, but the event that made them eligible for prevention took place probably in emergency care, and was coded with I46 (cardiac arrest) or I472 (ventricular tachycardia). <p>These are the core elements that would define our data, as there are no other particularities regarding sex, age or clinical context.</p>
Comparison	Structuring the definition:
	<p>Current treatment options are:</p> <ul style="list-style-type: none"> • Optimal pharmacological treatment • Cardiac resynchronization therapy (CRT) • Cardiac pacemaker
	<p>Constructing data requirements:</p> <p>Patients with indications for prevention of SCD may be treated with drugs or medical devices.</p> <p><u>Drugs</u> will be probably used in patients with diseases that increase the risk of SCD (primary prevention). Beta-blockers and ACE inhibitors are often used for ischemic heart disease, and heart failure patients often also resort to diuretics. These agents may be prescribed any moment during the patients' clinical history, the prescriptions being coded with Anatomical Therapeutic Chemical (ATC) codes like C07 (beta blocking agents), C09 (agents acting on the renin-angiotensin system), and C03 (diuretics).</p> <p><u>Devices</u> may be used for secondary prevention, probably soon after an episode of cardiac arrest. Also for primary prevention, but the process in which all the available treatment options are assessed will be probably longer. In any case, CRT or pacemaker insertions take place surgically on a scheduled basis. Thus we need to look into hospital intervention surgery registries with ICD-10 procedure codes like 0JH607Z (Insertion of Cardiac Resynchronization Pacemaker Pulse Generator), and 02H40JZ (Insertion of Pacemaker Lead into Coronary Vein). Devices may also be adjusted, replaced or even removed after the insertion, so we could also look for these codes.</p>

Outcomes	Structuring the definition:
	Most relevant outcomes are: <ul style="list-style-type: none"> • Sudden cardiac death • Death by any cause • Adverse events • Use of health resources • Costs
	Constructing data requirements:
	<p><u>Death by any cause</u> is something easy to know. When a person dies, a series of administrative steps are initiated that result in a change of status in that person's health electronic records (it becomes inactive due to death). As such, we just need to ask the health system if this change took place. We are probably interested in knowing when it happened too, to see which treatments and factors are associated with a longer or shorter survival, and compare it with what we find in the literature.</p> <p><u>Sudden cardiac death</u>, meaning death specifically because of a cardiac cause and specifically sudden/unexpected might be, on the other hand, difficult to know. This would be something apparent for researchers that regularly follow up participants in a planned study. But what about the real world? We might need data coming from medical certificates of death, or we might keep alert for any algorithmic or computerised definition we find while exploring cohort studies during the literature review.</p> <p><u>Adverse events</u>: two types are expected; drug and medical device-related adverse events. In a research study, patients are encouraged to report any adverse events, from mild to severe ones, and researchers register them meticulously. In the real world, a variety of notification systems exist for adverse events. We may ask for any adverse event appearing in these systems, only those linked to the drug codes and patients we previously defined. The insertion of CRT or pacemaker devices have a risk of postsurgical and long-term complications that may need hospital admission or consultation, due to infection (codes A40 and A41), pneumothorax (J9581), or mechanical complications (T821).</p> <p>Finally, for modelling the <u>use of health resources and costs</u>, we need information like the number, reason and duration of admissions, visits, procedures, and prescriptions along a defined time period. Other events, like the number of blood tests a patient undergoes, don't seem so relevant after the exploratory assessment.</p>

With the provided information, we may define an initial set of variables and entry criteria in HTA protocols that incorporate the use of RWD. An example of this is presented in *Annex IV. RWD section in HTA protocol for Direct Oral Anticoagulants (DOAC) quantification*. However, it is important to acknowledge that uncertainties will arise regarding which variables to

include, as well as how and when they should be measured. Therefore, our data requirements at this stage should be considered preliminary. It might be useful to establish some “reading objectives” during the full-text review of the studies focused on specific details that can help us clarify any doubts related to data variables and measurements.

4.1.2. Data extraction

After the data extraction phase of studies, HTA analysts have a better understanding of the scientific literature pertaining to the technology’s efficacy, security, costs, and other relevant outcomes. This familiarity enables us to further define our data requirements, including the following aspects:

- Common distribution of population sociodemographic characteristics within the population under study.
- Typical baseline parameters and their respective units of measurement.
- Prevalent comorbidities among the participants.
- Common clinical pathways, including the temporal sequence of events and the professionals involved during patient care.
- Attrition rate of participants and the frequency of deviations from the treatment protocol.
- Common variables used for subgroup analysis.
- Specific details regarding the measurement of efficacy outcomes, including the scales, instruments and units employed, as well as the timing of these measurements.
- Documentation of adverse events, including the monitored events, classification of mild versus severe events, and how they are attributed to the treatment.
- Information related to the use of health resources and associated costs, including existing economic models and the variables considered in those.

Having all this information at hand allows us to precisely delineate the scope of our data requirements and define the specific results we aim to obtain through the utilisation of RWD. The subsequent section demonstrates how we incorporated this information into the data requirements for the ICD-SCD use case.

Box 2. Definition of specific data requirements for the ICD-SCD use case

The DANISH clinical trial³⁰ was one of the most important studies identified during the HTA of the ICD for people with primary prevention of SCD. This study described the following baseline variables:

- Sociodemographic characteristics: age, sex.
- Measures: systolic and diastolic blood pressure (mm Hg), body-mass index (kg/m²), NT-proBNP level (pg/ml), QRS interval duration (ms), LVEF (%), glomerular filtration rate (ml/min/1.73 m²), New York Heart Association (NYHA) class (from II to IV).
- Disease cause: idiopathic, valvular, hypertension, or other.
- Comorbidities: hypertension, diabetes, permanent atrial fibrillation.
- Drugs: ACE inhibitors, angiotensin II receptor blocker (ARB), beta-blockers, mineralocorticoid-receptor antagonists, amiodarone.
- Devices: CRT, pre-existing pacemakers.

We judged it would be helpful to know how these variables are distributed in our population and how similar or different they are compared to the included population of published clinical trials. Thus, we included them in our list of variables.

As for the outcomes, the trial's main result was death from any cause. The absolute measure for this was events per 100 person-years, which is a mortality rate (incidence rate). Other results were described for:

- Efficacy (proportion of total sample and events per 100 person-years): cardiovascular death, sudden cardiac death.
- Adverse events (proportion of total sample and events per 100 person-years): bleeding events requiring intervention, pneumothorax, inappropriate shocks.

Additionally, we found that included studies reported the following drug classes as part of the optimal medical treatment: beta blockers, digitalis glycosides, angiotensin-converting enzyme (ACE) inhibitors, ARB, angiotensin receptor neprilysin inhibitor (ARNI), and diuretics. Inside diuretics, some studies differentiated a subgroup of patients with aldosterone antagonists. We may ask for all these data in our patient RWD cohort and include them as independent variables.

By systematically identifying all relevant variables, we can construct more comprehensive decision models that incorporate the key factors influencing clinical outcomes, use of health resources, and direct costs. This meticulous approach enables us to generate more precise estimates and obtain generalizable results, enhancing the reliability of preadoption decision-making processes.



Focus on relevant measures that provide a clear understanding of the outcomes. Clinical trials will often report association measures like relative risk (RR), odds ratio (OR), and hazard ratio (HR), while our primary interest lies in absolute frequency measures. Check full-text results, tables and supplementary materials of studies to find metrics like cumulative incidence and incidence rate.

At this stage, we have successfully identified key studies that provide valuable sociodemographic, clinical and economic variables necessary for real-world estimation. By comparing these variables with those obtained from clinical trials, we can achieve a certain level of comparability between the two sets of results. This level of detail allows us to integrate the gathered data into a decision model effectively. Now, it is crucial to consolidate our findings and proceed by requesting the data in a structured and detailed manner. This approach ensures that we obtain the necessary information to further refine our analysis and make informed decisions based on the available evidence.

4.2. Working with data experts

In research studies, particularly clinical trials, researchers gather baseline and follow-up data of the participants with the help of customised case report forms (CRFs). These CRFs outline the specific data to be collected, and the methods and timing of data collection. Consequently, analysts are already aware of the types of data that will be processed for statistical analysis. Furthermore, using CRFs guarantees a pre-processing of the data, minimising variability during their collection.

This is different when working with RWD since researchers or HTA analysts have limited control over how data flows into health information systems. In routine medical practice, patient health information is primarily collected for clinical, legal and management purposes. Therefore, it is crucial to establish our data requirements and methods in advance. This situation forces us to think about how the data are registered into the system and which available data are fit for purpose to answer our assessment questions. Collaboration with data holders or data experts is often required since these individuals possess the most knowledge about the structure and storage of RWD.

Working with data experts aims to develop a **data model specification (DMS)** that serves as the foundation for implementing the **extract, transform, and load (ETL)** process required to obtain the data essential for HTA. The adoption of a common data model (CDM) endorsed by the scientific community may be a good starting point, as they provide readily available data standards to enable efficient analyses. During the development of the DMS, data experts may assist in understanding which data are available for assessment, including the relevant tables, fields, and content. They can also help create the code mappings to retrieve the necessary data.

Keep in mind that while **data holders** may serve as data experts due to their familiarity with the structure and storage of RWD, they may not necessarily provide guidance on the best approaches for subsequent data visualisation or analysis. We might be interested in including additional profiles such as statisticians and data scientists in the HTA working group.

Future architectures, like the EHDS, envision a **data discovery phase**, where data users can search for the specific data they require, potentially leveraging **metadata catalogues** or other types of **search engines** for metadata. In this scenario, the input from data experts becomes crucial in assessing the feasibility of conducting the assessment based on the available data sources.

Overall, collaboration with data experts and data holders, along with establishing a solid data model specification, is vital in leveraging RWD effectively for HTA and ensuring the availability of appropriate data for analysis.



Ask for available data catalogues or data search engines before requesting a data set.

We highly recommend joining forces with data experts to assess what is possible and what is not. This collaborative effort typically begins by defining the specific data of interest, which is commonly referred to as **cohort definition**.

4.2.1. Defining cohort criteria

We have emphasised the importance of utilising a specific subset of all the available data for conducting assessments with RWD. After gathering background information and establishing data requirements, we should be able to imagine exactly how this subset of information should look like.

We need to consider data structures to effectively communicate this idea to data holders. Typically, computerised data is organised in tabular format, each row describing a distinct **entity** (e.g. a person, admission, measure, prescription), and each column representing a field of information shared by all entities. Health systems often store multiple interlinked tabular databases, connected by common variables or columns (relational database management system).

With this understanding, we can now focus on determining the types of entities that should be included in our final data. Usually, health

information systems' databases are **person-centric**, making persons or patients the primary units of analysis in HTA. Our objective is to identify individual-level results associated with different treatments and baseline characteristics. Therefore, we must define specific criteria for data holders to filter the target population related to a particular technology from the stored data – this process is known as cohort definition.

A **cohort** is a set of entities that satisfy one or more inclusion criteria over a specified period. The simplest approach to defining a cohort is by explicitly stating a set of rules that determine when a patient belongs to the cohort (**rule-based cohort definitions**, or also **computable phenotypes**). First, we shall define an initial filter to identify all relevant data, such as the occurrence of a specific disease or exposure to a device or drug. This event is referred to as the **cohort entry event**, which sets the **index date**, and the individuals who meet this criterion form the **initial event cohort**. Next, we further refine this cohort by applying additional **inclusion criteria**. When all criteria are satisfied, we obtain the **qualifying cohort**.

The entry events and inclusion or exclusion criteria may be defined by:

- Clinical conditions.
- Drugs.
- Procedures.
- Measurements.
- Observations.
- Visits.
- Socio-demographic characteristics.

Note that a cohort is not solely defined by a set of clinical conditions. The definition should also incorporate **temporal logic** to evaluate the relationship between an inclusion criterion and an event. A time frame for the inclusion/exclusion events is necessary, with careful consideration given to differentiating between incident and prevalent cases. When considering age, it is also important to specify whether it should be verified at a specific date (the start of the study) or at the time of inclusion in the cohort.

Box 3. Cohort entry events and inclusion/exclusion criteria in the ICD-SCD use case

ENTRY EVENT

According to the clinical uses of ICD and the treatment comparator(s) (i.e. primary and secondary prevention of SCD), we defined 5 possible entry events for the patients in the ICD-SCD cohort:

EVENT	CONDITION	TEMPORALITY
Initial event 1	Diagnosis of cardiac arrest	First diagnosis
Initial event 2	Diagnosis of ventricular arrhythmia	First diagnosis
Initial event 3	Diagnosis of chronic heart failure	First diagnosis
Initial event 4	Diagnosis of acute myocardial infarction	First diagnosis
Initial event 5	Diagnosis of dilated cardiomyopathy	First diagnosis

Thus, patients' data was collected when at least one of these 5 conditions was met, that is when a patient had one or more of these diagnoses anytime among their clinical records.

INCLUSION AND EXCLUSION CRITERIA

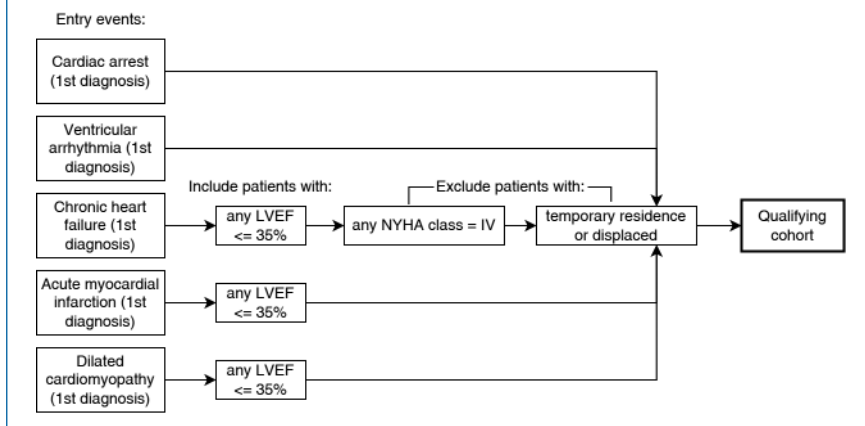
According to our background information, some patients were eligible for ICD treatment just after some of these initial events, particularly patients for secondary prevention of SCD (initial events 1 and 2). However, patients eligible for primary prevention of SCD had to meet additional criteria, as the technology was indicated only in severe patients (but, in the case of chronic heart failure, not the most severe). To account for this, we defined inclusion and exclusion criteria, each one linked to specific entry events:

TYPE	CONDITION	APPLY TO INITIAL EVENTS	TEMPORALITY
Inclusion criteria	LVEF \leq 35%	3, 4, and 5	Anytime during history
Exclusion criteria	NYHA Class IV	3	Anytime during history

While working on the case, data holders asked us about additional criteria we did not initially contemplate, related to administrative characteristics of the patients. Specifically, we defined whether there were:

- Restrictions by health coverage: if all subjects must have had health coverage or whether uninsured citizens would also be included
- Restrictions by usual place of residence: if only residents would be analysed, or whether temporary residents or displaced persons will also be included
- Restrictions by type of health provision: if only patients from the public health system would be studied, or also patients with private coverage would be included

Data experts were concerned about the data availability and reliability in records for uninsured citizens and displaced persons. Thus, we made the decision to exclude these populations from the cohort. In conclusion, our qualifying cohort was built in this way:



4.2.2. Defining research questions

Once we have clearly defined the qualifying cohort and identified the entities from which we will be collecting data, we need to think about the outcomes we want to evaluate using RWD. Similar to any research study, outcomes or results are assessed to answer pre-defined research questions.

Typically, **HTA questions** are systematic review questions, which allow some flexibility in terms of outcomes. This flexibility may include different scales to measure an outcome, different endpoints, and even indirect results. However, when working with RWD, it is of the utmost importance to precisely define how research questions will be answered. This serves as the foundation for the data model specification that will be shared with data holders. Among other things, this means **explicitly stating the variables and their units of measure**.

To facilitate this process, it is helpful to define a set of **indicators** linked to each research question. These are results that should be calculable immediately after retrieving the data. They provide data holders with a clear understanding of the primary information we are seeking and let them assess the feasibility of the assessment. Simultaneously, constructing a set of indicators allows us to share an analysis plan that can be executed by different assessment teams and settings. It also enables us to update the outcomes quickly if future assessments are required.

Indicators can be mathematically expressed as follows:

- **Proportions and percentages:** These indicators compare the numerator event (e.g. people who received a diagnostic test or treatment) to the denominator, which represents the number of persons at risk or eligible during a specific time period. Percentage indicators range from 0% to 100%, making it easier to compare across different groups.
- **Ratios:** Ratios describe the relationship between two numbers, indicating how many times one is contained within the other, e.g. the number of beds per 1.000 inhabitants or the incremental cost-effectiveness ratio (ICER) for assessing technology efficiency.
- **Rates:** Rates are a type of proportion that measure the frequency of an event over time, typically expressed as the number of events per unit of time (e.g. deaths per 1.000 person-years). Rates are useful when the denominator changes over time or when observation periods differ between groups.
- **Means and medians:** These statistics provide more detailed information about specific care aspects measured with quantitative variables, compared to proportion or percentage measures. Mean or median values can capture differences that may not be apparent using proportions or percentages that have a category-based nature. It is important to clearly explain the standards or thresholds for means and medians.
- **Counts:** Some indicators are reported as simple counts of patients meeting particular criteria, events or adverse outcomes. The population and criteria for counting events should be specified. Counts can be particularly useful for feeding cohort models.

For the ICD-SCD use case, a list of indicators is provided in Annex II. ICD-SCD use case indicators.

