HOW TO SET UP CLINICAL REGISTRY

Guidebook focused on multiple myeloma

Magda Bařinová, Vladimír Maisnar,
Lucie Brožová, Tereza Bohatá

HOW TO SET UP CLINICAL REGISTRY

A guidebook focused on multiple myeloma

Authors

Fourth Department of Medicine - Hematology, Charles University Hospital and Faculty of Medicine, Charles University, Hradec Králové, Czech Republic

prof. Vladimír Maisnar, MD, Ph.D., MBA (medical expert, guarantor)

Department of Haematology and Bone Marrow Transplantation, Poznan University of Medical Sciences, Poznań, Poland

Dominik Dytfeld, MD, Ph.D

Institute of Biostatistics and Analyses Ltd., spin-off company of Masaryk University Brno, Czech Republic

Magda Bařinová, MSc (team leader, project manager, specialist on haemato-oncology projects)

Marika Chrápavá, MSc (clinical registries specialist)

Petra Ovesná, PhD (senior analyst)

Lucie Brožová, MSc (analyst, specialist on haemato-oncology projects)

Jaroslav Koča, MSc (senior data manager)

Tereza Bohatá, MSc (data manager)

A guidebook focused on multiple myeloma is reworked edition of guidebook focused on chronic lymphocytic leukaemia written by:

Marika Chrápavá, MSc (team leader, clinical registries specialist)

Zlatuše Křístková, MSc (project manager, specialist on haemato-oncology projects)

Assoc. Prof. Daniel Schwarz, PhD (IT specialist)

Jaroslav Koča, MSc (senior data manager)

Petra Ovesná, PhD (senior analyst)

Michaela Škrabalová, MSc (quality manager)

Lucie Nekvindová, MSc (analyst, specialist on haemato-oncology projects)

Eva Koriťáková, PhD (analyst, specialist on haemato-oncology projects)

Jana Baranová, MSc (data manager, specialist on haemato-oncology projects)

Tereza Bohatá, MSc (data manager)

David Pavlík, MSc (data manager)

Vavřinec Havlíček, MSc (data manager)

Radomír Stáňa, BA (helpdesk specialist)

Petra Jobánková, MSc (data manager)

Martin Komenda, PhD (IT specialist)

Petr Brabec, PhD (clinical registries specialist)

1. INTRODUCTION

1.1 Preface

The aim of this guidebook is to gather the know-how of experts with long-term experience in the field of healthcare systems and clinical registries management. This guidebook will guide the user through the whole process, from the idea of establishing a patient registry to the valuable outcomes of the registry, which might contribute to improvements in common clinical practice.

1.2 Multiple myeloma

1.2.1 Epidemiology

Multiple myeloma (MM) is a malignant plasma cell disorder characterised by an uncontrolled proliferation of monoclonal plasma cells in the bone marrow [1] [2]. It constitutes approximately 1% of all reported neoplasms and approximately 13% of haematological malignancies worldwide [1]. In the US, Canada and Western European countries, approximately 5 to 7 new cases of MM are diagnosed per 100.000 population every year [1], [3], [4].

According to latest data from the Czech National Cancer Registry, the incidence of MM in the Czech Republic was 4.8 cases per 100.000 population in 2014. In the last decade (2004–2014), there was a 26.9% increase in MM incidence and an 8.3% increase in MM mortality in the Czech population, which corresponds to a 74.4% increase in MM prevalence [5]. The overall epidemiological situation (i.e. incidence and mortality of MM) in the Czech Republic from the 1970s to 2014 is summarised in Figure 1.

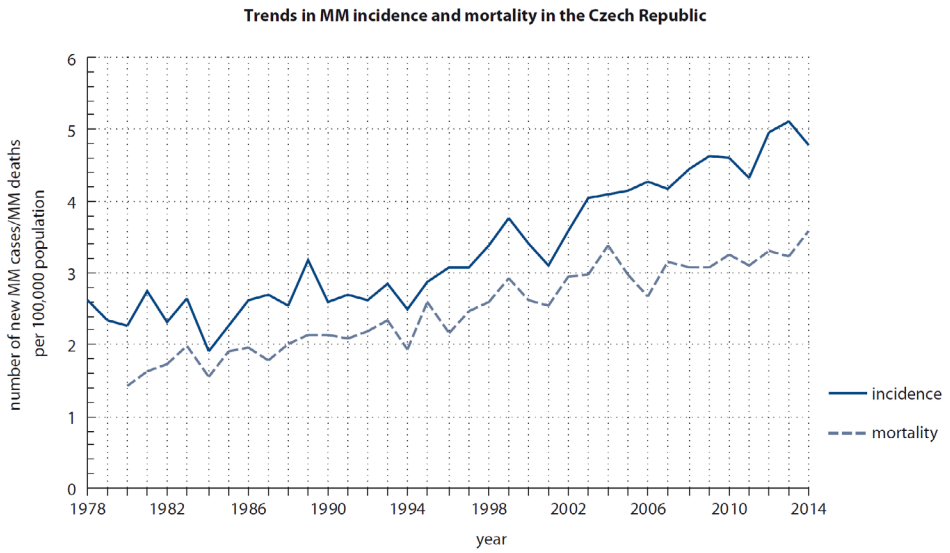


Figure 1 Trends in MM incidence and mortality in the Czech Republic

The risk of developing MM increases with age: the median age at diagnosis is 69 years [1]. Incidence rates are highest between the age of 62 and 76 years and age-specific incidence rates rise sharply from around the age of 45. At the time of diagnosis, less than 1% of all MM patients are under the age of 40 and 18.6% of all MM patients are under the age of 60. The ratio of men to women with MM is about 3:2.

Multiple myeloma is usually preceded by an asymptomatic premalignant stage referred to as the monoclonal gammopathy of undetermined significance (MGUS) [3], [6]. MGUS is present in roughly 3–4% of the population aged over 50 [3], [6]. The diagnosis of MGUS requires the absence of SLiCRAB symptoms: hypercalcaemia, renal failure, anaemia, bone lesions and myeloma defining events (percentage of clonal plasma cells >60, involved to uninvolved free light chains ratio >100 and at least 2 focal lesions (>5mm) seen in MRI serum) that can be attributed to an underlying plasma cell disorder [7], [8].

1.2.2 Clinical presentation

MM patients can have a variety of clinical presentations. Malignant plasma cells create abnormal antibodies called M protein (exactly it is one type of monoclonal immunoglobulin), which offer no benefit to the body. As multiple myeloma cells multiply, they crowd out normal plasma cells, which can lead to a number of signs and symptoms [8]. Multiple myeloma symptoms can vary from person to person. Some patients experience no symptoms at all. Others may experience some of the following symptoms: bone pain (the most common symptom present in 70% newly diagnosed MM patients), fatigue, loss of appetite and weight loss, mental changes, constipation or excessive thirst. Infections may occur more frequently in MM patients due to hypogammaglobulinaemia.

Diagnosis

The diagnosis of MM is based on the recommendation of the International Myeloma Working Group (IMWG) [8], with bone marrow examination, serum and urine electrophoresis and skeletal imaging techniques used as the main diagnostic methods.

Multiple myeloma evolves from a clinically silent premalignant stage referred to as the monoclonal gammopathy of undetermined significance (MGUS) [6], [9]. MGUS is a classic premalignant condition with a low risk of malignant conversion, but the risk of progression persists indefinitely. A small subset of patients has an intermediate clinical phenotype between MGUS and MM, and they are referred to as having smouldering multiple myeloma (SMM) [10]. MGUS is associated with a risk of progression to MM or related malignancy at a rate of approximately 1% per year, whereas SMM has a much higher risk of progression – approximately 10% per year [7], [11]. MGUS and SMM are typically asymptomatic and are typically diagnosed incidentally when a monoclonal (M) protein is detected during laboratory work-up of patients who have a wide spectrum of clinical conditions.

In the past, the diagnosis of MM required evidence of end-organ damage attributable to the neoplastic clone of plasma cells: hypercalcemia, renal failure, anaemia and osteolytic bone lesions, commonly referred to as the CRAB features [12]. In 2014, diagnostic criteria for MM were revised by the IMWG [8]. The diagnosis of MM now requires 10% or more clonal plasma cells on bone marrow examination or a biopsy-proven plasmacytoma plus the presence of one or more myeloma-defining events [8]. Myeloma-defining events include the presence of one or more CRAB features, or one or more myeloma defining

events [13]. They are clonal bone marrow plasma cells greater than or equal to 60%, serum FLC ratio of 100 or higher, provided involved FLC level is 100 mg/L or higher, or more than one focal lesion on MRI. Table 1 provides the revised IMWG criteria for the diagnosis of MGUS, SMM and MM. These three symptoms included in the definition of MM are associated with an approximately 80% risk of progression to symptomatic end-organ damage in two or more independent studies [8].

Table 1 Revised IMWG criteria for the diagnosis of non-IgM MGUS, SMM and MM [8]

Diagnosis	Criteria
Non-IgM MGUS	All three criteria must be met:
	Serum monoclonal protein (non-IgM type) < 3 g/dL
	Clonal bone marrow plasma cells < 10 %
	Absence of end-organ damage such as SLiMCRAB features that can be attributed to the plasma cell proliferative disorder
Smoulde-ring MM	Both criteria must be met:
	Serum monoclonal protein (IgG or IgA) ≥ 3 g/dL, or urinary monoclonal protein ≥ 500 mg per 24 h and/or clonal bone marrow plasma cells 10%–60%
	Absence of myeloma-defining events or amyloidosis
MM	Both criteria must be met:
	Clonal bone marrow plasma cells ≥ 10% or biopsy-proven bony or extramedullary plasmacytoma
	Any (one or more) of the following myeloma-defining events:
	Evidence of end organ damage that can be attributed to the underlying plasma cell proliferative disorder, specifically:
	Hypercalcemia: serum calcium > 0.25 mmol/L (> 1 mg/dL) higher than the upper limit of normal or > 2.75 mmol/L (> 11 mg/dL)
	Renal insufficiency: creatinine clearance < 40 mL/min or serum creatinine > 177 µmol/L (> 2 mg/dL)
	Anaemia: haemoglobin value of > 2 g/dL below the lower limit of normal, or a haemoglobin value < 10 g/dL
	Bone lesions: one or more osteolytic lesions on skeletal radiography, CT, or PET-CT
	Percentage of clonal bone marrow plasma cells ≥ 60 %
	Involved: uninvolved serum FLC ratio ≥ 100 (involved FLC level must be ≥ 100 mg/L)
	> 1 focal lesion on MRI studies (at least 5 mm in size)

Although the IMWG diagnostic criteria are not perfect, the intent and the implications of the updated version are to facilitate an earlier detection and earlier initiation of therapy with the aim to improve the overall survival of multiple myeloma patients. Indeed, population-based data support an earlier detection of multiple myeloma and an initiation of therapy [8].

1.2.3 Prognosis

There is a major variation in the survival of MM patients depending on host factors, tumour burden (stage), biology (cytogenetic abnormalities), and response to therapy [14]. Tumour burden in MM has traditionally been assessed using the Durie-Salmon Staging (DSS) [15] and newly by the International Staging System (ISS) [16], [17]. Both of these staging systems, however, have some limitations.

Disease biology in MM is best reflected based on the molecular subtype of the disease and the presence or absence of specific cytogenetic abnormalities [18]. For example, abnormalities such as t(4;14), t(14;16), t(14;20), gain(1q21), del(1p), and del(17p) influence the disease course, response to therapy, and prognosis in MM [18], [19]. The Revised International Staging System (RISS) combines elements of tumour burden (ISS) and disease biology (presence of high-risk cytogenetic abnormalities or elevated lactate dehydrogenase level) to create a unified prognostic index that helps in clinical care as well as in the comparison of data from clinical trials (Table 2) [20].

Table 2 Revised International Staging System (adapted from Palumbo et al.) [20]

Stage	Frequency (% of Patients)	5-Year Survival Rate (%)
Stage I	28	82
ISS stage I (serum albumin > 3.5, serum beta--2-microglobulin < 3.5) and		
No high-risk cytogenetics		
Normal LDH		
Stage II	62	62
Neither stage I or III		
Stage III	10	40
ISS Stage III (serum beta-2-microglobulin > 5.5) and		
High-risk cytogenetics [t(4;14) t(14;16), or del(17p)] or elevated LDH		

The RISS is important in the clinical practice in terms of counselling patients regarding the prognosis, as well as in clinical trials when comparing outcomes across clinical trials. The revised staging system also allows us to better predict outcome and tailor treatments accordingly in individual MM patients [18].

It is probably fair to say that our understanding of the heterogeneous biology of multiple myeloma and its implications to therapy is not very good. Emerging DNA sequencing data show a massive genetic heterogeneity across and within given multiple myeloma patients [21]. It has been found that, on average, a newly diagnosed multiple myeloma patient has mutations in a large number of genes and these mutations are unevenly distributed in parallel sub-clones present already at diagnosis [21], [22]. Despite all success

for the majority of multiple myeloma patients, yet today, approximately 25% of all newly diagnosed patients survive less than 3 years [23]. Therefore, prospective studies are needed to dissect tumour biology, to define multiple myeloma subtypes and, based on biological insights, seek to develop rational therapies for individual subtypes.

1.2.4 Therapy

At the beginning of the 21st Century, the survival time of patients with multiple myeloma was about 3 years on average [24]. Around that time, three drugs (bortezomib, lenalidomide and thalidomide) for the treatment of multiple myeloma were introduced to the market. Yet another MM drug – carfilzomib – was approved in 2012 and four other drugs (panobinostat, daratumumab, elotuzumab and ixazomib) were approved by the FDA in 2015 [8]. The increased availability of highly effective targeted agents with limited overlapping toxicities has shifted the therapeutic paradigm from less effective 2-drug combinations towards the use of modern, effective 3-drug combination strategies [25], [26]. Reflective of the fast-moving field, studies are already ongoing to evaluate modern 4-drug combinations, including monoclonal antibodies in combination with proteasome inhibitors (PIs), immunomodulatory drugs (IMiDs) and low-dose steroids. Nowadays, the overall survival of newly diagnosed multiple myeloma patients exceeds 10 years, according to data from available population registries [23], [27]. The use of modern combination therapy integrated with modern clinical management is expected to deliver an even longer overall survival for patients with multiple myeloma in years to come [24].

Response criteria in multiple myeloma were developed to use serum and urine assessment of monoclonal proteins and bone marrow assessment (Table 3) [28].

Table 3 International Myeloma Working Group uniform response criteria [28]

Category	Criteria
Stringent complete response (sCR)	CR as defined below PLUS: Normal SFLC ratio, AND Absence of clonal plasma cells in marrow by immunohistochemistry or flow cytometry
Complete response (CR)	Negative immunofixation on the serum and urine, AND Disappearance of any soft tissue plasmacytomas, AND ≤ 5% plasma cells in bone marrow
Very good partial response (VGPR)	Serum and urine M-protein detectable by immunofixation but not on electrophoresis, OR ≥ 90% reduction in serum M-protein plus reduction in 24h urinary M-protein by > 90% and to < 100 mg/24h
Partial response (PR)	≥ 50% reduction of serum M-protein, AND reduction in 24h urinary M-protein by ≥ 90% and to < 200 mg/24h, AND ≥ 50% reduction in the size of any plasmacytomas present at baseline (In place of M-protein criteria: If serum + urine M protein unmeasurable, ≥ 50% decrease in difference between involved and uninvolved SFLC levels, OR If serum + urine M-protein + SFLC assay also unmeasurable, a ≥ 50% reduction in marrow plasma cells is required, provided baseline bone marrow plasma cell percentage was ≥ 30%)

Category	Criteria
Stable disease (SD)	Not meeting criteria for CR, VGPR, PR or progressive disease
Progressive disease (PD)	At least ONE of the following: > 25% increase in serum M-protein in 3 months (absolute increase must be > 5 g/L), OR > 25% increase in urine M-protein in 3 months (absolute increase must be > 200 mg/24h), OR > 25% increase in the difference between involved and uninvolved SFLC levels (applicable only to patients without measurable serum and urine M- protein (absolute increase must be > 10 mg/dL), OR > 25% increase in bone marrow plasma cell percentage (absolute percentage must be > 10%), OR Development of new bone lesions or soft tissue plasmacytoma, OR Development of hypercalcaemia
Clinical relapse	Requires at least one of the following: Development of new bone lesions or soft tissue plasmacytoma, OR Increase in size of existing plasmacytomas or bone lesions, OR Any of the following attributable to myeloma: Development of hypercalcaemia, Development of anaemia (drop in Hb > 2 g/dL), Rise in serum creatinine
Relapse from CR	Requires at least one of the following: Reappearance of serum or urine M-protein by immunofixation or electrophoresis, OR Development of > 5% plasma cells in the bone marrow, OR Appearance of any other sign of progression (e.g. new plasmacytoma, new lytic bone lesion)

Given the high rates of complete response seen in patients with multiple myeloma with new treatment approaches, new response categories need to be defined that can identify responses that are deeper than those conventionally defined as complete response. Data from a recent meta-analysis show that MRD negativity is associated with longer progression-free survival and overall survival [29]. In 2016, the updated IMWG response criteria included MRD as the deepest level of treatment response in multiple myeloma. The research has therefore recently focused on the identification of residual tumour cells in the bone marrow using flow cytometry or gene sequencing [30], [31]. Based on the 2016 IMWG criteria, MRD negativity can be defined by either multiparameter flow-cytometry-based or molecular-based assays with a sensitivity of at least 1 in 100,000 cells (10-5 or even 10-6) to determine the MRD status; this status, however, should only be determined in patients who have achieved a conventional CR. Furthermore, sensitive imaging techniques can be used to detect the presence of residual disease outside of the bone marrow [32]. Combining these new methods, the International Myeloma Working Group has defined new response categories of minimal residual disease negativity, with or without imaging-based absence of extramedullary disease, to allow uniform reporting within and outside clinical trials [28].

1.2.5 Future challenges

MM patients should be treated in clinical trials. In this way, it might be possible to achieve better results and also to verify the optimal use of promising new agents for improvements of treatment outcomes.

1.3 Why collect MM data

Data from diagnosis and therapy of MM patients provide evidence for good decision-making for all potential stakeholders (clinicians, physician organisations, patients and patient advocacy organisations, payers, drug or device manufacturers, policymakers or decision makers). There are more different ways of collecting MM data.

For a clinician, "the patients' needs come first". To deliver the best possible treatment to our patients, the physicians' decision should be based on evidence [evidence-based medicine (EBM), the fundamental basis for clinical practice]. Physicians always have to determine the trade-off between the evidence, benefits, risks and costs of alternative treatment and consider their patients' values and preferences.

Randomised controlled trials (RCTs) are considered number one in the hierarchy of EBM. Patient registries resemble and differ from RCTs in many essential ways; they both have important and complementary roles in evaluating patient outcomes.

Well-designed patient registries collect comprehensive data from all patients with a specific condition (with only few excluded, if any) and they evaluate care as it is actually provided (and not assigned or recommended by a protocol). Data from a MM registry may be more generalisable and representative of what is achieved in real-world practice and the broader patient population.

Therefore, the overall purpose of a MM registry is to assure high-quality patient care and diagnostics in real haematological practice, to assess the adherence to national guidelines for MM, and to also provide source data for research purposes. However, to provide this information, it is necessary to ensure that as many patients as possible from the entire country are captured into the registry. Data from a single centre or even several haematological centres may be inaccurate.

Patient registries, RCTs, other study designs and other data sources are all useful in the development and evaluation of evidence, each with its own advantages and limitations [33]. Patient registries provide valuable information, contribute to EBM and fill the gaps in EBM that cannot be addressed by RCT. Properly designed and maintained patient registries can provide a real-world view of clinical practice, patient outcomes, safety and comparative effectiveness [34].

1.4 How to collect MM data

General definition of a registry

„A patient registry is an organised system that uses methods employed in observational studies to collect uniform data (clinical and other) in order to evaluate specified outcomes for a population defined by a particular disease, condition or exposure, and that

serves one or more predetermined scientific, clinical or policy purposes. A registry database is a file (or files) derived from the registry." [34]

1.4.1 Registry types and purposes

Registry types: Registries are classified according to definition of their populations [34].

- Patient registries (clinical registries, clinical data registries, disease registries, condition registries and outcomes registries) are defined by patients having the same diagnosis, such as MM.
- Product registries include patients who have been exposed to the same biopharmaceutical products, drugs or medical devices.
- Health services registries consist of patients who have had a common procedure, clinical encounter or hospitalisation.
- Population registries (based geographically and not based on a disease, condition or exposure) are created for public health reporting without tracking outcomes (e.g. vaccine registries).
- Listing registries are used solely to identify patients with particular diseases and not for evaluation outcomes.

Registry purposes: Patient registries may serve one or more purposes and are established for clinical, scientific and health policy reasons [34].

- To describe the natural history of disease
- To determine the clinical effectiveness or cost-effectiveness of healthcare products and services
- To measure or monitor safety and harm, and/or
- To measure quality of care

1.4.2 Registry of monoclonal gammopathies specifics

A registry of monoclonnal gammopathies (RMG) is a disease registry, i.e. it contains data of patients who have been diagnosed with one of the diseases belonging to monoclonal gammopathies. The registry typically enrols patients at the time of a routine healthcare service. Depending on the scope of the collected data, the registry may serve one or more purposes (see above) and provide valuable information for different stakeholders.

The data are collected in a 'real-life' clinical practice, the patient care is determined by the clinician and the patient (i.e. not assigned by the registry protocol).

The typical properties of a RMG include:

- The data are collected in a 'real-life' clinical practice, the patient care is determined by the clinician and the patient (i.e. not assigned by the registry protocol).
- The RMG is one of the few registries monitored similarly to clinical trials.
- The scope of data in the registry and the registry design is defined to fulfil specific purposes (data collection is purpose-driven) and/or there is a broader scope of data to answer possible new questions (purpose/question being data-driven).

- The RMG collects data with the same standardised data definitions.
- Data for each patient are collected uniformly; this applies to data type, scope and frequency.
- The registry can be matched with external databases (e.g. a mortality registry) or with other data sources to explore new questions.
- Studies derived from the RMG resemble traditional observational cohort studies (the collection of detailed clinical and longitudinal follow-up data). New studies (e.g. cluster-randomised studies or case-control studies) may be nested within the ongoing registry.
- Data from the registry are used by experts in the preparation of national guidelines.
- The RMG is also newly used in the assessment of the effect of new drugs by health insurers in the process of determining the reimbursement.

2. PHASE 1: WHAT TO DO BEFORE START OF THE REGISTRY

2.1 To make a good plan is the key to build up a registry

Making a good plan is very important before the actual start of the registry project. This phase of the project covers all steps that clarify how to set up the project, such as project description, scope of the project, project timelines, roles and responsibilities, communication methods and plan among all project participants, project budget and cost management throughout the entire project life cycle, reporting, quality standards the project should comply with, definition of risks and the risk management plan.

2.1.1 Project management plan

A project manager (PM) is responsible for the coordination and complete management of the whole project. The project management plan (PMP) represents a tool describing the main purpose(s) and scope of the project (the establishment of a new registry). The PM is obliged to build up the internal project team, to organise external cooperating members, to assign their responsibilities given for particular phases of the project and to set a sequence of goals, tasks and procedures needed for successfully meeting the project goals. Furthermore, the PMP includes a registry budget, timelines and schedule of critical milestones in the registry throughout the registry's duration.

A very useful tool for setting up a registry is a checklist for the project containing the basic steps of the PMP that should be done before starting a registry. We recommend creating the project checklist (see Table 4) as a useful guide for the initiation phase of the registry. The checklist should include the fundamental characteristics of the planned registry in terms of what type of study is under consideration: a short-term non-interventional epidemiological study (typically NIS with a duration from several months to 3–5 years) or a registry with a long-term prospect of data collection (e.g. the Czech registry RMG). If safety data of a new product or device are also gathered, the clinical registry can serve as a PASS (post-authorisation safety study). Likewise, data on treatment efficacy of a product may become the basis of a nested post-authorisation effectiveness study (PAES).

An example of clinical registry initiation checklist is shown in Table 4 below.

The basic operational elements of the initiation of a registry are:

Nominating an internal project team

The key project team members are the project manager, data manager, data analyst, pharmacovigilance manager, data entry coordinator, registry support: helpdesk, quality manager (for detailed information see the chapter Registry providers team and roles to manage the project).

Table 4 Clinical registry initiation checklist

PROJECT CHECKLIST				
PROJECT NAME	long name			
PROJECT NUMBER	generated			
PROJECT ABBREVIATION	123			
SPONSOR	XY			
CRO	XX			
PROJECT MANAGER	PM name			
LANGUAGE	English			
START DATE OF THE PROJECT	FPI estimated date			
STEP	**SPECIFICATION**			**COMMENTS**
	YES	NO	N/A	
Type of project				
Non-interventional epidemiological	☐	☐	☐	
Non-interventional scientific (academic)	☐	☐	☐	
Non-interventional pharmacoeconomics	☐	☐	☐	
Was the study ordered by state authority? (If "yes", then in most cases it is PAES)	☐	☐	☐	
Definition of target population	☐	☐	☐	
Definition of primary hypothesis (main project objective – scientific target)	☐	☐	☐	
Definition of secondary hypothesis (What other questions should the analysis answer?)	☐	☐	☐	
Definition of the number of subjects to correctly verify the primary hypothesis or clinically significant difference / Power Analysis	☐	☐	☐	
Definition of basic endpoints (variables) including the choice of appropriate methodology	☐	☐	☐	
Budget	☐	☐	☐	
Contract with sponsor (specify the type in comments): gift / work contract / grant	☐	☐	☐	
Contract/contracts with investigator/ investigators	☐	☐	☐	
Protocol of the study/registry	☐	☐	☐	
Informed consent form	☐	☐	☐	
Ethics Committee approval	☐	☐	☐	
LEC	☐	☐	☐	
MEC	☐	☐	☐	
Notification to regulatory authority	☐	☐	☐	

Establishing a scientific board for the registry

The engagement of key healthcare specialists and scientists, so-called "opinion makers", into the project, setting up a scientific board for the registry and involving a professional guarantee are critical steps for defining and fulfilling the registry's goals (for detailed information see the chapter Stakeholders and experts of the registry).

Main objectives of the registry

As a starting point for a new registry programme, it is critical to "begin with the end in mind". In other words, what are the basic research questions the registry should answer. In the registries, these questions are present as primary and secondary objectives. They should be transferred into a measureable set of variables with well-defined outcomes, including the assessment of a minimal sample size of collected data. The help of a medical expert is necessary. Another very important person in evaluating whether the defined variables can be analytically processed to meet the objectives is a statistician, who should validate the data set prior to the build-up of the complete case report form (CRF) (for detailed information see the chapter Scope of the registry).

Data elements (definition of basic endpoints)

A key step consists of setting up the collected variables for the registry. For a proper and long-term operation of the registry, it is undesirable and unhealthy to collect all the data with negligible research potential, as there is a risk of fatigue on the side of small centres and their gradual reluctance to enter the data into the registry. Therefore, it is necessary to compile a balanced data set with regard to the expected outcomes from the registry, the long-term sustainability of the registry, the cost of data collecting and the needs of potential sponsors. This data set should represent mandatory variables in the registry. In addition to these mandatory items, other interesting data could be collected, but they will not be essential ones. It is recommended to perform a pilot test of the main data set prior to the start of the registry in selected centres that shall be considered as participating centres in the future. Another topic of collected data involves patient-reported outcome measures (PROMs) taken directly from patients without interpretation by clinicians. PROMs can contribute to a better understanding of patients' experiences and outcomes during therapy, may depict real-life information on the patients' quality of life, form the feeling of personal involvement and facilitate joint decision-making with the patient during the follow-up visit [13]. The means of collecting PROMs data (paper-based→ digitisation, electronic, web-based) should be considered, as well as the frequency of collecting them and the type of questionnaire used to measure the quality of life. Many types of those questionnaires are available and validated; one of the most frequently used ones is SF36, for example.

Planning follow-ups based on the registry objectives

According to the main objectives of the registry, the frequency, length and extent of the follow-ups shall be planned in such a way as to retain the patients in the registry and to set adequate, but not unnecessarily frequent, periods of follow-up to address the main objectives.

Statistical Analysis Plan (SAP)

A statistical analysis plan (SAP) is critical in each research project, including registries. The SAP should contain defined guidelines for analysis and data presentation at various stages of the registry's life. The SAP identifies how the outcomes will be measured, defines the statistical tests that will be used for data analysis, provides key variables, endpoints and covariates, specifies the sample size and patient subgroups of interest etc. (for detailed information see the chapter Statistical analysis plan (SAP)).

Methods for data collection and control

Once the data set and source of data (clinicians, patients) are defined, the most suitable instrument for data collection should be selected. The most preferable means of data collection is through the Electronic Data Capture (EDC) system into eCRF (electronic case report form) with accessible via an Internet portal. The EDC system allows the collection, cleaning, storing, monitoring and reporting of registry data; therefore, it is the determining factor for meeting the goals of the registry. In some cases, a paper CRF can also be useful, e.g. PROMs in elderly patients with limited computer skills or with reluctance to use modern technologies. Registries may also integrate patient diaries, mail and telephone surveys with automatic communications and a centralised call centre to collect all of the requested information in a timely manner.

The validation tools are supposed to be included in the EDC system to cover the data quality (completeness, consistence, no contradiction). We recommend including tools for controlling the completeness of the primary data collection to assess and report the extent of missing data at the key points of follow-up for data elements that are critical to addressing the primary study questions (for detailed information see the chapter EDC system).

Target sites, physicians, patient population

It is necessary to get an overview of the healthcare facilities (hospitals and physician practice/sites) for a given disease/group of diseases in the target area. This is related to another parameter, namely the geographic scope of a registry (local or international). With regard to the disease in question, it is important to define a target patient group or, alternatively, to divide patients into cohorts according to the severity of the disease or therapeutic regimens. It is also necessary to consider whether the registry should have a short-term course or whether continuous data collection is planned. Patients should be enrolled systematically and followed-up using similar procedures in all participating centres. The loss of a follow-up should be monitored and minimised so that the realisation of the main objectives remains possible (for detailed information see the chapter Definition of target population).

Adverse events reporting

Will the pharmacovigilance data be involved in the registry? Registries can function as a post-authorisation safety studies for sponsoring pharmaceutical companies participating in the project. It is necessary to define the sort of pharmacovigilance data, frequency and form of reporting for adverse events / severe adverse events (AE/SAE) and the subjects that will be addressed. The settings of AE/SAE reconciliation in the project have to be made (for detailed information see the chapter Pharmacovigilance).

Protocol – essential document of the registry

The study/registry protocol forms an essential part of a registry project. It is a full description of the project and will act as a "manual" for members of the research team to ensure adherence to the methods outlined. It is important to document all critical changes in the registry and to inform all stakeholders about these changes. If any modification is needed, the amendment has to be prepared and justified (for detailed information see the chapter Protocol).

Contracting process

For its existence, the registry needs financial support. The conditions of sponsorship are given in a contract with the sponsors of the project where the exact conditions of the financial relationship are set. The process of revising and approving the contracts with the assignment of responsibilities to the personnel must be set. This demand is also applied to the investigator's contract, including settings of remuneration with regard to a fair market value (for detailed information see the chapters Cost management plan and Funding sources).

Regulatory and ethical requirements

The registry is supposed to be registered by the national regulatory authority in case that national legal regulations apply. The essential precondition for patient data entry to the registry is a signed informed consent form, which symbolises the patient's voluntary decision to participate in the project and to provide his/her personal and clinical data. The informed consent form is mandatory and has to be approved together with the study protocol by a multicentre/local ethics committee. In case that an international registry is being organised, approval of the registry before its launch has to be done in each participating state and must meet the local legal requirements (for detailed information see the chapter Regulatory and ethical feasibility).

Registry budget, cash flow

The Project manager is to prepare the overall budget of the registry, which has to take into consideration the scope and complexity of the registry, requested outcomes, numbers and arduousness of analyses, pharmacovigilance requirements, geographical character of the project and all of the above-mentioned activities (for detailed information see the chapter Cost management plan).