Cohort characterisation questions

Cohort characterisation entails examining the characteristics of individuals within a cohort both before and after a specific point in time. These kinds of questions will be answered by the count of observations, sociodemographic information, and the presence of conditions and comorbidities among cohort members.

By adopting this approach, we can gain a comprehensive overview of the cohort. Additionally, it enables us to conduct a thorough exploration of the data, identifying any variations and addressing any missing values (methods for addressing missing values will be explained later).

Table 2. Example cohort characterisation questions and indicators

QUESTION	INDICATOR
What is the size of the target population for mRNA human papillomavirus detection tests for cervical cancer screening?	Rate of target patients (per year)
What is the average age of adults diagnosed with specific phobia?	Mean and standard deviation of age (last 5 years)
What proportion of individuals eligible for pulmonary electrical impedance tomography (EIT) are smokers?	Percentage (last year)

Health outcomes questions

Some of the most relevant inquiries we can make of a cohort pertain to their health outcomes. By obtaining this information, we gain insights into the current state of the population under clinical practice circumstances. Subsequently, we can carefully compare these outcomes with those derived from assessing a novel technology, which we obtain by reviewing the scientific literature. Hence, it is optimal to devise indicators that employ the same units of measurement as those found in the research evidence.

Table 3. Example health outcome questions and indicators

QUESTION	INDICATOR
What is the incidence of thromboembolic and haemorrhagic events in patients treated with direct oral anticoagulants (DOACs)?	Incidence rate (events per 1.000 person-years)
What is the average health-related quality of life of people treated for non-neurogenic overactive bladder (OAB)?	Mean and standard deviation of EQ-5D-5L results in primary care (last 5 years)
How many females have tumour spread after sentinel lymph node biopsy for endometrial cancer?	Proportion (last 3 years)

Treatment pathways (comparator arm)

RWD enables pathway analysis to summarise the treatments, drugs, devices or processes received by the target patient population. This analysis allows us to determine whether these treatments occur at a specific point in time, in a static manner, or involve a treatment process that unfolds over time and encompasses the use of various health technologies (such as surgical

approaches, pharmacological treatments, and follow-up procedures). By examining these pathways, we gain valuable insights into the sequence and patterns of healthcare interventions experienced by patients, enabling a comprehensive understanding of their healthcare journey.

Table 4. Example treatment pathways questions and indicators

QUESTION	INDICATOR
How many patients with hip osteoarthritis have received complete conservative treatment before surgical joint replacement?	Proportion of drug use (during last year)
What is the point of access to hospital care for patients with systemic sclerosis?	Proportion of initial hospital consultations by clinical service (during last year)
What is the waiting time for a mental health appointment for individuals diagnosed with severe depression in primary care?	Mean and standard deviation of the number of days between primary care referral request and first mental health appointment (during last year)

Use of resources

One of the dimensions of RedETS that can be more adequately informed with RWD is efficiency. With the availability of RWD, we can now pose questions to the cohort that will help us identify and quantify the resources utilised in routine clinical practice.

Table 5. Example use of resources questions and indicators

QUESTION	INDICATOR
What is the frequency of primary care visits in people with acute gastroenteritis?	Rate of visits (per year)
What is the frequency of drug use in patients eligible for cardiac rehabilitation at home?	Proportion of drug use by ATC class (during last year)
What is the frequency of emergency department visits in people with chronic primary pain?	Rate (per patient and year)

4.2.3. Constructing a data model specification

A data model specification (DMS) is an accurate and comprehensive representation of our data requirements, covering content, structure and constraints. Its primary objective is to ensure a clear understanding of our data needs, allowing us to share this information with data holders,

health technology assessment units, researchers, and stakeholders interested in replicating our results or conducting critical reviews of our assessments.

The development of a DMS can be a costly procedure and not always feasible in an assessment. Alternatively, a common data model (CDM) specification such as the Observational Medical Outcomes Partnership (OMOP) CDM may be used instead, maximizing speed while keeping the reproducibility and transparency of the process. However, one must take into account that existing CDMs may not be suitable or precise enough for all HTA questions. For this reason, we provide guidance tailored for the development of a DMS below.

Currently, there is no standardised approach for developing a DMS, leading to variations with differing levels of complexity and completeness. However, at a minimum, a DMS should include essential components such as **cohort entry criteria**, dataset **entities**, and **variables**.

In sections 4.2.1 and 4.2.2, we reviewed concepts and procedures for defining cohort entry criteria, research questions and indicators. It is imperative to incorporate this information into the data model specification. For guidance, Annex III shows the ICD-SCD use case data model specification, which can serve as a template for this process.

The next crucial step involves defining the final list of entities and variables necessary to answer research questions and calculate indicators. Up to this point, we should have gathered sufficient information to draft this list, through conducting a systematic review of the literature. Still, working with data experts is vital to develop the DMS, especially for variable definition. Their insights into the availability of variables in the data sources, as well as their assistance in comprehensive mappings and crosswalks for diagnosis, procedure and drug codes, will greatly enhance the quality of our DMS.

Data model specifications are handy for establishing the computational definitions of variables with the required precision to avoid confusion during the data extraction process and facilitate the replicability of analyses. When defining variables, it is important to consider the following aspects:

- **Encoding:** The specific standards for diagnostic and procedure codes (e.g. ICD-10, ICD-9).
- **Format and type:** Whether the variables are represented as strings (e.g. “cardiac failure”), numeric values (e.g. “1.25”), or logical values (e.g. “TRUE”/“FALSE”).

- **Units:** In the case of numeric variables, the units of measurement (e.g. %, ml, mmHg).
- **Requirement level:** This communicates to data holders whether the variable is essential for our assessment (i.e. necessary to estimate a primary outcome) or optional.
- **Validation rules:** It is possible to set preliminary conditions to filter out extreme or implausible variable values in advance. However, be aware that this may limit our ability to assess data quality during analysis and hinder the possibility of partially recovering data through simple transformations (e.g. some extreme values might simply be missing a decimal point).
- **Transformations at the origin:** We may request some transformations for variables, such as dates following a particular format or encoding (e.g. ISO 8601 for dates in YYYY-MM-DD).
- **Property:** Variables may be defined as “observed”, meaning their values are as registered at the source, possibly with some transformations. Alternatively, we may request “calculated” variables that do not exist at the source level but can be generated using simple rules. For example, instead of requesting all beta-blocker prescriptions for patients in the cohort, we can ask the data holders to generate a TRUE/FALSE variable indicating the prescription of a beta-blocker by checking if there is any prescription for ATC codes related to beta-blockers in the electronic records for each patient and including this variable in the dataset.
- **Possible data sources:** Some variables may come from multiple data sources. For instance, when considering death, we can ask data holders to include only hospital admissions that ended in death or point to primary care administrative sources that include deaths occurring outside of a hospital context. The choice of data sources depends on the context of our assessments and research questions.

Considering these aspects will ensure that our data model specifications are comprehensive and accurately represent the variables of interest.

Table 6. List of entities and variables in the ICD-SCD use case

ENTITY	VARIABLES	
patient	Pseudoidentification	patient_id

ENTITY	VARIABLES	
patient	Sociodemographic	age_nm, sex_cd, socecon_lvl_cd, health_ar_cd, health_zn_cd, birth_place_cd
	Disease	cardiac_arr_bl, ventricular_arr_bl, ischemic_hf_bl, diagnosis_dt, nyha_cd
	Comorbidities	diabetes_bl, hypertension_bl, copd_bl, ckd_bl, cld_bl, cancer_bl, fibrillation_bl, smoking_bl, obesity_bl
	Devices	icd_bl, icd_dt, crt_bl, crt_dt, pacemaker_bl, pacemaker_dt, device_bl, device_dt
	Medications	beta_blocker_bl, digitalis_bl, ace_bl, arb_bl, arni_bl, diuretics_bl, aldosterone_anta_bl
	Resources	pc_visits_nm, em_adm_nm, hospital_adm_nm, cardiology_visits_nm
	Mortality	death_dt, scd_bl
	Adverse events	infection_bl, bleeding_bl, pneumothorax_bl, shocks_bl
	Follow-up	follow_up_nm, time_risk_nm
measure	patient_id, measure_cd, measure_nm, measure_unit_cd, measure_dt	
admission	patient_id, prescription_code_cd, prescription_dose_nm, prescription_unit_cd, prescription_dt	
medication	patient_id, admission_dt, discharge_dt, admission_diagnosis_cd, discharge_type_cd	

Note: a convention is used for different data types; ‘cd’ for categorical vars, ‘nm’ for numerical vars, ‘bl’ for binary/logical vars, and ‘dt’ for date vars; ‘id’ is reserved for the primary (and secondary) key of the entity



Think of entities as different tables, each one with different columns (variables) specific for that entity.

4.3. Implementing the data extraction process

Once an agreement has been reached on the data model specification and it has been shared with data experts, a technical professional should assume responsibility for implementing the extract, transform, load and (given the current scenario) export process. From this point onward, the assessment team should establish and maintain regular communication with data holders for **troubleshooting** when something does not work as expected. Despite specifying all aspects of the required dataset in the DMS, some common problems may still arise, including:

Table 7. Common problems and solutions during data extraction

PROBLEM	POSSIBLE SOLUTIONS
Multiple possible data sources for a variable	Prioritisation of data sources based on careful consideration of the data input context (e.g. prioritising primary care data over emergency care data for establishing patient baseline data)
Implausible outliers and extreme observations present in the values of a variable	<ul style="list-style-type: none"> • Including variable validation rules in the data model based on plausible values according to scientific literature, followed by a new data extraction • Using imputation methods to replace implausible values (based on the mean, median, mode, or using more advanced techniques) and performing sensitivity analysis around cap values • Transforming affected values when a clear registry error is suspected (e.g. values lacking decimal separators, or in other units of measurement) • Capping extreme values at a minimum and/or maximum threshold (e.g. the 5th and 95th percentiles of the distribution)
Unavailable data sources/ variables	<ul style="list-style-type: none"> • Looking for proxy variables • Exploring computer case definitions available in published literature • Designing and testing a probabilistic definitions for unavailable variables
Unstructured variable records	Implementing natural language processing (NLP) techniques to extract information from unstructured sources, such as patient discharge summaries

Given the factors mentioned above, the data request should be viewed as an **iterative** process. The DMS may require adjustments to accommodate any limitations identified during the data extraction phase. HTA analysts, data experts and technical personnel should at all times **weigh** the need to obtain some data against the costs required to do so. Probabilistic approaches followed by a more conservative stance during data analysis may be preferable to lengthy decision-making periods aimed at obtaining exact data.

Looking ahead, improvements in health information systems are expected to foster greater integration and organisation of data sources. These advancements will simplify the processes involved in working with RWD. It is also worth noting that recent initiatives tend to avoid all data extraction or transfer process, instead running all analyses locally in a secure environment using federated approaches. As such, this step is prone

to change as RedETS fosters its participation within these initiatives and the projected health data spaces.

Box 4. A problem of unavailable variables in the ICD-SCD use case

In the ICD-SCD use case, we requested the following patient variables to assess the outcome of sudden cardiac death:

- Immediate cause of death.
- Intermediate cause of death.
- Initial cause of death.
- Other significant pathological conditions that contributed to the death.

We pointed out [the mortality registry](#) as a possible data source. This was included in the first version of the data model specification. Later, data holders informed the assessment team that this source was not part of the BIGAN platform. They recommended [designing a probabilistic definition](#) of SCD based on short hospital admissions ending in death and accounting for diagnosis codes for the admission.

The assessment team looked for [computer case definitions](#) in PubMed, and found an article by Chung et al.³¹, which tested a computerised definition for SCD based on three criteria:

- No evidence of a terminal hospital admission/nursing home stay in any of the data sources.
- An underlying cause of death code consistent with sudden cardiac death.
- No terminal procedures inconsistent with non-resuscitated cardiac arrest.

The authors reported a positive predictive value of the computer case definition of 86.0% in a development sample and 86.8% in a subsequent validation sample. Looking at the reference list of this article, we found that this definition was later used in a study of mortality caused by ventricular arrhythmias by Viles-Gonzalez et al³². The authors observed a mean stay of 7.4 days (SD +/- 0.13) in admitted patients that ended with SCD.

Using crosswalks of ICD-9 codes provided by Chung et al. and the mean length of stay in the study by Viles-Gonzalez, we defined the following [criteria](#) for SCD in our use case:

- The patient died during the observation period.
- The patient had an emergency department admission or hospital admission during the last 8 days before the death (the day of death is included in this period).
- The patient had a diagnosis compatible with sudden cardiac death as the reason for the emergency department or hospital admission (see code list in the full specification provided in Annex III).

Limitations for this approach were discussed with data experts, mainly limited predictive value and lack of data for patients that did not die in a hospital context.

Box 5. A problem of unstructured records in the ICD-SCD use case

In the ICD-SCD use case, not all patients diagnosed with chronic heart disease were eligible for primary prevention of SCD. Only those with a high risk for this outcome due to left ventricular dysfunction (LVEF \leq 35%) would receive drug treatment or a medical device for prevention. Thus, we included an LVEF \leq 35% as part of the cohort inclusion criteria in the data model specification.

During the data extraction process, this variable turned out to be unavailable, as measures coming from an echocardiogram are not currently stored in a systematic or structured format in Aragonese health information systems. Not accounting for LVEF during the data extraction process resulted in a much bigger target patient population than expected. As the data set would probably include a high proportion of patients with normal LVEF, mortality estimation during data analysis was expected to be lower than it ought to be, thus limiting possible this and other comparisons between the RWD and the data published in the literature.

Data holders explained that, although this was not a variable that could be immediately extracted from the records, clinicians usually manually introduce it in hospital discharge summaries. As LVEF was considered necessary for adequate analysis, data holders offered to implement natural language processing of patient emergency care discharge summaries to retrieve this information.

This was a feasible solution because the way of recording LVEF is very homogeneous among healthcare professionals. Data experts described that in cases like this, in which clinical terms and concepts are very standardised among clinicians, extracting the associated values is very simple; it is usually enough to use algorithms based on regular expressions (text patterns). Furthermore, they had already worked with data from a previous research study of a cohort with congestive heart failure in which it was also necessary to extract information on LVEF, and in which an algorithm for extracting this value was developed and trained.

5. Data description and analysis

5.1. Exploratory data analysis

Exploratory data analysis (EDA) is a crucial step in the process. The main goal of EDA is to gain insights and understand patterns in the data that can be used to answer our research questions. RWD can be complex and messy, and EDA can help to identify potential issues and highlight important characteristics of the target population and inform the subsequent modelling and analysis. The EDA should be guided by objectives, which are materialised in research sub-questions that we will want to formulate to real-life data in order to solve certain unknowns or needs. Here we present some of the main objectives in EDA and which sub-questions would be the most appropriate in our use case.

Table 8. Objectives and questions in exploratory data analysis

OBJECTIVES	RESEARCH SUB-QUESTIONS
<p>Understanding data availability and quality: EDA allows for a comprehensive understanding of the available RWD and its quality. It helps identify missing data, outliers, inconsistencies, and potential biases. By exploring the data distribution, summary statistics, and data completeness, we can assess the data's reliability and potential limitations for our decision model.</p>	<ul style="list-style-type: none"> • Are there any missing or incomplete data fields that may impact the analysis? • What is the reliability and accuracy of the data sources used? • Are there any data quality issues, such as unstructured records?
<p>Cohort description: Cohort characterisation involves defining the study population and describing its key characteristics. It also helps identify the relevant patient subgroups and assess their representativeness compared to the target population.</p>	<ul style="list-style-type: none"> • How many patients with an indication for primary or secondary prevention of SCD are captured in our RWD? • What are the demographic characteristics of the cohort (age, gender, etc.)? • What are the clinical profiles of the patients (e.g., underlying cardiac diseases, comorbidities)? • How are the patients currently managed in terms of pharmacological treatment, cardiac resynchronization therapy (CRT), and cardiac pacemakers?

OBJECTIVES	RESEARCH SUB-QUESTIONS
<p>Assessing generalisability: By characterising the cohort, we could assess the generalisability of the RWD to the broader population of interest. It involves comparing the cohort's demographics and clinical profiles to the established clinical guidelines, previous clinical trials, or reference populations. This step helps understand the external validity of the RWD and identify potential sources of bias or heterogeneity.</p>	<ul style="list-style-type: none"> • To what extent does our RWD cohort represent the target population for the ICD? • Are there any systematic differences between our RWD cohort and the whole population of interest? • How similar or different are the characteristics and outcomes of our RWD cohort compared to previous clinical trials or systematic reviews that analyse interventions for primary or secondary prevention of SCD?
<p>Identifying confounding factors: EDA allows for the identification of potential confounding factors that may impact the outcomes. Confounders are variables associated with both the exposure and the outcome, and they can introduce bias in the analysis. By identifying and adjusting for confounders, we could be able to mitigate bias and improve the considerations arising from our decision models.</p>	<ul style="list-style-type: none"> • What potential confounding factors (e.g., age, comorbidities) may influence the outcomes of interest? • Are there any differences in baseline characteristics between patients receiving different treatments (optimal pharmacological treatment, CRT, pacemaker)?
<p>Supporting modelling and analysis: EDA and cohort characterisation provide valuable insights for subsequent modelling. They inform the selection of appropriate statistical methods, modelling techniques, and adjustment strategies.</p>	<ul style="list-style-type: none"> • What are the event rates (mortality, arrhythmia-related events) in the cohort over a specified follow-up period? • How can the transition probabilities between different health states (e.g., stable patients, hospital admissions) be estimated using the available data? • What are the resource utilisation patterns and costs associated with different treatment strategies?

Additionally, data description and visualisation techniques are utilised to ensure data accuracy, consistency, and completeness. These techniques are essential for ensuring the reliability and validity of conclusions drawn from the data. Various techniques can be employed for EDA depending on the type of data being explored³³⁻³⁵.

Categorical data

Nominal variables are those in which each category or value corresponds to a characteristic or quality that a person in the cohort possesses. The possible values are mutually exclusive (such as sex, smoking status, history of sudden cardiac arrest, etc.). Nominal variables include all variables whose value can simply be a yes (1) or a no (0), or corresponds to more than two non-hierarchically orderable classes.

Nominal variables can be:

- **Dichotomous or binary:** These variables have only two categories. It is recommended to assign 0 as “no” and 1 as “yes” to indicate the presence of a condition. For instance, a variable for sudden cardiac arrest (`cardiac_arr_bl`) would have a value of 1 if the patient in the cohort has experienced cardiac arrest. Similarly, a variable for the prescription of a beta-blocking agent (`beta_blocker_bl`) would be assigned a value of 1 if the patient has been prescribed any of the ATC codes corresponding to beta-blocking drugs.
- **Polychotomous:** These variables have multiple categories. For example, the variable “substance prescribed” (`prescription_code_cd`) can have different values based on the ATC codes prescribed to the patients in the cohort.

We use nominal scales to measure them, where values are identified with words. A nominal scale only allows for classification, but not ordering or ranking.

Ordinal variables have values that are ranked and ordered. The scale used to measure them is called an ordinal scale. These variables enable hierarchical ranking of different values. Examples include the maximum education level or socioeconomic status attained by patients in the cohort (`socecon_lvl_cd`).

Numerical data

Numerical variables are those expressed in numbers that accurately represent the actual data. There are two types of data:

- **Discrete variables** can only take on isolated numerical values, which are finite and coincide with whole numbers. Examples include age in years, the number of primary care visits for each patient during the observation period, etc.

- Continuous variables are also numerical, but they can take on values with a number of decimals that tend towards infinity. Examples include weight, height, blood pressure, etc. When working with RWD, it is possible that some theoretically continuous data ends up being treated as discrete because measurement instruments are limited.

Patient characteristics are collected in **variables**, which are qualities or quantities that have been collected in the information systems regarding a particular patient characteristic. When we put a group of patients together, according to a common set of rules (see chapter on Defining the cohort), we will find that there is **variation** in the values that a particular characteristic takes across the entire cohort, and we will also find that there is **covariation** between two or more variables whose values move in a related way. In the next boxes, we are going to use the RWD we have extracted for our use case and do a descriptive analysis as an example.

5.1.1. Analysing variation

The approach to exploring variation within a single variable for all patients in our cohort depends on the variable type. This step involves summarising the data using percentages or counts, measures of central tendency (such as mean, median, and mode) and measures of dispersion (such as range, variance, and standard deviation).

To summarise **categorical data**, we can use frequency tables:

- A **frequency table** displays the frequency of each category or value in our dataset. It provides an overview of how many times each category or value occurs in the entire cohort, offering a summary of the data distribution. Frequency tables are useful for identifying missing or invalid values, as well as outliers or unusual observations in the data. For example, Table 6A in Box 6 presents the count and percentage of observations for the “Sex” variable, while Table 6B shows the count and percentage for the “Hypertension” variable.

Box 6. Summarising categorical data

Let's consider our interest in determining the proportion of patients based on their sex (sex_cd in our data model) and the number of patients in our cohort who have hypertension as a baseline condition (hypertension_bl in our data model).

To analyse the proportion of patients by sex, we can calculate the frequency or count of each sex category (e.g., male and female) and express it as a percentage of the total cohort. This will provide an understanding of the distribution of sexes within our dataset.

Table 6A. Frequency table for "Sex" variable

SEX	COUNT	FREQ.
Men	12908	54.73%
Women	10675	45.26%

The composition of our cohort reveals that the proportion of men is slightly higher, accounting for 55% of the total cohort.

Furthermore, to determine the volume of patients with hypertension as a baseline condition, we can count the number of patients who have a positive indication for hypertension in the dataset. This will give us an estimate of the prevalence of hypertension within our cohort.

Table 6B. Frequency table for "Hypertension" variable

HYPERTENSION	COUNT	FREQ.
High blood pressure	19077	80.89%
No high blood pressure	4506	19.11%

In addition, it is notable that a majority of our patients have hypertension. Specifically, 81% of the patients in our cohort have information indicating the presence of a code compatible with hypertension.

Measures of central tendency are used to describe the central or typical value of a set of **numerical variables**. The most commonly used measures of central tendency are the mean, median, and mode (Box 7, Table 7A).

- The **mean** is the sum of all values divided by the total number of values. It is affected by extreme values or outliers and is most useful for normally distributed data.
- The **median** is the middle value when the data are ordered from smallest to largest. It is less affected by outliers than the mean and is more useful for skewed data.
- The **mode** is the most frequently occurring value in the data set. It is useful for identifying the most common value or values in the data set for discrete values.

Measures of dispersion are used to describe how spread out the data is. The most commonly used measures of dispersion are the range, variance, and standard deviation (Box 7, Table 7B).

- The **range** is the difference between the maximum and minimum values in the data set. It provides a rough estimate of the spread of the variable in the dataset.
- The **variance** is the average of the squared differences from the mean. It measures how much the data varies from the mean.
- The **standard deviation (SD)** is the square root of the variance. It also provides a measure of how spread out the data is, but expressed in the same units as the variable.
- The **coefficient of variation (CV)** is the ratio of the standard deviation to the mean and can be useful in comparing between data sets with different units or considerably different means.
- **Interquartile range (IQR)** is the distance between the first and third quartiles of the data. It is not affected by extreme values or outliers.

Measures of shape are used to describe the symmetry and peakedness of the distribution. The most commonly used measures of shape are skewness and kurtosis (Box 7, Table 7C).

- **Skewness** measures the degree of asymmetry in the distribution. A positive skewness indicates that the distribution is skewed to the right, while a negative skewness indicates that the distribution is skewed to the left. A symmetric distribution has skewness equal or close to zero.
- **Kurtosis** measures the peakedness of the distribution. A high kurtosis indicates a sharp peak, while a low kurtosis indicates a flatter distribution than the normal distribution (which is used as the reference).

Measures of position are used to describe the position of a value relative to the rest of the data.

- **Quantiles** indicate where a given value of a variable ranks in the ordered set of its data. This position is expressed as the proportion of the data that fall below that value (percentile). The median, being the central value of the ordered data, occupies the 50th percentile (P50) because below it lay 50% of the subjects.
- **Quartiles** are other important cut-off points are the 25th (P25) and 75th (P75) percentiles, which, together with the median, divide the ordered variable into four equal parts (quartiles).

Box 7. Summarising numerical data

We are now going to work with the variable “Hospital admissions” (hospital_adm_nm in our data model). This variable collects the total number of hospital admissions that each patient has had during the entire follow-up period of our cohort. It can be useful, for example, to estimate the overall resource use of our cohort. We show the measures of central tendency, dispersion, shape and position for this variable.

Table 7A. Central tendency measures

HOSPITAL ADMISSIONS (DURING THE OBSERVATION PERIOD)	
Mean	3.89
Median	3

The mean of 3.89 hospital admissions suggests that, on average, patients in the cohort experienced approximately 3.89 episodes requiring hospitalisation. Considering that the observation period is 6.5 years, this indicates a relatively moderate utilisation of healthcare services. The median of 3 implies that half of the patients had three or fewer admissions. This suggests that the majority of patients had a relatively low number of hospital episodes, while a smaller portion had a higher number of admissions.

Table 7B. Dispersion measures

HOSPITAL ADMISSIONS (DURING THE OBSERVATION PERIOD)	
Range	1-70
Variance	11.19
Standard deviation	3.35

The wide range of 1 to 70 admissions implies substantial variability in healthcare utilisation among the patients. This variation could be due to differences in health conditions, severity, and treatment needs. The high variance of 11.19 indicates that the number of hospital admissions is dispersed widely around the mean, indicating heterogeneity among the patients’ utilisation patterns. This suggests the presence of subgroups with different levels of healthcare needs. The standard deviation of 3.35 highlights the average amount of deviation from the mean, indicating the typical spread of the data.

Table 7C. Shape measures

HOSPITAL ADMISSIONS (DURING THE OBSERVATION PERIOD)	
Skewness	3.29
Kurtosis	30.22

The positive skewness of 3.29 suggests that there is a group of patients with a higher number of hospital admissions, causing the distribution to be skewed towards the right (also implied by the mean being larger than the median). This may indicate a subset of patients with more complex medical conditions or chronic illnesses requiring frequent hospitalisation. The kurtosis of 30.22 indicates a heavily tailed distribution with a distinct peak. This suggests the presence of outliers or extreme values in the dataset, potentially representing a small group of patients with an exceptionally high number of admissions.

Table 7D. Position measures

HOSPITAL ADMISSIONS (DURING THE OBSERVATION PERIOD)	
Q1	2.00
Q2 (median)	3.00
Q3	5.00
IQR	3.00

The quartiles and IQR provide insights into the distribution of hospital admissions. The first quartile (Q1) value of 2 indicates that 25% of the patients have 2 or fewer admissions, representing a relatively low level of hospital utilisation. The median value (Q2) of 3 suggests that 50% of the patients have 3 or fewer admissions. The third quartile (Q3) value of 5 indicates that 75% of the patients have 5 or fewer admissions, highlighting a subset of patients with higher hospital utilisation. The IQR tells that the middle 50% of patients have admissions ranging from 2 to 5. The presence of patients with up to 70 admissions suggests the existence of a small subset with exceptionally high utilisation. These outliers could represent patients with complex medical conditions, multiple comorbidities, or prolonged hospital stays.

Visualisation

Visualisations play a crucial role in presenting information in a concise and comprehensible manner, allowing for a quick understanding of the data. By employing effective graphical methods, we can convey a wealth of information in a visually appealing way. Well-designed visualisations establish a strong connection and provide immediate insights. Simple yet appropriate graphs can effectively describe large volumes of data, enhance the overall understanding, and facilitate the exploration of data distribution. They also aid in error detection and the identification of missing data and outliers

- **Pie charts** are particularly useful for visualising nominal categorical variables. Each category is represented by a slice of the pie, with the area proportional to its frequency. The percentage of the 360-degree circumference corresponds to the relative frequency of each category. Pie charts offer a high-level overview, illustrating how a set of cases is distributed. In Box 8, Figures 8A and 8B showcase pie charts demonstrating the proportion of patients by sex and hypertension at baseline in our cohort.
- **Bar plots** are the preferred choice for representing ordinal categorical variables. They consist of bars or rectangles, where the height is proportional to the number of observations in each category. Categories are displayed along the horizontal axis, while the vertical axis represents the frequency or count. The

rectangles are distinct and separated from one another. Only the observed values in the data are considered, so the axis does not necessarily include consecutive values. Figure 8C displays a bar plot for the “Socioeconomic group” variable (socecon_lvl_cd in our data model).

Box 8. Visualising categorical variables

Using the frequency tables we created earlier, we can visualise the percentage distribution by sex and by the existence of hypertension at baseline in our cohort.

Figure 8A. Pie chart for “Sex” variable

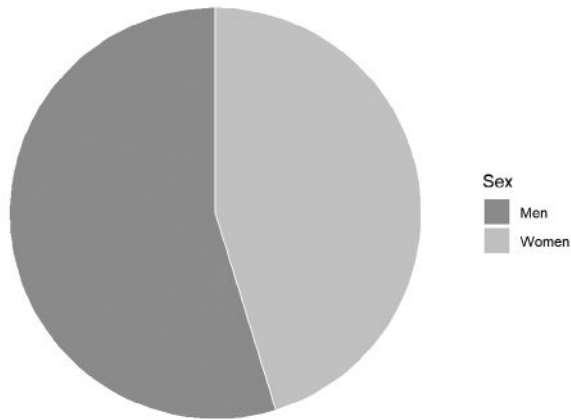


Figure 8B. Pie chart for “Hypertension” variable

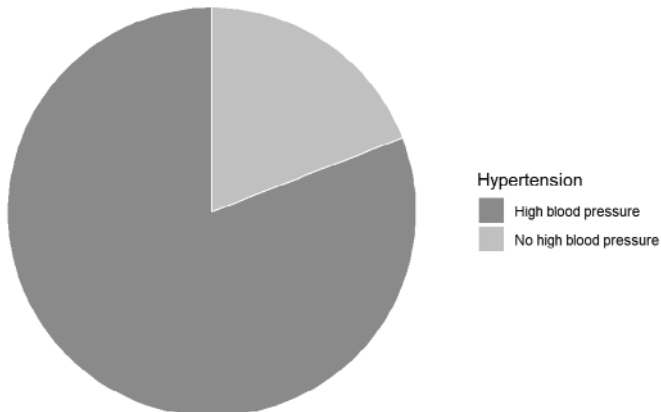
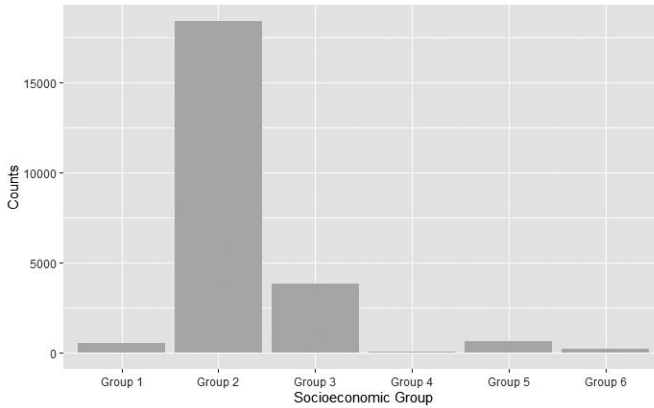


Figure 8C. Bar plot of counts for “Socioeconomic groups” variable



Since the bar chart is the most appropriate representation for ordinal categorical variables, we use an ordinal categorical variable (the ones used above cannot be considered ordinal). In this example, we plot the “Socioeconomic group” variable. This variable indicates the socioeconomic group for every patient in our cohort, which has been proxied with the categories defined for pharmaceutical co-payments (TSI codes). As we can see, most of the patients in our cohort belong to Group 2, who are retired persons.

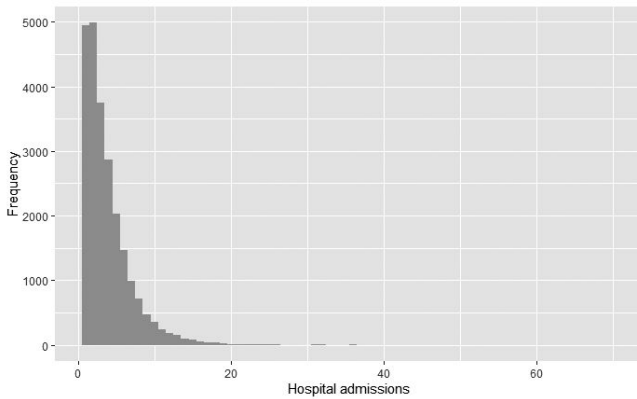
- **Histograms** are well suited for representing numerical variables, ideally continuous ones. Unlike bar charts, histograms consider all possible values within a specified range and present the rectangles together. As continuous variables can have decimal values, each rectangle is represented by the midpoint of the interval. The horizontal axis includes all categories, even those that are empty. Figure 9A in Box 9 illustrates the histogram for the “Hospital admissions” variable.
- **Density plots** provide a graphical representation of a variable’s distribution. They estimate the probability density function of a random variable by smoothing the observed data. Density plots are useful when working with continuous variables, as they offer a smooth representation of the distribution. They provide a clear overview of the data’s distribution, facilitating the identification of outliers, skewness, and other abnormalities.
- **Box plots (or box-and-whisker plots)** are frequently used visualisations due to their descriptive properties. They consist of a rectangular box and vertical extensions called whiskers. The box’s boundaries represent the 25th and 75th percentiles (interquartile range), with a line indicating the median (50th percentile). The whiskers extend from the 25th and 75th percentiles to the adjacent minimum and maximum values (quartiles ± 1.5 times

the interquartile range). However, outliers beyond the whiskers may exist. Box plots are commonly presented vertically, as depicted in Figure 9C in Box 9.

Box 9. Numerical data visualisation

Now we can produce an informative visual summary of the distribution of hospital admissions, allowing for a quick assessment of the frequency of different admission counts and an understanding of the general pattern of hospital utilisation in our cohort.

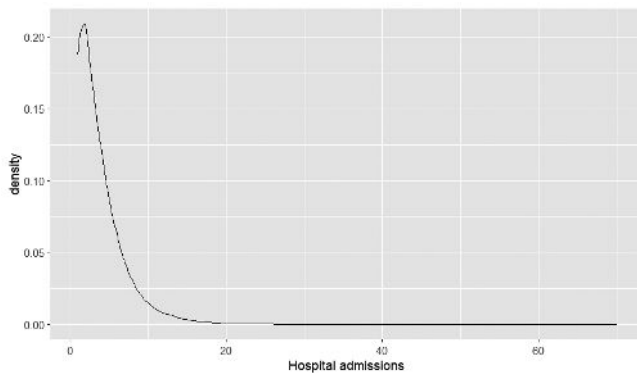
Figure 9A. Histogram for the “Hospital admissions” variable



The histogram for the Hospital admissions variable, which has a range of 1 to 70, a variance of 11.20, and a standard deviation of 3.35, provide a visual representation of the distribution of hospital admissions across the observations. This histogram exhibits a right-skewed distribution with a peak near the lower end of the range. This suggests that a significant proportion of observations have low hospital admission counts, with fewer observations having higher counts.

By examining the density plot, we can gain insights into the central tendency, variability, and shape of the “Hospital admissions” variable’s distribution. It provides a visual summary of the data and allows us for comparisons and further exploration.