Registry graphics

The requirements for processing the project may comprise a global project visual identity, including graphical elements such as the project logo, colour scheme, fonts, templates for MS Office, official project website, analytical reports and summary documents, posters, newsletters etc. All these elements should be described in detail by visual guidelines, where the usage rules of the communication elements aimed at promoting the visual identity should be introduced. It is crucial that the visual appearance reflects the global communication and marketing strategy of the given project. Firstly, a logo (the key element of each project), primary colour palette (for printing and digital media) and typography (comprehensive set of fonts) must be designed in order to ensure the following harmony of the project's visibility. The final visual identity manual contains the approved standard graphic elements and their general usage in practice. Based on

that, all other items, such as a standard stationery system of MS Office templates, can be easily created. Finally, a project website should be designed, developed and implemented. A typical website shows project activities and outputs in the form of basic information (static HTML text and PDF reports) together with interactive presentations of the available data, preferably using modern data mining and visualisation technologies.

2.1.2 Staff and communication management plan

The purpose of the communications management plan is to define the communication requirements for the registry and how information will be distributed. The communications management plan defines the following:

- What information will be communicated–to include the level of detail and format, e.g. reports of patient enrolment to the registry, reports of data completeness etc.
- How the information will be communicated–in meetings, email, telephone, web portal etc.
- When information will be distributed–the frequency of project communications both formal and informal
- Who is responsible for communicating project information
- Communication requirements for all project stakeholders
- How any sensitive or confidential information is communicated and who must authorise this
- How the communication process and the changes in communication are managed
- The flow of project communications
- Any constraints, internal or external, which affect project communications
- Any standard templates, formats or documents the project must use for communicating
- An escalation process for resolving any communication-based conflicts or issues

The project managers should give considerable thought to how they want to manage communications for the registry with all participating stakeholders. By having a solid communications management approach, they will soon find that many project management problems can be avoided. Most projects consist of a broad range of stakeholders, all of whom may have differing interests and influences on the project. As such, it is important for project teams to determine the communication requirements of these stakeholders in order to communicate project information as effectively as possible. There are a number of methods for determining stakeholder communication requirements; however, it is imperative that they are completely understood in order to effectively manage their interest, expectations and influence and to ensure a successful project.

As mentioned in the chapter Project management plan, an internal project team for the registry must be established. A very powerful tool for communication within the internal registry team is the use of professional project management software. It can manage the estimation and planning of tasks, scheduling, cost control and budget management, resource allocation, collaboration, communication etc. You can enter all the steps of initiation for the registry into the software, add the appropriate members of team to tasks, estimate the planned time and monitor the fulfilment of individual tasks. Under the same circumstances, you can follow the registry during its entire life cycle.

A critical step is the kick-off meeting where all prospective stakeholders discuss all issues necessary for the successful and reasonable initiation of the registry (see the chapter Stakeholders and experts of the registry). The stakeholders of the registry should define the design of the registry, research objectives, rules for patient recruitment, clinical centre recruitment, funding, data privacy and data access at the beginning of the registry. This broader advisory board should meet on a regular basis and more frequently upon the start of the registry.

One very important step is to make collective training for all the participating centres at the start of the project and pay attention to any user discomforts with entering data into the registry. Each new centre joining the registry has to be individually trained personally or by video teleconferencing (for detailed information see the chapter Training).

Last but not least, we recommend organising regular meetings of the advisory board twice per year at a minimum to exchange the first experiences, to evaluate the EDC system and the CRF, to assess patient enrolment and time consumption and to approve sources of funding and the budget of the registry. After some time, it will be the right moment to evaluate the first results from the registry.

The Steering Committee (see the chapter Professional society) may be in narrow contact with the registry project team. The connecting person is the project manager, who coordinates most activities linked to the registry. This governance group of the registry is obligated to communicate in any case of data access (see the chapter Data privacy and confidentiality).

Guidelines for meetings

Meeting agenda

The Meeting agenda should be distributed 10–15 business days in advance of the meeting. The agenda should identify the presenter for each topic along with a time limit for that topic. The first item in the agenda should be a review of the action items from the previous meeting.

Meeting minutes

Meeting minutes should be distributed within 5 business days following the meeting. Meeting minutes should include the status of all items from the agenda along with new action items and the list of attending participants.

Action items

Action items are recorded in both the meeting agenda and minutes. Action items should include both the action item along with the owner of the action item. Meetings should start with a review of the status of all action items from previous meetings and end with a review of all new action items resulting from the meeting. The review of the new action items should include the owner identified for each action item.

Meeting chairperson

The chairperson (mainly the project manager) is responsible for distributing the meeting agenda, facilitating the meeting and distributing the meeting minutes. The chairperson will ensure that the meeting starts and ends on time and that all presenters adhere to their allocated time frames.

Notetaker

The notetaker is responsible for documenting the proceedings of all meeting points and taking notes of anything else of importance during the meeting. The notetaker will give a copy of his/her notes to the chairperson at the end of the meeting, as the chairperson will use the notes to create the meeting minutes.

There are many ways to present registry data by means of different media, including abstracts, posters, podium presentations for scientific meetings and symposia, manuscripts for publication in clinical journals, dossiers for reimbursement authorities and speaker bureau/physician-support activities. The overall goal of any registry programme is to continually produce reports and other communications that keep the registry outputs fresh and relevant to the participating physicians and patients, thereby encouraging their continued engagement in the registry programme. In addition, registry findings circulated through publications, presentations or word of mouth may sway the prescribing behaviour of clinicians or the formulary decision-making of managed care organisations or reimbursement authorities which have never had any direct contact with the registry.

2.1.3 Cost management plan

The cost management plan clearly defines how the costs on the project are managed through the entire project life cycle, as well as the responsibilities for managing costs and cost changes approvals. Moreover, this includes the set up of the reporting, its frequency and presentations.

Regarding the clinical registry cost management plan, the responsible person is project manager, who manages and reports on projects costs and manages the project activities for the whole team to keep the project on budget.

There are several possibilities when making a budget for the clinical registry. The most transparent and easily manageable is a budget based on the estimated time and hourly rate for the activities and required deliverables. To create this type of budget, the cooperation of the whole team is needed as precision is very important for budget estimation. To make a clear and precise budget, a budget template can help. The budget template should be complex and should specify all possible individual activities that may appear in the sponsor's requirements.

The budget template may contain a section with detailed specification of the activity. To avoid misunderstanding regarding the scope of activity It is advisable to have a description of the actions.

An example of clinical registry activities specification is shown in Table 5.

Table 5 Budget template - Registry activities specification

Study/registry specification	**Project name**
Activity	Activity description
Clinical project management	
Study design	Definition of registry/study type, therapeutic area, definition of duties and responsibilities of each study side, definition of study budget. If a sample size is required, the analyst will count the power analysis.
Protocol	Protocol creation. Cooperation with the department of data analysis and the department of data management (DM).
Clinical data management	
Data management documentation	Preparation of documentation (list of requests to DM, data management plan, data validation plan). Sponsor approves DMP and DVP.
Database preparation and setup (eCRF) including validation rules	eCRF preparation, setup of validation rules, DM testing in advance to the final release of eCRF. Plan of the medical coding process, list of coded variables, setup of format, mean and frequency of data transfer. In advance of the final release, the sponsor is to test and to approve eCRF.
Helpdesk	Technical support for the entire project life cycle (problems with access, data entry, communication with centres etc.)
Statistics and programming	
Statistical programming	Programming in statistical software – data structure and derivation, outputs (tables, figures, listings).
Preparation of reports	Preparation of results. Prepared codes are used. Statistical report including the description of statistical methods, results and statistical interpretation.
IT and EDC system	
Maintenance of the EDC system	Regular upgrade of SW and release. Incidents and problem solving (SQL server, web server, operational systems, web browser).
Hosting and backup of the EDC system	Upgrades and maintenance of HW. Server hosting. Regular backup of data.
Project special activities	
Remote patients' access	Remote patients' access to enter data (QoL questionnaire filling, patients diaries completion etc.)

Precise budgeting at the beginning of the project is crucial, because increasing the budget during the project might be quite difficult. Another very important aspect is to keep the project on budget throughout the entire registry life cycle and to avoid going over the budget.

The main reasons for going over the budget may involve incorrect budgeting at the beginning of the project, inappropriate resourcing, both internal and external miscommunication, slow execution, incorrect deliverables and following corrections, which increase time spent as well as the costs of the activity [36].

During the budgeting process, it is very important to delegate people into the project team and to plan the number of executive persons for the entire project life cycle. Having an insufficient number of people faces the risk that the project goes on slowly and the interim and total timelines are not met. On the other hand, if the project is overstaffed, it is more likely to go over the budget. See more on project risks in the chapter Risk management plan.

Keeping the project on budget also depends on the project manager and his/her capabilities and experience in leading a project team, especially to accurately plan the team's activities to be well-executed with respect to the timeline and deliverables. It should be done in combination with correct and regular communication within the project team. Effective communication always saves time and resources. The project manager organises regular project meetings with the internal team and with the client in order to keep communication as effective as possible. It is very useful to use an internal software tool to manage the project, to see the current status and progress of the individual project activities, to have evidence of time spent on the activities and related budget, to share documents and records, to communicate issues etc.

An integral part of cost management is reporting, its frequency and presentation. It is suitable to revise the project's budget continuously and make internal reporting on a regular basis, for example once a month. In case of a budget risk, the detailed revision of processes should start to reveal the cause. Detailed reporting to the sponsor is done upon the sponsor's request or regularly, for example on a yearly basis or at the end of the contract with the sponsor. The contract with the sponsor is usually undertaken for one year. Revision of contractual relationships often leads to reviews of processes within the registry, such as a new form of reporting etc. (for more information see the chapter Risk management plan).

The clinical registry budget consists of the operational costs and costs spent on outsourced work, the so-called pass-through costs. Operational costs are costs spent on the project management, data management activities, data analyses, reporting setup, IT solution and EDC system, pharmacovigilance (if required) and the quality management of the registry. Operational costs vary depending on the project phase. The initial phase is the most costly because it covers all of the settings of the registry, from the plan and feasibility to database setup for the first data entry. The subsequent phase, registry operation, covers the maintenance of EDC system and common agenda within the project, data management, data analyses, regular reporting, communication etc. Pass-through costs are costs covering payment to investigators or data managers who enter the data into the registry database. Pass-through costs cover also costs for other outsourced work, for example to a translation agency, ethics committee fees or other fees.

An example of budget template for operational costs is shown in Table 6.

Table 6 Budget template for operational costs

Workload and costs		**Project name**		
Activity	Number of units	Unit	Unit cost (EUR)	Total cost (EUR)
Protocol	60	hour	80.00	4,800.00

An example of a budget template for pass-through costs is in Table 7.

Table 7 Budget template for pass-through costs

Pass-through costs		**Project name**		
Activity	Number of units	Unit	Unit cost (EUR)	Total pass-through cost – estimated (EUR)
Cooperation with medical expert/medical writer	40	hour	95.00	3,800.00
Translation agency cooperation		page		
Fee for EC approval		approval		
Fee for RA approval		approval		
Investigator's fee		fee / valid form		
Costs relating to the payment of the investi-gator's fee (bank charges, exchange losses, postage etc.)		per payment		
QoL questionnaire		licence		
Medical coding dictionary		licence		

The division of the funding between operational and pass-through costs is shown in the following scheme in Figure 2.

Operational costs in the first year (initial costs) are paid only once per clinical registry. Operational costs to keep the registry running are paid every year.

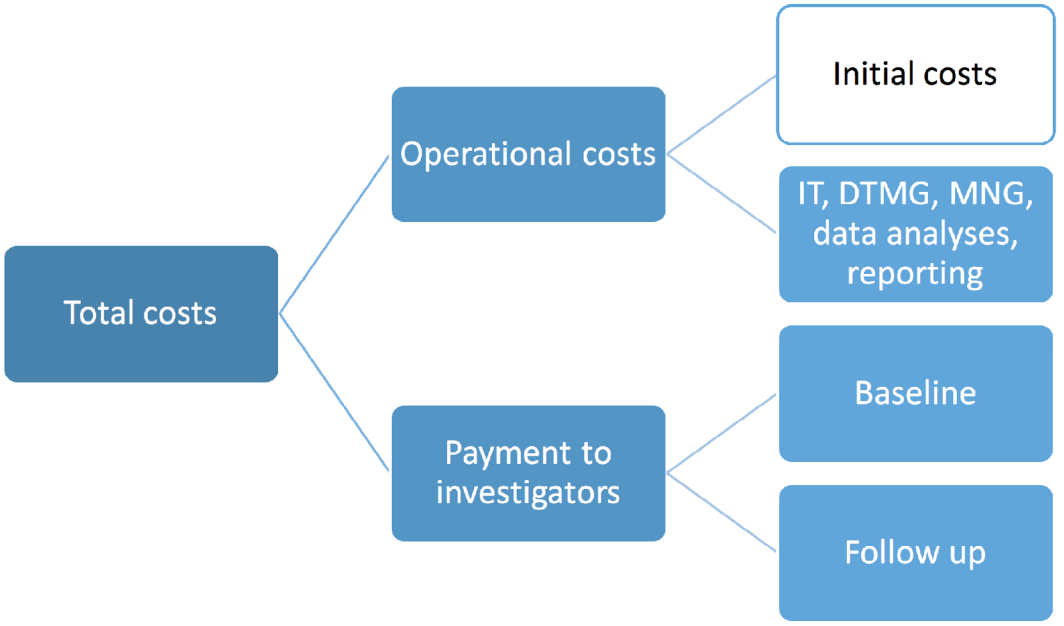


Figure 2 Budget structure

Payments to investigators are designed according to the registry structure and the number of forms that are filled with the patient's data. Each patient has a baseline form that contains data about the patient's history, status, current medication, comorbidities, concomitant medication etc. Each individual patient visit in the clinical centre is recorded in the follow-up form that contains data about their status, medication, switch to the other medication, medication interruption etc. The fees for an investigator for filling out the baseline form and depend on how difficult and time-consuming it is to fill out the form.

The baseline form fee is counted once per every individual patient entering the registry (fee per baseline form × number of patients recruited by investigator). The follow-up form fee depends on the number of visits in defined period of time (number of patients × number of follow-ups per period of time [quarterly, semi-annually, annually]).

The setting of rules for determining the remuneration of investigators in accordance with the fair market value process is one of the key areas of project setup and management. The investigator or data manager who enters data into the registry is responsible for fully completing all forms of the registry (eCRF) according to the preset validation criteria and shall be remunerated for this service. The remuneration provided to the investigator or data manager must be adequate to the services provided according to the so-called fair market value (FMV). A clinical registry must not serve as an incentive to prescribe or to recommend a medication. The remuneration of the investigator or data manager must be adequate to the work performed. When data are entered into the registry, the remuneration for filling out one form or multiple forms of a patient (according to the rules set up in the registry) must correspond to the time needed to properly fill out the form while also taking into account some other criteria.

It is the project manager's responsibility to determine the final remuneration of the investigator or data manager for filling out the form(s). This remuneration is specified in the contract concluded with the centres / investigator / data manager (for more information see the chapter Site feasibility).

The specific value corresponding to the fair market value is not given anywhere, as it cannot be determined globally. In the case of clinical registries, FMV includes only financial compensation for time and costs and must not promote the use of the medicinal product or prompt any such promotion.

When determining the amount of remuneration corresponding to the FMV, the project manager takes into account:

- Measured time needed to fill out the form(s)
- Average hourly wage of the investigator or data manager
- Customary practices in a particular therapeutic area
- Time needed to look up data in the patient's medical records (according to the length and scope of the eCRF, this section may be consulted with the expert guarantor of the project)
- Place of activity and customs (e.g. country)

It is essential that the amount of the remuneration is fair, is in accordance with the legitimate and expert knowledge and services provided by the investigator or data manager.

The project manager consults the determined value with the sponsor or the registry guarantor and requests the approval of the amount. The project manager documents the FMV determination process in the registry file.

The clinical registry is a long-term project, so regular finance resources are needed for setup at the beginning of the project with an outlook towards long-term financing. This should be considered when the contracts with sponsors are negotiated. The most advantageous option is a multi-financial approach when different financial sources are used: grants from the government, insurance companies, pharmaceutical companies, other industries or patient organisations (for detailed information see the chapter Funding sources).

2.1.4 Quality management plan

The quality management plan (QMP) is conducted component of the project management plan that describes how the organisation's quality policies will be implemented. It describes how the project management plans meet the quality requirements set for the project [37].

The quality management plan describes which quality control and quality assurance activities concerning the NIS have to be conducted. Thereby all quality relevant tasks are defined, concrete quality objectives are described and responsibilities for single tasks are named. Beyond this, the resources which need to be at disposal are defined accordingly [38].

The purpose of this plan is to:

- Ensure that quality is planned
- Define how quality will be managed
- Define quality assurance activities
- Define acceptable quality standards

Quality must always be planned into a project in order to prevent unnecessary rework, waste, costs and time.

The QMP is a written document that describes the system in place to implement the quality management programme. The preparation of the QMP is a multistep process involving a number of people participating in the project.

The quality management plan should include

- Implementation of the relevant standards
- Process documentation
- Process implementation
- Quality objectives
- Resources identification and their verification
- Project documentation
- Process for risk identification and risk management
- Internal audits – plan of internal audits
- Metrics settings, conducting measurement and evaluation of the results
- Responsibilities
- Time plan

Quality management

The first step in quality management is to define all processes that are relevant for registry management throughout the entire life cycle of the registry. We have to manage process settlement by the relevant rules and requirements for registry management.

In general, clinical studies have strict rules referred as GCP (Good Clinical Practice), whereas non-interventional studies have more general rules referred as GPP (Good Pharmacoepidemiology Practice). The first guidelines for GPP were issued by the ISPE (International Society for Pharmacoepidemiology) in 1996. The latest revision (number 3) was published in June 2015.

The organisation suitable for registry management introduces standards that ensure a high quality of provided services, final data and results.

The relevant standards for quality and security management in the field of registries execution are

- ISO/IEC 9001:2015 Quality Management System (QMS) – basic international standard describing quality management. The main objectives of this standard are to enhance customer satisfaction based on the understanding of customer needs and customer-specific requirements, risk-based thinking and the continuous improvement of provided services.
- ISO/IEC 27001:2013 Information Security Management System (ISMS) – the application of this international standard ensures data confidentiality, integrity and availability.
- ISO/IEC 20000-1:2011 Information Technology Service Management (ITSM) – the application of this standard ensures the quality of provided services in the field of information technologies.

All of the above-mentioned standards, including GPP, are based on process approach and management.

Process management

Process approach is a term that comes from the new, updated standards of ISO 9000 and the subsequent standards. It means that all steps and aspects leading to the outputs must be identified, described and documented. Processes are managed and controlled, and the interactions between these processes are identified and managed, too.

The specification and objectives of the registry are defined in the project documentation / project protocol accordingly.

Process management is illustrated in the following scheme in Figure 3.

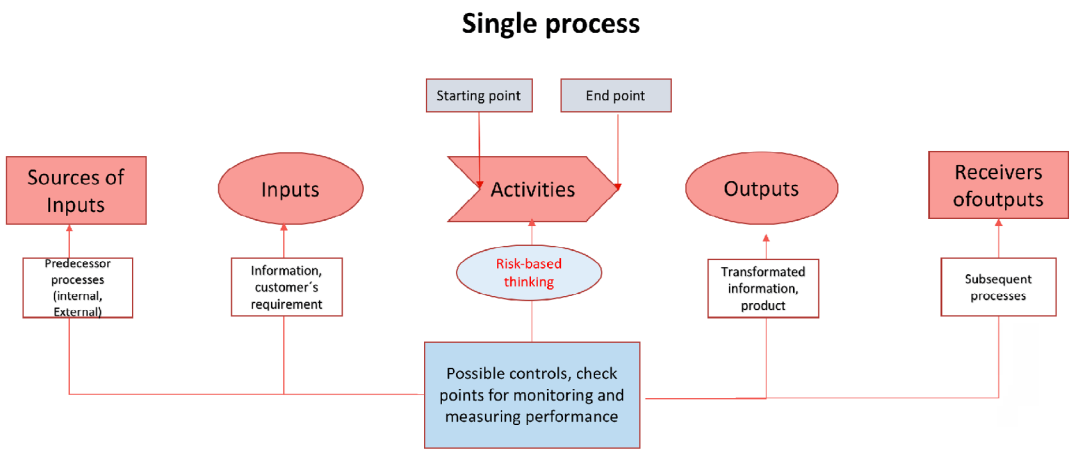


Figure 3 Schematic representation of single elements in a single process

These processes are divided into main and supported processes and are shown in the following scheme in Figure 4.

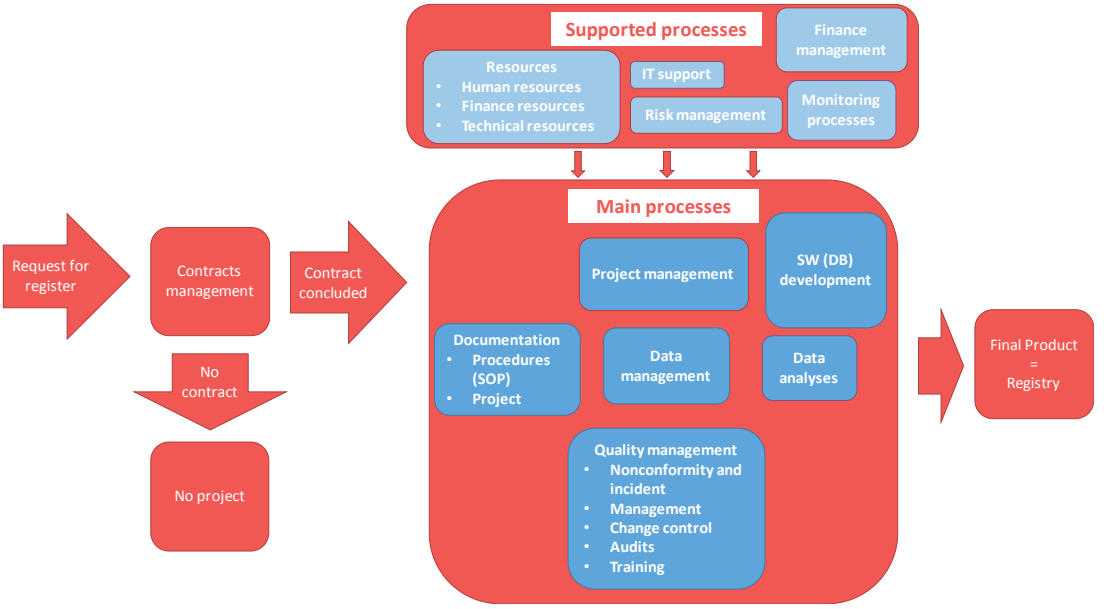


Figure 4 Main supported processes in registry management

All processes must be stated and documented in the internal standards (standard operating procedures) to ensure that each process will be done the same way every time by everyone, leading to the same results. The registry operation consists of many processes such as contractual management, database development, data management, data analyses, training of employees and other involved parties, monitoring of deviations, nonconformity and incident management, change control etc. All sub-processes are interconnected and interrelated. These interrelationships are represented in the process map.

According to ISO 9001, it is necessary to apply the PDCA cycle (or Deming cycle) of continual improvement for every process.

According to PDCA, the cycle consists of four basic parts (or phases), which are linked to each other:

P = plan phase

During this phase, the objectives and processes necessary to deliver results in accordance with the expected output (the target or goals) are established.

D = do phase

This phase covers the implementation of the plan, execution of the process and making the product. The collection of data for subsequent checks, monitoring and further action steps are also included.

C = check phase

During this phase, the results (data, information from previous phase) are checked and compared against the expected results (targets, goals from plan phase). Part of this phase consists of finding and identifying deviations and their sources.

A = act

This phase covers introduction of actions that help to improve the process. Actions will be planned, done, checked and so on.

It is clear that the PDCA cycle enables project team members to ensure that processes are adequately resourced and managed and that opportunities for improvement are determined and acted on.

All processes should be monitored and measured to provide objective results on the process effectiveness and to determine if there is any place for improvement. The PDCA cycle is illustrated in Figure 5.

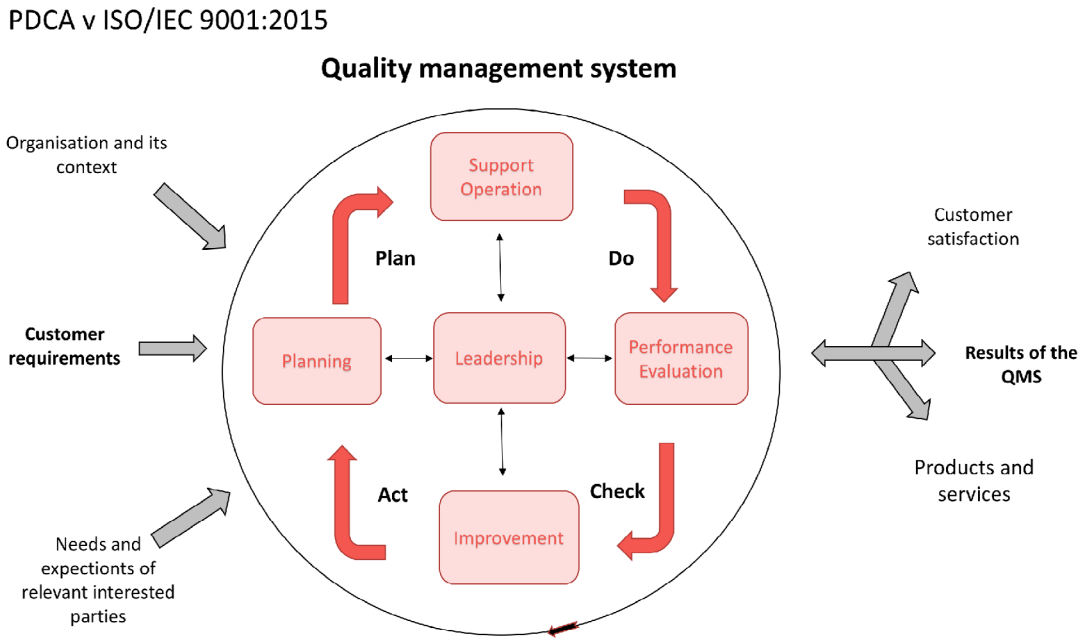


Figure 5 PDCA cycle (Deming cycle) of continual improvement

It is very important that the quality management system considers both real and potential risks. For more details, see the chapter Risk management plan.

GPP requirements

The GPP proposes the essential practices and procedures that should be considered to help ensure the quality and integrity of pharmacoepidemiological research and to provide adequate documentation of the research methods and results. The GPP do not prescribe specific research methods, nor will adherence to guidelines guarantee valid research [39].

The guidelines for GPP principles are applied to a wide range of studies

- Validation studies
- Comparative effectiveness research

- Benefit and risks studies
- Descriptive studies
- Post-authorisation safety studies
- The reasons for introducing GPP rules to the quality system into an organisation which operates registries consist of the following
- To assist researchers in adhering to good pharmacoepidemiological research principles, including the use of pharmacoepidemiological studies for risk management activities
- To promote sound pharmacoepidemiological research by encouraging rigorous data collection, analysis and reporting
- To provide a framework for conducting and evaluating pharmacoepidemiological studies
- To facilitate the appropriate utilisation of technical resources by promoting careful study/registry design and planning of study conduct
- To facilitate transparency and ethical integrity in research conduct

The GPP Guidelines address six fundamental areas. These areas are displayed in Figure 6

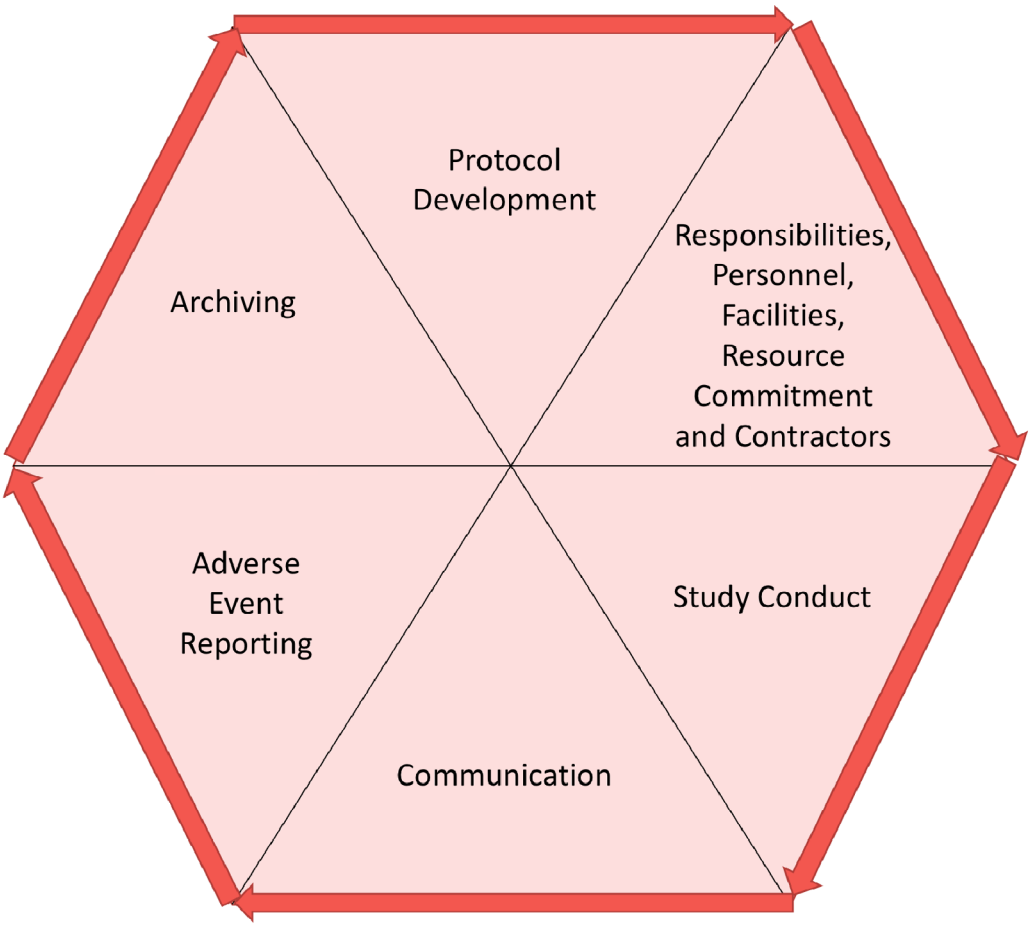


Figure 6 Six fundamental areas according GPP guidelines

1. Protocol development

At the beginning of each study/registry project, the protocol should be written. During the study life cycle, the protocol is amended or updated as needed. Some jurisdiction, such as the EU through EMA or ENCPePP, provides guidance on the content or the format of the protocol (for detailed information see the chapter Protocol).

2. Responsibilities, personnel, facilities, resource commitment and contractors

a. Responsibilities: The organisation(s) and individual(s) conducting and sponsoring the research shall be fully responsible for the research. The relationship, roles and responsibilities of the organisations and/or individuals conducting and sponsoring the study should be described. In the case of collaboration with a CRO, a contractual arrangement has to be signed and should include the timeline for study completion and potential action to be taken if the timeline cannot be met.

b. Personnel: Personnel participating in the study should have appropriate and verifiable education, training and experiences. The organisation should maintain updated training records (see the chapter Registry provider's team and roles in managing the project)

c. Facilities – adequate facilities for conducting research and storing records should be ensured.

d. The study sponsor should have the right to inspect the contractor's facilities to ensure conformance with the GPP. This possibility should be stated in the contract.

3. Study conduct

The principal investigator shall be responsible for the overall content of the research project, including the day-to-day conduct of the study, interpretation of the study data, and preparation and publication of the final report.

The study conduct section is divided into 4 parts:

a. Protection of human subjects: before human subjects will be included in the study, approval by an Institutional Review Board (IRB) and Independent Ethics Committee should be obtained. An explicit informed consent form must be signed by each human subject and personal data protection should be ensured. Investigators shall ensure that personal identifiers will be removed from any study files that are accessible to non-study personnel in accordance with the applicable laws and regulations and code keys will store the data separately (for more information see the chapter Regulatory and ethical feasibility).

b. Data collection, management and verification: All data collected for the study should be recorded accurately, promptly and legibly. All procedures used to obtain, verify and promote data quality and integrity should be recorded in sufficient detail so that others can replicate them. A historical file of these procedures shall be maintained, including all revisions and the dates of such revisions. A control

system should be established and adequate data backup should be ensured (for complete information about the data management process see Phase 2: How to create the registry and Phase 3: How to keep the registry alive).

c. Analyses
   - A statistical analyses plan with the statistical procedures should be clearly defined (for detailed information see the chapter Statistical analysis plan (SAP)).
   - All data management and statistical analysis programmes and packages used in the analyses should be documented and archived.
   - The analysis should be directed toward an unbiased estimation of the epidemiological parameters of interest (e.g. risk or rate differences, risk or rate ratios). The precision of the effect estimates should be quantified using confidence intervals (for complete information about the data management process see Phase 4: How to analyse MM data).

d. Project Reports: during the project's life cycle, interim reports can be issued (it must be stated in the protocol). Every project has to be summarised in a final report.

The final report should contain:

- A descriptive title
- Abstract
- General purpose of the study (according to the protocol)
- The names, titles, degrees, addresses and affiliations of the principal investigator and all co-investigators
- Name and address of each sponsor
- Name and address of CRO (if applicable)
- Important milestones of the project (dates for the beginning and completion the study)
- Specific aims and objectives of the study (according to the protocol)
- Methodology description (selection of study subjects, data management methods, data transformation, statistical methods) a description of circumstances that may affect data quality or integrity
- A discussion of strengths, limitations and possible bias of the study, including direction and magnitude of bias, if known
- A statement of the conclusions drawn from data analyses
- A discussion of the project results (prior research versus present findings; possible biases and limitations etc.
- Acknowledgements
- References

4. Communication

The organisation should establish procedures under which communications of the intent, conduct, results and interpretation of an epidemiological study will occur. The means and scope of communication should be stated in the internal standards, study protocol or contractual agreement (for detailed information see the chapter Staff and communication management plan).

5. Adverse event reporting

Adverse events can occur whenever during pharmacoepidemiological study. The sponsor should define procedures for reporting adverse events at the time of protocol development (for detailed information see the chapter Pharmacovigilance)

6. Archiving

Procedure archiving should be established. Secure archives must be maintained for the orderly storage and expedient retrieval of all study-related materials. Access to the archives shall be controlled and limited to authorised personnel only to ensure the confidentiality of the archived information.

The length of the retention time for study-related materials depends on the national or local requirements. If not stated otherwise, the study documentation should be archived for at least five years after the final report or the first publication of study results.

The minimum requirements for archiving are the following:

- Final report
- All source data, a printed sample of the master computer data file(s) with reference to the location of the machine-readable master
- Documentation adequate to identify and locate all computer programmes and statistical procedures used, including version numbers where appropriate
- Copies of electronic versions of analytic data sets and programmes, computer printouts, if feasible, including the relevant execution code, which forms the basis of any tables, graphs, discussions or interpretations in the final report. Any manually developed calculations shall be documented on a work sheet
- Correspondence pertaining to the study, standard operating procedures, informed consent releases, copies of all relevant representative material, copies of the signed institutional review board and other external reviewer reports, as well as copies of all quality assurance reports and audits. The communication of study results to the sponsor, regulators and scientific community should be documented
- Documentation relating to the collection and processing of study data, including laboratory/research notebooks, training and reference documents for abstracts, interviews and coders [39]

2.1.5 Risk management plan

The risk management plan describes why it is essential to manage risks in order to effectively manage the project. A risk is an uncertain event or condition that, if it occurs, has a negative impact on the achievement of the project goal.

Risk management can be explained as a set of techniques, actions, plans and synchronised activities designed to help assess, monitor, prevent or mitigate any risks. It is important to carefully review the risk and find out if it is within the acceptable levels. Another key process is to share information across departments and teams [40].

A guidance was released from the European Medicines Agency (EMA) regarding risk-based approaches to clinical research titled "Reflection paper on risk based quality management in clinical trials". The basic steps to risk management include the following: identify risks, analyse risks, evaluate risks and treat and review risks. Additional steps that are required throughout the process include the communication of risks and documentation activities [41].

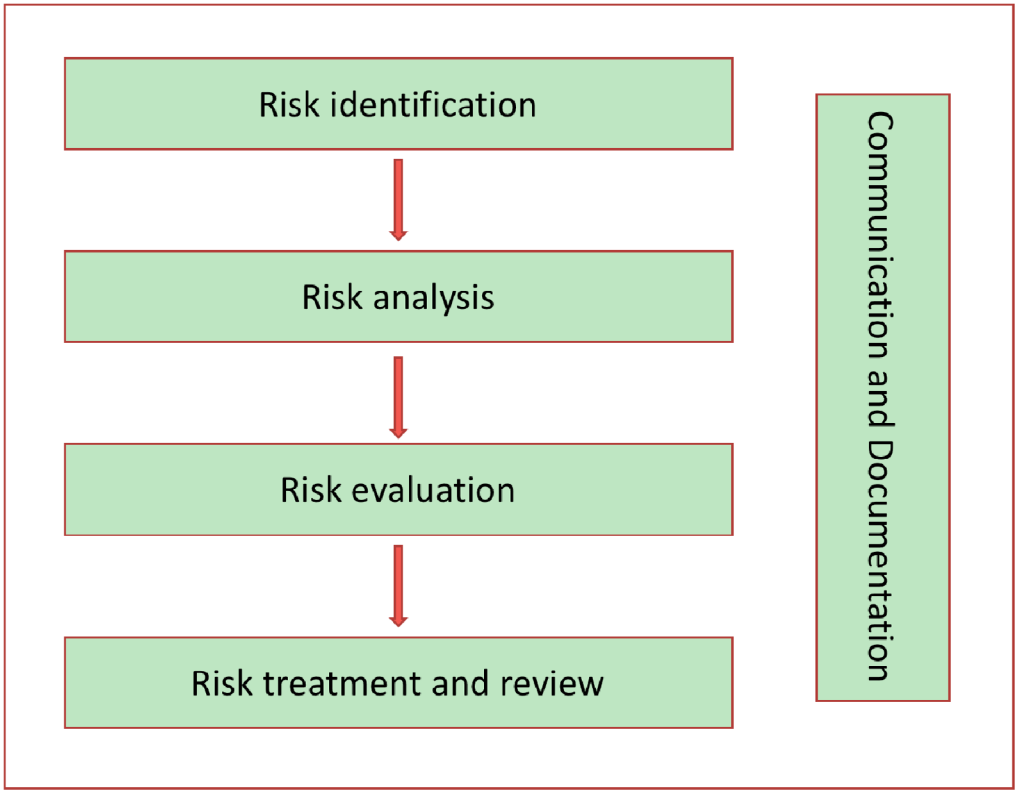


Figure 7 Flowchart of the risk management process [42]

Risks accompany each component of the registry establishment and maintenance process, from design, preparation, conduct, analysis and the use of results.

It is recommended that an extensive planning process be undertaken under the guidance of experts familiar with the process of registry design and with stakeholders [43].

The sponsor should include all stakeholders in the risk management process and define the rules, responsibilities and accountabilities of the different stakeholders beforehand.

Risk identification is the process of finding, recognising and recording risks. Risk identification addresses the question: “What might go wrong?” Risk identification relies on the brainstorming of stakeholders [43].

Risk analysis is "the process to comprehend the nature of risk and to determine the level of risk". Risk analysis consists of determining the causes, the consequences and their probabilities for identified risk events in a specific study [43].

Risk evaluation is needed to decide which risks need treatment and to prioritise for treatment implementation [43]. Based on the risk evaluation, a decision is made on whether a risk needs treatment, whether a risk needs priority in treatment, whether a treatment activity should be undertaken and which path of treatment should be followed.

The purpose of risk treatment is to reduce the risk to an acceptable level. Once a risk treatment is implemented, there may still be residual risks. Therefore, the risk treatment must be assessed in order to decide whether residual risk levels are tolerable. If they are not tolerable, a new risk treatment must be generated and its effectiveness assessed. This is a cyclical process [43].

Selecting the most appropriate risk treatment option involves balancing the costs and efforts of implementation against the benefits derived [43].

The review of risks is a continuous process throughout the project’s cycle. It is important to involve the different stakeholders and experts in the selection of treatment actions because the acceptability of each one may be different.

When the organisation already has a quality management system implemented according to the standard ISO 9001, risk management documentation is usually integrated to the quality management process, usually in the process of continuous improvement.

In a clinical registry setup process like in other projects, it is important to be able to identify risks. To do so, some historical information from similar projects are needed. When planning and building up the clinical registry, we can identify the following risks:

- Technical risks
- Safety risks
- Costs risks
- Schedule risks
- Contractual risks
- Human resource risks
- Quality risks

Good security is essential for legal, ethical and operational considerations. It should not be underestimated. Security is easier to maintain on a dedicated server, but that also requires a backup strategy and an intrusion detection system on the server itself to avoid hackers [44].

In order to maintain the privacy of participants enrolled in a registry and the data confidentiality, security measures should be implemented. All security measures should be contained in a document that describes in detail data security risks, policies and procedures specific to that registry. Physical and technical safeguards should be incorporated in the collection, storage, transmission of and access to data. These include data encryption, restriction of data access, data backups, methods (software) for the de-identification of local data during potential transmission and storage etc. [43].

The Table 8 illustrates the examples of technical and security/safety risks in clinical registries and how risk management can be applied.

Table 8 Examples of technical and security risks

Technical or security risks	Risks identification	Risks mitigation plan
EDC system setup	EDC system is not easy and flexible to be adjusted according to the client´s and registry needs	Before searching for EDC system, it is necessary to evaluate all requirements for the features of the EDC system
EDC system robustness	EDC system is not designed for collection data in a health environment (i.e. to collect a large amount of data)	To create a must-have list of conditions for EDC system properties in order to quickly evaluate the EDC system and immediately rule out those systems that do not meet your conditions
EDC system validation	EDC system is not validated and does not meet the regulatory requirements for data collection or the validation process is not properly documented	Before decision-making, it is necessary to ensure accuracy, reliability, intended performance and the appropriate valid documentation
EDC system maintenance	There is no guaranteed maintenance of the system on a regular basis	To close a good contract with the EDC system provider and to ensure that the provider has a development plan in place to maintain and regularly update the system
Quality data management in EDC system	EDC system does not have built-in quality edit checks and queries to maintain the data quality	Define all core functionalities you require from the system and make sure that the provider offers all additional modules or integrations that may be needed now or in the future
Quality standard of EDC system	EDC system does not meet ISO standards and certifications, there are no rules and standard operation procedures	Before decision-making, it is necessary to audit that the provider has implemented and follows ISO standards and that standard operation procedures are setup in place

Technical or security risks	Risks identification	Risks mitigation plan
EDC system setup to collect personal data	EDC system is not capable of operating with personal data (GDPR compliance)	Define all core functionalities you require from the system and make sure that the provider offers all additional modules or integrations that may be needed now or in the future
EDC system pricing model	EDC system is expensive from the long-term perspective of the clinical registry operation	To close a good contract with an EDC system provider and to ensure that there is a clear and transparent pricing model setup for looking several years ahead, as the clinical registry is a long--term project
Data security in place	There is no proper data confidentiality, integrity and availability and no security plan to prevent hardware failure, viruses and human errors	To make sure that EDC system implementation is provided properly, that data are accessible and available only to authorised personnel, that the data is complete. There is spare equipment, backups, regular reviews, antivirus updates etc.
Safety data collection	There is no reporting system for adverse events	One of the clinical registry objectives is safety data collection. In this case equipment for AE reporting must be part of the EDC system and must be included in the must-have list of requirements
Safety data collection	There is no reporting system for adverse events	One of the clinical registry objectives is safety data collection. In this case equipment for AE reporting must be part of the EDC system and must be included in the must-have list of requirements

Cost risks are risks that the project costs more than the estimated budget. Exceeding costs can lead to a reduction in scope or quality. Cost risks can also lead to schedule risks if the project is prolonged in order to complete the project in time. The Table 9 illustrates examples of cost risks.

Schedule risks are risks that the project takes longer than planned. This can lead to an increase in costs (costs risks), because longer projects always cost more and lack performance if the projects are completed too late.

The Table 10 illustrates examples of schedule risks in clinical registries and how risk management can be applied.

Table 9 Examples of costs risks

Costs risks	Risks identification	Risks mitigation plan
Budget overrun	Budget is exceeded in the initial phase of the process of building the clinical registry	Detailed activity and cost planning at the beginning of the project

Costs risks	Risks identification	Risks mitigation plan
Budget overrun	Budget is exceeded in the initial phase of the process of building the clinical registry	Appropriate communication in the project team and good project management
Budget overrun	Budget is exceeded in the initial phase of the process of building the clinical registry	Conclude good contracts with sponsors and vendors
Budget overrun	Budget is exceeded in the running phase of the clinical registry	Good project management and reporting system
Budget overrun	Budget is exceeded in the running phase of the clinical registry due to high internal operational costs	Appropriate planning of people working on the project, not to overstaff the project
Budget overrun	Budget is exceeded in the running phase of the clinical registry	Multiple and appropriate resourcing overlooking several years, as the clinical registry is a long-term project
Budget overrun	Budget is exceeded in the running phase of the clinical registry due to high variable costs	Qualified guess of the number of patients entering the registry in the first year and in each subsequent year. All registry stakeholders (professional society, sponsor etc.) must be included in the discussion

Table 10 Examples of schedule risks

Schedule risks	Risks identification	Risks mitigation plan
Timelines and deadlines are not met, the project part or the whole project takes longer	Schedule is planned as optimistic (best case) rather than realistic (expected case)	Experienced project manager who sees potential obstacles in the project life cycle
Timelines and deadlines are not met, the project part or the whole project takes longer	Schedule was based on the use of specific team members, but those team members were not available at the start of the project	To have backup team members with the required skills

Schedule risks	Risks identification	Risks mitigation plan
Timelines and deadlines are not met, the project part or the whole project takes longer	A delay in one task causes cascading delays in dependent tasks	Project needs an experienced and skilful team to be able to plan and work efficiently
Timelines and deadlines are not met, the project part or the whole project takes longer	Inefficient team structure reduces productivity, lack of team members	Team structure must be defined, QMS setup in place, backup team members
Timelines and deadlines are not met, the project part or the whole project takes longer	Third-party task takes longer than expected (translations, resolving legal issues, EC/RA project approval etc.)	Experienced project manager who sees potential obstacles in the project life cycle
Timelines and deadlines are not met, the project part or the whole project takes longer	Internal personnel work slower than expected	Have experienced and well-trained personnel, efficient project management
Timelines and deadlines are not met, the project part or the whole project takes longer	Requirements are poorly defined from the project manager or sponsor (misunderstandings) or additional requirements are added	To have an experienced and skilful project manager to clear the requirement with the sponsor and setup rules of communication within the team
Timelines and deadlines are not met, the project part or the whole project takes longer	There is a chaos in communication and the handover of deliverables, the appropriate setup for procedures does not exist	To have implemented QMS and working instructions. All team members must be well trained
Timelines and deadlines are not met, the project part or the whole project takes longer	The budget is insufficient, lack of money to continue, bad budgeting	Detailed activities and cost planning at the beginning of the project, an experienced project manager
Timelines and deadlines are not met, the project part or the whole project takes longer	The budget is insufficient, lack of money to continue, bad budgeting	Detailed activities and cost planning at the beginning of the project, an experienced project manager

The Table 11 illustrates examples of contractual risks in clinical registries and how risk management can be applied.

Table 11 Examples of contractual risks

Contractual risks	Risks identification	Risks mitigation plan
Budget overrun in the initial phase of the project	Contract with the sponsor is inappropriately designed, does not cover all activities made within the project to meet project objectives	Experienced project management to define the scope of the project and specify all project activities before the contract is signed
Budget overrun in the initial phase of the project	Contract with the sponsor is inappropriately designed, wrong estimation of number of patients for recruitment, late communication with the sponsor to renew the contract for the following period of time	Experienced project management, proper communication with the sponsor regarding the long-term project
Schedule is not met, the project is prolonged	Contract signed with the vendor with inaccurate conditions, definitions of timelines and specification of activities	Experienced project management, proper communication with vendors
Bad data quality in the clinical registry to meet the study/registry objectives	Contract with the EDC system provider does not define the functionalities of the system	Experienced project management and team members to specify the requirements for EDC system functionalities
EDC system is insufficient for the clinical registry (poor functionality), EDC system does not meet quality standards	Contract with the EDC system provider does not define the functionalities of the system	Experienced project management and team members to specify the requirements for EDC system functionalities
EDC system is insufficient for the clinical registry (poor functionality), EDC system does not meet quality standards	Contract with the EDC system provider does not define the functionalities of the system	Experienced project management and team members to specify the requirements for EDC system functionalities

The Table 12 illustrates examples of human resources risks in clinical registries and how risk management can be applied.

Table 12 Examples of human resources risks

Human resources risks	Risks identification	Risks mitigation plan
The project takes longer than planned, the project team is not skilled enough	Lack of needed personnel specialisation increases defects and rework, personnel need extra time to learn	To have well trained personnel for all positions in the project team. Backup for each of team member
The project takes longer, not enough team members	Understaffed team is a risk to project timelines, permanent employees le ave before the project is complete	To have backup for each member of the project team, project procedures setup in place, QMS setup
The project takes longer, not a good environment between the members of the team	Team members do not work together efficiently, conflicts between team members result in poor communication and extra rework	Project must be led by a strong and experienced project manager, frequent setup of project meetings
The project takes longer, chaos in leading the project	Project manager with a lack of experience to lead a project, no procedures in place	Project must be led by a strong and experienced project manager, project procedures setup in place, QMS setup
The project takes longer, chaos in leading the project	Project manager with a lack of experience to lead a project, no procedures in place	Project must be led by a strong and experienced project manager, project procedures setup in place, QMS setup

The Table 13 illustrates examples of quality risks in clinical registries and how risk management can be applied.

Table 13 Examples of quality risks

Quality resources risks	Risks identification	Risks mitigation plan
The objectives of the clinical registry are not fulfilled, loss of money or increasing the budget to monitor the sites, schedule delay.	Data in the registry is of bad quality or is missing.	Contract signed with setup conditioned for investigators or sites. Selected EDC system with requested functionalities. Validation criteria in the registry well setup.
The project does not follow the plan	No good project management, no procedures, no QMS setup in place	Experienced project management and team members, implemented QMS with standard operation procedures in place

Quality resources risks	Risks identification	Risks mitigation plan
The objectives of the clinical registry are not fulfilled	Lack of definition of primary and secondary aims, incorrectly setup eCRF structure to meet study/registry objectives	Experienced team members, close cooperation within the members of the internal project team (project management, data management, data analysts) and sponsor and client
Legal problems in the project	No good project management, no procedures, no QMS setup in place, regulatory and legal duties are not followed	Experienced project management, knowledge of regulatory affairs and legal duties included in the project plan
Hardware and software incidents	Provider of the EDC system does not have setup procedures in place, no implementation of ISO for data security system.	The detailed selection of the EDC system provider, audits, well designed contract with the EDC system provider

2.2 Scope of the registry

The registry design, objectives, extensiveness of its structure, duration of data collection and definition of the target population may reflect the scope of the registry and all parties that are interested in the data. Besides the investigators and sponsor, it could be healthcare providers, government, health authorities, patient organisations and other researchers or pharmaceutical companies.

2.2.1 Purpose and goals of the registry

The mission of most registries is improving patient health by improving the quality of patient care. Monitoring and evaluating patient care are therefore often the primary goal. This goal may be achieved in several ways. For example, clinical registries are a very useful tool for EMA to gain insight into the risks of a product through real-world experience. Clinical registries can also provide information on the appropriate use of a product, its effectiveness, costs and the cost-effectiveness of common clinical practice. Furthermore, clinical registries can bring essential information on patient-reported outcome measures (PROMs) in case data that are prospectively collected. Moreover, patient registries can provide information for public health planning [45].

The epidemiological studies of MM improve knowledge of the outcome for MM patients treated in "real-world" practice. Pending research identifies e.g. new treatment approaches, treatment-related issues etc.. Some of these questions may be answered using data from the MM registry

2.2.2 Objectives when defining the structure of the registry

The objectives of the registry should be as accurate and simple as possible. The goal is to obtain valid information on the objectives for all patients in different centres. Therefore, basic objectives should be determined, which are further exactly defined by endpoints, easily evaluated during common clinical practice, measured in all centres and in the same way.

The main objectives of the MM registry should be:

- To monitor and to assess the quality of care and diagnostics provided to MM patients
- To standardise the treatment procedures for MM patients in the participating centres
- To evaluate the compliance of treatment procedures and long-term benefits of therapy with national guidelines for MM
- To evaluate the PROMs
- To provide source data for research purposes

To reach all of the above-mentioned goals of the registry, researchers and clinicians – in close cooperation with the analysts – have to put together a set of variables to be collected and to embed them into the CRF as mandatory entries.

After everyone agrees on the main objectives, endpoints that answers the objectives are derived. The endpoint consists exactly of a specified variable collected directly in the registry or derived from the collected variables. Each objective is usually described by one primary endpoint and, if needed, by one or more secondary endpoints. For example, for the objective "to assess high quality patient care for patients with multiple myeloma", possible endpoints could be progression-free survival, overall survival, 10-year survival, time to any event, best response rate, duration of response etc.

In analyses of MM data, the most important endpoints are usually closely connected with survival analysis (see the chapter Descriptive analysis). The main endpoints obtained from the survival analysis include: overall survival, progression-free survival, time to progression, duration of response, time to next therapy, time to treatment failure or event-free survival (see the chapter Descriptive analysis for definitions). All of these outcomes are usually studied within different subgroups of patients to discover the risk/protective factors connected with shorter/longer survival (or lower/higher mortality). Other relevant endpoints in analyses of MM data include, for example, the safety of treatment.

In addition, a MM registry also includes baseline demographic characteristics, clinical stage and/or the World Health Organization performance status and laboratory examination results. They should reflect the occurrence of significant events that appeared during the patient's therapy (treatment, progression, transplantation or death). The variables have to be grouped into the data required on entry into the registry and those required during the planned follow-ups of appropriate frequency.

The variables to be collected include, for example:

- Demographic characteristics:
  - ID, date of birth, sex, initials, diagnosis, clinical centre
- Diagnostics:
  - Date of diagnosis, performance status, previous history of MGUS, SMM diagnosis
  - Laboratory examination (M-protein type, quantity etc.), Bone marrow examination, Biochemistry results, Cytogenetics results, Flow cytometry results, Medical Imaging Modalities, Extramedullary mass, Stage
- Treatment line (1st – Xth)
  - Line of treatment number, Reason for the start of therapy
  - Before treatment: Performance status, Laboratory examination, Bone marrow examination, Biochemistry results, Cytogenetics results, Flow cytometry results, Medical Imaging Modalities, Extramedullary mass, Stage
  - Treatment: Treatment regimen, start date of treatment, Drugs overview (dosage, number of administrations and cycles etc.), total cumulative dosage, dose adjustment
  - After treatment: Laboratory examination, Toxicity (before/after treatment)
  - Anticoagulant treatment, Granulocyte-colony stimulating factors
  - Radiotherapy
  - Transplantation
  - Treatment withdrawal (date, reason)
  - Response to induction treatment
  - Consolidation treatment
  - Maintenance therapy
  - Response after treatment line
- Current status
  - Patient status (alive, deceased, lost to follow-up), date of last contact with the patient, date of death, cause of death
  - MM associated amyloidosis (yes,no)

Other variables used in data analyses are derived from the collected data; for example, age at diagnosis is counted from the date of birth and the date of diagnosis, time to progression is calculated from the start date of treatment and the date of progression etc.

When working with an enthusiastic physician board of advisors, it may happen in reality that they believe everything about the patient should be collected, because it may become important.

Yielding to this temptation may produce a registry programme that is too cumbersome to implement and manage effectively, especially among physicians who are less research-savvy than the board.

It is recommended to collect only data whose use is determined before starting to collect them. Indeed, a registry is a dynamic tool for monitoring care and treatment outcomes reflecting current guidelines and attitudes towards treatment, thus the structure of the registry can develop. In any case, the registry should remain user-friendly.

2.2.3 Definition of target population

A more general population of patients can participate in the registry, in contrast to a clinical trial's strict inclusion/exclusion criteria. The MM registry is observational, with all interventions and scheduled visits performed solely at the discretion of the treating physician in accordance with their standard care. Participation in the registry is voluntary; patients can withdraw at any time without affecting their subsequent medical care.

The MM registry should include data on all patients diagnosed with MM (eventually MGUS). Patients are required to be aged ≥18 and provide written informed consent. Each patient is to be followed up for the duration of the disease until early discontinuation (i.e. due to death, withdrawal of consent, loss to follow-up). Current status data are to be collected approximately every 6 months or at each change of the patient's status. In each case, the setup of follow-ups should reflect common clinical practice.

Although it is preferable that data on all patients undergoing treatment for MM are collected, analyses can also be focused on more homogeneous subgroups of patients. That is a big advantage of a clinical registry compared to a clinical trial.

2.2.4 Duration of the registry

Longitudinal patient data collected in clinical registries on rare disease are important for studies on the natural history of disease. Typically, the primary goal of registry programmes is to publish long-term patient outcomes, including product effectiveness, safety and eventually patient-reported outcomes. Clinical registries in general are thought of as continual projects and the most difficult task is to keep such a registry alive.

It is highly advisable to keep motivating the clinicians and patients not only financially, but mainly by constant reminders of the principal vision of the registry, i.e. to improve the prognosis and treatment of MM patients. Clinicians should be getting regular and constant feedback in the form of analytical outcomes on the current situation, which may enable them to update and adjust the treatment strategy in their centres and can represent a valuable source of information for their publication activity.

For patients, the registry should represent a tool for improving their chances in fighting the disease not only for themselves, but also for the people with the same diagnosis yet to come.

It is critical to keep reminding participants of the vision, general idea and successes won so far, either in the form of scientific achievements (abstracts, presentations, posters, articles, manuscripts accepted for publication in peer-reviewed journals) or informative materials available to the patients and other registry stakeholders (newsletters). Presentations of the results are also crucial for keeping the interest and flow of sponsor funding. It is advised to take advantage of the fact that most subjects are likely to take part in the project with a vision and clearly defined, positive goals that are regularly reinforced by informing subjects of the current results and achievements on a regular basis.

The Czech registry RMG is an example of a long-running MM registry. Detailed information about the purpose, goals, endpoints, target population and duration of the RMG registry is provided in the next chapter (Goals of the CMG Registry of Monoclonal Gammopathies).

2.2.5 Goals of the CMG Registry of Monoclonal Gammopathies

The Registry of Monoclonal Gammopathies (RMG) is one of the main projects of the Czech Myeloma Group (CMG; http://www.myeloma.cz). The registry is run by the Institute of Biostatistics and Analyses Ltd, a spin-off company of the Masaryk University in Brno, Czech Republic (http://www.biostatistika.cz/), hereinafter referred to as "IBA". IBA is also responsible for the technological solution of the registry.

RMG was established by the Czech Myeloma Group in 2007 [46]. The registry was established in order to collect data on patients with monoclonal gammopathies in order to monitor the disease incidence, how individual treatment modalities are used, what are their results including adverse effects, and to monitor the patients' survival in the long term. Monoclonal gammopathies are a collective name for a heterogeneous group of diseases that are characterised by the proliferation of one or more clones of differentiated B cells producing a monoclonal immunoglobulin, which is sometimes referred to as the M protein (paraprotein) [12]. The registry was originally focused on the collection of data on two main representatives of monoclonal gammopathies: monoclonal gammopathy of undetermined significance (MGUS) and multiple myeloma (MM). The registry was extended in 2014 when also forms for the collection of data on patients with Waldenström macroglobulinaemia (WM) and primary or MM-associated AL ("amyloid light chain") amyloidosis were added.

Definition of RMG target population (inclusion/exclusion criteria)

The main target population of RMG are the patients with multiple myeloma. All MM patients diagnosed, treated or followed in one of the participating haematology centres in the Czech Republic and Slovakia are included in the registry; there are no exclusion criteria (as compared e.g. to clinical trials). Each year data of approximately 350 new MM patients are recorded in the registry. Some centres also perform the retrospective registration of patients diagnosed prior to 2007.

The technological background of the registry

The technological background of the registry has been built on the CLADE-IS platform (Clinical Data Warehousing Information System), which is software specifically designed for a repeated generation of electronic data capture (EDC) systems used in research management in clinical practice and in healthcare in general. Apart from classic randomised control trials (RCTs), CLADE-IS also supports the design and running of observational non-interventional studies, recently also referred to as "RWD/RWE" (real-world data / real-world evidence) studies [47].

Structure and principles of data collection in the RMG

RMG is an online web-based database suitable for parametric monitoring of patients. Co-operation of important treatment centres is key to success, because an adequate amount of representative data needs to be collected. A total of 23 treatment centres in the Czech Republic and in Slovakia currently contribute to the registry (Figure 8). The main Czech centres involve University Hospital Brno, University Hospital Hradec Kralove, University Hospital Olomouc, University Hospital Ostrava, University Hospital Plzen, University Hospital Kralovske Vinohrady and General University Hospital in Prague. Each centre has access to its data at any time; summary analyses are based on data from those centres which have agreed to provide their data for analysis. The project is open for cooperation to all treatment centres in the Czech Republic, Slovakia and other European countries. The project has been designed as a study with both retrospective and prospective patient recruitment. All patients are asked to sign an informed consent approving the registration of their clinical data into the registry. All patient data are anonymised.

Figure 8 Participating centres from the Czech Republic and Slovakia

The registry currently contains data of approximately 10,000 patients; specifically, data on more than 6,000 patients with MM and 3,500 patients with MGUS. RMG is therefore one of the largest and most comprehensive registries collecting data on patients with monoclonal gammopathies in Europe. As one of the few registries is regularly monitored similarly to clinical trials. RMG has been designed as a purely observational, research-focused and epidemiological study. Treatment with all available therapies is recorded into the registry, thus making RMG a valuable source of data from real clinical practice. Visit https://rmg.healthregistry.org to find more information on the registry itself, the list of centres with respective numbers of diagnosed patients, registry entry for authorised users, online visualisations of data, a form requesting data analysis, list of publications based on data from the registry etc.

Data are entered into the registry by data managers from individual centres. Collection and updates of data in the registry are controlled by the registry coordinator. Apart from these tasks, the registry coordinator arranges training of new registry users as well as current users (aiming to inform them of new recommendations by CMG on data collection), helping users with entering data into the registry, having new user accounts created or existing user accounts deleted, etc. A detailed online visualisation of data on MM patients from selected centres is available for representatives of individual centres as well as for top representatives of CMG. Online visualisations provide a well-arranged view of required information, together with interactive elements and animated effects. As part of the RMG project, an interactive browser has been implemented over data that had been collected via the CLADE-IS platform. With regard to the nature of published information, the visualisation is divided into a public part and an authenticated part, revealing only relevant information to a given user, according to his/her level of access rights. For more information about data visualisation, see the chapter Types of outputs from the registry.

Summary to CMG Registry of Monoclonal Gammopathies

RMG represents an international database designed for the collection of data on patients with monoclonal gammopathies. The Registry of Monoclonal Gammopathies (RMG) was established in 2007. Data originally stored in a former database system (based on a modified version of the TrialDB system) were converted to a new CLADE-IS system in 2017. Revision of form structure and contents was performed together with database conversion. New validation criteria were set in order to minimise the error rate in data. Data are recorded into the system by data managers, who are supervised by the registry coordinator.

A basic overview of data on individual diagnoses from all Czech or Slovak centres together is publicly available on the registry website. A detailed online visualisation of data on MM patients from selected centres, which can be also exported to a PDF file, is available for representatives of individual centres as well as for top representatives of CMG. Data from RMG are utilised in the preparation of national guidelines for treatment of monoclonal gammopathies, serve as a basis for negotiations with the State Institute for Drug Control (SUKL), and are also used for subsequent negotiations on reimbursements of new drugs with healthcare payers.

2.3 Feasibility

The feasibility of the project is a detailed analysis of how successfully a project can be completed. In the clinical registry, it is important to consider feasibility from the technical and technological, legal and regulatory, and site point of views. The project manager makes a feasibility study to determine the potential positive and negative outcomes of a project before investing a considerable amount of time and money into it.

2.3.1 Technical feasibility

Conventional data collection for clinical trials traditionally is focused on paper-based case report forms (CRF) followed by double data entry into a relational database. Electronic data capture (EDC) systems represent an intriguing alternative, which allow investigators to enter data, to review data in real time and to implement online data validation checks, thus assuring data quality more effectively at the point of entry. The recent trend to per-

form clinical studies with the use of EDC systems transformed the paper-based CRF into an electronic CRF (eCRF), bringing multiple advantages such as faster and more efficient data entry, higher security in terms of data integrity and better care for the environment.