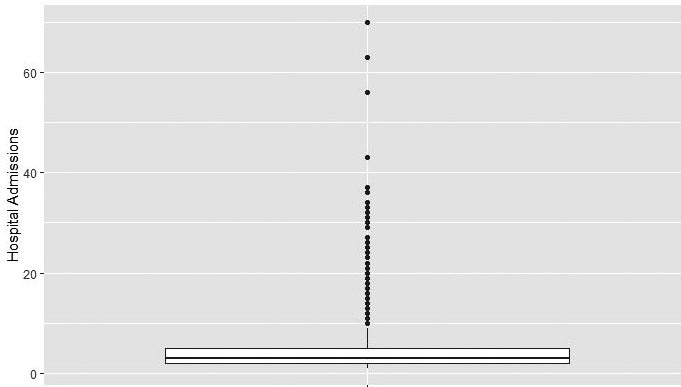
Figure 9B. Density plot for the “Hospital admissions” variable



Note: a smoothing bandwidth based on the biased cross-validation rule was specified within the density plot parameters.

The density plot displays the estimated probability density function of the variable. It provides a curve that represents the distribution of “hospital_admissions” values. This plot shows a peak around the median value of 3 hospital admissions. The curve is slightly skewed to the right, indicating that the distribution is positively skewed. This means that there are relatively more observations with lower hospital admissions and a few outliers with higher values.

Figure 9C. Box plot for the “Hospital admissions” variable



The box represents the interquartile range (IQR) and encompasses the middle 50% of the data. The bottom boundary of the box represents the first quartile (Q1), which is 2, and the top boundary represents the third quartile (Q3), which is 5. The IQR is the range between Q1 and Q3, indicating that half of the data falls within this range. The horizontal line inside the box represents the median, which is 3. It divides the data into two equal halves, with 50% of the observations falling below the median and 50% above. The individual points that fall within 1.5 times the IQR (here, 9.5 hospital admissions) are considered outliers, which are values that fall significantly outside the typical range of the data. Values within 3 times the IQR (here, 14 hospital admissions) may be considered extreme values.

5.1.2. Analysing covariation

Covariation refers to the relationship or association between two or more variables. It involves examining how the values of one variable change in relation to the values of another variable.

Two categorical variables

- **Contingency tables** provide a useful tabular representation of their joint distribution. In a contingency table, the rows represent one categorical variable, while the columns represent another categorical variable. Each cell in the table contains the count or frequency of observations that belong to each combination of categories. To analyse the covariation between these variables, we can derive several measures from the contingency table (Box 10, Table 10A).
 - Conditional proportions or row/column proportions offer insights into the distribution of one categorical variable within the categories of another categorical variable. Row/column proportions can be calculated by dividing the counts in each cell by the sum of counts within the corresponding row/column.
 - Marginal proportions represent the overall distribution of each variable individually. They can be obtained by summing the counts within each row or column and dividing them by the total count of all observations.

The statistical significance of the association between two variables in a contingency table may be assessed with a chi-squared test (χ^2 test). The most widely used of these tests is the Pearson's chi-squared test, which is a p-value based test. In addition, Fisher's exact test might be used instead when working with small sample sizes.

These measures allow us to examine the relationships and patterns between categorical variables, providing valuable insights into their covariation. By analysing the joint distribution and calculating various proportions, we can gain a deeper understanding of how these variables are related within our cohort.

Box 10. Summarising two categorical variables

Contingency table allows us to examine the distribution of high blood pressure across gender categories. By analysing row proportions, column proportions, conditional proportions, and marginals, we would gain insights into the prevalence of high blood pressure within each category and the overall relationship between the variables.

Table 10A. Contingency table for “Sex” and “Hypertension” with row proportions

GROUP	HIGH BLOOD PRESSURE	NO HIGH BLOOD PRESSURE	TOTAL HYPERTENSION
Men (n [%])	9916 (76.82%)	2992 (23.18%)	12908 (100%)
Women (n [%])	9161 (85.80%)	1514 (14.20%)	10675 (100%)

Row proportions indicate the distribution of each row category as a proportion of the total count in that row. For example, among men, the row proportion of high blood pressure indicates that approximately 76.82% of men have high blood pressure. Similarly, among women, the low proportion of high blood pressure indicates that around 85.8% of women have high blood pressure.

Table 10B. Contingency table for “Sex” and “Hypertension” with column proportions

GROUP	HIGH BLOOD PRESSURE	NO HIGH BLOOD PRESSURE
Men (n [%])	9916 (51.98)	2992 (66.40%)
Women (n [%])	9161 (48.02%)	1514 (33.60%)
Total Sex	19077 (100%)	4506 (100%)

Column proportions represent the distribution of each column category as a proportion of the total count in that column. For instance, for the “High blood pressure” column, the column proportion for men indicates that approximately 51.9% of individuals with high blood pressure are men. Likewise, the column proportion for women suggests that about 48.1% of individuals with high blood pressure are women.

Marginals represent the totals for each category of the variables. In this table, the marginal totals represent the total counts for each level of “High blood pressure” and “Sex.” For instance, the total count for high blood pressure is 19.077, while the total count for men is 12.908. In percentages, these constitute the 80.89% and 54.73% of the cohort population, respectively.

Numerical vs. categorical variables

- **Summary statistics**, such as mean, median, standard deviation, minimum, and maximum, mentioned before, for the numerical variable within each category of the categorical variable. This would allow us

to observe the central tendency and variability of the numerical variable separately for each category (Box 11, Table 11A).

The statistical significance of the difference between the means of two groups may be assessed with a t-test when the data's normality assumption holds. For more than two groups, a one-factorial analysis of variance (or one-factor ANOVA) might instead be used.

Box 11. Summarising numerical vs categorical variables

We can obtain a table showing the mean, median, standard deviation, minimum, and maximum values of the “Hospital admissions” variable for each category of “Hypertension”. This allows us to compare the summary statistics across different categories.

Table 11A. Summary for “Hospital admissions” and “Hypertension”

	MEAN	MEDIAN	SD	MIN.	MAX.
High blood pressure	4.14	3	3.45	1	70
No high blood pressure	2.84	2	2.60	1	34

For individuals with high blood pressure, the mean hospital admissions is approximately 4.14, indicating that, on average, they have a higher number of admissions, compared to those without high blood pressure. The median value of 3 suggests that the middle value of hospital admissions is lower than the mean, indicating a right-skewed distribution. The standard deviation of 3.45 indicates a relatively large variability in hospital admissions for individuals with high blood pressure. The minimum value of 1 suggests that some individuals with high blood pressure had very few admissions, while the maximum value of 70 represents the highest number of admissions observed.

On the other hand, for individuals without high blood pressure, the mean number of hospital admissions is approximately 2.84, which is lower than the mean for individuals with high blood pressure. The median value of 2 suggests that the middle value of hospital admissions is lower than the mean, indicating a right-skewed distribution. The standard deviation of 2.60 indicates a relatively lower variability in hospital admissions for individuals without high blood pressure compared to those with high blood pressure. The minimum value of 1 suggests that even individuals without high blood pressure had some hospital admissions, while the maximum value of 34 represents the highest number of admissions observed in this group.

It is important to note that in this analysis, we have excluded missing values (NA) from the calculations. Therefore, the results are based on the available valid data points. However, when analysing RWD in HTA, we need to carefully consider how to handle missing data and choose appropriate imputation methods, depending on the type and extent of missing information (see Section 5.2. Addressing data availability and quality).

Two numerical variables

- **Summary statistics** provide valuable insights into the individual distributions of the two variables. By calculating measures such as

means, medians, ranges, and standard deviations for each variable, we can understand their central tendency, spread, and variability. Comparing these statistics between the two variables (if they are measured in the same scale) allows us to observe their relationship and identify potential patterns or differences (Box 12, Table 12A and Table 12B).

- **Correlation coefficient:** To further explore the strength and direction of the relationship between the two variables, we can utilise the linear correlation coefficient. This coefficient ranges from -1 to +1. A positive value indicates a positive linear relationship, suggesting that as one variable increases, the other tends to increase as well. Conversely, a negative value suggests a negative linear relationship, indicating that as one variable increases, the other tends to decrease. Values closer to -1 or +1 represent stronger associations, while values closer to 0 indicate a weaker or no linear relationship.
- **Covariance** also assesses the direction and strength of the linear relationship between two numerical variables. It can take both positive and negative values, indicating the direction of the relationship. However, the magnitude of the covariance is challenging to interpret directly as it depends on the scales of the variables involved. Therefore, it is often more informative to focus on the correlation coefficient, which standardises the covariance to provide a clearer measure of association between the variables.

Box 12. Summarising two numerical variables

For this example of analysis of covariance of two numerical variables, we will select the variable “Age” and “Hospital admissions” analysed before. This analysis will give us an idea of how these two variables move together.

Table 12A. Summary for “Age” and “Hospital Admissions”

	MEAN	MEDIAN	SD	MIN.	MAX.
Age	82.25	85	12.08	2	111
Hospital admissions	3.89	3	3.35	1	70

Analysing the age variable, we find that the mean age of the patients is approximately 82 years. The median age, 85 years, indicates that half of the patients are 85 years or older, highlighting the prevalence of older individuals in the cohort.

The standard deviation of 12.08 signifies that there is some variation in the ages of the patients within the cohort. This variability of approximately 12 years suggests that the ages are not tightly clustered around the mean, but rather spread out to some extent. This could

be attributed to factors such as different disease progression rates or varying responses to treatment among individuals.

Examining the minimum and maximum ages, we observe that the youngest patient in the cohort is 2 years old; this would be consistent with the target population of the device. Given that children may be the primary recipients of ICDs or treatment for secondary prevention of sudden death.

We previously summarised the “Hospital admissions” variable, observing a wide variation in the number of hospital admissions among individuals. It would be interesting to assess if this is somehow related with the age of individuals.

Table 12B. Correlation coefficient and covariance for “Age” and “Hospital Admissions”

COVARIATION COEFFICIENT	COVARIANCE
0.003	0.138

The correlation coefficient between “Age” and “Hospital admissions” is 0.003. This value suggests a very weak correlation between the two variables. It indicates that there is almost no tendency for higher age to be associated with slightly higher hospital admissions. This implies that age alone does not serve as a reliable predictor of hospital admissions in this cohort of heart failure patients.

The covariance between age and hospital admissions is 0.138. This value indicates a positive covariance, suggesting that as age increases, there tends to be a corresponding increase in hospital admissions. However, the magnitude of the covariance does not provide a clear indication of the strength of the relationship between the variables. Covariance can be influenced by the units and scales of the variables, making it difficult to interpret directly.

Visualisation

Two categorical variables³⁵

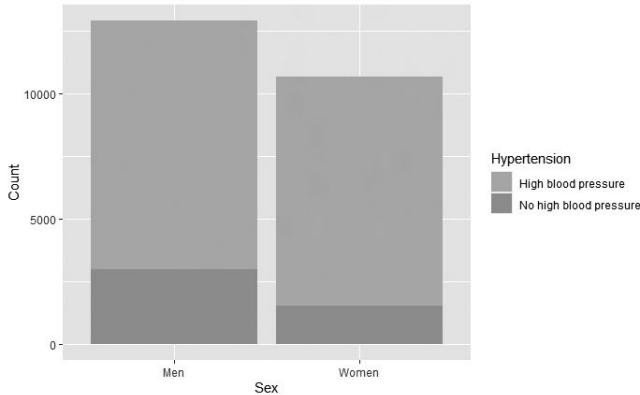
- **Bar plots** are effective for visualising the distribution of a categorical variable. When examining the covariation between two categorical variables, a grouped bar plot can be created. Each bar represents a category of one variable, and its height corresponds to the frequency or proportion of observations in that category. The bars for different categories of the second variable are grouped together, enabling comparison of the distribution of the first variable across the categories of the second variable. This facilitates the identification of any associations or patterns between the two variables.
- **Mosaic plots** are graphical representations that depict the joint distribution of two categorical variables. They display the proportion of observations in each combination of categories as rectangles within a plot. The width of each rectangle represents the proportion of observations for a specific category of the first

variable, while the height represents the proportion of observations for a specific category of the second variable. Mosaic plots are particularly valuable for visualising the associations and dependencies between two categorical variables, aiding in the identification of relationships and patterns.

Box 13. Visualising two categorical variables

In these steps, we are going to visualise a bar plot for “Sex” and “Hypertension” variables; this would help us identify patterns, trends, and differences across different groups or levels of the “Sex” variable.

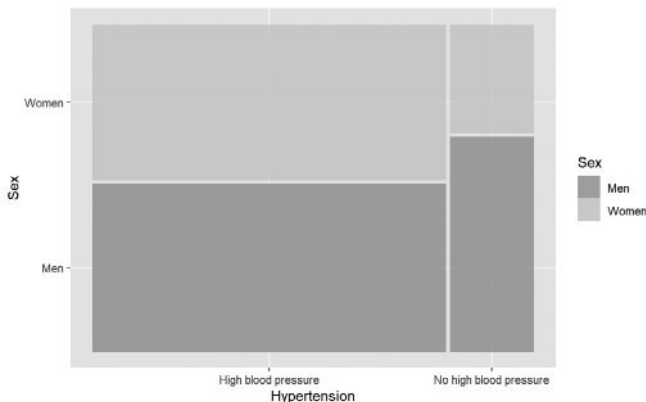
Figure 13A. Grouped bar plot for “Sex” and “Hypertension”



We can observe that there are more individuals with “High blood pressure” compared to “No high blood pressure” in both the Men and Women categories.

Now let’s visualise the association between these categorical variables using a mosaic plot to provide insights into the distribution and relationship between different categories.

Figure 13B. Mosaic plot for “Sex” and “Hypertension”



The plot is divided into two sections, one for each level of the “Sex” variable. The grey rectangles within each section represent the proportion of individuals with a specific combination of “Sex” and “Hypertension”. The darker shade of grey represents men, and the lighter shade represents women. We can observe that the majority of individuals in both sex categories have “High blood pressure.” However, among those with “No high blood pressure”, men seem to have a slightly higher proportion compared to women.

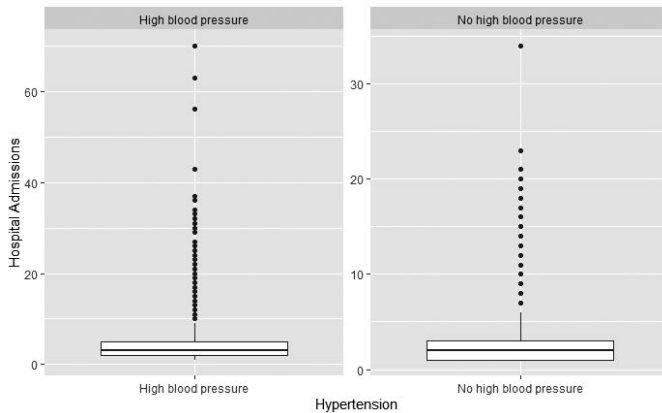
Numerical vs categorical variables

- **Box plots with facets** for each category of the variable. By incorporating facets, we are able to create a grid of box plots, where each facet represents a category of the categorical variable. This arrangement allows us to easily compare the distribution of the numerical variable across different categories.

Box 14. Visualising numerical vs. categorical variables

Creating this box plot with facets allows us to compare the distribution of hospital admissions between individuals with high blood pressure and those without, providing a visual representation of the differences in medians, quartiles, and potential outliers between the two groups.

Figure 14A. Box plot with facets for “Hospital admissions” and “Hypertension”



Facets³⁴ are the technique of creating multiple small panels or subplots within a larger plot, each representing a subset or category of the data. By visually separating the data based on a categorical variable, facets provide a clear and concise way to observe patterns, trends, and variations within each category. They could help us to enhance the interpretability and depth of analysis by providing a more detailed view of our data. Note that a useful category to compare might be the “missing values” category.



Two numerical variables

- **Scatter plots** are a powerful tool for visualising the relationship between two numerical variables. Each data point represents an observation and is plotted according to its values on the x-axis

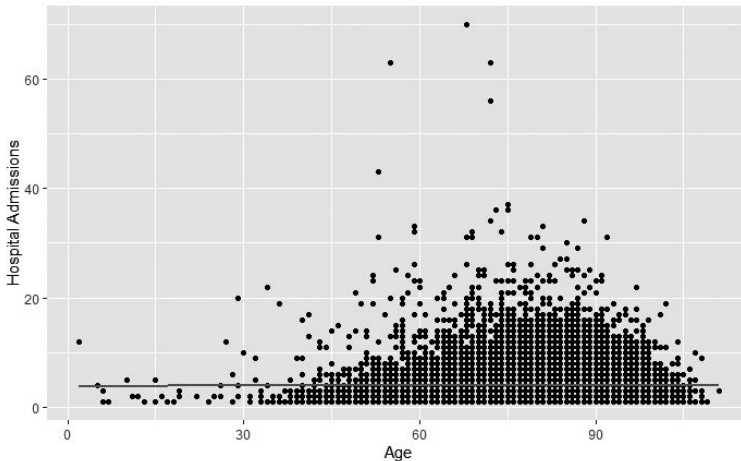
and y-axis. By examining the pattern of the data points on the scatter plot, we gain insights into the nature of the relationship between the variables.

- **Trend lines** are a complement of scatter plots to visualise the overall trend or relationship between the two numerical variables. Trend lines can be helpful in identifying the direction and strength of the relationship.

Box 15. Visualising two numerical variables

By creating this scatter plot, we can visualise how age can be associated to the number of hospitalisations, which we have previously calculated with the correlation coefficient and covariance values.