Besides various EDC systems and tools for the facilitation of data management, there are also other IT solutions termed as Clinical Trial Management Systems (CTMS), which help pharmaceutical companies and medical research institutes to manage their projects in clinical research. CTMS are focused on the operational side of the projects by monitoring non-clinical workflows. More precisely, the tools that CTMS offer are to visualise and to solve operational problems such as investigator communication, site payment tracking, patient scheduling etc. Often, CTMS provide data to a business intelligence system, which acts as a digital dashboard for trial or project managers. Ideally, EDCs and CTMS should be applied in an integrated manner, thus allowing all involved users and stakeholders to achieve more success and accomplishments in the sphere [48].

2.3.2 Technical solution and requirements

There are multiple approaches to characterise EDC/CTMS systems. For instance, a review by Gazali et al. [49] focuses on data formats and key features of eight various clinical data management software packages, including SAS Clinical Software, Oracle Clinical software, COGNOS 8 Business Intelligence Software and Akaza's OPENCLINICA Software. Unfortunately, the authors of this review do not aim to set a clear boundary between EDC and CTMS. In addition, any evaluation criteria applicable to decision-making about a particular EDC/CTMS systems are missing as well.

The most repeated requirements and features of EDC/CTMS software include:

- Hosted solution which refers to the easy online access of data
- Delivery in a software as a service (SaaS) operational mode, which refers to no need for dedicated servers
- Role-based user permissions, which refers to secure access to data
- eCRF design which allows the optimisation of forms for data entry and enables the configuration of data flow and various form states
- Data entry that sets the data into the appropriate eCRF, which refers to data integrity
- Query management that provides streamlined communication between monitors and other involved users (investigators, clinical site managers, data managers etc.)
- Audit trails that support validation in accordance with 21 CFR part 11 [50]
- Data exports

Deciding on an appropriate system will be definitely influenced by the scale of a clinical trial; the evaluation criteria must be different for low-, medium- and large-scale studies. Electronic data capture in small-scale clinical trials is investigated in [51], where the authors present five clear evaluation criteria extracted from their previous concept map of EDC system features that facilitate translational research [29]:

- Availability of training materials and documentation
- Ease of designing CRFs, including automated error detection
- Ability to create a patient visit schedule for data entry

- Presence of site and user roles and permissions
- Efforts taken in exporting or importing data in a standards-compliant way

The list of investigated systems is narrowed by selecting only EDC software that is:

- Open source, free or very low-cost
- Web-based, preventing the need to install software
- Offering an application programming interface (API) for integration with other data sources

The preselected list of EDC systems contains seven systems. Of these, Visitrial and TrialDB are discarded as they had no recent project activity, and Caisis and LabKey do not offer a way for the study team to create their own CRFs. The remaining three EDC systems are Catalyst Web Tools from the University of Washington Learning and Scholarly Technologies, OpenClinica from Akaza Research and REDCap from Vanderbilt University and the REDCap Consortium. The review concludes by deciding on REDCap, pointing out its very clear advantage due to extensive tutorials and online training materials.

Integration with other data sources can be an issue when trying to bridge traditional RCTs with RWD/RWE trials. For instance, population-based registries are combined with electronic health records from the public health domain in a pa per by Kibbelaar et al. [52], providing high-quality data for observational studies and support for the development of best practices.

The experience from [48], [49], [51], [52] can be generalised into the following lessons learned: more functionalities in an EDC/CTMS do not warrant a higher value for clinical trials, whereas comprehensive documentation and tutorials delivered in non-technical language may contribute to a higher value of these trials. Stakeholders involved in medium- and large-scale clinical trials prefer to have EDC and CTMS in a single integrated web-based solution. In any event, finding competent admins for such complex EDC/CTMS platforms could be an issue.

2.3.3 Information security and data protection

There are two major data regulations worldwide: (i) the Health Insurance Portability and Accountability Act (HIPAA), applicable in the United States, and (ii) the General Data Protection Regulation (GDPR), applicable in the European Union. While HIPAA is focused on personal health information, GDPR is focused on personal data information in general. There is no authority who can "certify" that an organisation is HIPAA-compliant, and the same is true for HIPAA-compliant EDC/CTMS software. A similar situation can be expected regarding GDPR, which is enforceable from 25 May 2018. Anyway, all regulations in general require performing periodic technical and non-technical evaluations to make sure that the security policies and procedures meet the security requirements.

A formal compliance audit must be performed in case of the implementation of an organisation's information security management system that is based on ISO/IEC 27001 (the international best practice standard for information security) and further standards within the ISO 27k family [53]. These standards are certifiable standards and encompass the three essential aspects of a comprehensive information security regime: people, processes and technology.

By implementing measures to protect information using this three-pronged approach, the company is able to defend itself not only from technology-based risks, but also other, more common threats, such as poorly informed staff or ineffective procedures [54].

The information security as a technical term encompasses the so-called C.I.A. triad (confidentiality, integrity, and availability):

1. Availability of information
2. Its integrity (completeness and accuracy)
3. Its confidentiality

Many managers believe that confidentiality of information is the most important or even the only aspect of information security, but it only forms a minor part of the issue.

While administrative safeguards (policies and procedures that have to be in place to ensure the proper employee management, training etc.) are explained in this guidebook (see the chapter Quality assurance in pharmacoepidemiological research), the purely technical part of the information security can be divided into physical safeguards and technical safeguards.

Table 14 exemplifies and categorises measures implemented in an organisation dealing with information security management. Figure 9 outlines an organisation's IT infrastructure that hosts an EDC/CTMS system in a data centre that consists of two independent distant server rooms.

Table 14 Physical safeguards and technical safeguards related to data security and data safety – an example illustrating measures within an implementation of an information security management system in an organisation that deals with personal health information

	Physical safeguards	**Technical safeguards**
Data security	The data centre encompasses two independent distant server rooms	
	Physical access to the data centre is strictly limited to authorised personnel	The network provides services towards cybernetic security by analysing all network packets in real time with the use of machine and human resources
	Electronic ID cards	
	Non-stop camera monitoring	
	Fire detectors	Servers are placed in a separated subnet dedicated only for server traffic
	Temperature detectors	Firewall on the subnet and dedicated firewalls on each server
	Movement detectors	
	Alarm/detectors inform reception with non-stop service and system administrators via SMS.	Fragmentation detectors

Data safety		Server operating system (Linux – Ubuntu) is updated regularly
		Servers are backed up before any system changes are carried out
	High-quality servers with HW accelerated disc arrays	Encrypted communication
	Servers equipped with a backup power supply and redundant power supplies	HTTPS trusted certificates SSL encryption
	Both the system configuration and the data stored are subject to a regular backup procedure. The entire system including data can be restored promptly	Authenticated users (login, password) Authorised users (data access management based on roles and permissions) Automatic logoff after a predefined period of user inactivity

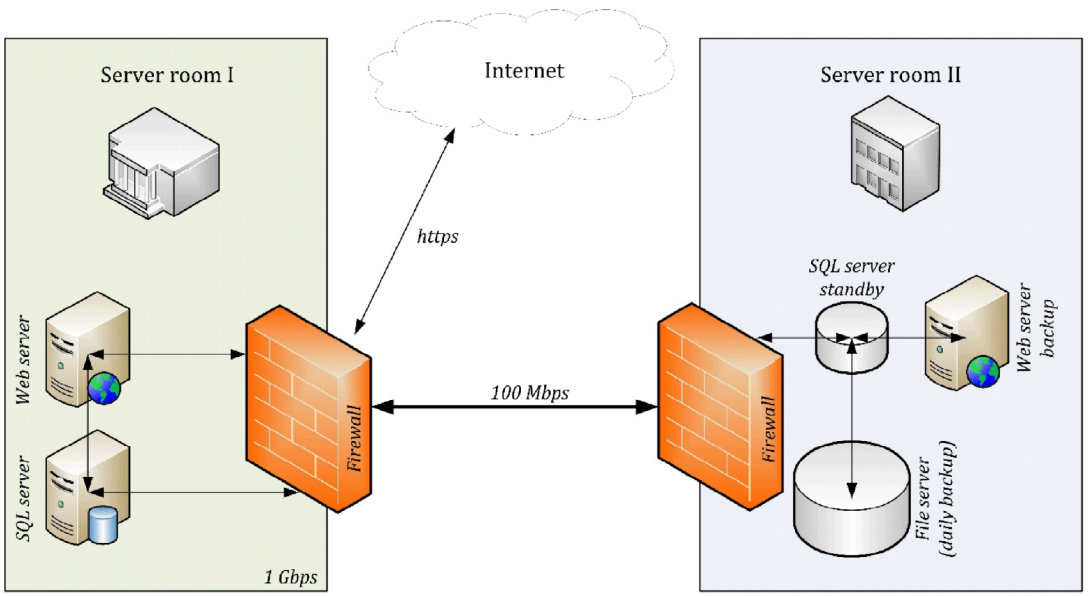


Figure 9 An exemplary IT infrastructure that hosts an EDC/CTMS system (web server, SQL server) in data centre that consists of two independent distant server rooms

2.3.4 Legal feasibility

Serious ethical concerns have led to the creation of legal requirements for using health information for research purposes. The creation and use of a clinical registry for research purposes constitute "research involving human subjects". Legal requirements exist at the regional, national and European level, but are not always the same. Many clinical

registries have difficulties in consolidating this information and fulfilling the various obligations. The ethical principle of respect for persons supports the practice of obtaining individuals' consent to the use of their health information for research purposes. Individuals (including children and their parents) should be informed with respect to the type of the research that will or might be carried out, the arrangements for access to or the sharing of stored information and the duration of storage (for more see the chapter Informed consent) [44].

2.3.5 Data privacy and confidentiality

Two primary issues for participants often override all others: "How do you maintain patient confidentiality?" and "Who owns the data?"

Patient data, and data on cancer patients in particular, are highly confidential. An improper disclosure of these data could result in emotional, psychological and financial harm to patients and their families.

Here is the answer to the first question: it is crucial to ensure patients' privacy, in particular for patient identifiers. This includes the anonymisation or pseudonymisation of data to ensure that information entered into the registry cannot lead to an individual patient identification by third parties. We have to realise that in case of the need for follow-ups (updating of the records) or combining different sources of data, anonymisation may hamper these specific registry functionalities [45]. Pseudonymisation involves replacing identifying items by artificial identifiers or pseudonyms (personal ID number) at the moment of entering patient and its data into the registry. Using personal ID numbers allows the registry data to be shared while confidentiality is ensured. The key for re-identifying a patient for the purpose of follow-ups or additional queries is only at the site of the treating centre that the patient belongs to [33][55].

If the data collection is undertaken by a group from a different country, regulations from both the country of origin of the data and of the country where the data is stored should be respected in order to maximise the protection of the participating patients. Furthermore, in case of indirect identifiability in the registry, there is a threat of re-identifying the person suffering from a rare disease and living in a defined place. Such a person can be easily identified as being the only one. In this situation it could be helpful to revise the data set in the CRF to try to prevent such a combination of data leading to threating the patient's data privacy. In addition, we also recommend consulting the above-mentioned situation with the appropriate national authorities in charge of data protection to ensure compliance with the requested standards. Another issue is whether existing databases can be used for secondary analyses, i.e. a use not originally planned when the data were collected without a specific informed consent. Ideally, researchers would inform each subject in the study and obtain his/her informed consent. This is hardly feasible when using linked records, especially if the data were collected years back in the past and the number of subjects is large. As a result, the requirement for individual consent would preclude the conduct of many important epidemiological studies. Moreover, this type of study imposes no medical risk to the patients. In contrast, in these studies, contacting an individual for consent would compromise patient confidentiality, since obtaining consent would require access to their name, address or telephone number.

For research studies based on linked data, investigators and other researchers generally have no interest in the patients' identities. However, they must have assurance that the records of the same person can be joined, and the assurance that those records are protected at the source with suitable safeguards. For such secondary data analyses, the

Steering Committee of the registry should review the importance of the research and the adequacy of privacy protections, and determine whether a secondary analysis without specific informed consent is appropriate [34].

General principles of confidentiality and data privacy in a registry:

- The informed consent must be acquired prior to the entering of patient data in a registry.
- We recommend providing evidence that the de-identified data of a patient in the registry are accessible according to their ID number only for the treating physician of the patient.
- No disadvantage, harm or distress may be caused to the patient by any release of data in any way.
- Appropriate safeguards must be in place to preserve the confidentiality of the information in the registry.
- Reports or analytical outcomes from the registry may not contain information that might destroy the anonymity of either a patient or a physician.
- The registry represented by the Steering Committee has a duty to maximise the use of information in its possession to the benefit of all patients.
- The personnel of the registry must be trained in the field of data protection, patient privacy and confidentiality set in the registry and they have to comply to them.
- Data in the registry must be protected by passwords and encoding.
- Data in the registry have to be stored securely on dedicated servers with limited access by authorised registry personnel only (for more information see the chapter Information security and data protection). The rules of data protection must also be set up for data transfer (e.g. sharing the data with international database/registry).
- Physical security: every written document needed for archival purposes must be scanned and saved on the dedicated server with authorised access and the paper version of these documents must be held within a locked storage cabinet with limited and defined access; the room with access to the registry is obligated to be under electronic alarm control.
- Communication: all requests for the data from the registry should be referred to the Steering Committee, and reports or analytical outcomes from the registry may not contain information which would destroy the anonymity of either a patient or a physician.

2.3.6 Data ownership and data access

There may be a couple of answers to the second question of "Who owns the data?" (see also Data privacy and confidentiality).

Various parties, such as healthcare providers, patient advocacy organisations or national law officials, often dictate the answers to this question according to their own interests and interpretations. Sometimes the healthcare providers feel a proprietary ownership of the data because of the work they have done to acquire the data about the state of patient health. In addition, they often have concerns that they may be violating the patients' privacy rights by relinquishing control of the data to a third party of registry.

The most widespread opinion is that data belongs to individuals in the patient community. There is no doubt that the patient definitely has the right to this information and can provide this information directly to the registry or allow his/her treating physician to enter his/her data to the registry after acquiring the patient's informed consent. The use of patient data in a national registry is justified because it benefits the entire community of patients, including the individual. When patients understand this principle (principle of beneficence), they provide their consent and can strongly support the application of their data, especially if they know that their data can be made anonymous and that the analytical outcomes will be done on an aggregated data set. We can also consider the registry providers' request for patient consent as the representation of the respect for persons, which is one of the most important ethical principles in medical science [55].

The Steering Committee (for detailed information see the chapter Professional society) has input on the strategic direction of the registry, acts as a data access committee and ensures that the registry acts in the best interest of the patients and meets the registry objectives. If there is a request from participating centres or even from outside the registry on either the data themselves or on an analytical outcome (secondary data use), the procedure for registry data release should consider the following steps: In case of data analysis, the request should state a clearly detailed description of the analytical hypothesis, the data items to be used in the analysis and the outputs needed.

- In case of data request, the subset of the data required has to be defined.
- The request must be in a written form, delivered by email, post or fax addressed to the project manager of the registry.
- A vague, non-specified request will be rejected and returned.
- The project manager is obligated to acquire the agreement with sharing the data from all centres participating in the registry (by voting) if there is a requirement to use either overall data from the registry or from particular centres of interest; after voting, the project manager will contact the Steering Committee of the registry and ask them to approve the request with respect to the results of the centres' votes.
- All members of the Steering Committee are obligated to issue their statement within the time frame specified in the operating instructions of the registry concerning data access; it may be useful to adjust the rules on how to proceed if there is disagreement on whether or not data should be provided (full consensus or majority consensus specified by the ratio of voting members).
- The final decision of the Steering Committee must be recorded in a written form with all comments and objections, and it must clearly state whether the request was accepted or not.
- Approval of the analysis by the Steering Committee is required before providing this outcome to the requester.

2.3.7 Personal data protection (GDPR)

In registry projects, the personal data and special categories of personal data of the patients are collected and processed.

These data can be anonymised (meaning there is no possibility to identify a specific data subject) or usually pseudonymised (personal data can be attributed to a specific subject with the use of additional information).

In case that data are only pseudonymised, the data have to be processed in compliance with the Regulation (EU) 2016/679 of the European Parliament and of the Council of

27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation – hereinafter referred to as the "GDPR").

The effective date of the regulation was 25 May 2018. Each organisation processing personal data should implement appropriate and sufficient processes for personal data processing into its quality system to ensure data confidentiality, availability and integrity.

Basic dictionary relevant to GDPR

Personal data – any information relating to an identified or identifiable person (data subject); an identifiable natural person is one who can be identified, directly or indirectly (identifier can be a name, an identification number, location data, an online identifier etc.) [56].

Special categories of personal data – data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs or trade union membership, genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation [56].

Processing – means any operation or set of operations which are performed on personal data or on sets of personal data, whether or not by automated means (processing is collection, recording, storage, disclosure etc.) [56].

Consent of the data subject – means any freely given, specific, informed and unambiguous indication of the data subjects wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her [56].

Controller – means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data [56].

Processor – means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller [56].

The basic rules for Personal data processing are the following:

1. Legality – meaning the processing is lawful only in case there exists a legal basis for processing

    a. The data subject has given consent for the processing of his/her personal data

    b. The processing is necessary for compliance with the legal obligation

    c. The processing is necessary for the performance of a contract

    d. The processing on the base of the legitimate interest pursued by the controller or by a third party

    e. The processing of personal data for the protection of vital interests of the data subject

    Note: the special categories of personal data in the registry project can be processed only on the basis of explicit consent with this processing, which was given by the data subject for this purpose.

2. Purpose limitation – personal data can be collected only for the stated specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes
3. Minimisation of processed data – the processed personal data are adequate, relevant and limited for the specified purpose and for the necessary long-term period
4. Accuracy – personal data have to be accurate, and if needed, updated. Accuracy is ensured by regular verification
5. Transparency – the data subject must be informed or notified about the processing of his/her personal data by responsible controller
6. Confidentiality and integrity – confidentiality means the personal data are accessible only to the authorised person (access is managed); integrity means the assurance of the accuracy and consistency of data
7. Responsibility – the organisation is responsible for the personal data and is able to demonstrate compliance with GDPR

Implementation of the GDPR in the organisation

The basic principles of GDPR data integrity, data confidentiality and data availability are the same with ISO 27001. It is therefore possible to say the implementation of ISO 27001 requirements ensures data security and at the same time compliance with GDPR.

The GDPR have some extra requirements that must be implemented to this system.

The implementation of GDPR consists of the following steps:

1. Mapping of personal data processing. The organisation should identify all processed personal data, define their classification, purpose for processing, legality, preventive technical and organisational actions, processing period and methods of processing.
2. Identification of resources for personal data processing (SW, HW, employees, facilities...)

   Note: Secured personal data as resources are marked in the information security management system as protected assets.
3. An internal standard for personal data processing should be written, implemented and trained to ensure the data will be processed in the safest way. The internal standard should describe how the data subject can apply her/his rights. The subject rights are the following:

   a. Right of access
   b. Right to rectification
   c. Right to erasure (right to be forgotten)
   d. Right to the restriction of processing
   e. Right to data portability
   f. Right to object

4. The organisation should execute risk analyses and evaluate the risks with an impact on the data subject. In case that the processing of personal data proves a high risk to the rights and freedoms of natural persons, the processor shall implement appropriate technical and organisational measures to ensure that this risk is decreased. The implementation of preventive actions depends on their technical and financial severity. If there is a very high risk to the rights and freedoms of natural persons and it is not possible to implement a preventive action, it is necessary to communicate the processing with the relevant supervisory authority.
5. The organisation/controller shall designate a data protection officer (DPO). The organisation operating the registries fulfils one of the GDPR conditions to designate a DPO. This condition ensures the core activities of the controller or the processor, consisting of processing on a large scale of special categories of data [56].

2.3.8 Regulatory and ethical feasibility

The use of health information on patients for research purposes is subject to strict ethical rules at both national and international level. These rules may vary significantly. Because a clinical registry collects sensitive health data about patients participating in the study, it is necessary to ensure the safety and security of collected data, considering compliance with the local laws and regulations. It is often advisable to consult the relevant local or multicentre ethics committee and regulatory body. These legislative requirements can significantly influence the choice of collected parameters and the subsequent use of collected data for secondary data use for research purposes. The description of data collection in the registry and the nature of data, including other project information, is available in the protocol of a given registry (for more information see the chapter Protocol).

Project approval procedures resulting in a clinical registry may be required by more than one managing authority; for example, when participants are recruited from multiple locations or the registry is intended to collect the data from more countries with different legislative rules. Therefore, it is important to consult the requirements of the local governing bodies prior to the start of the project. Data collection can only start after all necessary approvals are in place. When data are collected for research purposes, an ethics review committee will oversee the process and will be required to assess and to ensure that the benefits of the goals of the research outweigh the risks of participation [57].

If data collection is undertaken by a group from a different country, regulations valid in both the country of origin of the data and in the country where these data are stored should be respected in order to maximise the protection of the rights of the investigated group of patients [57].

A tool for ethical access to registry participants is obtaining an informed consent from individual patients before their entry into the registry to use their health data for research purposes.

2.3.9 Informed consent

Informed consent and data protection apply to all registries where individual patient data are collected and stored.

According to GDPR, data collected in a registry belong to a special category of personal data and can only be processed on the basis of an explicit patient consent (for more information see the chapter Personal data protection (GDPR))

The informed consent should be obtained from the patient in a written form. The patient's informed consent to provide his/her to the clinical registry must contain several particularities as stipulated by the legislation. The following is a list of the most important ones:

- Registry purpose and objectives
- Information that the registry will only collect data from common clinical practice. There is no intervention during the treatment
- The nature of data (anonymous / pseudonymous / personal) as well as the possibility and ways of identification in case of anonymised or pseudonymous data
- Who is the controller and who is the processor of personal data in case that the registry will collect personal data; the administrator determines the purposes and means of such data processing
- Information about data safety and security
- Reference to the legislation governing the handling of personal data
- Form and method for the publication of collected data
- Patient's name and date of birth
- List of patient data that will be collected in the registry
- The period of time for which the informed consent is given
- The purpose for which the informed consent is given
- Statement that the patient has been properly instructed
- Statement that the patient is aware that his/her participation in the registry is free of charge
- Information that the patient may withdraw his/her informed consent at any time by his/her own will and without giving any reason

Other important information concerning the patient's right in relation to the informed consent are:

- The right to ask the controller to provide information about personal data processed in relation to the patient, the purpose, timing and nature of personal data processing and the recipients of personal data
- The right to ask the controller to provide a copy of the processed personal data
- The right to request an explanation from the controller, if necessary, to temporarily stop the processing and to remove the defective condition if the patient finds or believes that the processing of his/her personal data is contrary to the protection of the his/her private and personal life or in conflict with the legislation
- The right to ask the controller to delete the patient's data
- The right to obtain from the controller personal data relating to the patient in a structured, commonly used and machine-readable format and the right to pass this information to another controller without hindrance
- The right to contact the Office for Personal Data Protection in case of threats to the patient's personal data with a request for remedial measures

The patient should provide his/her informed consent freely, without pressure or persuasion, and based on information provided by trained personnel. If the patient removes his/her consent, his/her data can no longer be used [57].

Registry participants under the age of 18 cannot sign an informed consent by themselves. The informed consent shall be signed by his/her legal guardian or parent. After the age of 18, a new informed consent from the patient should be obtained for their continuation in the registry. The patient has to sign the informed consent before their participation in the registry at the time of recruitment. The informed consent should be regularly reviewed and updated according to the legislative changes applicable to the personal data protection and data safety and security. Consent obtained from patients should be clear about data sharing and access for stakeholders other than the registry holders, with appropriate consent in place for levels of data sharing; for example, aggregated anonymised data versus individual patient data [58].

If there is an intention to merge data from one registry to another registry, this should also be included in the informed consent.

If legislative conditions regarding data privacy protection changes or if there are any other changes which influence the text of the informed consent form, this requires obtaining re-consent from the patient. The new (updated) informed consent form must be approved by the ethics committee before its use.

2.3.10 Site feasibility

Determining site feasibility is one of the first steps in planning a clinical registry. The identification of suitable centres and physicians is a complex process managed by the team consisting of the sponsor, guarantor of the registry, professional society and contract research organisation providing services for the sponsor.

2.3.11 Site recruitment and selection

The process of site recruitment and selection and number of sites participating in the registry are described in the protocol, which is presented to the centres during the site recruitment procedure. The protocol also contains other key information such as the registry purpose and objectives, research aims, data safety and confidentiality, data and results sharing, conditions of data use and publication, motivation for patient recruitment etc. (for more information see the chapter Protocol).

It is important to consider the number of sites needed to recruit and to retain enough patients in order to achieve a reasonably informative number of person-years for analysis. Many factors are involved in estimating the number of sites needed for a given study, including the number of eligible patients seen in a given practice during the relevant time period, the desired representativeness of sites with regard to geography, practice size or other features, and the time frame within which study results are required, which may also limit the time frame for patient recruitment [33].

Site selection also should reflect the target population. The purpose of a registry is to provide information or to describe events, and often to generate hypotheses about a specific target patient population to whom study results are meant to apply. In case of selecting sites with different treatment patterns (e.g. complex centres) than unselected sites, the results cannot be generalised to the entire population, but only to patients treated in those types of centres.

The usual methods for site recruitment can be:

- Invitation letter which describes the purpose and objectives of the registry
- Recommendation made by the professional society or the guarantor of the registry
- Interest of physicians from the centre

Among the most frequent motivation factors for a centre's participation in the registry, the following ones are notable:

- Interesting scientific objectives of the registry with research potential
- Obtaining data and analytical results about the centre's performance
- Offered remuneration

The main rights and requirements of the participating centres are defined in the contract between the centre and either the sponsor or the contract research organisation as a registry coordinator. The contract should specify:

- Confidentiality of the collected data in the centre
- Policies regarding the ethics committee approvals and the informed consent form
- Voting rules about data sharing and reporting
- Rights of the centre regarding self-assessment data and the quality of care (analysis based on the data provided by the centre)
- Remuneration for the data entry into the registry based on the fair market value rules (for detailed information see the chapter Cost management plan)
- Service for the centre from the sponsor or from the contract research organisation (helpdesk, training, providing the centre with enrolment reports, data analysis for the centre for research purposes, presentations at conferences, publications)

The contract research organisation usually helps centres and investigators solve their problems by providing a helpdesk service.

Recent news from the registry are presented at the registry website, along with other information about the registries such as its objectives, participating centres, publications, EDC system used etc.

2.3.12 Identification of physician (centre lead)

The centre lead is the main person responsible in a given centre for data entry and data quality. The centre lead is selected either by the centre management, by the professional society or by the guarantor of the registry. It is very often the same physician who has initiated the centre's participation in the registry.

Among the most frequent motivation factors for a physician's participation in the registry, the following ones are notable:

- Interesting scientific objectives of the registry
- Research interest and publication potential
- Well-respected guarantor in the professional society
- Offered remuneration

2.3.13 Recruitment of patients

Patients' participation in most clinical registries is usually voluntary. That is why it can be quite a challenge for the physician to recruit eligible patients. The main motivation for participating in a clinical registry can be the patients' desire to help other patients or to help himself/herself to achieve better treatment results and to have a better communication with the treating physician.

Focusing on the patient population and recruitment strategy is key to success. Inclusive and exclusive criteria for patient selection in the registry are broad and could be practically only limited to diagnosis, age or treatment. Very often, all patients are enrolled in the registry and no selection is made; when using this approach, the results are expected to reflect real patterns of practice and not to describe a pre-selected group, which always hides risk selection bias. Some registries can collect data specific to treatment of selected individuals, such as the registries on targeted biological therapies, which can only involve patients with a specific diagnosis and treated with a defined therapy. Participation in some clinical registries can be limited by age; for example, only patients over 18 years old can be involved.

These broad inclusion and exclusion criteria exist because the clinical registries are non-interventional studies in their design and a patient's entry into the registry does not influence his/her treatment.

The target population is specified in the protocol (for detailed information see the Definition of target population and Protocol).

Regarding the registry design, registries are often designed as prospective studies. Some registries can be retrospective–prospective, when additional data about patients' history and previous treatment are collected.

Registries are long-term projects; recruitment of patients is therefore continuous and is not limited by time.

The recruitment and retention of patients as registry participants, and of providers as registry sites, are essential elements in the design and operation of a registry as well as in the success of a registry [33]. Registries are often intended to be representative of a certain population of patients and reflective of the practices of certain providers and geographic areas [33]. The easier the recruitment to the clinical registry that collects anonymous data is, the better the patient's privacy is ensured. This data can be either anonymised or pseudonymised and therefore cannot be traced to identify an individual patient. Regarding pseudonymised data, only the treating physician of the patient can trace the patient and re-identify him or her.

Methods of patient recruitment:

- By the treating physician of the patient (the most common method)
- By the activity of a patient organisation

The physician who enrols the patients into the registry provides the patients with all the information about the purpose and objectives of the registry and instructs them about their rights. The patient is asked to carefully read the informed consent form and to sign it (for detailed information see the chapter Informed consent).

It is good to set up reports that are distributed to all centres participating in the clinical registry. These enrolment reports may be distributed on a monthly basis, providing updated information about the number of recruited patients in each centre, the number of follow-ups etc.

Patient recruitment can be supported by newsletters by which the centres are informed about news, rules of filling patients' forms in the registry, notification of filling errors and recommendations etc.

2.4 Stakeholders and experts of the registry

The role of the contract research organisation and its internal project team is to ensure that the key stakeholders are on the board and to make them familiar with the principles demanded for developing and operating the registry. This working group should include the representation of all key stakeholders that are expected to utilise, benefit or influence the progress of the registry. This is a critical step, as the clear understanding of the needs and uses of the data will directly affect the success of the registry [59].

The main stakeholders are:

- Professional society
- Clinicians
- Molecular testing experts, laboratory experts, researchers
- Contract research organisation (CRO)
- Patient organisations and representatives
- Pharmaceutical industry representatives
- Regulatory authorities, healthcare providers, healthcare payers

Pharmaceutical companies often get a conditional market authorisation for their product that orders them to conduct a post-marketing surveillance about the safety or effectiveness of its new product. Another reason could be the interest of manufacturers in accessing data on the natural history of the disease for which they have a product in development. The registry could be also a source of volunteers for clinical trials.

Patient organisations may consider gathering of data about the studied disease as pivotal for a better understanding of their disease and a better prognosis for patients to be diagnosed in future.

A very important step in starting up the registry is a kick-off meeting of this working group as we mentioned in the chapter Staff and communication management plan. However, such a broad advisory board can lead to difficulties in the daily management of the registry. Therefore, a flexible governing body of the registry should be established in order to resolve various issues that might occur in everyday practice.

2.4.1 Professional society

A professional society, comprised of the top representative clinicians and researchers, is very often the initiator of a clinical registry and acts as an expert guarantor of the registry. Professional societies need the data for evaluating the recommended national treatment guidelines that they are responsible for, as well as for assessing the new trends in diag-

nosis and possible methods for new interventions that can lead to changes in the patient prognosis. A further goal could be to coordinate the level of treatment care across all centres in a given country. Data about the effectiveness of high-cost therapy might be crucial for negotiation with healthcare payers.

Each registry should have a clearly outlined governance structure that will define the management of the registry, oversight roles and responsibilities. As we mentioned in the chapter Project management plan, the project manager could be a registry coordinator, but the subordinate role of the project manager is generally determined by the governing body of the registry, the so-called Steering committee. The Committee is composed chiefly of members of the professional society, CRO (project manager) and possibly representatives of the patient organisation. The Steering Committee should meet on a regular basis, either by conference call or in person. Meetings are more frequent in the early stages of the registry formation, focusing on achieving a consensus about the registry's design and goals.

In summary, the Steering Committee should cover the following tasks in the registry:

- Governance of the registry
- Data access and use by researchers
- Database content and research objectives
- Data ownership
- Sharing the data with other sources
- Alignment of interests (which might possibly be conflicting)
- Funding the registry
- Recognising legal and other obligations towards stakeholders
- Recognising risks and managing difficulties
- Cooperation with the broader audience or other stakeholders of the registry

The committee can delegate some of these tasks to the CRO, e.g. funding, legal and ethical obligations, communication with stakeholders, training of centres, data analyses and reporting.