Figure 15A. Scatter plot and trend line for “Hospital admissions” vs “Age”



In this scatter plot, each point represents an individual with their corresponding age and hospital admissions. The trend line fits a linear regression line.

We can see that the points are scattered without a clear linear pattern. The regression line also appears almost horizontal, further suggesting the lack of a strong relationship. The covariance value of 0.14 calculated before confirms that there is some positive association between the variables, but it doesn't provide a clear indication of the strength or direction of the relationship.



Using R, specifically R Markdown (Rmd) files, along with the tidyverse and dplyr packages, can be good practice for developing an EDA. See Annex V for more details on the advantages of using this format and download the notebook that contains the EDA examples we have shown.

5.2. Addressing data availability and quality

Data quality is generally defined as the fitness for purpose of the data for users' needs in relation to health research, policymaking, and regulation⁴⁹. It also refers to the extent to which the data reflect the reality they aim to represent. Addressing data quality issues is crucial to ensure the reliability and validity of the findings from EDA and subsequent modelling approaches. From a regulatory perspective, five dimensions may be distinguished⁵⁰:

Table 9. Data quality dimensions according to the EMA Data Quality Framework⁵⁰

DIMENSION	QUESTION ADDRESSED	SUB-DIMENSIONS
Reliability	To what degree are data accurate or correctly representing an observed reality?	Accuracy
		Precision
Extensiveness	How much data do we have?	Completeness
		Coverage
Coherence	Is the data analysable as a whole or are additional steps needed like linkage of multiple datasets?	Format coherence
		Structural or relational coherence
		Semantic coherence
		Uniqueness
		Conformance
Timeliness	Are the data reflecting the intended reality at the point of time of its use?	Currency
Relevance	Does the dataset present the values that are needed to address a specific question, using a specific method?	N/A

The reliability of the data mostly depends on the systems and process for primary collection and curation of data. The adoption of a set of validation rules that guarantee the plausibility/accuracy of the relevant variables in the data model specification may prevent possible errors or imprecision. In the EDA phase, we can assess data reliability through some validation and verification actions:

Table 10. How to address data reliability

IMPORTANCE	DEALING WITH IT
It is important to ensure that the data used for the analysis are reliable and accurately represent the relevant phenomena; otherwise, it can lead to biased conclusions and compromise the validity of the findings.	Conducting data validation checks and data cleaning procedures is important to identify and correct inconsistencies. This may involve: <ul style="list-style-type: none">● Cross-referencing data from different sources or data elements.● Describing missing data, outliers, performing data profiling techniques (look for distributions, patterns) and visualisations

When dealing with extensiveness, it is important to assess the level of completeness of the data through metrics that quantify the number of missing values. As missing data can lead to bias and cause problems during analysis in some statistical packages, we can deal with this with a series of techniques (Table 11):

Table 11. How to deal with missing or incomplete data

IMPORTANCE	DEALING WITH IT
Missing data can introduce bias and affect the representativeness of our cohort. It can also impact the estimation of treatment effects, as the missingness may be related to certain patient characteristics or outcomes of interest.	Various techniques can be employed, such as imputation methods to estimate missing values. These methods are: <ul style="list-style-type: none">● Mean/median imputation: Replaces missing values with the mean or median of the observed values for that variable.● Regression imputation: Uses regression models to predict missing values based on other variables in the dataset.● Multiple imputation: Generates multiple imputed datasets using advanced statistical techniques and combines the results to obtain more robust estimates.

Overall, to deal with data quality issues effectively when performing the EDA we can use different strategies common in HTA, which will be adapted to the RWD particular considerations:

- Collaboration with experts with different backgrounds: We should look for input from clinicians, researchers and also with data experts to understand the context, potential data limitations and biases inherent in the dataset.

- **Data cleaning and processing documentation:** It is highly recommended to proceed with a transparent record of the data cleaning procedures undertaken, including any imputation methods used or decisions/assumptions made.
- **Uncertainty analysis:** Performing sensitivity analyses would help us to assess the robustness of the findings by testing different assumptions, imputation methods or excluding certain data subsets to understand the potential impact on the results.
- **Findings validation:** If possible, we should compare the findings from the EDA with data from other sources, such as clinical trials used to document the intervention arm, to assess consistency.

6. Modelling

Finally, the decision model serves as a crucial artefact that consolidates and integrates the information we have regarding patients, resources, and outcomes. Initially, this information may exist in separate forms and sources, making it challenging to comprehend the overall picture and draw meaningful insights. The model acts as a unifying framework allowing us to simulate and analyse the complex interactions between our research questions.

6.1. Building and executing the model

In previous sections, we have explained the keys to defining our decision model. Now it is time to get down to work to implement it and feed it with data that we have explored in previous phases of the evaluation process.

6.1.1. Calculate patient-specific transition probabilities

In traditional models, transition probabilities are often derived from aggregated data or published literature, representing the average behaviour of a cohort. However, when using RWD, we have the opportunity to go beyond average estimates and incorporate individual patient characteristics to estimate personalised transition probabilities³⁶.

Patient-specific transition probabilities capture the unique dynamics of each patient's disease progression or treatment pathway. By considering factors such as demographics, clinical characteristics, treatment history, and comorbidities, we can gain a more nuanced understanding of how these variables influence the likelihood of transitioning between different states or health outcomes. Two steps are key to this:

- 1. Data analysis: We should explore the RWD to identify factors influencing state transitions. It is useful to employ statistical techniques, such as logistic regression or survival analysis⁴⁷, to model the relationship between predictors and transition probabilities.

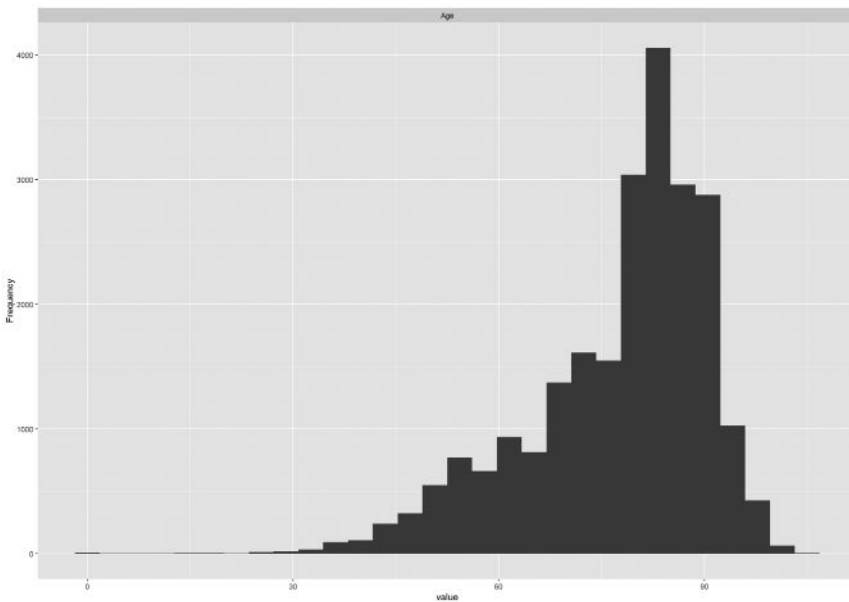
Box 16. Baseline patient characteristics

In our use case, patient characteristics were based on a hypothetical cohort of patients aged 60 years, with chronic heart failure, NYHA class II or III and Ejection Fraction LVEF < 35% with or without a history of ischaemic heart disease. These characteristics were selected based on a patient registry and the literature reviewed.

However, our extracted patient cohort represents how real patients are in the healthcare system. In contrast to the hypothetical cohort with a mean age of 60 years, our patients have an age of 77.44 years (SD =13.22) at the time of cohort entry (Figure 14A).

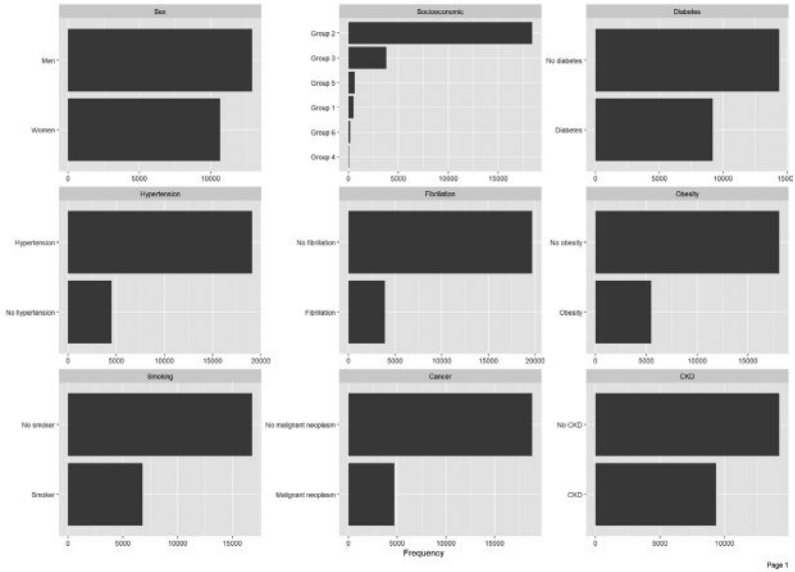
Similarly, the hypothesis maintained in the evaluation report is that patients have an LVEF < 35%. We do not have that data; however, we can describe some cardiac disease characteristics of our patients: 80% have hypertension, 16% suffer from permanent atrial fibrillation, 2% have had ventricular arrhythmia, 0.03% have had a sudden cardiac arrest averted event (Figure 14B).

Figure 14A. Histogram for Age



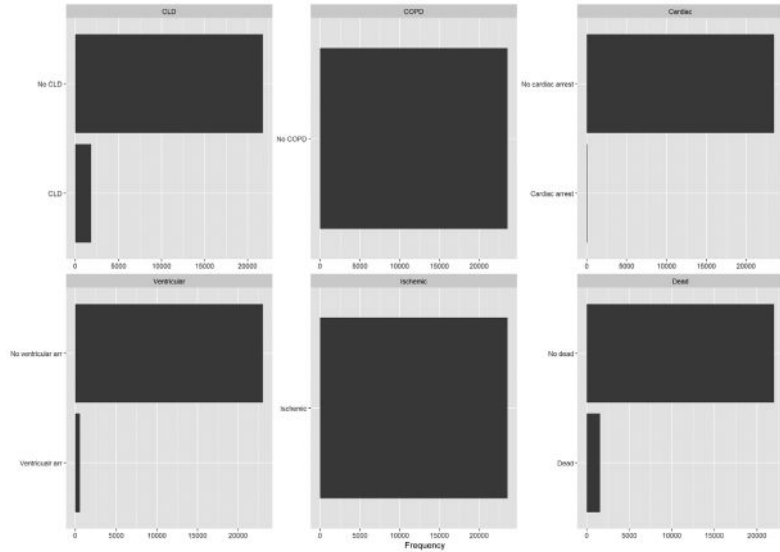
Additionally, the use of RWD gives us much more power over the knowledge about our cohort. For example, we will be able to know that 23% of our patients had obesity at cohort entry, 27% were smokers, 20% had a diagnosed malignancy at cohort entry, 40% had CKD, 7% had CLD.

Figure 14B. Bar plots for categorical variables



Page 1

Figure 14B. Bar plots for categorical variables (cont.)



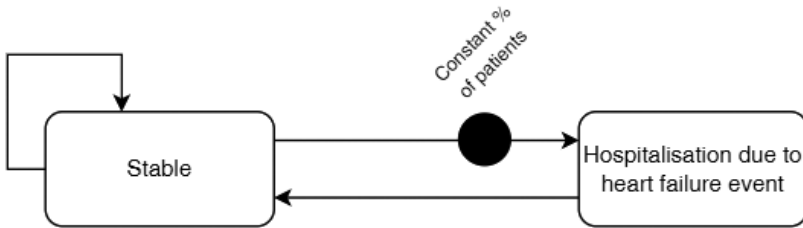
Page 2

More information on baseline characteristics is available [in this report](#)

- 2. Patient-specific transition probabilities calculation: We can use the modelled relationships and patient characteristics to estimate personalised transition probabilities for each patient.

Box 17. Transforming a Markov model into an individual patient simulation

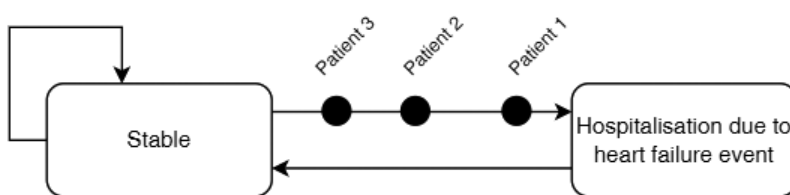
The modelling technique used in the example is a Markov model. Its structure and simulation follow the traditional approach. Transition probabilities are derived from published literature, representing the average behaviour of a cohort. This approach assumes homogeneous transition probabilities across individuals and does not capture individual-level variation.



Initial state definition: In the Markov model, we have a state representing “Stable” patients and a state representing those admitted to the hospital due to a cardiac event. In the simulation approach, we can derive the initial state of each patient based on their individual data. In our use case, patients may start in a “Stable” state, generally those entering the cohort with a chronic heart disease diagnosis, but may also start in a “Hospitalisation” state, especially those with conditions usually diagnosed in an emergency setting (e.g. cardiac arrest, myocardial infarction, ventricular arrhythmia).

Model transitions: In the Markov model, transitions between states are defined using constant transition probabilities. In the simulation approach, we would model transitions based on the patient’s individual characteristics. For example, we could use the rate of hospital admissions to determine the probability of transitioning from the “Stable” state to the “Hospitalisation” state or vice versa. We can use statistical techniques such as logistic regression or machine learning models, as explained before, to estimate these transition probabilities based on the individual patient data.

Simulate trajectories: Using the estimated transition probabilities, we can simulate every patient trajectory over time. We start with the initial state of each patient and simulate their transitions between the “Stable” and “Hospitalisation” states based on the calculated probabilities. Then we can aggregate all patients to have a full picture of the simulated trajectories for the cohort, as in the diagram below.



6.1.2. Incorporate patient-specific resource use

Traditional average cohort models often assume uniform resource utilisation across the entire cohort, neglecting the heterogeneity in individual patient characteristics and treatment pathways. Incorporating patient-specific resource use allows us to simulate the utilisation of healthcare resources for individual patients and evaluate the associated costs and outcomes. Some statistical methods can help us in integrating and accounting for this patient-specific variability (Table 12).

This personalised approach enables a more comprehensive assessment of the economic implications of different interventions, considering the variation in resource utilisation across patients. It allows us to explore the impact of individual patient characteristics on resource use, identify high-cost subgroups, and assess the cost-effectiveness of specific treatment strategies.

- 1. Relevant resource identification: It is important to determine the resources associated with each state or transition, such as healthcare visits, surgical procedures, medication usage, or diagnostic tests.
- 2. Assigning patient-specific resource utilisation: We can merge the resource utilisation data explored in the previous steps to each patient and their corresponding states or transitions to reflect individual variations in resource use.

Table 12. Statistical techniques for populating models with RWD

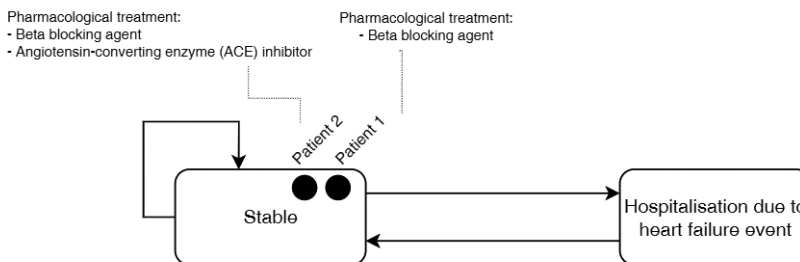
Regression models³⁷	Regression models can be used to estimate the probabilities of transitioning between different states, outcomes or other binary variables (logistic regression), and the association between patient characteristics and resource utilisation (generalised linear models, Poisson regression). In this way, they can identify significant predictors and quantify their impact, allowing us to calculate patient-specific parameters.
Machine learning algorithms	We can also employ machine-learning algorithms to predict patient-specific resource use based on a wide range of patient characteristics, treatment factors, and other relevant variables. These algorithms can capture complex relationships and interactions, enabling accurate estimation of individual patient resource use patterns.
Matching methods	We can match patients with similar characteristics and treatment profiles by a series of statistical techniques aiming to reduce bias due to confounding. Propensity score matching (PSM) is one of the most commonly used matching techniques to control for confounding factors and estimate the causal effect of specific interventions on resource use.

Time-to-event or survival analysis ^{38,39}	Using survival analysis or time-to-event analysis methods, we can model the hazard risk of transitioning between states over time and know more about the timing and likelihood of resource utilisation events, such as hospital admissions or procedures. Additionally, some of these techniques, such as the Cox proportional hazards model, account for varying follow-up duration and censoring.
Bayesian methods ⁴⁰	Bayesian techniques provide a valuable approach for modelling with RWD, allowing for the integration of prior knowledge, complex modelling, and incorporation of uncertainty in a coherent framework.
Further explanations and indications on how to implement these techniques will be covered in detail in the toolkit.	

Box 18. Simulating resource utilisation

Parameters related to resource use and costs in the use case were derived from the literature, analytical accounting from a single hospital and the opinion of professional experts.

Resource utilisation estimation: In our simulation approach, we should analyse RWD to estimate resource utilisation for each patient in the different states of our model. We can calculate the frequency and duration of resource use based on the indicators defined. We can calculate the rate of primary care visits in a period of time, the length of hospital stays, and the number of cardiology visits for each patient and assign them to different states or transitions. For example, Patient 1 may have a different use of resources in the “Stable” state compared to Patient 2 due to pharmacological treatment.