2.4.2 Clinicians (centres)

Clinicians are integral to the registry and should be involved, informed and empowered where possible. They are the main contributors to the registry and they can benefit from the registry by a better care provided to their patients. The proposal of participating centres is a critical point of the initiation process. The subsequent phase involves addressing the proposed centres' representatives with an initial explanation of the idea and goals of the registry, stating the advantages of involving their centre, explaining what are the obligations and workload extent of the centre after becoming part of the registry and how the feedback will be maintained. Another very important point of the initiation phase, following recruitment, is the meeting with all representatives of all candidate centres with the advisory board member, where the overall organisation of the registry and suggested CRF will be presented and discussed so that the centres know that they are respected partners and that they would play an integral role in the system. The CRF and other duties must be put together concisely, shortly and in a way easy to understand so that the clinicians, already being overwhelmed by their everyday clinical duties, will not be further burdened by demanding data entries which would inevitably devolve into a loss of

motivation and compromised data quality. An important role in maintaining motivation is played by regular meetings of the participating centres, which involve information and updates on recent activities and results, setting new challenges with time frames and a clear distribution of tasks. Any member of the team is thus well informed of milestones and outcomes with a strong sense of being part of a successful team.

The time and effort spent with entering data into the registry must be fairly rewarded. The reward should correspond to fair market values set for a particular registry.

2.4.3 CRO – provider/academic institution

As we mentioned in the chapter Professional society, the advisory board of the registry might delegate some critical steps and tasks in the initiation of a registry to a contract research organisation (CRO). CROs have thorough experience with organising and managing interventional and/or non-interventional studies and are the most suitable candidates for providing some of the activities mentioned in the chapter Professional society.

The main responsibilities of CRO in the registry involve:

- Coordination of the registry initiation
  - Preparation of CRF in cooperation with the registry stakeholders
  - Protocol development
  - Centres recruitment and training
  - Providing an Electronic Data Capture (EDC) system
  - Arranging ethical and regulatory requirements (e.g. informed consent approval)
- Coordination of running the registry
  - Reporting, data management, data quality control
  - Analytical outcomes, interim analyses
  - Pharmacovigilance
  - Organising advisory board meetings, cooperation with the Steering Committee is a critical connecting factor of the registry
  - Contract management with sponsor (funding) and participating centres (remuneration of investigators in the centre)

There are other activities concerning the termination of a registry, but this topic is not relevant for registries such as the RMG registry, i.e. registries that are regarded as tools for continual data collection. The CRO can be also included in organising collaboration among another MM registries or international haematological registries.

2.4.4 Funding sources

Registries enable healthcare providers, researchers but also regulators and sponsors to observe clinical responses in real time in a broader patient population and geographical region. Clinical trials conducted for a narrow sample of the target patient group with strict inclusion and exclusion criteria are not able to provide information on e.g. the proper use of a product (new medicine, device, diagnostic product) or expectations of treatment response in “real-world” practice. The answers to these questions are concealed in real-world studies including clinical registries. This information is crucial for

pharmaceutical companies in successful product launch and long-term product growth, but also for global regulatory authorities as a means to monitor the safety of a medicinal product after approval. Reimbursement agencies will be interested in how the clinical outcomes (effectiveness and safety outcomes) of patients treated with a new product compare with the outcomes of patients still receiving the current standard of care. In this case, collecting appropriate data as well as patient-reported outcomes may provide the information needed to demonstrate the value of the new product in the target patient population. The increasing value of these post-marketing data significantly reinforces the potential and importance of the registries for all of the above-mentioned subjects and pharmaceutical, biotechnological and diagnostics companies are often willing to support such registry programmes.

Another issue is the fact that in an existing clinical registry, clinical data represent treatment results from a range of products available to physicians for treating their patients. Data from the registry can provide evidence about the actual position of a new drug on the market. Furthermore, registry data related to a new product can assure the sponsor of the proper use of the product in real practice. In addition, over time, registry data can suggest some more off-label use of the product or changes in the recommended dosage performed/tested because of specific patients' needs.

One of the determining elements in the establishment of a new registry is how to raise funds not only for starting a new registry, but also for maintaining a longitudinal collection of patient data into the registry. The amount of expected costs is defined by the scope of the registry in terms of the number of participating centres, number of patients, extent of eCRF depending on the scope of data collected, complexity and frequency of required outcomes (reports, analyses), geographical location etc.

There are many potential funding sources for registries, and funding models may vary significantly. Potential sources of registry funding include:

- Government: state regulatory authorities, health plan providers (Ministry of Health, regional health agency and reimbursement agency) may be interested in a registry to determine the long-term outcomes of medicines, devices or procedures and to gain a powerful tool for providing evidence for health coverage and healthcare decisions.
- Patient advocacy group: a patient organisation may also be willing to support rare disease registries.
- Pharmaceutical/product manufacturers, who are often interested in: studying the natural history of the disease for which they have a new product; monitoring the effectiveness and safety of their new approved products in real life; data mining in the registry in terms of new information or consequences for improving the quality of care of treated patients or for their own research and developing intentions. Also, worth mentioning as a source of funding is secondary data usage for third parties that could not be regular sponsors of the registry.
- Professional societies and public/private science foundation organisations: healthcare professional associations are directly participating and partnering in the establishment of new registries, setting up their goals and evaluating outcomes. Granting is one of the most important forms of registry funding.

Funding needs should also be examined in terms of the projected life of the registry and/or its long-term sustainability. Registries, when designed effectively and integrated with a sponsor's market research activities, can provide a steady stream of new information welcomed and desired for communication between physicians and other healthcare providers, researchers, private and public payers and regulatory authorities [43] [33].

2.5 Registry provider's team and roles to manage the project

Building up and keeping the registry alive are very complex processes. Prior to the start of any registry, a team that will keep the registry ongoing has to be set-up. Each team member has his/her single role, which is irreplaceable and needs special knowledge and skills.

2.5.1 Project manager

The project manager is responsible for planning and managing the process and system setup for the clinical registry.

The project manager's responsibilities include:

- Preparing and maintaining updated all project-related documentation such as the protocol, reports, protocol revision in case of changes etc.
- Negotiating with the sponsor and centres regarding contracts
- Coordinating and monitoring the activities of all members of the project team (data manager, analyst, medical expert, etc.) to ensure compliance with the protocol and overall clinical objectives
- Providing service to the sponsor personnel, professional society and centres participating in the registry
- Training investigators on the protocol and eCRF and instructing investigators in informed consent procedures and documentation procedures
- Maintaining budgets and timelines
- Maintaining the regulatory and legislative compliance of the project
- Being aware of all communication in the project – as the primary contact for the client, the project manager maintains an updated contact list
- Ensuring communication within the project team
- Organising internal and external meetings and trainings, providing minutes
- Preparing or validating all project documents (data management documentation, quality management system documentation, eCRF, ...)
- Preparing for and participating in quality assurance audits conducted by study sponsors, regulatory agencies or other review groups
- Reviewing scientific literature, attending conferences and seminars

2.5.2 Data manager

The data manager is one of the key members of a clinical project team. He/she is responsible for all of the actions related mainly to database construction and data validation. Critical responsibilities of a data manager include:

- Proposing the structure of the CRF with concrete parameters, forms and visits
- Suggesting database structure and validation rules corresponding to the project's design and available documentation (study protocol, data management plan etc.)

- Setting up the registry database including all relevant validation checks in order to ensure smooth data entry and data validation process
- Looking after the registry database during the data entry process
- Updating the database according to the needs determined during the project duration
- Exporting and reporting data from the database according to the project setup rules
- Locking the database

The data manager is a very important member of the project team over the entire project duration. He/she starts working on the project in the phase of project designing and budgeting. The main and critical part of his/her work is greatly appreciated during the setup of CRF and related database, including all relevant checks. The data manager also plays an important role during the live part of project when checking the entered data. The last steps of data manager's work for any given project involve exporting the data and locking the database.

2.5.3 Data entry specialist

In some clinical projects or registries, it may happen that some data need to be entered into the database beyond the standard process, i.e. differently than by the investigator on site (e.g. QoL questionnaires data). For this purpose, it is necessary to have a data entry specialist on the team who is able to perform this data entry work. The main responsibilities of a data entry specialist involve:

- Entering relevant project clinical data into the project database (eCRF)
- Performing the control of data entered into the database by another data entry specialist
- Preparing a potential list of discrepancies to be sent to the investigator for data correction
- Processing answered queries
- Performing quality control

The data entry specialist does not need to part of the project team, as he/she works on the project only in very specific parts, especially when some data need to be entered into the database.

2.5.4 Data analyst

The data analyst is involved in the procedures and requirements for statistical activities during the planning/designing, conduct and closure of the clinical registry.

These activities may include performing, contributing to or reviewing the following:

- Support for planning and the optimisation of the clinical registry (including the definition of objectives and endpoints, selection of retrospective or prospective designs, determination of appropriate analysis methods and hypotheses with respect to the objectives and endpoints/variables, sample size determination and the set up of interim as well as the final analysis)

- Writing a statistical part of protocol, reviewing the whole protocol from the statistical point of view
- Development of a case report form (CRF)
- Writing a statistical analysis plan (including examples/templates/shells of tables, figures and listings)
- Data analysis processing
- Statistical reporting (including participation on writing publications)
- Providing analytical services in terms of processing administrative and operational data obtained from healthcare facilities, data-related and analytical support to projects, such as the Health Technology Assessment (HTA)
- Other support (statistical input/contribution at discussions and meetings)

2.5.5 Quality manager

The quality manager is an important control body throughout the entire life cycle of the project. He/she is responsible for establishing the quality system its management, improvement and control systems to ensure a high quality of provided services and products by meeting specific customer requirements.

Responsibilities of a quality manager include:

- Ensuring that relevant legal and normative requirements are established into the quality system of the organisation (GPP, ISO/IEC 9001, ISO/IEC 27001, ISO/ IEC 20000-1, GDPR and other relevant local legal requirements for registry management)
- Ensuring that all processes are defined and clearly described in internal standards (Standard Operating Procedures, Working Instructions, ...)
- Management of control documentation (i.e. ensuring that there is integrity, availability and confidentiality of control documentation)
- Execution of internal audits and CAPA management
- Ensuring that employees participating in the project are trained properly
- Management of the incidents and non-conformities
- Risk management
- Change management

2.5.6 Project support: helpdesk

Managing complicated clinical projects and having an electronic system for data capture means that at least part of the users will need support, especially during the project's operation. For this reason, we recommend having an established helpdesk team who will be able to receive and partly resolve eventual issues and help users to work properly according to the project's needs. The main responsibilities of the helpdesk team are:

- Dealing with incoming emails and phone calls
- Resolving simple issues
- Helping users online

- Distributing requirements to the project team according the nature of request
- Replying to emails within a short period of time (24–48 hours are recommended, according to the type of project)

The key period when the helpdesk is beneficial is during the project setup, when many new users need help with a new project and a new system. The helpdesk is also very helpful over the entire project duration because it is very convenient for users to be able to contact someone who could help them with unexpected issues or just advise them once they forget or make a mistake.

2.5.7 Pharmacovigilance manager

The pharmacovigilance manger is responsible for safety management, adverse event reporting and the reconciliation procedure.

The pharmacovigilance manger's responsibilities consist of:

- Monitoring, recording and reporting of adverse events reported by investigators via clinical registries, follow-up processing towards marketing authorisation holders
- Ensuring the reconciliation of adverse events
- Processing and administration of internal documentation, SOP
- Compliance of the registry's setup with pharmacovigilance guidelines
- Collaboration with project manager when setting up studies (CRF revision, preparation of relevant pharmacovigilance part in contracts with the sponsor)
- Setting regular training sessions for investigators and all project collaborators in the area of pharmacovigilance
- Preparing for and participating in quality assurance audits conducted by study sponsors, regulatory agencies or other review groups

2.5.8 Medical expert

In a clinical registry, the medical expert plays the role of an expert guarantor. The medical expert's responsibilities include:

- Defining project objectives
- Protocol approval and CRF approval
- Guaranteeing compliance with the interests of the professional society
- Project coordination at the highest level, communication with investigators
- Presenting clinical registry results at centres' meetings, conferences etc.
- Publishing results, writing research articles
- Participating in professional society meetings and conferences
- Negotiating with representatives of the regulatory authorities and healthcare payers
- Incorporating results into common clinical practice

2.6 Quality assurance in pharmacoepidemiological research

Quality assurance (QA) is the control body of relevant processes in clinical research projects. The aim of quality assurance is to make valid, scientific statements based on the results in NIS, in hopes to avoid the possibility of bias in the results by using an appropriate study design and an adequate data analysis – assuring data authenticity, completeness and validity and identifying and resolving deficiencies at an early stage [38].

Quality assurance checks and verifies that:

- All study-related activities were appropriately planned and documented, the study is performed in compliance with the study protocol and the other relevant documents and all activities are adequately documented
- Potential and real risks were considered and evaluated and preventive actions to decrease these risks were introduced
- Responsibilities for particular actions were defined
- The study is processed in compliance with the relevant legal and normative requirements

2.6.1 Registered documentation (or documented information)

It does not matter whether the documentation is available in printed or electronic form; it must be controlled in both cases.

Controlled/registered documents are summarised in Figure 10.

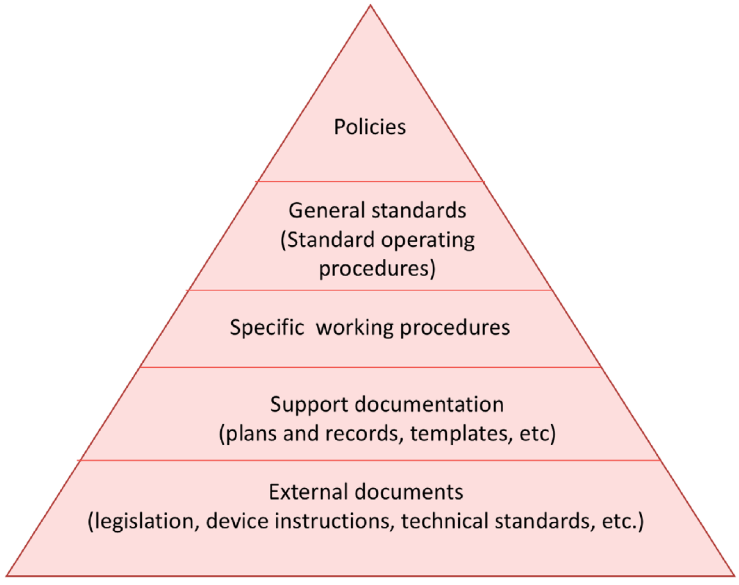


Figure 10 Controlled documentation

Controlled documentation should meet the following requirements:

- The documents should include a predefined identification (title/name, date, author, version)
- Suitable format – determined by the organisation (language, graphics, paper-based vs electronic etc.)
- A person should be appointed to review and to approve the documents (according to the kind of documentation)
- Controlled documentation should be available and suitable for use when and where it is needed
- Controlled documentation should be adequately protected (integrity, confidentiality)
- The organisation should manage distribution, access and use
- The organisation should manage changes to the documents (changes are recorded, documents are versioned and all versions are archived)
- The organisation should manage the storage, archiving, backup and disposition of controlled documents, especially documents with confidential data

Individual quality standards require different types of documents. Some documents are the same for all standards, but some documents are required only by specific standards. An overview of documents is shown in Table 15.

Table 15 Overview of documentation relevant to a particular standard

Document	GPP	ISO/IEC 9001	ISO/IEC 27001	ISO/IEC 20000-1	GDPR
Policy	x	Quality Policy	ISMS Policy	ITSM Policy	Yes
Process documentation (SOP, WI, internal standards)	Yes	Yes	Yes	Yes	Processes relevant to personal data processing
Project documentation (specification of product, customer requirements, protocol, ICF, CRF, interim and final reports, contracts agreement, ...)	Yes	Yes	Yes	Yes	Explicit informed consent
Plans and records relating to processing (records from audits, records of nonconformities and incidents, records about changes, records from meetings and communication, metrics, ...)	Yes	Yes	Yes	Yes	Records about the processing of personal data

Document	GPP	ISO/IEC 9001	ISO/ IEC 27001	ISO/IEC 20000-1	GDPR
Context of organisation	x	Yes	Yes	Yes	x
Risk analyses	x	Yes	Yes	Yes	Yes
Statement of applicability	x	x	Yes	x	No
Registry of protected assets	x	x	Yes	x	Yes

2.6.2 Supported processes managed by QA

Internal audits

An internal audit is the main control tool for an organisation.

An internal audit is a systematic, independent and documented process for obtaining objective evidence and evaluating it objectively to determine the extent to which criteria are fulfilled.

Internal audits are usually performed by an employee of the organisation, but it is possible to perform this kind of audit by utilising an independent organisation on behalf of the organisation.

Internal audits should be conducted in regular intervals according to the internal audit plan.

According to ISO 9001, an organisation shall:

1. Plan, establish, implement and maintain an audit programme including the frequency, methods, responsibilities, planning requirement and reporting, which shall take into consideration the importance of the processes concerned, changes affecting the organisation and the results of previous audits
2. Define the audit criteria and scope for each audit
3. Select auditors and conduct audits to ensure the objectivity and impartiality of the process
4. Ensure the results of the audits are reported to relevant management
5. Take appropriate corrective actions without undue delay
6. Retain documented information as evidence of the implementation of the audit programme and the audit results [60]

The sectional points of an internal audit are displayed on the following scheme (Figure 11).

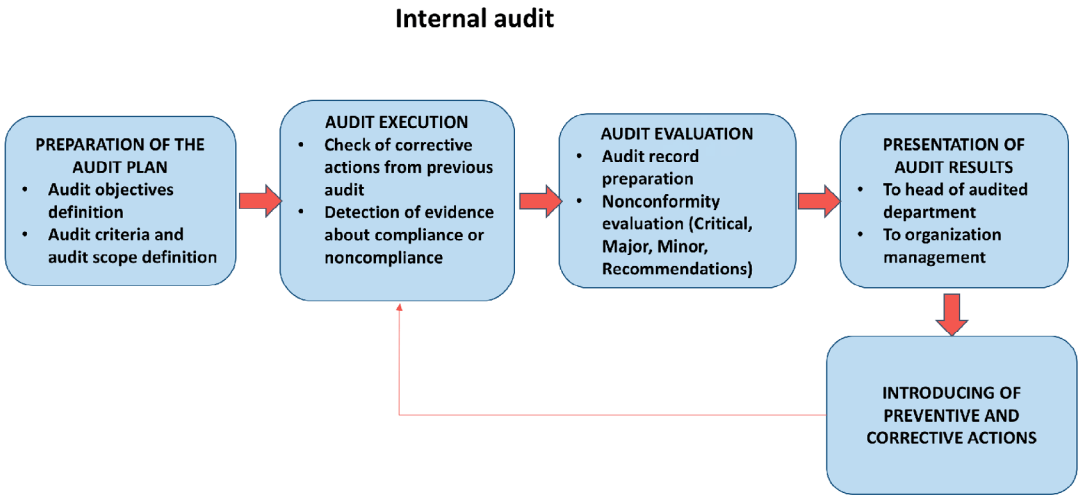


Figure 11 Sectional points of internal audit

Top management shall regularly review the QMS to ensure its continuing suitability, adequacy, effectiveness and alignment with the strategic direction of the organisation. The outputs from this QMS review are inputs for other areas of improvement.

External audits

The CRO may be audited at any time by the study sponsor. The conditions for the audit should be defined in the contract agreement.

The organisation may be also audited by a regulatory authority.

Risk management

Risk management is a necessary part of every project. All relevant standards are based on risk management. ISO/IEC 31 000 Risk management is a special standard describing risk management. The organisation should manage the risks, which means identifying, analysing, evaluating and modifying the risks by treatment (technical or organisational actions) in order to minimise the risks to acceptable criteria.

Risk analysis is the main tool for risk assessment. Risk analyses involves considering the causes and sources of a risk, their positive and negative consequences and the likelihood that those consequences can occur [60].

As a first step, the risk sources / protective assets should be identified (resources – employees, finance, SW, HW, facilities etc.)

As a second step, the possible threats should be identified (natural disasters, failure of the technique or employees etc.)

As a third step for each asset and threat, the value of the probability, consequence and vulnerability should be established. According to GDPR, processes with an influence on the data subject should be included into the risk analyses due to the impact of risk on the data subject (for more information see the chapter Personal data protection (GDPR)).

The final risk value is = Probability × Consequence × Vulnerability (× Impact of risk on data subject)

The organisation should set up a limit of acceptability. According to the final value, the risk can be:

- Acceptable – the limit of acceptability was not exceeded, no actions are required
- Not acceptable – the limit of acceptability was exceeded, corrective and preventive actions are required (for more information see the subchapter Corrective and preventive actions)

For detailed information about risks relevant to registry management, see the chapter Risk management plan.

Nonconformity management

The process of managing nonconformities describes all activities from the moment of detecting nonconformities until their remedy.

The objectives of nonconformity management are to:

- Settle the nonconformities which have arisen
- Prevent the risk of new nonconformities
- Maintain the image of the organisation
- •Remove the nonconformities which have arisen
- Limit the extent of consequential damages

Scheme of nonconformity management

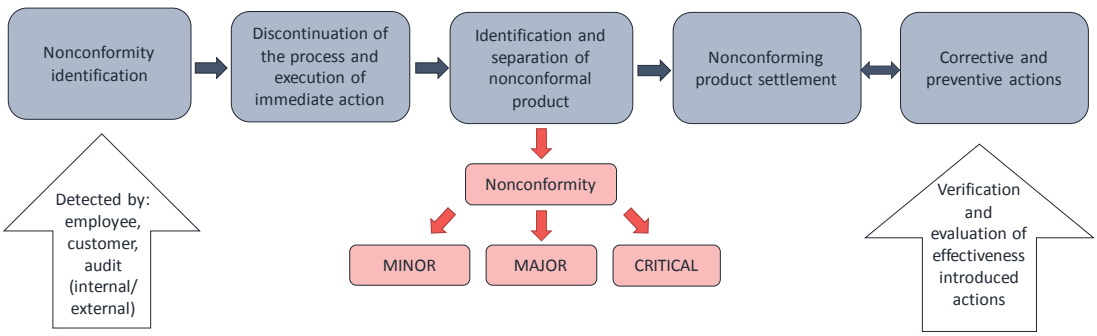


Figure 12 Process of nonconformity management

The process consists of the following steps (see Figure 12):

1. Nonconformity identification (The nonconformity can be identified and subsequently announced by the employee, customer or detected during the internal or external audit)
2. Discontinuation of the process and execution of immediate actions (The nonconforming product should be immediately marked after detection to prevent its use. There should be accepted immediate actions to prevent other nonconforming products)

3. Detection of whether the nonconformity is a process or system error and determining the root cause. After this, the source of the nonconformities should be removed from production
4. Introducing corrective and preventive actions to ensure the nonconformity will not be repeated
5. Verification and evaluation of the effectiveness of introduced actions (After the implementation of corrective and preventive actions, it is necessary to track and evaluate if the introduced actions are sufficient and effective or if it is necessary to execute other actions)

All steps through the process of nonconformity management must be appropriately documented. The form of documentation (record of nonconformity, registry of nonconformities) is fully dependent on the organisation.

Incident management

Information security is a term from the ISO/IEC 27001 standard. It means maintaining the confidentiality, integrity and availability and other features of information such as authenticity, responsibility, non-repudiation and reliability.

An incident is a specific and identified event, which is usually associated with the failure of provided services or their parts or with a decrease in their quality.

A security event is an identified state of the system, service or network, pointing to possible violations of the security policy or the failure of security measures; this can also include different situations not seen before, that may be important in terms of information security.

Incidents are classified into three categories according to their severity and extent:

Standard incidents

Standard incidents mostly have a short-term character and occur due to elementary errors of the users or a temporary operating fault of the ICT; they are usually solved through a routine intervention of a helpdesk worker, department of data administration or ICT administrator. In some cases, however, these standard incidents may require more complex methods of management (removal or elimination) and will subsequently fall into the category of serious incidents or the category of security incidents.

Serious incidents

Serious incidents are events or system messages that indicate that the respective service is clearly not provided at all or shows shortcomings that are materially inconsistent with the guaranteed parameters of the service (outages are repeated or their duration greatly exceeds the tolerated and guaranteed service downtime, or the service provided does not include all qualitative and quantitative elements).

Security incidents

Security incidents (information security incidents) are unwanted or unexpected security events associated with a considerable likelihood of compromising the security of protect-

ed information. This category also includes incidents involving privacy violations. In order to ensure the least amount of damage, the following three points are important:

- Early detection
- Fast reaction
- Sufficient prevention

The whole process for incident management is displayed in Figure 13.

Scheme of incident management

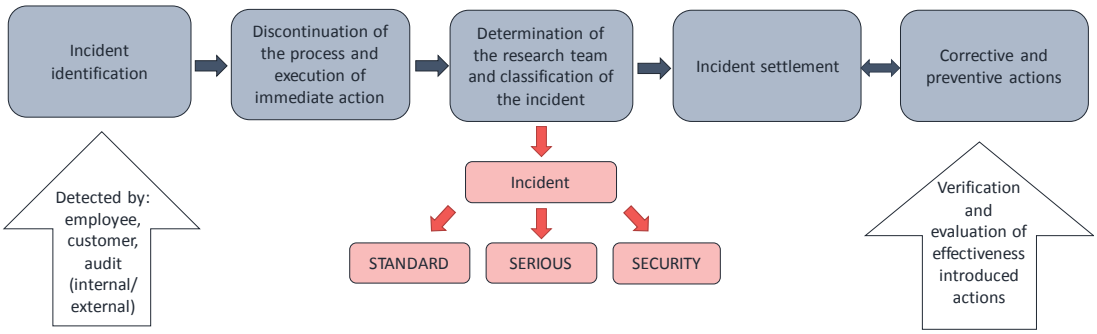


Figure 13 Scheme of incident management

All actions must be thoroughly documented. The method of documentation depends on the organisation.

Corrective and preventive actions

Corrective and preventive actions, also called CAPA, are instruments for the improvement of the organisation's processes. CAPA are part of the processes for ensuring the quality of the provided services and products. CAPA are the reactions to lacking items/ nonconformities found during the audits, nonconformity/incident management and risk analyses.

If noncompliance that significantly affects or has the potential to significantly affect human subject protection or reliability of trial results is discovered, the sponsor should perform a root cause analyses and implement appropriate corrective and preventive actions [61].

During the implementation of CAPA, the following aspects are usually considered:

- The likelihood that the nonconformity will be repeated if its causes are not eliminated
- The complexity (labour consumption) of potential corrective actions
- The costs for the eventual implementation of the corrective action
- Customer satisfaction (complaints)
- Nonconformity and incident frequency
- etc.

All steps in CAPA management must be recorded and documented. After the implementation of CAPA, the actions must be monitored to evaluate if the implemented actions are sufficient or if it is necessary to introduce other actions.

Change control

The change control process combines and interconnects the subprocesses of initialisation, collection of information, evaluation, decision, planning, resources (people, finance, technical aspects), management, implementation, testing and the documentation of changes.

Change control consists of the following steps (see Figure 14):

1. A specific request for a change is made (quality increase, efficiency improvement, security increase, customer requirement, new technical possibilities, changes in legal or directive regulations etc.).
2. Each request for change has to be considered and assessed. It is necessary to define what benefits the change will bring and what are the potential risks. After that, resources should be defined and an implementation plan stated (partial steps and responsibilities).
3. On the basis of available information, the final decision is taken if the change is approved or rejected.
4. Change execution should be performed according to the implementation plan. All steps should be documented.
5. After the implementation of the change, the control systems should be introduced to verify the influence of the change on the system.

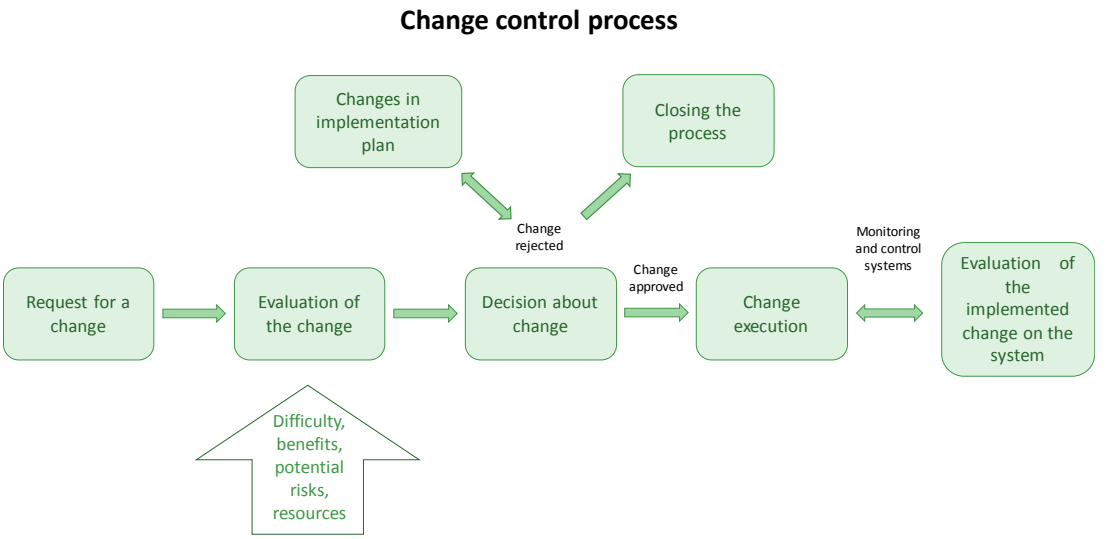


Figure 14 Scheme of change control process

It is necessary to determine a responsible person. Each process should be appropriately recorded and documented. The form of documentation depends on the organisation (change control form, change registry).

Training of employees

Training of new employees as well as the continuous training and education of current employees is a key point for ensuring a high quality and continuous improvement of services provided by each organisation.

Therefore, it is important to implement a training process in the organisation.

This process begins by the selection of employees. Within the selection procedure, high demands on education, knowledge and the experience of the potential employee are required.

Each new employee has to be properly familiarised with the internal standards / internal processes. The familiarisation has to be properly documented.

Each employee has to be familiar with updates to the relevant internal standard(s).

If it is required by the customer/sponsor of the study, all employees participating in the study have to be trained on the sponsors SOPs and other relevant documents.

2.7 EDC system

CLADE-IS (Clinical Data Warehousing Information System) belongs to the class of the most modern and progressive EDC systems developed on the grounds of web technologies. It comes with the Entity–Attribute–Value model (EAV) for the database. The EAV model is also known as the vertical database model or open schema and it allows for the encoding of data structures or entities with sparse characteristics. This is exactly the case of data entities in clinical research databases where the number of parameters or attributes used to describe an entity is potentially large, but the actual number of the applied attributes with assigned values is relatively low. At a high level, the openness of the data model means that CLADE-IS is in fact a robust platform for the design and deployment of standalone EDC systems for data collection and management in clinical studies of various types, across a wide range of clinical fields.

The design capability of the platform is provided by its original and complex online form builder and form generator. These components allow the defining of all the necessary for clinical data collection, such as study arms, phases, forms, question groups and individual questions with all possible types of data in the expected answers, including categorical data. The efficient deployment capability of CLADE-IS is achieved with the inherent data access model, which allows countless configurations of user privileges, roles and data flow. In the default configuration, the following user roles are recognised inside CLADE-IS: (i) investigator, (ii) site manager, (iii) regional coordinator, (iv) data manager, (v) monitor, (vi) administrator.

CLADE-IS can be used in two different setups: (i) multicentric data collection, as shown in Figure 15; (ii) ICT infrastructure for clinical data collection and harmonisation in the centralised data warehouse, as shown in Figure 16. The latter setup represents an optional way of collecting personal data and storing them outside the central repository of the system: personal data, if collected, are stored in a data warehouse run as part of an IT infrastructure of the respective healthcare facility or other GDPR compliant institution. (For more information about GDPR see the chapter Personal data protection (GDPR))

Only authorised users – based on their login and password – can access the online interface of CLADE-IS. Data in the registry are anonymous: for each patient/case, a unique ID is always generated. The communication between CLADE-IS and its users is encrypted with the use of SSL (Secure Sockets Layer).

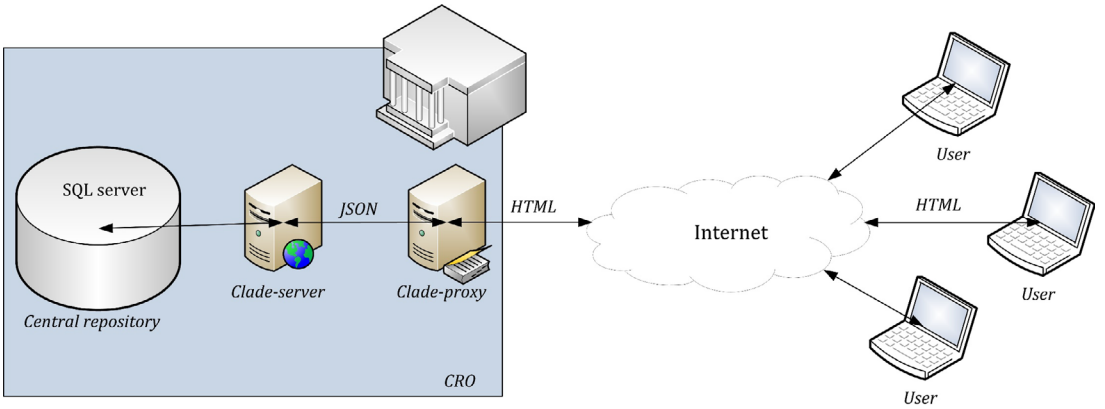


Figure 15 A typical setup of CLADE-IS providing solution for multicentric data collection

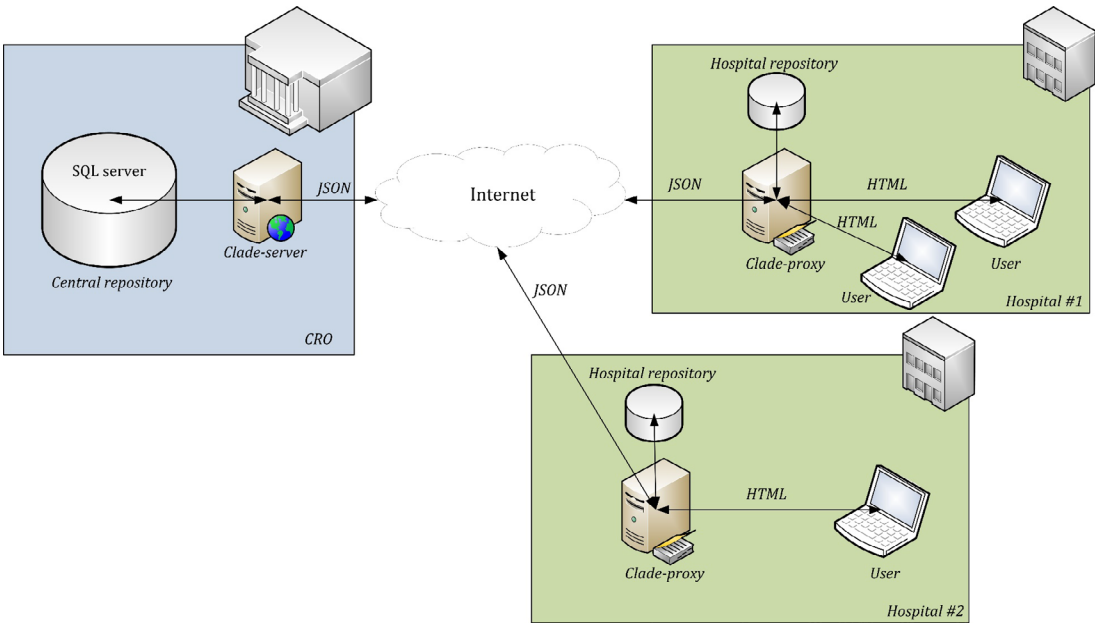


Figure 16 The optional de-centralisation of CLADE-IS, with the CLADE-proxy component implemented inside each clinical site (or other GDPR approved institution), which is useful for personal identity data filtration. In this setup, CLADE-IS provides an ICT infrastructure for clinical data collection, harmonisation and management in the centralised data warehouse

With the use of responsive webdesign, CLADE-IS provides its users with an easy and ergonomic interface. Navigation and reading require only a minimum of resizing, panning and scrolling; a wide range of devices can be used: from desktop computers to tablets and smartphones. CLADE-IS works in the majority of available web browsers, and therefore does not require any desktop application to be installed by the user (i.e. investigator). It is recommended to use only up-to-date browsers running on up-to-date operation systems.

CLADE-IS basic components and modules

Although the CLADE-IS platform encompasses many modules and components, not all of them have to be included in each individual deployed EDC system. There are, however, four basic components that are involved in each particular CLADE-IS installation. Note that one CLADE-IS installation can be used to deploy multiple EDC systems and each of them can support more than one clinical study.

Server – a component that has no interface viewable in a web browser. It supports the storage of study data with the assurance of its integrity and adherence to eCRF structures and a configuration of data access management and data flow (user permissions, form states etc.).

Proxy – the only component that is directly in contact with the end-user's web browser through which data are entered into forms. It communicates with the CLADEIS-Server via REST API (JSON Data Exchange Format), builds a web page based on the received structured data and sends the generated HTML code to the end-user's browser. It is the only component that works with personal patient data and therefore should be deployed in the IT infrastructure of a GDPR-approved institution (e.g. clinical sites where the data are collected). In case of anonymous data collection, it can be deployed in the same infrastructure as the CLADE-IS-Server (e.g. inside the CRO providing the data management services).

Adminer – a module for the configuration of data access management and data flow within one single study. It also provides data managers with data import features.

Designer – a module for building and generating eCRFs. It provides an online interface for the definition of all entities and data structures required by a particular study or a data collection, i.e. study arms, phases, forms, question groups and individual questions with all possible types of data in the expected answers, including categorical data (drop-down lists or combo boxes for making a choice among a list of mutually exclusive values). It also provides data managers with features for skip logic definitions, interactive calculations, validation procedures and instant help tools.

CLADE-IS optional components and modules

These modules and components enrich the CLADE-IS platform with additional features. All of them may run as standalone web applications which are capable of serving multiple CLADE-IS installations.

Reporter – a module for composing operational reports and providing their automated delivery. It also provides data managers and analysts with features towards complete study data export and data transformations that enable them to build data views for analysts and online visual analytics.

Central Administration Tool (CAT) – a component that extends the basic functionality of the Adminer module. Besides the capability of managing multiple CLADE-IS installations it also provides an authentication service for all CLADE-IS components and modules and thus enables the SingleSignOn property. The investigators who are enrolled in a series of studies (e.g. in the same field of medicine and in one particular region) do not have to use different usernames or passwords and seamlessly sign on the studies.

Labyrinther – a module for the creation of virtual scenarios for the facilitation of scenario-based learning – a safe way for training in competency, clinical reasoning and decision-making. Virtual scenarios encourage users (e.g. the investigators in lifelong training courses) to use their knowledge base for exploring simple management decisions as they work through a case that mimics a real-world problem. In addition, virtual scenarios can be used to extend the common study protocol and parameterised data collection by enabling the entering of unstructured text data with subsequent extraction and analysis using natural language processing and machine learning algorithms. Datavisor – a module for scientific calculations and online data analytics over primary data warehouses transformed from the collected datasets. It enables online data browsing according to the principles of modern visual analytics. The openCPU platform is employed here to mediate between the web server and mathematical computing platforms such as R, Matlab etc.

CLADE-IS technologies

In the long-term perspective, software development within the CLADE-IS platform is oriented towards open-source technologies and products. The CLADE-IS developers prefer open-source technologies with long-term support from other developer communities or software houses which release their codes to the public. The most recent stable versions of all included packages are used. No additional proprietary software licences are needed to obtain the full functionality of the online EDC system capable to securely collect, validate, process and analyse clinical study data.

Table 16 An overview of the technologies used inside CLADE-IS and the underlying ICT infrastructure.

Database	PostgreSQL	PostgreSQL is a powerful, open source object-relational database system. It has more than 15 years of active development and a proven architecture that has earned it a strong reputation for reliability, data integrity and correctness.
	Doctrine ORM	Object Relational Mapper that is a powerful part of the database abstraction layer – this provides developers with a powerful alternative to SQL that maintains flexibility without requiring unnecessary code duplication.
Application	PHP	Server-side scripting language designed primarily for web development but also used as a general-purpose programming language.
	Symfony	The leading PHP framework to create websites and web applications.

Communication	REST API	CLADE-IS employs RESTful Web services, in order to provide interoperability between computer systems on the Internet. For instance, the Canadian WAPPS-Hemo project communicates with CLADE-IS through REST API when pharmacokinetic analyses are computed in the Czech National Haemophilia Programme (CNHP) registry project.
	JSON	Data exchange format used for internal as well as external (REST API) communication.
	HTTPS	HTTP over SSL – the protocol used for server communication. The main motivation for HTTPS is authentication of the visited website and protection of the privacy and integrity of the exchanged data.
User interface	ZURB Foundation	Foundation is a family of responsive front-end frameworks that make it easy to design responsive websites and applications. Responsive Web Design is about using CSS and HTML to resize, hide, shrink, enlarge or move the content to make it look good on any screen – smartphones, tablets, personal computers etc.
	jQuery	Feature-rich JavaScript library which is used to enrich the front-end of CLADE-IS with easy-touse features (online calculations, online validation procedures, skip logic etc.)
	Knockout	JavaScript library enabling CLADE-IS to provide dynamic user interface.
Operation system	Linux Ubuntu LTS	The underlying servers use one of the most stable Linux distributions Ubuntu – LTS stands for long-term support – which means five years of free security and maintenance updates.
Virtualisation	KVM	Kernel-based Virtual Machine is a full virtualisation solution for Linux on x86 hardware. The use of KVM results in faster recovery procedures in case of hardware failure or disaster.

	R	R is a free software environment for statistical computing and graphics. CLADE-IS employs R for sophisticated data analytics. The OpenCPU system is used to provide a reliable and interoperable HTTP API for data analysis based on R.
Data & Visual analytics	D3.js	Online analytics are composed with the use of D3.js – a JavaScript library for manipulating documents based on data. D3's emphasis on web standards gives the full capabilities of modern browsers, combining powerful visualisation components and a data-driven approach to web site manipulation.

2.8 Protocol

The protocol is the basic document containing all information about the registry. Each registry should have its protocol before the eCRF is set up.

The protocol is usually written by members of the project team along with the guarantor of the registry. The protocol should be approved by all participating sides, i.e. guarantor, study lead, CRO and sponsor.

Once the protocol is internally approved, it is submitted to the ethics committee and regulatory authority for evaluation and official approval. The procedure depends on the country's regulatory and legislative rules. The approval procedure must follow all local legislative requirements or regulations and medical practice. Local legislative requirements of the countries in which the study takes place must be applied.

In case of any changes in the protocol during the registry's life cycle, the protocol should go through the approval procedure again.

The protocol should be distributed to all centres before data collection starts. The project manager should train all investigators mentioned in the protocol.

The project manager's responsibility is to keep the protocol updated. Once there are some changes, the project manager has to inform all registry participants about the changes and updates to the protocol. Changes have to be approved by the guarantor and sponsors. The protocol belongs under the registered documentation and should follow specific rules (for more information see the chapter Registered documentation (or documented information).

The protocol should contain the following items:

- Descriptive title, version identifier, date of issue, the registration/protocol number
- Names, titles, degrees, addresses and affiliations of all responsible parties, including the principal investigator and co-investigators, as well as a list of all collaborating primary institutions and other relevant study sites

- Name and address of each sponsor
- Name and address of the contract research organisation
- Project design – basic information about the kind of project (multicentre, national / international, prospective / retrospective, observational, epidemiological ...)
- What are the primary and secondary research objectives, specific aims and rationale
- Investigated population (number and information about subjects included to the project); inclusion and exclusion criteria; details of sampling methods if applicable
- The proposed study tasks, milestones and timeline
- Information about the evaluated medicinal product or group of products
- Information on how the investigators and other participants of the project will be trained
- Procedures for data management (data collection, data preparation, methods for data correction and validation, reporting, data cleaning; description of storage of the collected data, intervals of collection, frequency of follow-ups, types of forms, types of patient cohorts etc.)
- Statistical methods and data analyses (statistical analyses plan (SAP), power analyses, methods of analysing and presenting of results, possible impact of biases, confidence intervals of measures
- Quality assurance and quality control procedures during the project's life cycle
- Regulatory and legislative requirements (informed consent form, ethics committee requirements and other local legal or international requirements)
- Information about the protection of the privacy of human subjects, such as information on whether study subjects will be placed at risk, confidentiality (what kind of personal data will be collected, how the data will be protected: anonymisation / pseudonymisation) or conditions under which the project will be terminated
- Pharmacovigilance – adverse event reporting
- Information about investigator's fee for data entry into eCRF
- Publication rules
- References

The protocol can have appendices such as the list of participating centres, overview of collected data, eCRF, informed consent form template, EDC system description etc. All appendices are integral parts of the protocol. If any changes in appendices are made, all registry participants, ethics committee and the regulatory authority have to be informed. The procedure depends on the approval process and the regulatory and legislative rules in the individual country.

2.9 Other project documentation

All processes of the registry's life cycle have to be precisely documented. Each department participating in the setup and processing of the registry has to make its own documentation, plans and records. The main reason for this is to keep records of each step. This chapter describes the documentation of the data management and analytical department.

2.9.1 Data management documentation

Except the study protocol as the critical document for whole project, there is a group of very useful documents related to the data management process. Some of these documents are very important and should be always applied, whereas some of them are optional and might be helpful in special cases. In this guidebook, we are talking mainly about long-term projects. In the long-term, there can be personnel changes in the project team, significant modifications of the database structure, changes in the legislation while the project is still running etc. In each case, it is very important to stick to the original idea behind the creation of this project, including the major project objectives. Also, the database including all checks is prepared according to initial assumptions and requirements. In order to maintain consistency, it is very important to document all plans and all settings including all partial updates. All documents are in line with the quality management system (see the chapter Quality management plan) in order to guarantee consistency of the whole process.

Data management plan

The so-called data management plan (DMP) is one of two most important documents. This document should be prepared during the initiation phase of each clinical project, once the protocol and basic project attributes are known and approved. DMP is created in agreement with other project plans (see the chapters Project management plan, Staff and communication management plan, Cost management plan, Quality management plan) to keep whole project consistent. The goal of the DM plan is to describe in detail a list of all activities planned to be performed during the operational part of the project. The document should be carefully written in accordance with the protocol and corresponding to all practical expectations regarding data collection and processing (e.g. reporting). The document should be discussed within the project team and finally approved by the sponsor in order to guarantee that the whole process, including all particular parts, will be performed in line with initial expectations. The DM Plan should consist of:

- Project scope. Even though the project scope is usually mentioned in the project protocol, some of its items are also useful to include in the DM Plan, such as the number of countries, number of sites and the estimated number of patients expected to participate in the project. The known number of users might help to set up the database interface correctly. What should not be mentioned in the protocol is the number of language mutations planned to be prepared for the project. In addition, this is very practical and important information.
- List of relevant SOPs. Data management in clinical projects (registries) is usually a complex process containing many particular activities. It should be clear before the project start how different activities will be managed and what processes will follow. Then, it is good to present a list of SOPs that will be followed during the project.
- Requirements on eCRF structure. There are many ways to construct an eCRF/ database. It should be clear that the protocol and paper CRF template will be followed, but we need to specify much more as part of the DM Plan:
    - Special request for the visual design of EDC web pages
    - Relevant form statuses
    - Relevant user roles

  - Request to use simple validation
  - Request to use coherence checks at the form level
  - Request to have complex validation crosswise through the entire CRF
  - Adverse events reporting
  - Enrolment reporting
  - Reports tailored according client needs
  - Translation/Coding
  - Medical review
  - Other optional setup (voting system etc.)
- Structure of collected data and data flow description. There can be more than just CRF/eCRF data in a given clinical project. We need to know all types of data that are planned to be captured, processed and analysed. Except for key CRF data, we can work, for example, with lab data or quality of life data and we must know the structure of the data, how often these data will be processed, by whom and what exactly will happen with these data.
- Information on whether some advanced data management activities is required. Activities such as translation, medical coding or SAE reconciliation are very useful but are not required in all projects. We therefore need to specify at the very beginning of the project whether we need them, and if so, we need to plan detailed processes for these activities such as the expected frequency, involved personnel, data flow, responsibilities etc.
- List of applicable documents. As stated above, there can be lots of useful documents, but not all of them are relevant and helpful for all projects. At the beginning, we should choose those documents that will be needed and helpful for our project. It is highly recommended to define all of these documents at the very beginning so that we know which processes will be described in which document; only then, we can be sure that nothing will be forgotten.
- Preliminary calendar of key milestones. Not only the question of "what", but also "when" is very important in the planning and realisation of a clinical project. We should incorporate this plan into the data management process and to write down key milestones and related timelines. The reason is clear: to be ready and able to manage all of the activities on time.

Data validation plan

A second very important document is usually called the data validation plan (DVP). The DVP should be linked to the DMP and describe in detail all checks and validations planned to be performed on the collected data. Similarly, the DVP should be discussed within the project team and finally approved by the sponsor in order to guarantee that all controls performed on the data are in line with initial expectations. The DVP should consist of:

- General rules on how to work with validations and queries. Overview of validation types used for a concrete project.
- List of simple validations, usually focused on the control of mandatory items, permitted ranges of entered values, and whether an entered value corresponds to the defined format. The goal of these checks is to alert the user that the entered value is wrong and that it should be corrected.

- List of coherence checks on the form level. These checks are based on comparing different values within one form. The goal is to alert the user that some of the entered values are probably wrong in order to correct the data immediately.
- List of complex checks across the whole database. Although complex validations are in principle similar to coherence checks, they compare values among two or more forms; they are therefore more complicated because particular forms are saved at different times so that we have to deal with form statuses and the delay of the identified discrepancy.
- List of manual checks. Still, some checks cannot be programmed. However, if we need what should be compared during the project, it is fine to define even these checks at the very beginning and to mention them in the validation plan.
- List of derived variables. It is useful to display the list of variables where the value is calculated from other values entered into the system (e.g., the derived variable of BMI is calculated from height and weight).
- List of self-evident corrections. SEC are special types of corrections. They can usually be used when a group of the same answers should be automatically coded/rewritten to another value. The original value must be kept, then the new variable is created to also keep this corrected value. This can be used if some investigators systematically fill in some value correctly, but differently than required.

Other relevant DM documents

Some of other documents that are typically useful, but not critical, consist of the following:

- DB validation protocol. This document confirms that the database has been created, tested and validated. It is important for internal purposes of the system provider to guarantee that the database was prepared properly.
- eCRF approval form. It is important to have this document approved before the database is switched to live mode. It approves that key members of the study team are in line with the prepared eCRF, including all functionalities and validations.
- Data entry guide. This document is very important, describing systematically how to log into the system and how to work with particular forms and functionalities.
- Request for DB modification. Unlike clinical trials, registries are long-term projects and there is a high probability that the database/eCRF will have to be updated. For these purposes, it is good to have document where all the requirements for updates could be specified in order to clarify all potential issues before proceeding with any modifications.
- External data. As stated above, if we are handling more than eCRF data, it might be useful to create a separate document to describe the details on external data (lab, QoL etc.). We are focused mainly on data structure, data format, frequency and the type of data flow. We are especially interested in the process of how the external data will be matched to CRF data.
- Translation process. This process is usually applied for international projects where local languages are used for capturing text fields. These items need to be finally translated before further processing. Translation process should be planned; the list of variables for translation should be mentioned.

- Medical coding. If we would like to analyse adverse events, medical history or concomitant medication data, the medical coding process needs to be considered. If applicable for our registry, the process needs to be carefully planned and a specific document could be used as well.
- SAE reconciliation. If safety data are captured into more than one database, we should consider the SAE reconciliation process in order to guarantee that correct and latest data will be analysed. If this process is applicable, it is useful to have a document to describe the whole process properly and to have it approved by the responsible parties.
- Database lock and archiving. Every project will end one day. For this reason, it is good to have a document specifying what will happen once the project comes to its conclusion. How should user accounts be handled? How to deal with the data? These and other questions should be answered in this document.

2.9.2 Statistical analysis plan (SAP)

According to ICH-E9 Glossary [62], the statistical analysis plan (SAP) is defined as follows: "A statistical analysis plan is a document that contains more technical and detailed elaboration of the principle features of the statistical analysis than is described in the protocol, and includes detailed procedures for executing the statistical analysis of primary and secondary variables and other data".

The SAP aims to ensure the integrity of the data analysis and confidence in the final results by the specification of statistical analyses a priori. The SAP comprises a comprehensive and detailed description of the analysis methods to be performed upon the completion of the project and the corresponding lock of the particular database.

In general, the SAP does NOT apply to projects such as long-term registries. However, if there are defined hypotheses conducted within a clinical registry, the SAP is applicable.

The SAP includes the following items:

- Specification of software to be used for the statistical analysis
- Brief description of the project (summary of objectives, design, specification of target population)
- Conditions for preparing the flow of patients (a flow diagram can follow, for example the Consort structure [63]).
- Research hypotheses and endpoints section containing detailed specification of the primary and secondary endpoints and corresponding hypotheses
- Sample size consideration regarding the primary (and secondary, if needed) endpoint
- The baseline characteristics section should outline the methods for the summary and description of the population. A list of variables used to assess similarity, and how these will be reported (e.g. means/standard deviations [errors], medians/range, proportions) should be included. If appropriate, a comparison of characteristics of patients included in the registry and those who did not give their consent should be also added

- Analysis sets (populations) should be defined in the SAP according to the protocol
- Usage and description of any external data (data not recorded in the clinical database)
- Definitions of baseline values or measurements (for example, if more than one baseline measurement is relevant)
- Details of any formulas for derived variables, rules, references or programmes for the calculation of derived variables
- Handling/classification of concomitant medication and adverse events
- A description of the methodology to be used to handle missing data, incomplete dates/times, repeated measurements, multiple observations, outliers, withdrawals and so on
- The efficacy and safety section should address the evaluation of the effectiveness/efficacy and safety of the treatment or intervention under investigation. Both p-values and confidence intervals should be provided with estimates of the effect size for the primary outcome(s). Moreover, the following issues should be discussed, if applicable:
  - An adjustment for multiple comparisons if there is more than one primary outcome
  - If appropriate, a secondary/sensitivity analysis for treatment efficacy that is adjusted for predefined, highly prognostic baseline characteristics, and variables used in stratified randomisation
  - An evaluation of treatment effectiveness or safety with respect to secondary outcomes. Confidence intervals should be provided with estimates of the effect size
  - Any planned subgroup analyses. Regression models with interaction terms would be typically used as well Significance level of the statistical tests, two or one-sided statistical tests
  - Level of validation of outputs, e.g. double programming or review by another statistician
  - A list of planned tables, figures and listings (TFL) and/or examples/templates/shells of TFLs, specifications of data presentation (e.g. sorting in listings, labels to be used for presentation of visits, treatments, time-points or abnormal values)

The process of writing the SAP is driven by a project analyst. He/she ensures an accurate and timely completion, review, approval and storage of the SAP as well as the validation documentation. He/she cooperates closely with the project manager, who should provide all relevant materials, such as the final version of the protocol and (e-)CRFs, and potentially access to registry data or a sample of data.

The SAP has to be written after the project protocol (or similar document) and (e-)CRFs are finalised, and the final version has to be completed and approved before the database lock for the analysis (or an equivalent point of time).

The SAP in the form defined above is usually not prepared for long-term registries because their objectives are more general, descriptive and not exactly specified to one endpoint. However, some activities must be done before each statistical analysis in any case.

The project data analyst has to check the following formal requirements and data properties (from the statistical point of view) and to ensure that they have been completed before the analysis is performed:

- All relevant data-management processes are finalised (approval by the project data manager and/or project manager)
- Confirmation by the sponsor that data are complete (i.e. include all patients relevant to the analysis)
- Any external data are checked for discrepancies
- Statistical hypotheses and endpoints are clearly defined
- It is clearly stated which analyses are planned for confirmatory and for exploratory purposes
- The analysed population (selection criteria) and subgroups are clearly defined
- Correct formulas for all derived variables are given
- Variables to be transformed and formulas for transformation are specified
- Reference ranges for laboratory and other relevant data are known
- Relevant publications with similar data are available, provided by the investigators or sponsor
- The coding of adverse events, concomitant mediation and medical history are finalised and approved

2.10 Pharmacovigilance

Clinical registries as primary data collection systems are also focused on assessing safety associated with the use of medication. The processes of adverse event (AE) detection and reporting should be established.

Pharmacovigilance is defined as the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem (WHO 2002).

2.10.1 Adverse event reporting

An adverse event is defined as a medical occurrence temporally associated with the use of a medicinal product, but not necessarily causally related.

The timely reporting and follow-up of adverse reactions is a requirement described in the Good Pharmacovigilance Practices (Module VI) for marketing authorisation holders in the case of post-authorisation studies based on primary data collection. This requirement cannot be fulfilled if the post-authorisation study is based on a registry where only de-identified drug data can be provided to a pharmaceutical company. In some registries, processes have been put in place to inform investigators about the spontaneous reporting of adverse events/adverse reactions, but there is no standardised procedure across registries. In some cases, this is due to a lack of knowledge about reporting requirements [43].

Clinical registries can often be used to monitor the safety profile of the drug under real-world usage.

The advantage of registries is that their observational and inclusive design may allow for the surveillance of a diverse patient population that can include sensitive subgroups and other groups not typically included in initial clinical trials, such as pregnant women, minorities, older patients, children or patients with multiple comorbidities, as well as those taking concomitant medications [33].

The decision to collect adverse events (AE) in the registry and AE reporting to the marketing authorisation holders is considered at the outset when planning a clinical registry and defining primary and secondary objectives. The collection and reporting of AE is then described in details in the protocol of the clinical registry. If the registry collects AEs, then there is a form in the eCRF devoted to AE and investigators are obliged to use and fill it in when an AE occurs during patient treatment.

The AE form collects the minimum dataset to be collected and reported by investigator:

- Date of AE report
- Adverse event term
- Date of AE onset
- AE duration (this could be automatically calculated from the data in the registry – start date of AE and end date of AE)
- AE first awareness date
- End date of AE
- Outcome
- Seriousness of AE
- Causal relationship with drug
- Measures taken to treat AE

Before the start of the registry, the process for the collection and reporting of AEs has to be established in collaboration with the sponsor and professional society. The process has to be consistent with national and international guidelines. Collection and reporting should be described as internal procedures related to the clinical registry and all investigators and the project team should be trained in these procedures.

The process for the collection and reporting AEs could be setup as follows:

1. The patient experiences any adverse event and reports this to the investigator during a visit. The investigator performs a causality assessment and enters the adverse event along with the other information directly into the registry within 24 hours.
2. The filled-in and completed adverse event form could be automatically generated as a report by the EDC system. This step depends on the EDC system setup and its functionalities.
3. The automatically generated AE report is directed to the predefined e-mail addresses of the pharmacovigilance department of the marketing authorisation holder and the pharmacovigilance department of contract research organisation.

Reporting form the registry does not replace the duty of the investigator to report AEs to the local competent authority according to the current laws and regulations. This fact is described in more detail in the protocol of each project.

2.10.2 Reconciliation

We speak about reconciliation especially in dealing with the issue of serious adverse events, but the process is also applicable more generally, i.e. for all ad verse events or potentially for any data of special attention being collected at two or more independent places.

Why reconcile safety data?

Especially due to regulatory and safety reasons, there is often special attention dedicated to adverse event data. These data are usually collected together with other project data but they are frequently reported to the pharmacovigilance department of the CRO or sponsor for further processing and pharmacovigilance purposes. During this processing, the PV department can ask the investigator additional questions for follow-up or to make the reported information more precise. An adverse event can be reclassified too. In practice, it might happen that this follow-up information is not reported into the clinical database anymore, and they are reported only to the safety database managed by the PV team. However, global statistical analyses are usually performed on data in the clinical database so it might happen that we will analyse obsolete data. To prevent this situation, we should consider the process of reconciliation to compare data in both databases in order to unify them and ensure that we are analysing the correct data.

When to reconcile data?

The goal is clear. Before planning the concrete steps of the reconciliation process, we should consider the project's agenda, key milestones and expected outputs. In projects where we have only one source of data and all data modifications are performed in one database, we do not need to perform the reconciliation process at all. If we work with two or more databases, we should integrate the reconciliation process into the agenda of particular steps. In light of the previous paragraph, it is critical to perform reconciliation before providing data for statistical analyses. Nevertheless, it is not the only time point to reconcile the data. It might be performed, for example, on a quarterly basis to maintain data accuracy in the clinical database. If the first reconciliation process is planned after a long period of time, it is recommended to perform it once after a few months of data collection for checking purposes, to see if the process is running correctly as part of the data review process.

How to reconcile data?

The first point is to plan the whole process, ideally during the initiation phase of the project. It is recommended to create a plan of reconciliation containing:

- Names of parties responsible for reconciliation
- Expected frequency of reconciliation process
- Format of reconciled variables
- List of variables to be reconciled
- Consistency rate, e.g. a description if the compared items must be an exact match or if they can be consistent only in their meaning
- Description of the queries process

It is necessary to compare exports from both databases, to check them field-by-field and highlight any identified inconsistencies. Inconsistencies should be resolved via the queries process and the database updated until there are no remaining inconsistencies.

The Figure 17 displays the common steps for the SAE reconciliation process.

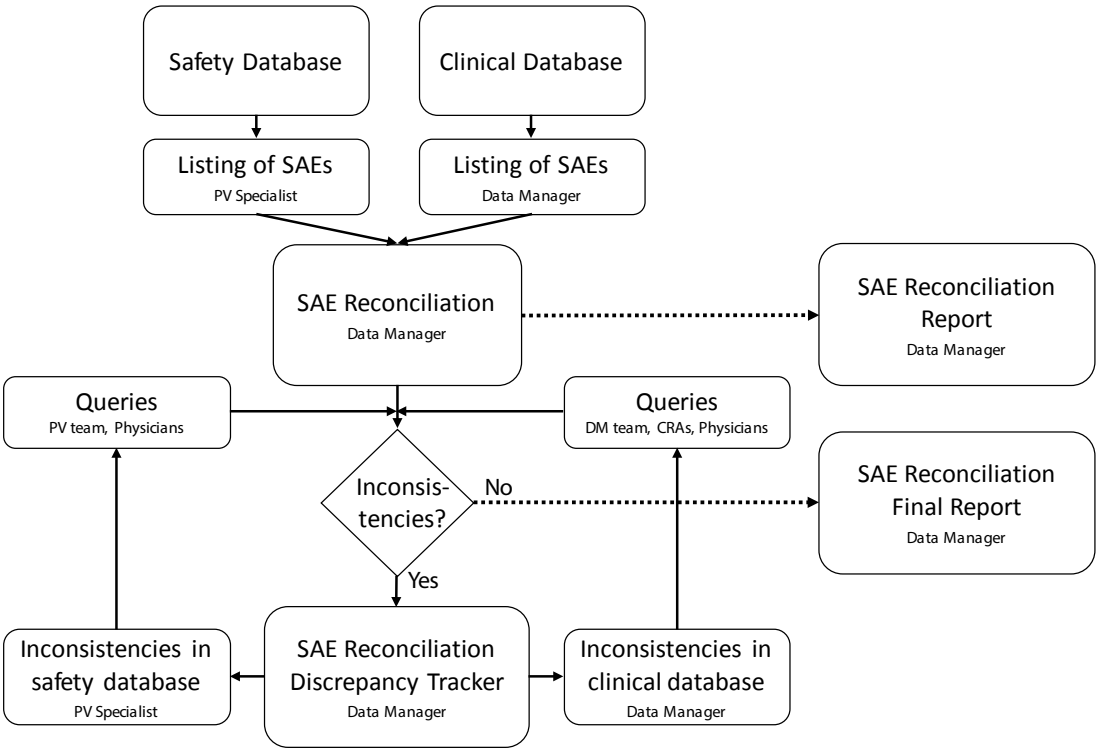


Figure 17 SAE Reconciliation process

3. PHASE 2: HOW TO CREATE THE REGISTRY

3.1 Introduction

After many preparation steps performed mainly under the responsibility of the project manager, the practical creation of the registry primarily falls under the responsibility of the data management team. As you can see in the scheme below, there are several steps leading to the live release of a registry. The scheme can be divided into three columns:

- The first column (final protocol to final CRF) represents preparation activities and the fine tuning of key documents before the start of database development.
- The second column (database setup to database final version) represents one of the crucial activities in building up the registry database.
- The final column (data entry guide to FPI) represents several minor but important steps allowing the registry database to go live.

The individual steps displayed in the scheme (Figure 18) are described in detail in the following chapters.