Simulation of resource use and costs: We can use the patient-specific transition probabilities derived before and resource use estimates to simulate resource use and costs for each patient over the desired time horizon.

6.1.3. Estimating patient-specific outcomes

Estimating patient-specific outcomes involves capturing relevant health outcomes, quality of life measures, or other patient-centred endpoints that reflect the impact of interventions on individual patients. Two steps are key to this:

- 1. Outcome data: We should assign patient-level outcome data from RWD extracted to the defined model variables and states.

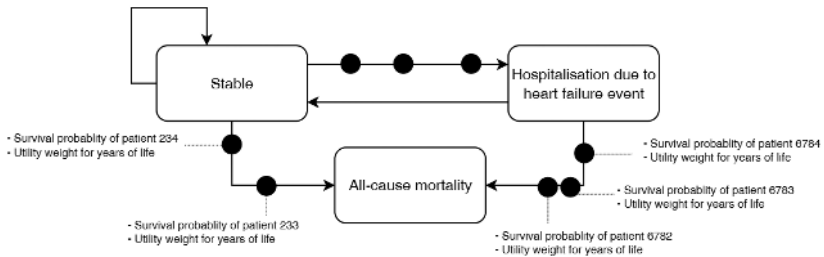
In this way, we will be able to identify the results of each patient and analyse possible dependencies with their baseline characteristics, treatments received or events they have had.

- 2. Assign patient-specific outcomes: Associate the outcome data with their corresponding states or transitions to capture individual variations in outcomes.

Box 19. Simulating patient-specific outcomes

In our example, the main outcome is mortality events, which helps to estimate the efficiency outcome, in this case the ICER, measured in € per Quality Adjusted Life Year (QALY) gained. The period that patients remain alive through the model is determined by constant transition probabilities between the different health states (“Stable” and “Hospitalised”) towards death.

Estimating individual survival probabilities: We can use RWD extracted regarding mortality to estimate patient-specific survival probabilities based on their health state and characteristics and other characteristics.



It is important to consider disease-specific parameters or utility weights to capture the impact of health states on patient outcomes. In this case, we have used mortality as our main outcome, in order to calculate life years gained and use utility weights drawn from the literature. In a different situation, we may use other outcomes relevant to patients.

6.1.4. Additional parameters in our decision model

In addition to the parameters mentioned focused on transition probabilities, resource use (and costs) or outcomes, there are several general parameters that need to be considered when developing any decision model. These parameters are associated with the economic and decision-making context in which we want to report. They are generally assigned by convention.

Discount rates are used to adjust future costs and outcomes to their present value, taking into account the time value of money and the time preference of individuals. The choice of the discount rate depends on various factors, such as the time horizon of the analysis, the opportunity cost of capital, and the country or region where the evaluation is being

conducted. According to the guidelines for developing economic evaluation analyses in the field of Health Technology Assessment in Spain, 3% is usually used, with this rate varying between 0% and 5% in the sensitivity analysis⁴¹.

Cost-effectiveness thresholds represent the maximum amount that a society is willing to pay for a unit of health gain. These thresholds can be expressed in different ways, such as cost per quality-adjusted life year (QALY) gained. The choice of the threshold depends on various factors, such as the disease or condition being evaluated, the level of health spending in the country or region, and societal values and preferences. Efforts have been made by RedETS to estimate this threshold, reaching the conclusion that the willingness to pay in Spain for an additional QALY could be between 22,000 and 25,000€⁴².

Variability in the deterministic sensitivity analysis involves testing the robustness of the model results to changes in the input parameters. This involves varying the values of one or more parameters at a time to see how the results change. This helps to identify which parameters have the greatest impact on the results and can inform future data collection efforts or guide decision-making. We need to indicate which variance thresholds will be used from the beginning of our model approach, typically $\pm 20\%$ of the point value entered in the model is used (see 5.1.1. Analysing variation section).

6.1.5 Model simulation

Model simulation focuses on the essential steps for accurately simulating the dynamics of the model.

- 1. Developing the simulation algorithm: This step involves the development of a computational algorithm for putting together all the dimensions and relationships that we have built up, individual transition probabilities, resource utilisation, and outcomes to simulate transitions over time.
- 2. Trajectories simulation: We should apply the simulation algorithm to each patient in the cohort, considering their initial state, individual characteristics, and the defined transition probabilities.
- 3. Aggregation and results: Finally, we should analyse the simulated patient trajectories to generate aggregate results, such as mean outcomes, average resource utilisation, or cost-effectiveness measures.

6.2. Assumptions

When working with RWD for modelling, the challenges of incomplete information or time-consuming data processing can be particularly pronounced. In these situations, assumptions become even more crucial for populating the model and enabling the analysis of healthcare interventions or policies^{43,44}.

RWD often comes from diverse and complex sources, such as electronic health records. While RWD provides valuable insights into real-world patient outcomes and healthcare utilisation, it may lack complete information on certain parameters necessary for modelling. This can include missing values, incomplete records, or variables that were not originally captured in the data sources.

Assumptions play a vital role in addressing these data limitations and enabling the construction of a comprehensive model. They act as a bridge to fill the gaps in the data by estimating or imputing missing values based on reasonable assumptions. For example, if data on a specific health outcome is missing for a subset of patients, an assumption can be made regarding the similarity of those patients to others with similar characteristics in the dataset, allowing for the imputation of missing values.

There are several reasons why assumptions are essential in populating a model:

- **Incomplete data availability:** Data collection for every parameter within a model can be challenging and resource-intensive. In many cases, certain data may be unavailable or unstructured, making it difficult to accurately estimate specific values. Assumptions enable us to extrapolate and estimate missing or unavailable data, providing a more comprehensive picture of the phenomenon under evaluation.
- **Future projections:** Modelling often involves forecasting of projecting outcomes beyond the timeframe covered by available data. Assumptions are necessary to make reasonable estimates of future trends and events, allowing decision-makers to assess the potential impact of interventions in the long term.
- **Simplification of complex systems:** Factors influencing the use of a technology, interventions and the whole healthcare system are multifaceted and interconnected. As we explained before, models help simplify these complex systems to facilitate analysis and decision-making. Assumptions aid in defining relationships and

interactions between variables, enabling the model to capture key aspects of the dynamics, behaviours or conditions.

- **Resources and time constraints:** Conducting extensive data collection for all model parameters may be time-consuming. Assumptions provide a more efficient and practical approach to populate the model, allowing us to focus resources on critical aspects of the technologies and conditions under assessment while still achieving a reasonable level of accuracy and validity.
- **Exploration of what-if scenarios:** Models are valuable tools for exploring different scenarios and assessing the potential consequences of alternative decisions or interventions. Assumptions allow for the creation of hypothetical scenarios by varying input values and parameters, enabling us to provide decision-makers with the information to understand the potential outcomes and trade-offs associated with different situations.

However, it is important to acknowledge the potential limitations and uncertainties introduced by assumptions. Assumptions should be carefully considered, justifiable, and transparently documented. Sensitivity analyses should be conducted to assess the impact of different assumptions on the model's results and conclusions. This allows decision-makers to understand the range of potential outcomes based on different sets of assumptions and make informed choices.

Furthermore, efforts should be made to improve data quality and completeness over time. By continuously working towards better data capture and reducing data gaps, the reliance on assumptions can be gradually minimised, leading to more accurate and reliable decision models.

6.3. Model validation

Model validation helps us ensure the model's accuracy and reliability. This process encompasses more than just programming errors. It also involves assessing whether the decision problem is well defined and whether the model adequately captures the intended decision-making process and the underlying reality of the patient population, as well as producing valid results^{36,37}.

- **Validation objectives:** It is important to determine what aspects of the model we want to assess, such as the overall model structure, parameter estimation, or the ability to reproduce observed outcomes.

- **Compare model outputs with other data:** Generally, it is advisable to replicate RWD analyses in more than one data source and to check the robustness of the model/s.
- **Documentation and transparency:** Document the data sources used, methods employed, assumptions made and findings obtained from different iterations. Transparently report the strengths, limitations, and uncertainties associated with the model. It is advisable to share all analytical code used to analyse RWD and to produce models, as to allow reproducibility in other settings and RWD data assets.



Involving domain experts, clinicians, and other stakeholders in the validation process would be a good practice. Their input and feedback on the model assumptions, structure, and outputs can provide valuable insights and enhance the credibility of the model.

6.4. Obtaining results

This phase involves analysing the output of the model to obtain meaningful results that can inform decision-making. By examining the outcomes, costs, and other features of different interventions, this phase aims to provide valuable insights into the potential impact and value of the technologies being evaluated. The results obtained from the model serve as a basis for assessing the comparative effectiveness, cost-effectiveness, and potential benefits of different interventions, enabling stakeholders to make informed decisions regarding the adoption and allocation of healthcare resources.

Through careful analysis, interpretation, and presentation of the results, this phase plays a vital role in translating complex modelling outcomes into actionable information for decision-making. This includes:

- **Group analysis:** We can utilise the available RWD to perform subgroup analyses based on relevant factors, such as cohort characteristics and treatment strategies. As in a clinical trial, subgroup analyses should be pre-specified in the assessment protocol to avoid overinterpretation and generation of false positive results.
- **Calculation of incremental results:** Calculate the incremental outcomes and costs by comparing the potential benefits of the technology under evaluation (obtained from a systematic review of the literature and used to populate the intervention arm) with the outcomes observed in the comparator arm derived from

RWD. In our specific example, to quantify the additional benefits and costs associated with adopting ICD.

- Graphical presentation: We should present the results graphically using appropriate charts and graphs. Visualise the incremental outcomes, such as improvements in survival rates or reduction in adverse events, alongside the comparator arm outcomes observed in the RWD.



Use the cost-effectiveness plane to display the incremental cost and incremental effectiveness (e.g., quality-adjusted life years gained) for each intervention compared to the next best alternative. The points on the graph represent different strategies for the main subgroups in our cohort, and their position indicates the relative cost and effectiveness. This plot helps visualise and identify dominant or dominated interventions.

- Calculation of the cost-effectiveness threshold: consider the incremental cost-effectiveness ratio (ICER) calculation based on the additional costs and outcomes derived from the model. Compare the ICER with the predetermined cost-effectiveness threshold to assess the value for money of implementing the intervention.



In this step, the cost-effectiveness acceptability curve illustrates the probability that an intervention is cost-effective compared to a specific threshold or willingness-to-pay value. It plots the probability of cost-effectiveness against different threshold values, demonstrating the uncertainty surrounding the cost-effectiveness estimates. The curve shows the likelihood that each intervention is cost-effective at various thresholds.

- Interpretation of results: Interpret the results in light of the RWD limitations, potential biases, and assumptions made in the model. Recognise that the lack of RWD introduces uncertainty in estimating outcomes and costs.

6.5. Dealing with uncertainty

Sensitivity analysis is used to explore the impact of uncertain inputs or assumptions on our model results²³. This process helps us understand the robustness and reliability of our findings. We can perform two main types of sensitivity analysis: deterministic sensitivity analysis (DSA) and probabilistic sensitivity analysis (PSA).

When conducting DSA, our goal is to examine the influence of individual parameters on the model outcomes. By varying one input or assumption at a time while keeping others constant, we can assess the impact of each parameter on the results. Steps needed for this are:

- **Identifying key parameters:** We need to identify the parameters that are likely to have a substantial influence on our model results. These parameters could encompass costs, clinical outcomes, probabilities, or utility values.
- **Definition of plausible ranges:** It is important to establish plausible ranges or values for each identified parameter. These ranges can be informed by the available literature, expert opinion, or previous studies.



When working with RWD, leverage the previous EDA to derive more accurate ranges and distributions for your parameters.

- **Parameter variation:** We systematically change the values of each parameter within its defined range and run the model. This allows us to observe how variations in parameter values impact the outcomes of interest.
- **Interpretation of results:** We should analyse the effects of parameter changes on the output of our decision model. To facilitate this analysis, we can employ visual representations.



Use tornado diagrams⁴⁸ to display the magnitude and direction of the effects of parameter variations on your outcomes.

6.5.1 Making our decision model probabilistic

By incorporating RWD, we can enhance the estimation of parameter values and their associated distributions. This can make it much easier for our models to be stochastic or probabilistic.

A stochastic model or a model with Probabilistic Sensitivity Analysis (PSA) enables capturing the joint uncertainty of multiple parameters simultaneously. By assigning probability distributions to each parameter and utilising techniques like Monte Carlo simulation, we can quantify the overall uncertainty of our model results.

Parameter distributions definition

We should assign probability distributions to the parameters of interest. RWD can play a valuable role in deriving realistic distributions and estimating distribution parameters (e.g., mean, standard deviation) based on observed data.

Box 20. Key distributions commonly used in stochastic modelling

Beta distribution: is a flexible distribution that can represent data bounded between 0 and 1. It is useful for modelling probabilities of events or proportions of patients experiencing certain outcomes.

Gamma distribution: is often used to model positively skewed data, such as healthcare costs or resource utilisation. It is also suitable for modelling time intervals, such as time to event data.

Normal distribution: commonly employed for modelling continuous data that follows a symmetric bell-shaped curve. It is appropriate for variables such as costs, utility values, or continuous clinical outcomes.

Log-normal distribution: is useful when dealing with continuous data that is positively skewed and does not have negative values. It is often applied to variables such as costs or utility values, which tend to have a skewed distribution with a long tail on the positive side.

Uniform distribution: represents a range of equally likely values. It is used when there is limited information or uncertainty about the parameter, and all values within a specified range are considered equally plausible.

DISTRIBUTIONS	PARAMETERS	TYPE OF DATA
Beta distribution	Shape parameters (α , β): estimated by fitting the distribution to proportions or probabilities observed in the data.	Probabilities, proportions, or rates
Gamma distribution	Shape parameters (α , β): we can estimate α and β by equating our data mean and variance to the theoretical mean and variance of the gamma distribution.	Costs, resource utilisation, or time intervals
Normal distribution	Mean (μ): estimated as the average of the observed data for our variable. Standard deviation (σ): sample standard deviation	Costs, utilities, or continuous clinical outcomes
Log-normal distribution	Logarithmic mean (μ): this can be estimated as the average of the logarithmically transformed data. Logarithmic standard deviation (σ): derived using the standard deviation of the logarithmic transformed data	Costs, utilities, or continuous clinical outcomes with positive skewness

DISTRIBUTIONS	PARAMETERS	TYPE OF DATA
Uniform distribution	Minimum value (a): can be derived as the smallest observed value in the data. Maximum value (b): can be derived as the largest observed value in the data.	Parameters with limited information or uncertainty

These distributions are just examples, and the choice of distribution should be based on the specific characteristics of your data and expert judgement. Some nonparametric statistical tests can aid in this choice by evaluating the goodness-of-fit of the data to a theoretical distribution, e.g. Kolmogorov-Smirnov or Shapiro-Wilk tests of normality.

Parameter uncertainty propagation

By propagating parameter uncertainty through Monte Carlo simulation, decision-makers can gain insights into the range of possible outcomes and their associated probabilities. This approach allows for a comprehensive assessment of decision uncertainty and supports robust decision-making by considering a broad spectrum of potential scenarios.

Monte Carlo simulation is a powerful technique used to incorporate parameter uncertainty into decision models. It involves sampling values from the previously defined probability distributions for uncertain parameters and repeatedly running the model to generate a distribution of model outcomes. To analyse the distribution of the results we can summarise key statistics (means, medians, or percentiles) and generate graphical representations (as histograms or probability density plots).

Probabilistic sensitivity analysis to decision making

Showing uncertainty from PSA to decision makers is essential for conveying the range of possible outcomes and the associated probabilities. Decision makers can better understand the risk and uncertainty inherent in the decision context. Key approaches for communicating uncertainty include:

- **Probability Distributions:** Presenting the results as probability distributions allow decision makers to visualise the full range of possible outcomes and their likelihoods. Histograms, density plots, or cumulative distribution functions can be used to represent the uncertainty in a concise and informative manner.
- **Cost-Effectiveness Acceptability Curves:** Illustrate the probability of an intervention being cost-effective across a range of willingness-to-pay thresholds (in our case, it would range from

22.000€/QALY to 25.000 €/QALY). They plot the proportion of iterations from the probabilistic analysis in which the intervention is cost-effective at each threshold. This representation provides decision-makers with insights into the trade-offs between costs and outcomes and aids in determining the cost-effectiveness of a health technology at different threshold values⁴⁵.

- **Value of Information (VOI) analysis:** VOI is a framework for evaluating the potential value of additional data to reduce decision uncertainty. It quantifies the expected benefits and costs associated with gathering more information. We can estimate the value of perfect or partial information²³.
- **Rank probabilities:** Represent the relative likelihood or order of preference among different interventions. They can provide decision-makers with an understanding of the relative uncertainty associated with different options. They can be presented as a matrix of ranking probabilities or ordering of interventions based on their likelihood of being the best or most effective option. This approach helps decision-makers identify the most promising alternatives and prioritise further analysis⁴⁶.