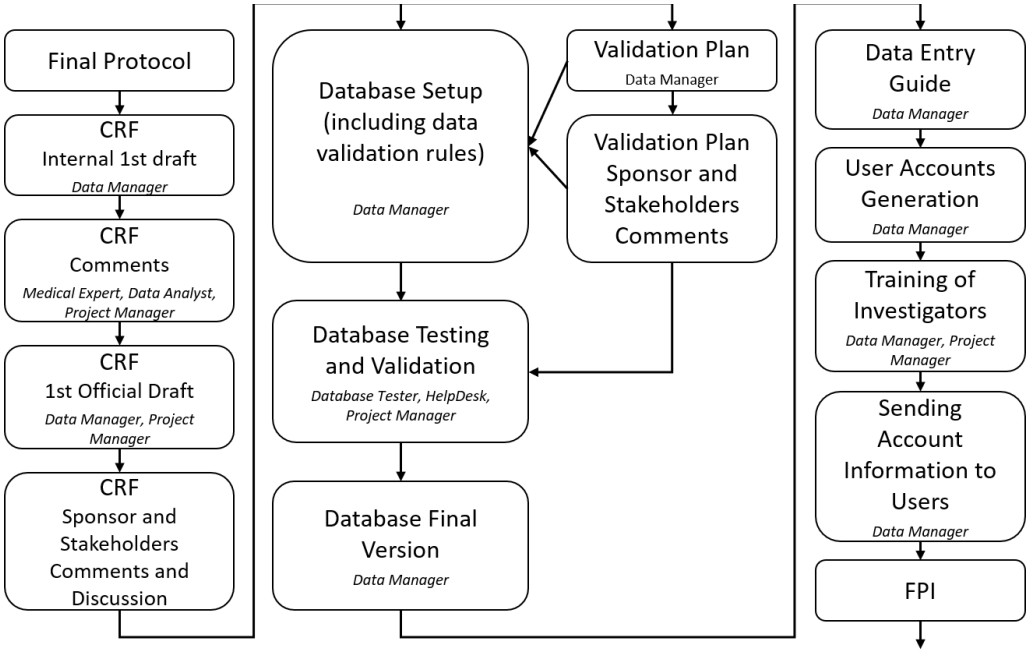


Figure 18 Data Management Process: Final Protocol to FPI

3.2 CRF creation

This chapter highlights the importance of the CRF existence and describes the personnel and stakeholders participating in CRF creation and their roles. A significant part of this chapter is on selecting the right parameters and CRF structure.

A case report form (CRF) is a specialised printed or electronic document that is designed to record all the information required by the project protocol (more details about the pro-

tocol are available in the chapter Protocol). It is a crucial document for clinical research and it should collect a sufficient amount of information from all participating subjects needed to answer questions based on both primary and secondary project objectives.

This form is used as a template for database creation; therefore, it is essential to think about the structure of visits from the very beginning. There may be a few short questionnaires following identical parameters, which are then compared over time or several-year patient monitoring in which a large number of parameters are collected and evaluated. In both cases – simple questionnaires and long forms – the main objective is to collect information that will be accurately analysed. For this reason, the data analyst, who is responsible for the analysis, should also comment on the CRF structure. To capture important information, it is necessary to focus on the objectives of the registry, considering the anticipated number of patients and the number of centres involved. A CRF is not intended to collect all available patient data, but to collect the data needed for a particular registry. Therefore, it is important at first to consider which pieces of information can really be useful and which pieces of information are important for answering the main research questions. Questions necessary for an accurate analysis can be marked as mandatory. For all questions, it is important to make sure they are undoubtedly unambiguous. If questions were ambiguous, it might easily happen that different people understand them in different ways and their answers could not be further processed. The easier and more comprehensible a CRF is for the person entering the data, the fewer problems will be encountered after launching the project. On the other hand, it is easier to collect some information and not to use them later than to find out that we are missing some of them during the data collection stage.

It is advisable to focus on the target group when preparing the study. At this stage, the clinical background of the project plays an important role. In case it consists of a disease mainly affecting men/women or older/younger people, it is appropriate to predetermine the target population; in particular, the application of exclusion criteria for the study will ensure that those who might cause outliers in the analysis will be excluded. These criteria should be involved in a CRF and may be required to be met when creating a record on a patient in the data capture system.

Typical parameters of CRF:

Any parameters needed for a specific project may be captured. Typical groups of parameters that are usually collected are given in the points below:

- Project identification, site identification, subject identification
- Current date (Date of project start / screening / first visit – it is useful to have a date which can be compared to all future dates)
- Date of signature of the informed consent
- Basic information about the subjects (age or year of birth, sex, height, weight)
- Subject characteristics – smoking habits, diet, medical history, pregnancy test, previous medication etc.
- Diagnosis, an indication of study medication
- Inclusion criteria, exclusion criteria
- Dosing schedule, compliance

- Treatment duration – list of visits
- Concomitant medication
- Adverse events record
- Premature discontinuation

Potential structure of CRF:

CRF is a complex document that should have a logical structure corresponding to the project design and enabling to develop an electronic version of CRF (eCRF) which will be understandable to database users, i.e. physicians who will be entering data. Examples of CRF structure components are:

- Parameters (variables)
- Paragraphs (question groups) – a paragraph consisting of a set of questions (parameters)
- Forms (pages) – a form is a set of paragraphs
- Visits (phases) – a visit consists of a set of forms
- Arms – an arm is a set of visits

3.2.1 Paper-based projects

In some special cases, data may be collected via paper questionnaires. These may be one-time anonymous questionnaires or a project for which no data collection database is created and no EDC system needed. Another possibility is that the physician writes the required data on a paper questionnaire and another person transcribes the data into a database. In all of these cases involving paper questionnaires, digitisation is required after the data are collected. This process involves converting data into an electronic form, which will make it possible to carry out a statistical evaluation later. Potential errors that can be made by the person transcribing the data present an obvious challenge of this approach. Another potential problem is the impossibility of setting validation, i.e. the person entering the data is not notified if the entered value is beyond the expected range. Another problem may also occur due to unreadability of data written on the paper; if the data is written illegibly and data entry operator is not able to verify the accuracy of the reading, the data cannot be used for analysis and must be labelled as missing. All these errors can lead to data distortion and therefore to a lower quality of statistical analysis. Another important aspect is the time and much slower progress involved in paper-based projects. Nowadays, for most projects, it is recommended to avoid data collection using paper questionnaires except for simple disposable data collection purposes.

CRF creation process

The CRF creation process should always be the product of teamwork.

- The sponsor along with the medical expert should provide the initial parameters which the CRF should meet, including the list of parameters to be captured, project design and medical scheme (all this information should be included in the project protocol).

- The data manager should then design the CRF details, including the structure of individual visits and basic information about the future database and planned manual checks to be applied on the data.
- The data analyst should review the CRF from the point of view of data analysis, especially if all key parameters are included in the CRF in a proper format, which will make possible not only data collection, but also their correct analysis.
- The project manager should review the CRF and make sure that it is in line with the protocol and sponsor expectations and intelligible for investigators.
- The medical expert should review the CRF and make sure that it makes sense from the clinical point of view, as well as that data are collected in a format corresponding to treatment in common practice and on sites participating in this project.

The CRF should be consulted during its creation by all of the above-mentioned team members; it is very important to conduct mutual consultations between departments, because each of these people sees CRF-related issues from their own point of view. For this reason, all of them should go through the final version of the CRF and sign it if they agree. This signature approves that they are able to perform their part of work on the registry. After the final version of the CRF is prepared and approved, it is possible to start creating a database (electronic CRF, eCRF).

3.3 Database setup

This chapter contains a detailed description of the entire process of registry database creation from the very beginning to its activation, including specific details and the explanation of different steps and various functionalities.

Once the final version of the protocol and CRF is obtained, we can start working on the development of the database. Database development is a very complex process involving many aspects, and only a careful approach can lead to demanded results. The following aspects should be taken into account:

- Protocol. Because the project protocol is the key document, we should obviously build a database based on the key information mentioned in it. We should focus on project goals, the list of required parameters or the treatment scheme.
- CRF. The CRF is another key document representing the structure of visits in more detail and is commented on by several members of the project team. We should focus on the structure of the CRF and the variable types of different parameters.
- Data management plan. Complete data management process for the whole project is defined in the DMP (more details about the DMP in the chapter Data management documentation).
- Data. All checks are defined and approved in the DVP (more details about the DVP in the chapter Data management documentation).
- Number of investigators and their PC skills. The more investigators and the worse their PC skills should mean a more transparent database, easier handling, available helpdesk contacts etc.
- Project design. It is not only the structure of the CRF that is important for database development. You should know the detailed design of the project, schedule of planned visits, flow of all collected data and consequences.

- Repeated visits, repeated forms, repeated questions. To be able to develop the database properly, you should know all possible scenarios of how a patient could proceed in the project.
- Translation and codelists. The database should contain only a minimum number of free text fields, otherwise data will be hard to analyse.
- Special requests for database functionalities. Even though we can expect that the majority of activities will be similar to most projects, there are often specific requests that have to be discussed, planned and implemented. Here are only a few questions that must be answered in the initial phase of database development:
  - Is some specific reporting expected?
  - Will safety data be automatically transferred to the PV department?
  - Shall some data be transferred between forms?
  - Should some values be calculated automatically?
  - Is the randomisation process planned?
  - Will the investigators be paid? What is the criterion for payment? Is it according to the number of completed forms? Which forms?
  - Will some data be imported into the clinical database (e.g. from laboratories)?
  - What will be the access policy to different forms? Will all investigators have the same access rights? What about rights for other project team members? What about access rights for patients (e.g. QoL questionnaires)?
  - Will some data be manually entered by someone other than investigators (elsewhere than on site)?
  - Will the number of patients be limited for the project arm/site/region/country?
  - Will some automatic alerting be required (e.g. once the agreed number of patients are enrolled or once a screening failure form completed etc.)?
- Reporting process. Some reports are required from the beginning. But we should think about potential future reports and adapt the database accordingly.
- Adverse events management. Maybe no action with adverse events is required. But at least basic questions about handling with safety data should be asked and consequently answered to develop the database properly.
- Other relevant information. Constructing the database is a big game taking place in the playground of the data manager. The data manager can follow a lot of aspects as sketched out above, but he/she should know the best system in which the database is developed and how data entry screens will look for final users, so he/she should figure out all important information, ask questions and arrange the database in the best possible way.
- There are several possible approaches for how to grasp the process of developing the database. Just keep the process logical and consistent. The following paragraphs will present one of the possible scenarios for developing a clinical database.

3.3.1 Database structure

As described above, the database development with the creation of project arms, particular phases, visits, forms, variables and special functions is one of the key parts of a clinical project's duration. It strongly affects all future steps, data collection, data cleaning

and validation, communication with sites, data quality, data analyses and global results of projects.

The first steps in creating the database are based on the study protocol. The protocol includes information about the name of the study, the website address and the primary language of the project. This information is usually the first practical step in establishing the database. It is also needed to specify the requirement for randomisation or pharmacovigilance reporting if relevant for the study, whether queries will be applicable or not, or if payments for investigators will be connected to form statuses.

Subject form

When creating the forms, we firstly need to create the subject form, which has a specific role. The form needs to be completed for each patient when enrolled in the project. It should contain basic patient data, including the data needed to find the patient in the database (searching options). If inclusion/exclusion criteria are involved the study, the question of the fulfilment of these criteria should be included in this form. After entering the data completely and saving the form, a new patient is created in the database and it is possible to start filling in other forms, as defined in the CRF. If patient randomisation is involved, the patient could be assigned to a specific group based on the information contained in the subject form. On the other hand, the patient could be randomised later; alternatively, a specific randomisation form could be prepared.

Once the subject form is created, the patient ID is automatically generated. It might be useful to insert a country code or site code directly into the ID as it is clearer for further activities in the system. An example of a subject form is shown in Figure 19.

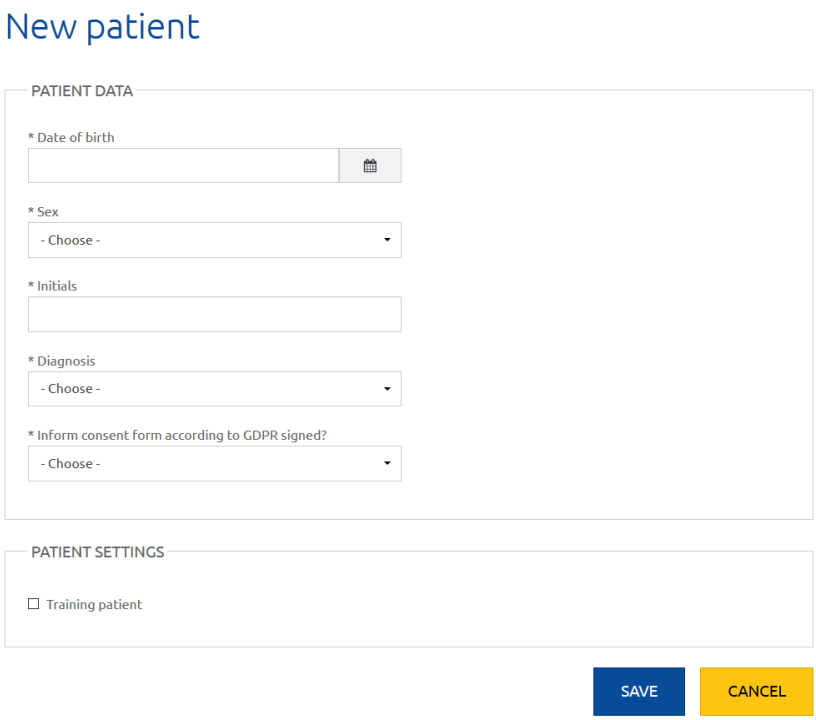


Figure 19 Example subject form

Site form

In addition to completing the form of the subject, the site form may also be required. This form is mainly used in multicentre projects and helps to fill in the patient data that are identical for the centre, e.g. the name of the hospital or the physician's name. Site forms can be also used for reference laboratory values in different sites. There can be a request to compare the entered laboratory values of enrolled patients with reference values that might be different for different sites. Site forms may help us to meet these requirements.

Arms and phases

The first step that has to be developed for data collection in a project is the creation of project arms and phases. There can be one or more arms as well as one or more phases. Different arms are typically used for different diagnoses; different phases are used for particular parts of one subject during a given amount of time. One phase contains one or more forms that are logically composed together (e.g. in the same time period); individual phases can represent, for example, a baseline visit, a treatment visit and follow-up visits.

Phase form

The largest amount of the collected information is mostly gathered in phase forms that record pieces of information from individual visits. The preparation of these forms is based on the CRF. The structure of individual forms and their inclusion in phases should be obvious according to the CRF. Each form consists of individual questions; these are arranged in question groups. The first part of the validation process is to distinguish among mandatory questions. These questions are usually marked with an asterisk and it is impossible to save the form as completed until these questions are answered. Each question should have its place in the structure and its data format should be clearly known from the CRF. The most commonly used data formats are listed in the Table 17.

Database field types

Table 17 and the following pictures (Figure 20 to 24) describe specific data types that have to be set up for all variables involved in the project database. Each field type has some attributes and affect the entire process of data entry and data validation.

Table 17 Data types

Field type	Description	Detailed field type
Text	Short text (usually less than 256 characters)	String
	Text that is expected to have a length of more than 256 characters, such as a comment box, note	Text
	This is used for any purpose that does not require data collection	Help text

Field type	Description	Detailed field type
Number	Number without decimal places (used, e.g. for a count)	Integer
	A real number with one or more places behind the decimal point	Real
Code list	List of more options where only one can be chosen	Radio button
	List of more options where only one can be chosen	Drop-down list
	List of more options where more than one can be chosen	Multiselect
Checkbox	This control is a simple box that EDC users can mark with a check by clicking or leave blank	
Analog scale	Analog scale where the desired value is chosen according to the patient´s feeling	
Picture map	Picture divided in parts from which the desired value is chosen according to the patient´s feeling	
Date	Entering date in pre-defined format yyyy-mm-dd	Date
	Entering time in pre-defined format hh:mm	Time
	Entering combination of date and time in pre-defined format yyyy--mm-dd hh:mm	Date time
	Entering partial date	Partial date

Figure 20 Database field type – Text

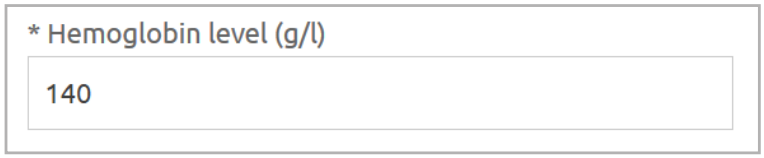


Figure 21 Database field type – Number

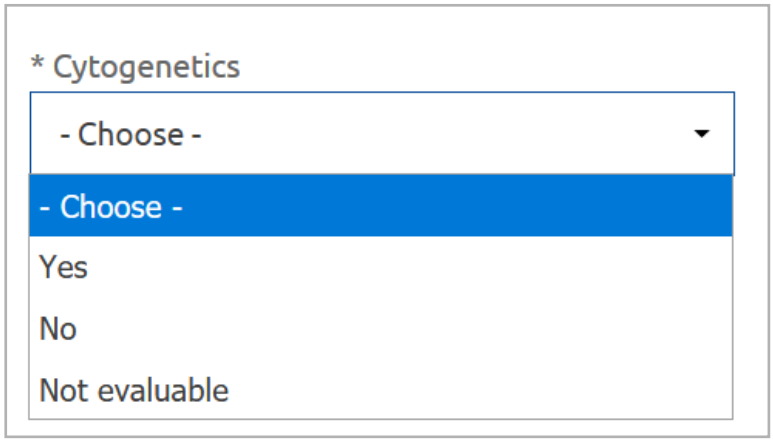


Figure 22 Database field type – Drop-down

Figure 23 Database field type – Checkbox

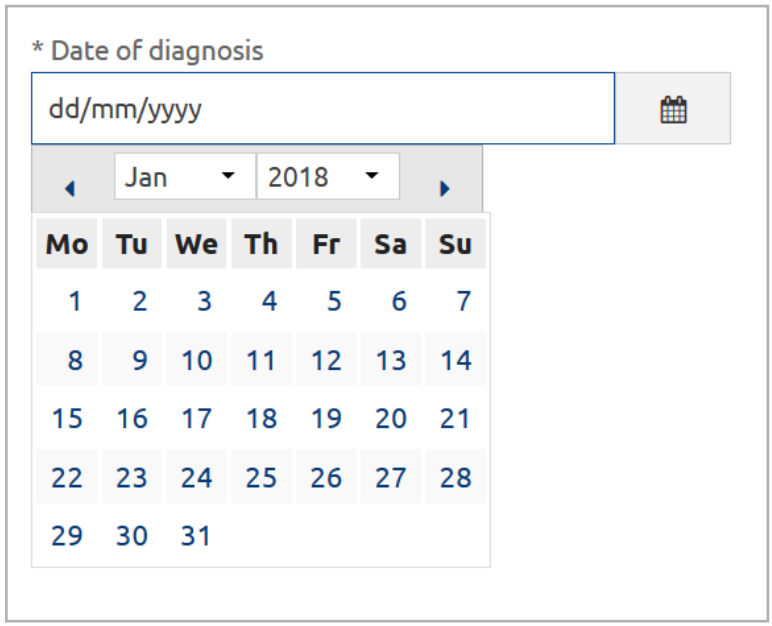


Figure 24 Database field type – Date

There are also special data types that can be configured to suit a specific registry. Finding a combination of drugs from a predefined list is one example; after writing several letters contained in the drug name, a rollout selection is displayed in which the combinations are contained in the written text (Figure 25).

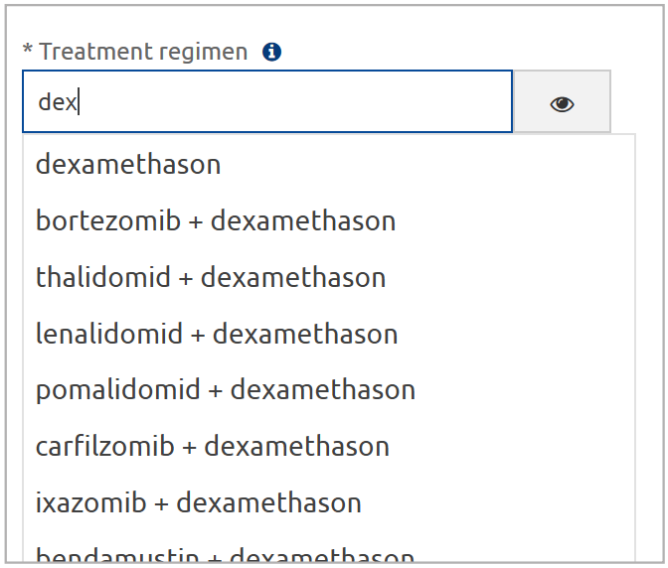


Figure 25 Rollout selection

In addition to data types, it is also important to consider the attributes of data fields, which might be:

- Mandatory – these must be always filled in
- Hidden – these are not visible on the form, but are saved in the database and can be used for calculations
- Read-only – it is not possible to enter data into these fields (Figure 26)
- Automatically calculated (e.g. automatic filling of the score based on the calculation from previous questions or refilling the values from other forms)

The example given in the Figure 26 shows a read-only question for Polyclonal PC (%), which is calculated from the previous two questions – Plasmocyte count FC (%) and Clonal PC (%).

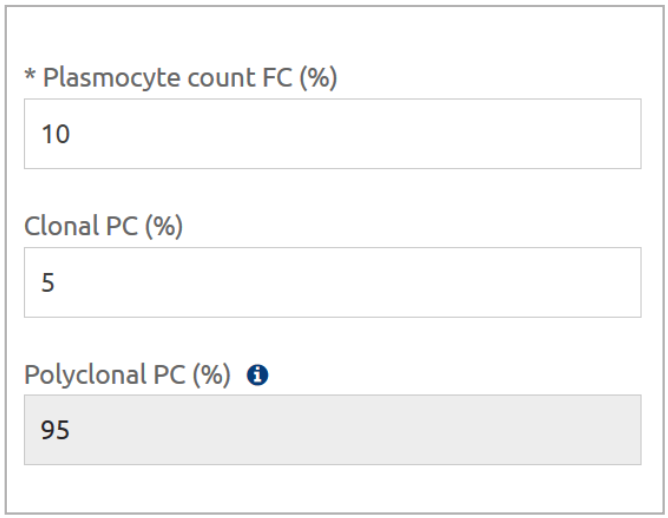


Figure 26 Automatically calculated read-only question

Question groups

Questions related to a specific examination (a monitored parameter or a logical unit) are usually placed in the same question group. It contributes to an easier orientation in the form and to the creation of relationships between questions. If a question group concerning an examination that may be repeated, it is possible to set the repeated input of answers. Repeating groups can also be used to write concurrent data or data for which the maximum number of records is not specified, e.g. concomitant medication or treatment regimes. It is also possible to define a concrete number of repeating steps. Similarly, it is possible to define the number of groups that will appear on the default screen. This might be useful, for example, in cases where we know we will need at least three repeating groups; therefore, we can define that three repeating groups will appear on the initial screen. An example of a repeated group is shown in Figure 27.

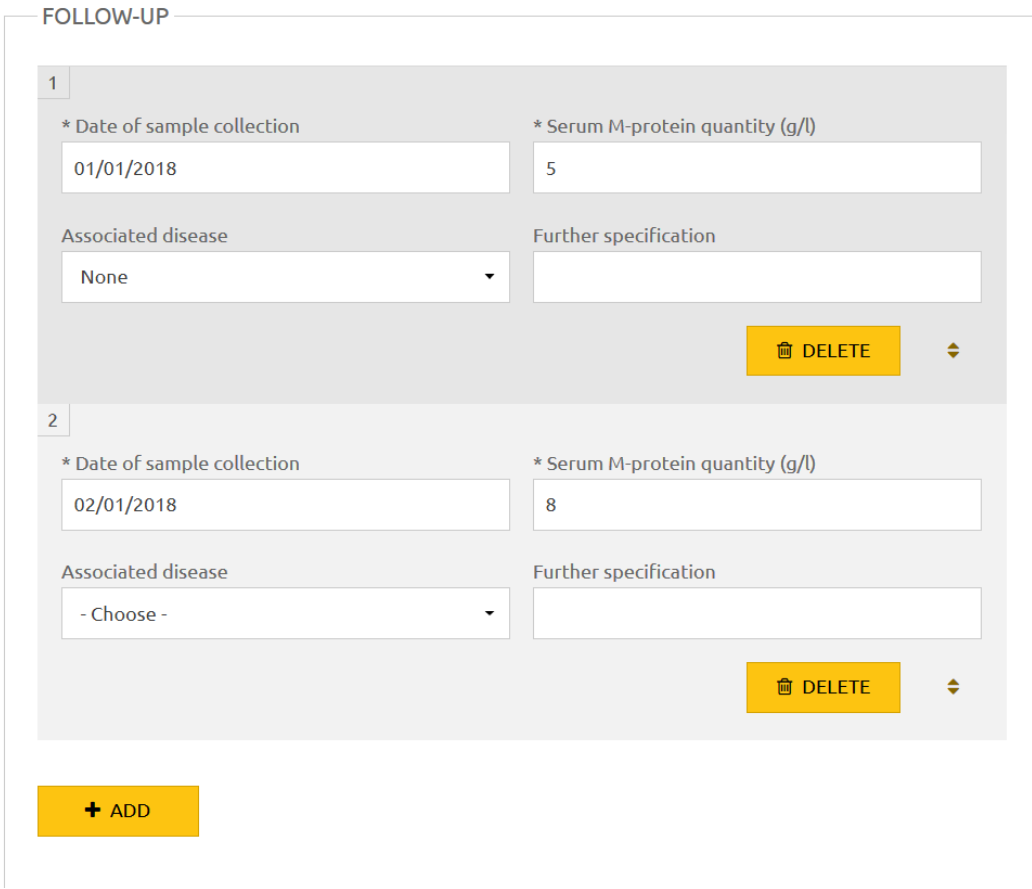


Figure 27 Repeat group

If a repeated medical examination with multiple data should be recorded, it is advisable to have a separate form for recording data from this examination. For this purpose, it is possible to allow a repeated creation of the form (similarly as in repeating forms, it is possible to define the limit for repeating – unlimited or to define the limit). In contrast to the repeat group, it is possible to ask for more questions on the repeat form without losing clarity. In addition, the repeat form may include some repeat groups of the question, such as medication. The adverse events (AE) reporting should also use the repeat form because the number of repetitions is not known in advance. The forms of predefined visits or the termination form can be set to only one completion.

Linked parameters

If some of the questions are identical for several forms and it is not expected that their values will vary greatly between the individual forms, it is possible to set the transfer values. This means that some values from the previous form are transferred when the follow-up form is created. This can be used especially for parameters that are not expected to change markedly. A typical example is the medication during individual patient visits; if the answers are automatically transferred, the physician does not have to list all the medication and their dosages, but only checks and adjusts the values transferred from the previous form.

When entering information into the follow-up form, some information from the previous forms may also be helpful for the person entering the data. These may be answers to some questions about the patient or the number of days since the last visit. In both cases, it is a question whose answer should not be changed, and therefore it can be marked as read-only. In addition to transferring values from other forms, the read-only questions can also be a calculation based on previous answers.

In a typical example, the investigator filled in some data in the past (months, years ago) and would like to see these historical data on the current form for informational purposes. We just need to be careful with the setup of these values to ensure it will not affect a proper entering of the current data, i.e. that the investigator will not be influenced by those older data.

Skip logic

An important step in creating the eCRF is to set each form in a way that will enable a simple and convenient data entry. If the form is long and some questions in it are dependent on previous answers, it is unnecessary to display everything after opening the form. For this purpose, we use the skip logic, which allows us to hide questions that do not need to be filled in for all patients. It is a relatively simple step that will help the form become clearer. The use of skip logic can be seen in the Figure 28 and Figure 29: if there is a positive answer, other questions are uncovered; if there is a negative or unfilled answer, the questions remain hidden.

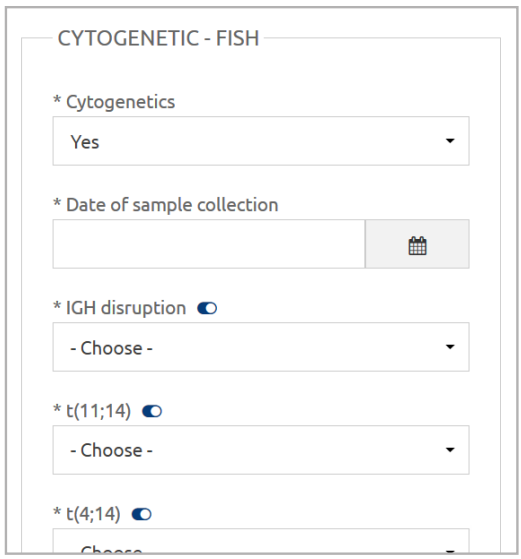


Figure 28 Skip logic – positive answer

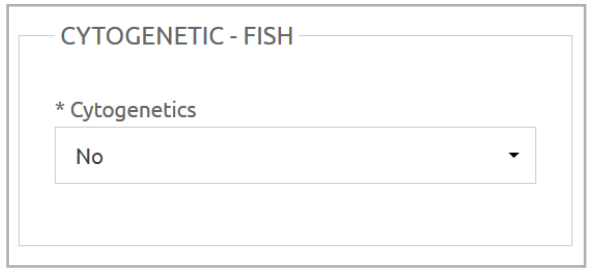


Figure 29 Skip logic – negative answer

Language mutations

The primary language was already mentioned in previous chapters. It is important to define the primary language for other setup processes. Nevertheless, besides the primary language, it might be useful to create the project database in more languages, especially for multi-country registries. Once you define the whole database, including all specific functionalities, you could use the translation module to translate all items into all required languages. Of course, you need to specify all terms in different languages but, at the end, the database will exist in more languages; that is, it will be the same database with the same functionalities, but with several language mutations. Moreover, if you plan to capture data in local languages, you should also plan the process of translation; ideally, you should translate all captured data into one primary language for clear data analyses. The translation process is applicable after the collection and validation of data.

Graphic design

It is nice if a specific project has its own design, project-specific colours or own logo. In the case of RMG, a colourful design is used and a specific logo is placed in the header. You can read more details about graphic design in general in the chapter Project management plan. Besides the global graphic design, a separate part of database development is linked to the functional design of particular forms. This involves:

- description texts displayed above or below particular variables and informing the user in more detail about this item,
- hidden information texts appearing only if you click on the relevant icon and displaying additional text,
- indentation of particular variables for better clarity of the entire form,
- a set of questions can be grouped into a table for better clarity as well (e.g. laboratory value, unit and ranges on one line).

Form statuses

Multiple statuses can be set up for an easier handling of forms. These statuses can be seen by the person entering the data in the form overview; in this way, this person will immediately see in which forms some data are still missing or which form requires an update. When opening a new form, its status is "Pending", which means that the values on the form are not final and can be changed. If all mandatory values are entered and the submitter wants to mark the form as finished, he/she can change its status to "Completed". At this point, validation starts and checks the entered data if they are within

the allowed range and meet all the defined conditions. This step is performed in order to highlight possible misspellings and omitted mandatory questions. If all values are in the allowed range and meet the conditions, the system automatically switches the status of the form to "Valid". This status means that the submitter has marked the data as complete and the system did not find any errors. If the system does find some error, it provides an error message and marks the form again as "Pending".

The use of form statuses can be used not only to alert the user to an error when saving the form, but also to allow the user to complete data later. If the value of a question is not known at the time of entering the data (for example, biopsy results are awaiting), the form is stored in the "Pending" status because it is not complete. Once the biopsy results are known, it is possible to search for patients who have pending forms and add the required information to their forms.

Besides the above-described statuses, there other statuses according to specific project needs can be defined. Examples of these statuses involve:

- "Monitored" – allowing to demonstrate the fact that the form has been checked by the monitor, if applicable.
- "Signed" – this could logically follow the status "Monitored", allowing the investigator to confirm that the form contains all data, is checked and no more updates will be performed on it.
- "Locked" – confirming that form is locked and no more actions can be performed on it.
- "Paid" – if payment procedures are applied in projects and connected to form statuses, it can be demonstrated that successfully filled in forms were paid to investigators.

In addition to searching for forms with the "Pending" status, it is possible to set the search according to data entered in the database or according to the form's name. It is good to think about setting up a patient search when creating a database because selecting the appropriate parameters to search for may be important in the further use of the registry. If we want to set up searching based on some data entered for the patient, the question should be marked as mandatory to include all patients in the search. In addition to the patient's personal data, it can be useful to search for information about the name of the person that entered the patient into the database or about the date of their entry. The proper setting of search criteria can be crucial for adding information to the forms or filtering patients meeting certain requirements.