7. Limitations

Currently, there are some limitations to consider when working with RWD during preadoption HTA. There are important challenges pertaining to the quality of the data itself, such as the data completeness or relevance. The following are some of the key ones:

- **Loss of data during quality control:** Clinical information stored in centralised repositories for secondary use suffers from quality problems that, despite being minimised through quality improvement procedures during data capture, should still be taken into account. For example, missing or out-of-range data can be removed during the loading processes. While this reduces noise, it also inevitably reduces the amount of available information. Concerns about the coverage of the data may be partially addressed through cross validation cross-validation with other sources of data or sensitivity analysis (see Section 5.2. Addressing data availability and quality).
- **Lack of sufficient and up-to-date data:** Health information systems rarely include detailed information with sufficient coverage of variables such as education level or job title. Even when these data are included, their coverage is usually limited, and their up-to-dateness is not guaranteed. Some clinical measures and results may not be readily retrievable, as they may not be registered in a structured or integrated source. Even if proxies exist, such as computerised definitions for sudden cardiac death, these variables have limited sensitivity and specificity, introducing uncertainty that may be challenging to address.
- **Underreporting of adverse events:** Adverse event reporting systems often face issues of underreporting, reporting bias, and inconsistency, particularly for medical devices and non-drug therapies. While there is a unified system for monitoring drug adverse events, the same does not apply to medical devices and non-drug therapies and interventions. Research trials tend to have closer follow-up, leading to better reporting of adverse events. However, in the real world, mild adverse events are more likely to go unnoticed.
- **Lack of qualitative data:** Currently, RWD is primarily useful for making decisions on quantifiable outcomes or events recorded in

systems. However, in some assessments, qualitative information may be more decisive. Despite advances in the implementation of unified syntaxes such as SNOMED or natural language processing, it is unlikely that valuable qualitative information will be consistently recorded in medical records.

- **Risk of misinterpretation:** RWD has been collected for purposes other than HTA. It is data to be interpreted within a particular context (the clinical context). When the same data is captured and processed in an automated manner, much of the clinical context, which often provides the semantics of clinical information, is lost. Automated interpretation of these data without contextual information can lead to misinterpretation of their value or consider information as true when it may not strictly be the case.

It is crucial to account for these limitations when interpreting results obtained from DVR modelling. Beyond these challenges, there exist limitations around data availability inherent in health systems which limit the potential of RWD to inform real-world decision-making. For instance, the current Spanish ecosystem of health data consists of different platforms (e.g. BIGAN) and programmes (e.g. PADRIS) for accessing health data. In some regions, there is no specific pathway, but rather a health information systems department that handles the data extraction process. The decentralised nature of the Spanish healthcare system poses a challenge for generalising HTA assessments.

While RedETS assessments aim to draw conclusions about a technology for the population at the national level, the agencies within this network can currently only access health data from their respective regions. Given the notable differences in population size, density, sociodemographic characteristics, health policies, and access to health resources among regions, it is difficult to predict the generalisability of findings based on regional population data and usability in national-level decision-making. The advent of the Spanish Healthcare Data Space (ENDS) and European Health Data Space (EHDS) will likely offer solutions to some of these problems, but these endeavours still require further regulation, development and implementation.

There exist other challenges pertaining to the methods used for generating evidence for decision-making from RWD. Some of them have been discussed in this manual (see specific section for references), but others, including methods developed specifically for evaluating causal relationships, lay outside the scope of this work and have not been discussed. These topics will be an integral part of the next phases of the methodological guidance for RWD in HTA (see also the next section).

8. Conclusions and further work

This document provides guidance specifically tailored to the most common case of HTA in RedETS production, which is assessing the inclusion of a new technology in the benefits/services basket during the preadoption phase.

We have outlined a workflow consistent with RedETS practices, highlighting the key points and milestones where RWD can add value. Additionally, we offer guidance on methods and tools to effectively incorporate RWD at each stage, from defining data requirements to analysis, with a particular emphasis on building decision models. While these instructions primarily apply to the preadoption phase, many of them are applicable to other stages of the technology's life cycle.

To illustrate the potential of RWD, we have replicated an existing HTA report, incorporating RWD at relevant points in the workflow. Through this exploration, we demonstrate how RWD provides valuable information about real patient profiles, their healthcare journeys, their actual health status (including comorbidities) that may impact health outcomes, and insights into the utilisation of healthcare resources. This demonstrates the potential of RWD to enhance HTA analyses, offering a more comprehensive understanding of the context-specific patient population and the potential impact of a new health technology.

We also acknowledge the limitations encountered regarding data availability and emphasize the need for further work to implement the modelling phase. In this spirit, this methodological handbook is intended to be continuously updated based on the cumulative experience gained from producing full HTA reports using RWD, as planned for instance within the RedETS 2023 work plan.

While we recognize that implementing the full workflow outlined in this handbook may be challenging at present, certain elements are already suitable for inclusion in RedETS assessments. For example, exploratory data analysis can be combined with systematic reviews and traditional modelling techniques to enhance the contextual value of RedETS assessments for decision-making.

The task ahead holds great hope but requires overcoming some challenges for full deployment of RWD-driven methods. This entails

fostering collaboration with health authorities and designated data holders at the national and regional levels to address data access challenges across RedETS agencies. In the short term, it is essential to include data scientists in assessment teams and provide appropriate capacity building, expanding the continuous training of HTA analysts to encompass RWD tools and modelling techniques.

It is worth noting that technology is continuously evolving, and we anticipate the emergence of new approaches, such as natural language processing, to leverage the information generated in the healthcare system, thereby expanding the data sources available to bridge existing gaps. Additionally, the development of the Spanish Healthcare Data Space and European Health Data Space may simplify some data access and analysis processes described in this document.

In conclusion, our exploration of using RWD for HTA in the preadoption phase has uncovered limitations but has also highlighted immediate options for better informing evidence-based decision-making. This exercise points the way forward by embracing advancements in RWD analysis and modelling techniques, fostering interdisciplinary collaboration, and building capacity to fully harness the potential of RWD in HTA. By doing so, we can strengthen RedETS' ability to provide robust and comprehensive assessments, ultimately improving decision-making.

Annexes

Annex I. Bibliography

1. Spanish Network of Agencies for Health Technology and Services Assessment of the National Health System (RedETS). Information needs of RedETS that can be covered by RWD. 2022.
2. Vivanco-Hidalgo RM, Blanco-Silvente L. Generación de evidencia con datos del mundo real en la Evaluación de Tecnologías Sanitarias. Guía metodológica. Barcelona: Agència de Qualitat i Avaluació Sanitàries de Catalunya. Departamento de Salud. Generalitat de Catalunya; 2022.
3. Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
4. Berger ML, Sox H, Willke RJ, Brixner DL, Eichler HG, Goettsch W, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol Drug Saf.* 2017;26(9):1033-9. <https://doi.org/10.1002/pds.4297>
5. Chinese National Medical Products Administration (NMPA). Technical Guideline on Real World Data (RWD) Used in Medical Device Clinical Evaluation (Trial Implementation) (No. 77 of 2020). 2020.
6. Dreyer NA, Bryant A, Velentgas P. The GRACE Checklist: A Validated Assessment Tool for High Quality Observational Studies of Comparative Effectiveness. *J Manag Care Spec Pharm.* 2016;22(10):1107-13. <https://doi.org/10.18553/jmcp.2016.22.10.1107>
7. Enrique Bernal-Delgado, Sarah Craig, Thomas Engsig-Karup, Francisco Estupiñán-Ronces, Nina Sahlertz Kristiansen, Jesper Bredmose Simons. TEHDAS WP6. European Health Data Space Data Quality Framework. 2022.
8. European Medicines Agency (EMA). Guideline on registry-based studies. 2021.
9. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I. Formulating causal questions and principled statistical answers. *Stat Med.* 2020;39(30):4922-48. <https://doi.org/10.1002/sim.8741>

10. Jaakko Lähteenmäki, Juha Pajula, Juan Gonzalez-Garcia, Carlos Telleria. TEHDAS WP7. Report on EHDS services users' expectations. 2021.
11. Japanese Pharmaceuticals and Medical Devices Agency. Guidelines for the Conduct of Pharmacoepidemiological Studies in Drug Safety Assessment with Medical Information Databases. 2014.
12. Jeonghoon Ahn, Dechen Choiphel, Anne Julienne Genuino, Anna Melissa Guerrero, Budi Hidayat, Yuehua Liu, et al. Use of Real-World Data and Real-World Evidence to Support Drug Reimbursement Decision-Making in Asia. 2021.
13. Juan González-García, Jaakko Lähteenmäki, Juha Pajula, Helena Lodenius, Carlos Tellería-Orrriols, Enrique Bernal-Delgado, et al. Options for the minimum set of services for secondary use of health data in the EHDS. 2022.
14. Lee KJ, Tilling KM, Cornish RP, Little RJA, Bell ML, Goetghebeur E, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *J Clin Epidemiol.* 2021;134:79-88. <https://doi.org/10.1016/j.jclinepi.2021.01.008>
15. Mina Tadrous, Christine Fahim, Kaley Hayes, Theresa Aves. Guidance for Reporting Real World Evidence; 2023. <https://www.cda-amc.ca/sites/default/files/RWE/MG0020/MG0020-RWE-Guidance-Report-Secured.pdf>
16. Murray EJ, Swanson SA, Hernán MA. Guidelines for estimating causal effects in pragmatic randomized trials. arXiv preprint arXiv:191106030. 2019.
17. National Institute for Health and Care Excellence. NICE real-world evidence framework. 2022.
18. OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics. OHDSI; 2019. <https://books.google.es/books?id=JxpnzQEACAAJ>
19. Orsini LS, Berger M, Crown W, Daniel G, Eichler HG, Goettsch W, et al. Improving Transparency to Build Trust in Real-World Secondary Data Studies for Hypothesis Testing-Why, What, and How: Recommendations and a Road Map from the Real-World Evidence Transparency Initiative. *Value Health.* 2020;23(9):1128-36. <https://doi.org/10.1016/j.jval.2020.04.002>

20. Pall Jonsson, Maximilian Salcher, Seamus Kent. *IMPACT HTA - Work Package 6. Methodological guidance on the analysis and interpretation of non-randomised studies to inform health economic evaluation*. 2022.
21. Wang SV, Pottegård A, Crown W, Arlett P, Ashcroft DM, Benchimol EI, et al. *HARmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force*. *Pharmacoepidemiol Drug Saf.* 2023;32(1):44-55. <https://doi.org/10.1002/pds.5507>
22. Aragonese Institute for Health Sciences (IACS). *Big Data in Health Care of Aragón (BIGAN)* [Internet]. [citado 2023 Jun 7]. Available from: <https://www.iacs.es/bigant/>
23. Briggs A, Claxton K, Sculpher M. *Decision Modelling for Health Economic Evaluation*. Oxford: Oxford University Press; 2006.
24. Haji Ali Afzali, H, Karnon J. *Specification and Implementation of Decision Analytic Model Structures for Economic Evaluation of Health Care Technologies*. 2014. p. 340-7.
25. Alarid-Escudero, F, Krijkamp E, Enns EA, Yang A, Hunink MGM, Pechlivanoglou P, et al. *A Tutorial on Time-Dependent Cohort State-Transition Models in R Using a Cost-Effectiveness Analysis Example*. *Med Decis Making.* 2023;43(1):21-41.
26. Vynnycky E, White RG. *An introduction to infectious disease modelling*. Oxford, New York: Oxford University Press, 2010.
27. Ribera A, Giménez E, Oristrell G, Osorio D, García L, Espallargues M, Ferreira I. *Desfibrilador automático implantable para prevención primaria de la muerte súbita cardíaca en España. Eficacia, seguridad y eficiencia*. Madrid: Ministerio de Sanidad. Barcelona: Agència de Qualitat i Avaluació Sanitàries de Catalunya; 2020. (Colección: Informes, estudios e investigación / Ministerio de Sanidad. Informes de Evaluación de Tecnologías Sanitarias).
28. Spanish Ministry of Health. *eCIE-Maps - CIE-10-ES Diagnósticos* [Internet]. [citado 10 may 2023]. Available from: <https://eciemaps.mscbs.gob.es/>
29. *Methodology WCCfDS. ATC/DDD Index 2023 - WHOCC* [Internet]. [citado 10 may 2023]. Disponible en: https://www.whooc.no/atc_ddd_index/

30. Kober L, Thune JJ, Nielsen JC, Haarbo J, Videbaek L, Korup E, et al. Defibrillator Implantation in Patients with Nonischemic Systolic Heart Failure. *N Engl J Med.* 2016;375(13):1221-30. <https://doi.org/10.1056/NEJMoa1608029>
31. Chung CP, Murray KT, Stein CM, Hall K, Ray WA. A computer case definition for sudden cardiac death. *Pharmacoepidemiology and Drug Safety.* 2010;19(6):563-72. <https://doi.org/10.1002/pds.1888>
32. Viles-Gonzalez JF, Arora S, Deshmukh A, Atti V, Agnihotri K, Patel N, et al. Outcomes of patients admitted with ventricular arrhythmias and sudden cardiac death in the United States. *Heart Rhythm.* 2019;16(3):358-66. <https://doi.org/10.1016/j.hrthm.2018.09.007>
33. Martínez González MA, Sánchez Villegas A, Faulín Fajardo FJ, et al. *Bioestadística Amigable - 4ª Edición.* Madrid: Elsevier España; 2020.
34. Wickham H, Golemund G. *R for Data Science.* 1st ed. [Internet]. Sebastopol, CA: O'Reilly Media; 2017. Available from: <https://r4ds.had.co.nz/>
35. Wilke CO. *Introduction to Modern Statistics: Statistical Inference and Data Science.* 1st ed. [Internet]. New York, NY: O'Reilly Media; 2017. Available from: <https://openintro-ims.netlify.app/>
36. National Institute for Health and Care Excellence Decision Support Unit. *NICE DSU Technical Support Document 15: Cost-effectiveness Modelling Using Patient-level Simulation.* [Internet]. London: National Institute for Health and Care Excellence; 2014. Available from: <https://www.publichealth.columbia.edu/research/population-health-methods/timeevent-data-analysis>
37. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R.* New York, NY: Springer; 2013.
38. Columbia University Mailman School of Public Health. *Time-To-Event Data Analysis.* [Internet]. New York, NY: Columbia University Mailman School of Public Health. Available from: <https://www.publichealth.columbia.edu/research/population-health-methods/timeevent-data-analysis>
39. National Institute for Health and Care Excellence Decision Support Unit. *NICE DSU Technical Support Document 14: Survival Analysis for Economic Evaluations Alongside Clinical Trials - Extrapolation with Patient-Level Data.* [Internet]. London: National Institute for Health and Care Excellence; 2013. Available from: <https://doi.org/10.1017/S0266462309990316>

40. McCarron CE, Pullenayegum EM, Marshall DA, Goeree R, Tarride J-E. Handling uncertainty in economic evaluations of patient level data: A review of the use of Bayesian methods to inform health technology assessments. *International Journal of Technology Assessment in Health Care*. 2009;25(4):546-54. <https://www.sheffield.ac.uk/nice-dsu/tsds/survival-analysis>
41. López-Bastida J, Oliva J, Antoñanzas F, García-Altés A, Gisbert R, Mar J, et al. Spanish recommendations on economic evaluation of health technologies. *The European Journal of Health Economics*. 2010;11(5):513-20. <https://doi.org/10.1007/s10198-010-0244-4>
42. Vallejo-Torres L, García-Lorenzo B, Serrano-Aguilar P. Estimating a cost-effectiveness threshold for the Spanish NHS. *Health Economics*. 2018;27(4):746-61. <https://doi.org/10.1002/hec.3633>
43. EUnetHTA. Methodological guideline – Methods for health economic evaluations. [Internet]. Place of Publication: Publisher; 2015. Available from: https://www.eunetha.eu/wp-content/uploads/2018/03/Methods_for_health_economic_evaluations.pdf
44. Haute Autorité de Santé. Choices in Methods for Economic Evaluation. Saint-Denis La Plaine: HAS; 2020.
45. Barton GR, Briggs AH, Fenwick EAL. Optimal Cost-Effectiveness Decisions: The Role of the Cost-Effectiveness Acceptability Curve (CEAC), the Cost-Effectiveness Acceptability Frontier (CEAF), and the Expected Value of Perfection Information (EVPI). *Value in Health*. 2008;11(5):886-97. <https://doi.org/10.1111/j.1524-4733.2008.00358.x>
46. Epstein D. Beyond the cost-effectiveness acceptability curve: The appropriateness of rank probabilities for presenting the results of economic evaluation in multiple technology appraisal. *Health Economics*. 2019;28(6):801-7. <https://doi.org/10.1002/hec.3884>
47. Lash TL, VanderWeele TJ, Haneuse S, Rothman K. *Modern Epidemiology*. Wolters Kluwer Health; 2020.
48. Eschenbach TG. Spiderplots versus Tornado Diagrams for Sensitivity Analysis. *Interfaces*. 1992;22(6):40-6. <https://doi.org/10.1287/inte.22.6.40>.
49. European Health Data Space Data Quality Framework, Deliverable 6.1 of TEHDAS EU 3rd Health Program (GA: 101035467) [Internet]. 2022. Available from: <https://tehdas.eu/tehdas1/app/uploads/2022/05/tehdas-european-health-data-space-data-quality-framework-2022-05-18.pdf>.

50. European Medicines Agency. Data quality framework for EU medicines regulation [Internet]. Amsterdam: European Medicines Agency; 2023. Available from: https://www.ema.europa.eu/system/files/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en_1.pdf.

Images

This document has been designed using images from [Flaticon.com](https://www.flaticon.com):

- [Employee Appraisal](#), [Interview](#) and [Big Data](#) icon packs by [Eucalyp](#)
- [Training and Coaching](#) icon pack by [Irfansusanto20](#)
- [Documentation](#) icon pack by [juicy_fish](#)
- [Costumer satisfaction](#) icon pack by [noomtah](#)

Annex II. ICD-SCD use case indicators

Some indicators for the preadoption assessment of the implantable cardioverter defibrillator (ICD) for the prevention of sudden cardiac death (SCD) are described below. These are presented under a generic name so that they can be extrapolated to evaluations of other similar technologies, adapting the variables to be used. The indicators are calculated using summary statistics coming from the data specification model described in Annex III.

QUESTION	HOW LARGE IS THE ICD TARGET PATIENT POPULATION?	
Indicator 1.1	Rate of target patients (per year)	
	Formula	$\frac{\textit{Total patients in the cohort}}{\textit{Observation time (years)}}$
Comments	<p><i>Observation time (years) = 6.5</i></p> <ul style="list-style-type: none"> The observation time in this assessment is ≈ 6.5 years (from 01/12/2011 to 31/05/2018). This rate would represent only the target population of Aragón. The national target population could be estimated by calculating a standardised ICD indication rate by age group and sex, and multiplying it by the 2018 Spanish population for each subgroup extracted from INE population figures. Conceptually, this rate is the same as the incidence of the diseases for which the ICD is indicated. 	

QUESTION	WHAT IS THE RATE OF SUDDEN CARDIAC DEATH (SCD) AMONG THE IMPLANTABLE CARDIAC DEFIBRILLATOR (ICD) TARGET PATIENT POPULATION?	
Indicator 2.1	Cause-specific mortality rate in target population (deaths per 100 patient-years)	
	Formula	$\frac{(\textit{Total patients with scd_bl = TRUE}) \times 100 \times 365.25}{\sum \textit{time_risk_death_nm}}$
Comments	$\sum \textit{time_risk_death_nm}$ is the sum of the variable defined in the data model containing the days of follow-up for each patient in the cohort.	

QUESTION	WHAT IS THE ALL-CAUSE MORTALITY AMONG THE ICD TARGET PATIENT POPULATION?	
Indicator 3.1	Mortality rate in target population (deaths per 100 patient-years)	
	Formula	$\frac{(Total\ patients\ with\ death_dt \neq empty) \times 100 \times 365.25}{\sum time_risk_death_nm}$
Comments	$\sum time_risk_death_nm$ is the sum of the variable defined in the data model containing the days of follow-up for each patient in the cohort.	

QUESTION	WHAT IS THE FREQUENCY OF ADVERSE EVENTS WITH CURRENT INTERVENTIONS FOR PREVENTION OF SCD AMONG ICD TARGET PATIENT POPULATION?	
Indicator 4.1	Incidence rate of infection in patients receiving current devices (events per 100 patient-years)	
	Formula	$\frac{(Total\ patients\ with\ infection_bl = TRUE) \times 100 \times 365.25}{\sum time_risk_ae_nm}$
Indicator 4.2	Incidence rate of bleeding in patients receiving current devices (events per 100 patient-years)	
	Formula	$\frac{(Total\ patients\ with\ bleeding_bl = TRUE) \times 100 \times 365.25}{\sum time_risk_ae_nm}$
Indicator 4.3	Incidence rate of pneumothorax in patients receiving current devices (events per 100 patient-years)	
	Formula	$\frac{(Total\ patients\ with\ pneumothorax_bl = TRUE) \times 100 \times 365.25}{\sum time_risk_ae_nm}$
Indicator 4.4	Incidence rate of inappropriate shock in patients receiving current devices (events per 100 patient-years)	
	Formula	$\frac{(Total\ patients\ with\ shocks_bl = TRUE) \times 100 \times 365.25}{\sum time_risk_ae_nm}$
Comments	$\sum time_risk_ae_nm$ is the sum of the variable defined in the data model containing the follow-up period of each patient who was implanted with a medical device.	

QUESTION	WHAT IS THE LEVEL OF HEALTH RESOURCES USE AMONG THE ICD TARGET PATIENT POPULATION?											
Indicator 5.1	Rate of primary care visits in target population (per patient and year)											
	Formula	$\frac{\sum pc_visits_nm}{Total\ patients\ in\ the\ cohort \times Observation\ time\ (years)}$										
Indicator 5.2	Rate of emergency department visits in target population (per patient and year)											
	Formula	$\frac{\sum em_adm_nm}{Total\ patients\ in\ the\ cohort \times Observation\ time\ (years)}$										
Indicator 5.3	Rate of medical specialty department visits in target population (per patient and year)											
	Formula	$\frac{\sum cardiology_visits_nm}{Total\ patients\ in\ the\ cohort \times Observation\ time\ (years)}$										
Indicator 5.4	Rate of hospital admissions in target population (per patient and year)											
	Formula	$\frac{\sum hospital_adm_nm}{Total\ patients\ in\ the\ cohort \times Observation\ time\ (years)}$										
Comments (for 5.1 to 5.4)	<p>Observation time (years) = 6.5</p> <ul style="list-style-type: none"> The observation time in the evaluation is ≈ 6.5 years (from 01/12/2011 to 31/05/2018). We use the sums of the variables defined in the data model containing the number of contacts with the health system. We calculate visits for any reason, but the variable counting the number of visits could be restricted in the data model to certain diagnostic codes (ICD, CIAP, APR-GRD or SNOMED). 											
Indicator 5.5	Proportion of drug use in target population (%)											
	Formula	$\frac{(Total\ patients\ with\ [var_bl] = TRUE) \times 100}{Total\ patients\ in\ the\ cohort}$										
Comments	<ul style="list-style-type: none"> var_bl is substituted for each of the drug variables, so that the Indicator is calculated for each of the following variables: beta_blocker_bl, digitalis_bl, ace_bl, arb_bl, arni_bl, diuretics_bl and aldosterone_anta_bl. The indicator would be presented in a table like this: <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>drug</th> <th>prop (%)</th> </tr> </thead> <tbody> <tr> <td>beta_blocker_bl</td> <td></td> </tr> <tr> <td>digitalis_bl</td> <td></td> </tr> <tr> <td>ace_bl</td> <td></td> </tr> <tr> <td>...</td> <td>...</td> </tr> </tbody> </table>		drug	prop (%)	beta_blocker_bl		digitalis_bl		ace_bl	
drug	prop (%)											
beta_blocker_bl												
digitalis_bl												
ace_bl												
...	...											

QUESTION	WHAT IS THE LEVEL OF HEALTH RESOURCES USE AMONG THE ICD TARGET PATIENT POPULATION?									
Indicator 5.6	Proportion of device use in target population (%)									
	Formula	$\frac{(Total\ patients\ with\ [var_bl] = TRUE) \times 100}{Total\ patients\ in\ the\ cohort}$								
Comments	<ul style="list-style-type: none"> • <i>var_bl</i> is substituted for each of the device variables, so that the Indicator is calculated for each of the following variables: icd_bl, crt_bl and pacemaker_bl. • The indicator would be presented in a table like this: <table border="1" data-bbox="472 551 918 697"> <thead> <tr> <th data-bbox="472 551 698 587">device</th> <th data-bbox="698 551 918 587">prop (%)</th> </tr> </thead> <tbody> <tr> <td data-bbox="472 587 698 624">icd_bl</td> <td data-bbox="698 587 918 624"></td> </tr> <tr> <td data-bbox="472 624 698 660">crt_bl</td> <td data-bbox="698 624 918 660"></td> </tr> <tr> <td data-bbox="472 660 698 697">pacemaker_bl</td> <td data-bbox="698 660 918 697"></td> </tr> </tbody> </table> • Although ICD is the technology to be adopted and only CRT and pacemaker are the comparators fully adopted in this period, ICD may already be present in several centres. It is for this reason that it is included in the indicator. 		device	prop (%)	icd_bl		crt_bl		pacemaker_bl	
device	prop (%)									
icd_bl										
crt_bl										
pacemaker_bl										

Annex III. ICD-SCD use case data model specification description

The full model specification is available here:

[icd_scd_data_model_specification_preadoption_0.2.1.xlsx](#)

Model structure

The data model specification for the ICD-SCD use case is described in a spreadsheet format. It contains four primary tabs describing the scope of the project, assessment questions, cohort definition, and description of variables at patient (individual) level. The rest of the tabs further define the variables described at the individual level.

TAB	CONTENT
model_metadata	A general description of the project, authors, conventions for the model (such as nomenclature of variables), and version control.
research_questions	A list of the research questions to be answered using the data requested in the specification.
cohort_definition	The general cohort description, initial events (including clinical codes), and inclusion and exclusion criteria.
model_description_individual_level	Entities and variables description, including format, type, units, requirement level, optional validation rules, possible data sources and comments.
variable (def)	A tab for each variable that requires a list of diagnostic or procedure codes to obtain its value, or a set of rules to be calculated.

Annex IV. RWD section in HTA protocol for direct oral anticoagulants (DOAC) quantification

In RedETS work plan for 2023, RWD for preadoption HTA will be tested in an assessment of the techniques for the quantification of direct oral anticoagulants (DOAC).

Below we show the research question and RWD sections we included in the protocol for this assessment, after gathering preliminary information about the technology.

Research question

DESCRIPTION	SCOPE
Population	Patients receiving direct-acting oral anticoagulants for the prevention or treatment of thromboembolic diseases.
Intervention	Methods for quantification of plasma concentrations of DOAC, including specific quantitative and chromogenic methods: dilute thrombin test (dTT), ecarin-based methods and specific chromogenic assays (anti-IIa and anti-Xa).
Comparator	<p>Usual practice without the use of quantitative and specific measurement assays based on chromogenicity.</p> <ul style="list-style-type: none"> • Dose adjustment based on clinical experience and clinical assessment of the patient, using clinical parameters such as age, weight, renal function, and presence of risk factors for bleeding or thrombosis. • Monitoring of indirect markers of coagulation, such as prothrombin time (PT) or activated partial thromboplastin time (APTT), although these are not specific tests for DOAC monitoring and do not allow direct measurement of drug concentration in the blood.
Results	<ul style="list-style-type: none"> • Efficacy/effectiveness: Analytical precision of quantitative and chromogenic assays specific for DOACs, including accuracy and precision of measurements. Clinical utility, ability to guide decision-making and its impact on clinical outcomes: reduction of thromboembolic events (e.g. DVT, PE or stroke) and/or reduction of bleeding. • Safety: Incidence of adverse events (e.g. bleeding, drug interactions, allergic reactions). • Efficiency: Use of resources required to perform diluted TT tests, ecarin-based methods and chromogenic assays in DOAC monitoring compared to standard practice. Cost-effectiveness of quantitative tests in DOAC monitoring compared to standard practice.

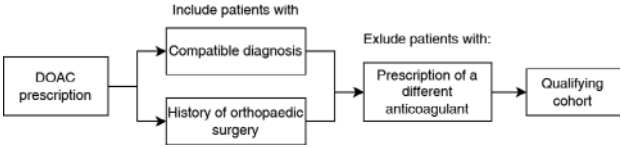
DESCRIPTION	SCOPE
Design	<ul style="list-style-type: none"> • Efficacy/effectiveness and safety: Randomised or non-randomised clinical trials (RCTs) and observational studies with comparison groups assessing the effectiveness of quantitative techniques with standard practice in DOAC monitoring.) If not enough RCTs are available, other designs such as observational studies without comparison groups (with sample size > 30 persons) will also be considered. If systematic reviews are identified, individual studies will be extracted from them. • Efficiency: Full economic evaluations, partial economic evaluations (cost studies) or budget impact analyses.

Real World Data

The development of this report involves the use of Real-World Data (RWD), i.e. data routinely generated by the different sources of health information.

To develop this analysis, an evaluation model will be constructed that represents in a simplified way the usual courses of action and results of the usual practice in the National Health System, compared with the incorporation of the new technology. The steps described below will be followed:

1. A cohort will be defined by different entry criteria and the characteristics (socio-demographic, clinical, health outcomes, resource use, etc.) of the patients throughout the follow-up period will be described, constituting the parameters of the comparison arm.
2. Information extracted from the systematic literature review on the characteristics and outcomes of the technology under evaluation will be entered into the intervention arm.
3. Simulation of the potential comparative results will be performed to guide decision making.
4. In order to cover the uncertainty as much as possible, deterministic and probabilistic sensitivity analyses will be performed, taking the dispersion measures of the different parameters obtained using DVR.

<p>Cohort entry criteria</p>	<p>Cohort description: patients receiving direct-acting oral anticoagulants (DOAC) for the prevention or treatment of thromboembolic disease.</p> <p>Entry event. DOAC prescription, according to ATC codes:</p> <ul style="list-style-type: none"> - Rivaroxaban (B01AF01). - Dabigatran (B01AE07) - Apixaban (B01AF02) - Edoxaban (B01AF03) <p>Inclusion criteria:</p> <ol style="list-style-type: none"> 1. Diagnosis compatible with indication for OACDs, according to ICD-10 Diagnostic codes: <ul style="list-style-type: none"> - Deep vein thrombosis (I80, I82) - Pulmonary embolism (I26) - Atrial fibrillation (I48) - Venous thromboembolic disease (I82, O22.5) 2. History of major orthopaedic surgery, according to ICD-10 Procedure codes: <ul style="list-style-type: none"> - Hip replacement (e.g., 0SRB, 0SR9) - Knee replacement (e.g. 0SRC, 0SRD) <p>Exclusion criteria:</p> <p>Prescription of other anticoagulant therapy, according to ATC codes:</p> <ul style="list-style-type: none"> - Heparin (B01AA) - Vitamin K antagonists (B01AE)  <pre> graph LR A[DOAC prescription] --> B[Compatible diagnosis] A --> C[History of orthopaedic surgery] B --> D[Prescription of a different anticoagulant] C --> D D --> E[Qualifying cohort] </pre> <p>The flowchart illustrates the selection process for the cohort. It starts with 'DOAC prescription'. This leads to two parallel inclusion criteria: 'Compatible diagnosis' and 'History of orthopaedic surgery'. Both of these lead to an exclusion step: 'Prescription of a different anticoagulant'. Finally, the remaining patients form the 'Qualifying cohort'.</p>
<p>Socio-demographic characteristics</p>	<p>For each patient in the cohort, socio-demographic characteristics will be collected at entry:</p> <ul style="list-style-type: none"> - Age - Sex - Socio-economic level - Health sector/basic health area - Nationality/Place of birth

<p>Characteristics related to the comparison technology</p>	<p>For each patient in the cohort, information will be collected on their previous history, risk factors and coagulation tests performed.</p> <ul style="list-style-type: none"> - History of cardiovascular disease (at cohort entry): atrial fibrillation, heart failure, coronary artery disease or peripheral arterial disease. - History of thromboembolic events (at cohort entry): deep vein thrombosis, pulmonary embolism or stroke. - Presence of risk factors (at cohort entry): obesity, smoking, BMI, atherogenic/thromboembolic risk index. - Patients undergoing major orthopaedic surgery, such as hip or knee replacement. - Non-specific global coagulation tests (during observation time): aPTT, PT and TT. - Comorbidities: renal disease, liver disease, diabetes, hypertension.
<p>Outcome measures</p>	<p>The incidence of thromboembolic and haemorrhagic events over the follow-up period will be calculated from information on admissions and visits associated with ICD-10 Diagnosis codes of:</p> <ul style="list-style-type: none"> - Deep vein thrombosis - Pulmonary embolism - Cerebrovascular disease - Venous thromboembolic disease - Major haemorrhage
<p>Pharmacological treatment</p>	<p>During the observation period, the data necessary to describe their pharmacological treatment, both with DOAC and with other prescribed drugs, should be collected:</p> <ul style="list-style-type: none"> - Treatment with DOAC: class of DOAC, dose, duration of treatment, changes in treatment during the observation period, interruptions. - Adherence to treatment: Morisky-Green test result in primary care history. - Adverse events related to treatment: bleeding, allergic reactions, etc. - Concomitant use of other drugs: active substances prescribed, dosage, duration of treatment.

Use of resources	The use of healthcare resources that each patient takes during the observation period will be modelled, collecting data on: <ul style="list-style-type: none">- Turnaround times for results- Frequency of clinic visits for monitoring- Hospital admissions- Emergency episodes
Mortality	Mortality events will be obtained for each patient in the cohort, over the observation period: <ul style="list-style-type: none">- Cardiovascular mortality- All-cause mortality

Annex V. Exploratory data analysis tools

There are a few reasons why we consider it good practice to use R, specifically R Markdown (Rmd) files, along with the tidyverse and dplyr packages. It helps us with reproducibility and transparency, we can have an organised workflow and there are extensive communities where we can share and get help for working with RWD.

- **Reproducibility:** Rmd files allow us to combine code, text, and visualisations in a single document. This promotes reproducibility as others can easily reproduce our analysis by running the Rmd file. It also allows us to revisit and rerun our analysis in the future, ensuring consistency and transparency.
- **Data manipulation:** The tidyverse and dplyr packages provide powerful tools for data manipulation. They offer a consistent and intuitive syntax for filtering, transforming, summarising, and visualising data. This makes it easier to clean and preprocess RWD, handle missing values, and derive new variables for analysis. As further work, we will develop a toolkit where we will recommend the main functions to be used when exploring the data with this package.
- **Workflow and project organisation:** Rmd files support a modular and structured approach to analysis. We can divide our analysis into sections, create reusable code chunks, and easily incorporate changes and updates. This promotes a more organised and efficient workflow, especially when working with large and complex RWD projects.
- **Community and resources:** R has a large and active community of users, including data scientists, statisticians, and researchers, who contribute to its development and create various packages and resources. This means we can benefit from a wealth of knowledge, tutorials, forums, and online communities to help us with our EDA tasks and RWD analyses. There are several platforms available for sharing R code, data, and analysis outputs. These platforms provide a means to disseminate our work, collaborate with others, and receive feedback. Some popular platforms include:
 - **GitHub:** A widely used platform for version control, collaboration, and sharing code and projects.

- Zenodo: A repository that enables researchers to share and preserve their work, including R code, datasets, and publications.
- Kaggle: A platform for data science competitions and sharing datasets, code, and notebooks.

The full example of exploratory data analysis is available here (.Rmd and .html format):

[EDA_Notebook.Rmd](#)

[EDA_Notebook.html](#)