The ability to create a form in different ways is a specific aspect of working with forms. The default option is usually to create forms in the same structure for all patients. Nevertheless, it is possible to define a specific algorithm of conditions regarding the creation of particular forms. One of the simplest examples could be creating a form containing questions related to ovaries that will be created only for female patients. Therefore, the condition could be the patient's sex, but also date or a concrete expected answer to a specific question. Nowadays it is possible to prepare and programme many specific functionalities, with the important aspect being to specify them during the preparation phase of the project.

Dashboard

A dashboard is a place where the logged-in user can see basic information about the project. Different information can be displayed there, and it is generally useful to present the following:

- Basic information about the project
- Current information about the project's progress
- Current versions of the technical and clinical manual
- Other project documents like informed consent form or questionnaires

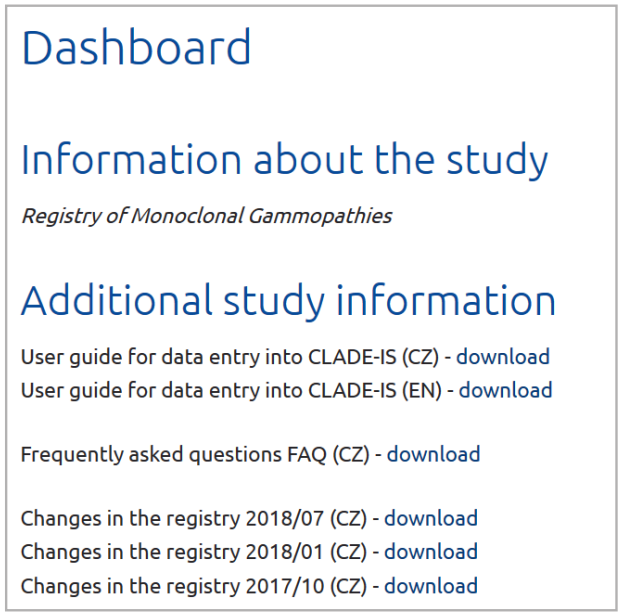


Figure 30 Dashboard in the RMG registry

A welcome system functionality could also be the ability to record a file into the system. Such a file can consist of an X-ray picture or some scanned documents that can be kept in the system as additional information.

Validations

Because of possible human and machine error, data entered into eCRFs are rarely completely accurate. The aim of a correctly set eCRF, in addition to clarity, is the effort to eliminate these errors as much as possible. Therefore, it is advisable during the eCRF's creation to think about potential problems and to minimise them. One way to minimise problems is to set minimum and maximum values for numeric questions and to set various validations that are described in more detail in the next section.

3.3.2 Programming of validations

One of the key aspects of data management is to ensure the validity and consistency of data in the registry. It is needed for subsequent reports, analysis and the statistic evaluation of the entered data.

Every system for collecting data (medical data in this case) should have validations implemented, as should any registry or clinical study. As every registry or clinical study usually has some analysis planned or even just a simple data output, there is a need for valid data. A great tool for establishing and negotiating the specifications of data validation

rules and data checks performed is a DVP – a document describing the process of data cleaning, specifications and details about the programmed checks by the data manager. When the DVP is being created, all interested parties (project manager, data manager, analyst etc.) should take part in this process and all parties should understand and agree on the validations described therein. All programmed checks have to be tested and validated. It is recommended that this process is performed along with the creation of the database/registry.

Particular checks can be divided into the categories described below.

Control of format

The format of the data is defined at the stage of database creation. Details for the setup of every defined variable are described in the Data Dictionary document. The most important part of this document could be printed directly from the EDC system. The Data Dictionary is a structured list of questions containing the following information:

- If a question is mandatory
- If a question refers to the validation rule
- If a question is linked to skip logic
- If a question is read-only
- If a question contains calculated variable
- Question datatype
- Setup question boundaries

Values entered into the database are compared with the defined format of variables in real time for the data entry process upon saving the page. Variables might be checked according to the following parameters:

- Data type: number, text, date etc.
- Mandatory items: must be filled in (otherwise an explanation is eventually required)

Coherence checks on the form level

These controls are planned and programmed to check for the coherence of variables within one form. The coherence check on the form level is processed at the same time as the check of the format, in real time during the data entry process, and eventually during saving the form. If a mistake is found, an error message appears on the screen. This mistake has to be corrected before the form is saved as complete and further steps of data processing can be continued. The following controls can be checked at this level:

- Length: maximal length of the entered text
- Boundaries: minimum and maximum permitted value
- Future dates: it might not be possible to enter a future date
- Special checks: compare two values within one form etc.

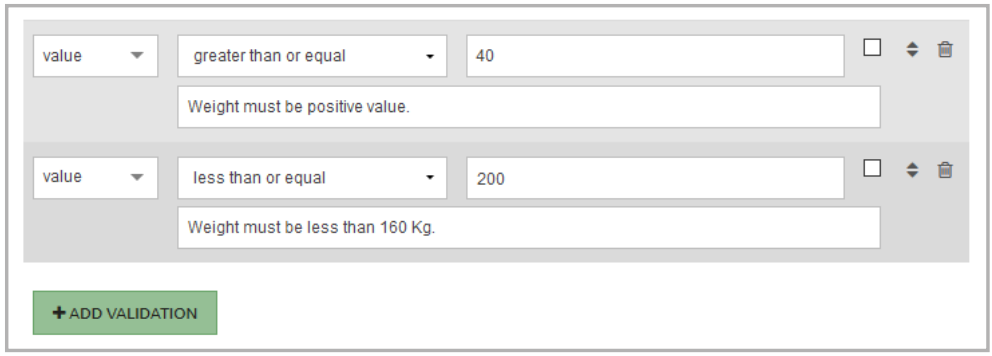


Figure 31 The setup of new validation on the form level is straightforward in the Clade-IS system

Coherence checks across the database

The controls of coherence are planned and programmed to check the coherence of variables across the whole database. These checks are programmed by the data manager according to specifications and details described in the DVP. The data manager tests all programmed controls by running them on dummy data. The combinations of possible options of data values are tested. These checks are run after the processing of format checks and coherence checks at the form level. These controls mean, for example:

- Date of informed consent signature cannot be later than the date of the first examination.
- Date of diagnoses cannot be later than the date of the baseline visit.
- The weight of patients should not differ more than 10% between two consecutive visits.
- Adverse event presence is ticked on the visit form but the adverse event form is missing.
- Date of patient's death cannot be earlier than the date of the last visit.

Manual checks

Manual checks can be performed on the data entered into the database. Any team member who has access to the database and relevant rights (e.g. data manager, monitor, safety officer, medical reviewer) can check the data and in case of any discrepancies, the query can be manually created in the system.

- Consistency between concomitant therapy and concomitant medication should be checked.
- Consistency between medical history and previous medication should be checked.
- Consistency between adverse events and concomitant medication should be checked.

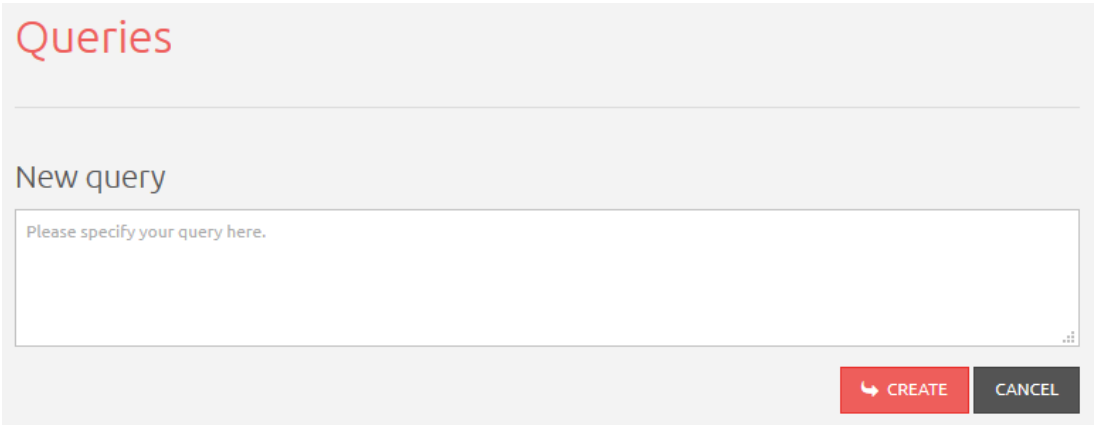


Figure 32 Possibility to create a new manual query in Clade-IS

Based on the used EDC system, when there are validations implemented we can describe the overall status of a study form. As described in the subchapter Form statuses, there are two basic form statuses in Clade-IS. The study form is either "Pending", meaning not yet completed with possible invalid data, or "Valid", meaning all data were added by the user and all were validated with a positive result. Depending on this form, the project analyst can then decide which forms to include into his/her statistical analysis or a particular report could be set to include only valid data.

Special checks

Special checks can be made for additional registry requirements. An example might be sending a monthly list of patients who are missing a form.

3.3.3 Setup of the reporting process

Well-designed registries can be an important source of information about healthcare patterns, decision-making and insight into the quality of patient care.

In order to present this type of information and as a tool to enhance the registry throughout its duration, specific reports can be set.

Depending on how complex the reports are and what type of information they contain, we can divide them into categories described below.

Enrolment report

We can set up an enrolment report as an example of a simple report that contains information about how data in general (patients and study forms) in the registry grow over time or how active each investigator is (i.e. how many patients and forms he or she enrolled in a specified time frame).

Report of entered data

Another example of a simple report would be an overview report of entered data that can be a good indicator of data quality. Investigators, to whom this type of report would be directed, can reflect and adjust their data entry process so their data quality improves. This report is very useful on the country level, on the site level and on the subject level.

Queries

A report containing the number of generated queries, number of resolved queries, and especially the number of pending queries, shows very well the quality of data entered into the database. In addition, an overview of the number of queries per sites (countries) might be very interesting.

Custom reports

The last category would be a custom report showing any type of information that can be obtained from the collected data. Typically, we can focus on adverse events or details about a given treatment. In more complex studies that are aimed at specific information, a report reflecting the study or registry's primary or secondary objectives can be set as a custom report, too.

Report processing

Reports can be processed in different ways. It is very popular to view some reports online and being able to view the situation in real time. Alternatively, reports can be automatically sent to predefined email addresses once per agreed period (weekly/monthly/quarterly). Another efficient way of report handling is a semi-automatic process when an automatically generated report is commented on by a data manager and then provided to other team members (sponsor, investigators etc.). In this wasy, it is safer to guarantee the provision of real and checked data and prevent working with non-validated reports. Reports can be provided in read-only format (pdf) as an official document; in these cases, it is recommended to review the data and validate the report before sending it. Alternatively, it can be sent in an updateable form (xlsx) for the purposes of further modification, combining with other data etc.

Each report type can be a part of the regular reporting process or as a per-request item.

It is recommended that more complex reports should have a longer interval than simpler ones in terms of the regular reporting process, simply because more complex reports require more time to be processed and analysed by their recipients and it can take more time to recommend further steps based on the report's content.

As mentioned above, reports can be used to improve the quality of such a registry, so not only the CRO (Contract Research Organisation), sponsor or client are the ones that can make good use of those, but the registry provider as well. For example, it can be used to modify specific parameters as well as to remove or add new ones to the eCRF. Patient criteria may be adjusted as well if they are too loose or too strict, depending on a specific case.

The whole reporting process should be a part of the DMP that both sides agreed on and signed. Frequent communication about the content of the report before its first generation is a good practice. All interested parties should fully understand what those allocated data represent and which parameters they are computed from.

All reports that are being programmed have to be properly tested as well. It is strongly recommended that a large time frame is allocated for this testing process because more complex reports require more time to get validated. Discovering an error or discrepancy once the report is being generated on a regular basis could be quite difficult or may just be a matter of coincidence.

Considering the practical preparation of various reports, it is very efficient if the required reports are defined before the project goes live, because the required report structure may affect the database's design and specific functionalities.

3.3.4 Setup of the pharmacovigilance process

As previously described in the chapter Pharmacovigilance, we must consider if safety data are collected in the project or not; in general, we can say that adverse event data are related to all clinical projects. An adverse event can arise from any use of the drug (e.g. off-label use, use in combination with another drug) and from any route of administration, formulation or dose, including an overdose.

If adverse event data are collected, we need to know several details about the AE data capture process in order to develop the database properly:

- If adverse event data should only be collected with no further processing and no reporting, then we can consider them as other data. The only thing we need to consider is that the safety data should be considered as critical data (for example when planning database coherence checks).

- If adverse events are classified within the project, we should consider creating a separate instance for serious adverse events; it is highly probable that these SAE data will be somehow reported and consequently processed, and it is much simpler and more efficient to have these data in a separate form.

- If the request for safety data reporting already exists, we should focus on the means, format and sending frequency of these data. The optimal format is having the database form corresponding to the form in the initial CRF and to the sponsor's standards and/or the professional society's requirements. As described below, the most efficient way is to directly generate an adverse event report from the database once the adverse event is entered in the database. If this report is sent via email, we just need to take care how to set up this functionality to prevent the system from sending too many emails without completed information. Then, adverse event reporting should be connected to such a form status ensuring the data will not be modified.

An adverse event is classified as "serious" if the patient died, involved or prolonged inpatient hospitalisation or involved persistent or significant disability or incapacity, was life-threatening or involved a congenital anomaly. All other AE are classified as "non-serious". As we are talking about safety data, it is obvious that special attention needs to be paid to these data. We must be sure to respect all project requirements in the management of adverse events.

The ideal solution is if the EDC system allows the automatic generation (and sending) of reports on any adverse event (AE) that have occurred. Such a system allows setting up the generation of an automatic email to the pharmacovigilance department of the CRO once the AE form is saved with a valid status. Therefore, the CRO is immediately informed that an AE occurred. Therefore, the AE questionnaire must be on a separate form. The drug safety department does not necessarily have to be comprised of a part of the CRO and may be outsourced to a different company. The automatic sending of an AE notice may be conditioned according to various settings. As an example, it can be mentioned that only serious AEs will be forwarded to the drug safety department. Another interesting example is that AE notices are distributed according to the country of the treatment centre to an appropriate national authority.

Reports for all AE can be set up with a certain frequency (i.e. monthly, quarterly) and provided to the CRO.

The reporting of AE depends on the local legacy of each country. Reporting by the database to the CRO does not stand for the investigator's obligation to report the AE to the local committee regarding the laws in his/her country.

3.3.5 Database testing

Testing of the database before its launch intends to provide objective feedback to the author of the database. The main purpose is to prevent any future errors in the database. Any modification, while the database is already running, is always more complicated than having the modification done before the start of data collection. Future modifications usually demand that the study stops and sometimes also involves the import of data. The functionality of the database is tested by entering the complete data of at least one patient into all language mutations.

Testing should be done by a data manager other than the one who created the database in order to bring a different point of view into the task. The database can be reviewed before launch also by another member of the project team if needed. If any errors are found, the author corrects the errors and provides the improvements proposed by the tester. The database is retested until the last feedback is implemented.

The testing process mainly consists of reviewing the database structure and checking the validations and verification of various other setups, including reports. All setups have to correspond to the CRF, data management and validation plans. An example of testing parts is described in Table 18. However, not all of them have to necessarily be part of every clinical project, depending on the requirements specified by each client and on the needs of a particular project.

Table 18 Example of points to be tested

Test according to a relevant document	Test focused on
Protocol	Table with list of visits and parameters
	Eligibility criteria
CRF	Database structure (arms, visits, forms, question groups, questions)
	Simple validations (mandatory items)
	Format of all variables (data type, code list limits, units)
	Functionality of all skip logics
DMP	Database URL address and technical setup
	Specific exports
	Reporting process
	Form statuses
	User roles and user groups created in the database and their permissions
DVP	Format of all variables (data type, code list limits, units)
	Transfer of values
	Coherence checks within one form
	Coherence checks across the database
	Correctness of calculated variables
Special functions	Randomisation, limit of enrolled patients, forms created according to specific conditions
General	Grammatical correctness
	Graphic design

Once the testing process is finished and all suggestions to improve the database are included in the settings, the tester writes the database validation protocol as a last step of his/her goal. Testing has to be done before the initial launch of the database and after each database modification.

3.3.6 Database validation

It was already mentioned that getting the database into production mode is one of the key milestones. It is a starting shot, meaning that the race (data collection for the clinical project) has begun and this step cannot be returned back. Then it is quite clear that not only internal testing by the data management team, but wider global revision is needed (or at least strongly welcome) in order to minimise potential future troubles that always make the project less efficient (meaning extra work, the worsening of investigator's work etc.). After testing the database (by the sponsor, investigators, project manager, analyst and tester from the data management team), it is necessary to approve the final version of the database and sign all important documents such as the DMP, DVP and eCRF approval form confirming the correct eCRF structure. These documents have to be signed by the entire project team – the project data manager, project data analyst, project manager and, of course, the sponsor (and CRO if relevant) – and they describe the setting

of the database, validations and plans or goals of the project. When all documents are signed, the database can be launched.

A good practice is to remove all test patients and their forms from the database before setting the database into production. This step cannot be performed later as all changes must be kept in the database because of the audit trail.

Subsequently, user access is sent – the login, password and web address of the registry. User access is created by the data management team or helpdesk upon the request of the sponsor or project manager. Access is also sent to all members of the project team. Users also will receive an email about the launching of the registry, basic project information, and manuals and information about the potential training of investigators. After these emails, investigators can enter data for the first patients.

3.3.7 Management of database users

Considering the database users, we first need to think about:

- Basic roles applicable for this project. Do we need only an investigator and data manager? Alternatively, do we need access permissions for the sponsor, medical expert, pharmacovigilance expert, medical coder etc.?
- Investigator. Will all the physicians have the same rights? Do we have a principal investigator with more rights? Do we have data entry coordinators on sites having less rights?
- Do we have different countries? Do we need regional coordinators having, for example, read-only rights to all the sites of this country?
- Considering the project team, what rights we need for the project manager, data analyst, sponsor or medical expert?
- Is the payment of investigators linked to database access permissions?
- Do we need form statuses because of some specific functionality (locking forms because of randomisation), specific form status for CRA, or to allow the investigator to sign pages that have been filled in by the data entry specialist?

These are some of the questions that should be answered before we create and set up the system of database users, roles and permissions. Typical users that are present in the clinical project may consist of:Typical users that are present in the clinical project may consist of:

- investigator – physician on site, primarily responsible for project on site,
- site coordinator – support on site, could be a nurse,
- sponsor – global permission, main responsible person for the project,
- sponsor – regional coordinator, applicable in multi-county projects, has limited access, can see data only for the sites in their region,
- medical expert (professional society) – representative of the professional society supervising the whole project,
- data manager – managing the database and all access,
- data analyst – biostatistician responsible for the project,
- medical coder – if medical coding is applicable,
- pharmacovigilance specialist – responsible for data safety,
- CRA – only if monitoring is applicable.

After all documents are signed, the sponsor should provide a list of users, their sites and roles. Only then is the data management able to create access for the users.

It is very important for each user to have his/her individual login, because only then can the database track all changes done by each user (creating patients, updating forms).

Each user will acquire his/her unique login and password, which is sent directly to his/her email address. The password is automatically generated and can be set to work only for the first login, after which the user is then prompted to create his/her unique password.

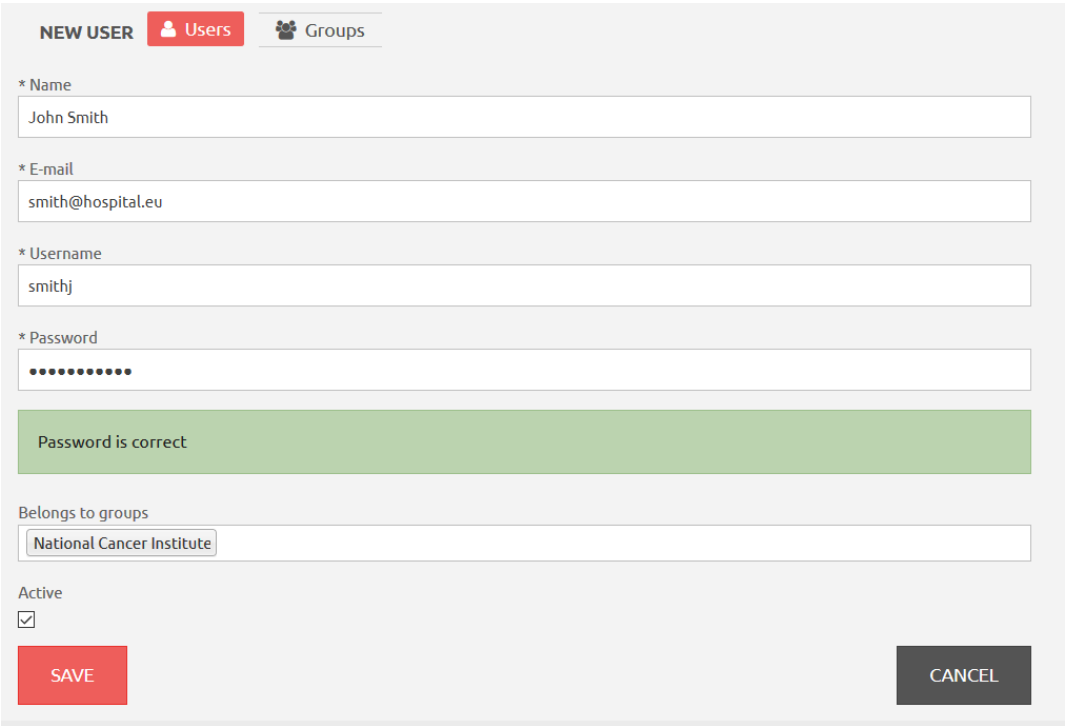


Figure 33 New user

Following data are usually defined for each user:

- name – name and surname of user,
- e-mail – user's email,
- username – user's login, most common is the user's surname and the first letter of their name,
- password – automatically generated (uppercase, lowercase letters and numbers),
- belongs to groups – user's defined centre or/and roles,
- active – user is immediately activated after creation.

After the user is saved, the system will send an automatically generated email with the username, password and the web address of the registry.

3.3.8 Groups – sites and roles

Groups are not only individual sites of users, but there is a wide range of settings for creating distinct roles for each user such as the project coordinator, national coordinator, helpdesk etc. The "investigator" role, which is added to almost every site, is the most commonly granted type of access to the database; in Figure 34, this role is added to group/centre "RMG IBA". This role means that every user in this centre can create, update and delete all records provided by this centre.

The "project coordinator" is another role, allowing the user to see all data in the database but only in a read-only view (i.e. nothing can be modified). This type of access is typically granted to the sponsor as they oversee the data. The role of "national coordinator" is similar, but data that are available to this user are limited to only his/her country.

Besides the permissions for adding new patients and viewing/editing their data, there permissions can be set up for query management (creating, answering, closing), for handling reports/exports and for designing the database and the management of user accounts. The more flexible the EDC system, the more complicated it is to set particular roles and their permissions, so a global permission schedule has to be carefully prepared and consequently set up.

If a user has more than one site/role, the database will let him/her choose under which role they want to work after logging in (only one role/site can be chosen at one time and the user can switch between them in the database).

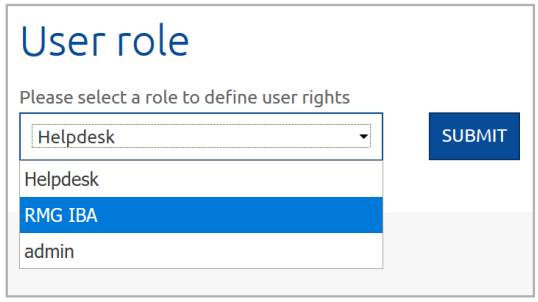


Figure 34 User role

3.3.9 Patient access

It is also possible to create access to the database for the patients themselves. This access is convenient when the patient is supposed to fill in certain forms such as the "Quality of life" or if they want to see the recorded data about themselves. There are two different settings for patient access:

- Without creating the patient's access details – the database uses only the patient's e-mail to send him/her an automatically generated e-mail with a web link to one specific form, which the patient can fill and save.
- Including the creation of the patient's access details – each patient obtains a personal username and a password that he/she can use to log into the database. This access is always limited to a particular patient, who can subsequently only view data on himself/herself. Other restrictions might apply, such as that all forms can be viewed in a read-only mode except those the patient is supposed to fill in, or that other forms – except the selected forms – can be even hidden for this user.

3.4 Technical manual

The technical manual is one of the key documents and basic tools for registry users (investigators).

The manual is usually created by the data management team because of the standardisation of rules across all included centres and users. There can be hundreds of users, which is an amount that is impossible to communicate and train individually. When activating each user's access, this manual is also included in the e-mail message and it serves as basic training for all users. For more details about different possibilities of training, see the chapter Training.

The manual contains all mandatory information about the database with illustrative examples and pictures. Also, users can be enrolled in the study with various levels of IT knowledge; for this reason, therefore, the manual is intended to be as illustrative as possible.

The user guide is divided into chapters that describe the process from the login page through form completions to the specifics of each registry.

3.4.1 Common chapters of the manual

The first chapter of the technical manual is usually dedicated to the login procedure. The first screenshots display the login page and then the basic application window.

Basic application window

Navigation to all the necessary functions is described here. This includes searching for submitted patients and their forms, creating new patients in the database, reading the structure of the registry, the option to search for queries and accessing contacts for the helpdesk.

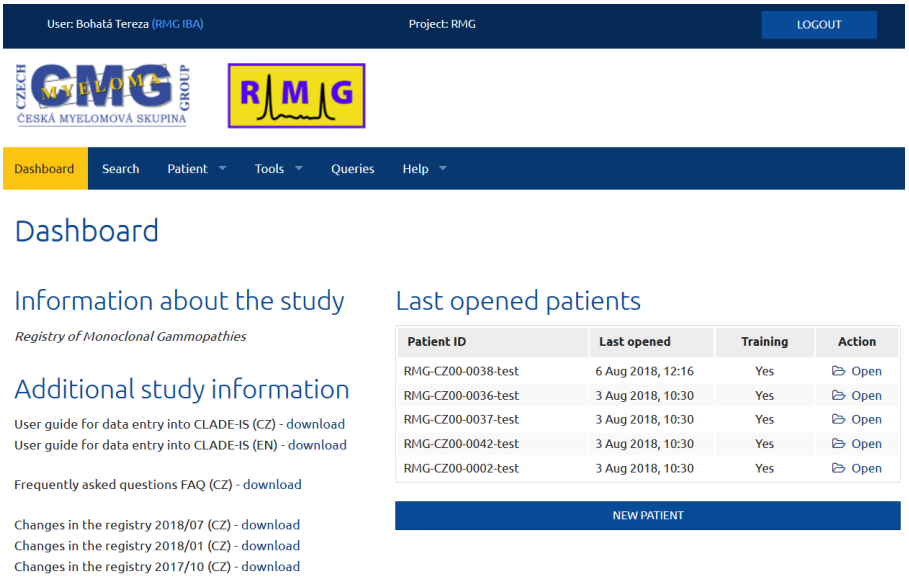


Figure 35 Dashboard

New patient registration

How to create a new patient in the project is obviously an important part of documentation, making it possible for the physician to find out how to enter a new patient into system and continue with data entering.

Patient search

This chapter describes how to search for patients by ID and by various other parameters that can be combined together for a more precise search.

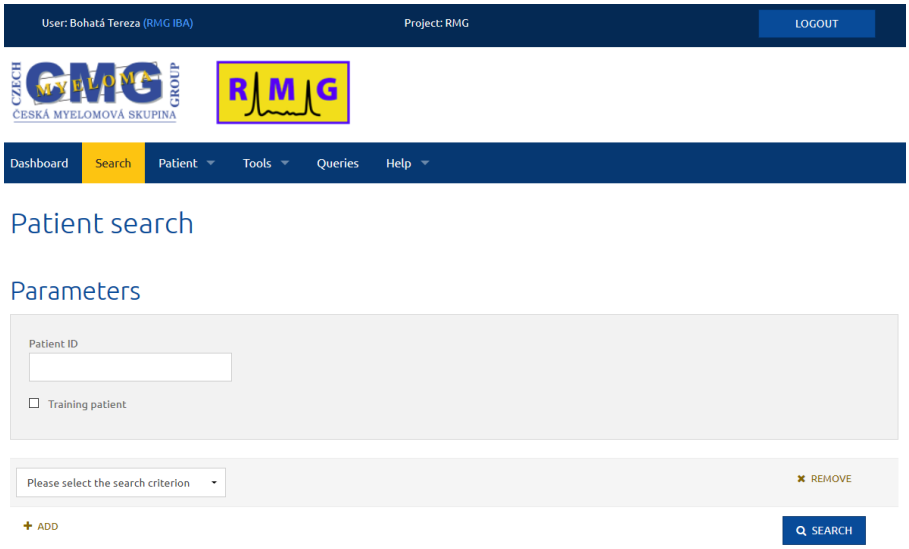


Figure 36 Patient search

The technical manual should have two parts: a general part describing the general functionalities of the system, and a specific part describing specific functionalities of the project, including:

- automatic reporting of adverse events,
- handling with payments,
- questionnaires filled in by patients.

An important part of the document is the contact list, which is needed in order to provide an easy navigation to direct support.

3.4.2 Working with the forms

For the most important chapters and subchapters describing how to create, update and enter data into the forms in the database, see the example below.

Form completion

The form contains questions and entry fields beneath the questions. There are several options for data input (see A-E, Figure 37 and Figure 38):

A. Direct writing

B. Selecting an option from a drop-down menu

C. Dates can be written (in the mandatory format – dd/mm/yyyy) or selected from the calendar

D. Checking a checkbox

E. Repeating a group

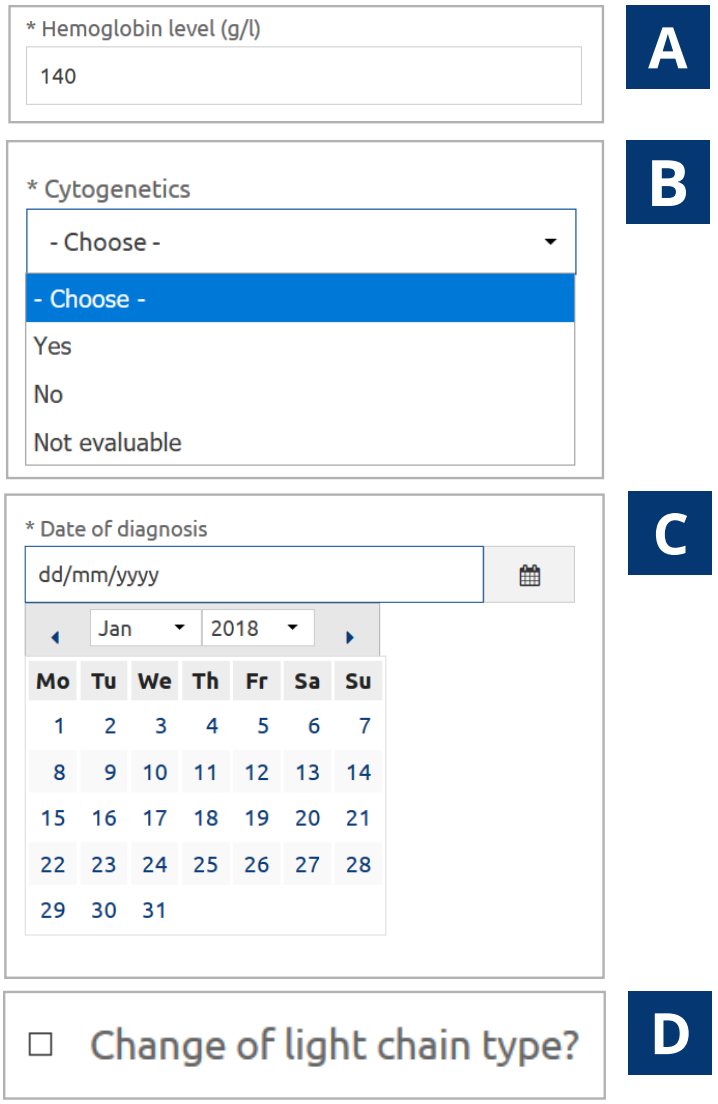


Figure 37 Field types

Some sections of forms consist of repeating groups of questions. You can add new items to this group by clicking the "Add" button (see 1 on the figure below). A row of blank questions appears to be filled in.

Clicking the "Delete" button (see 2 on the figure below) then deletes the selected group of questions.

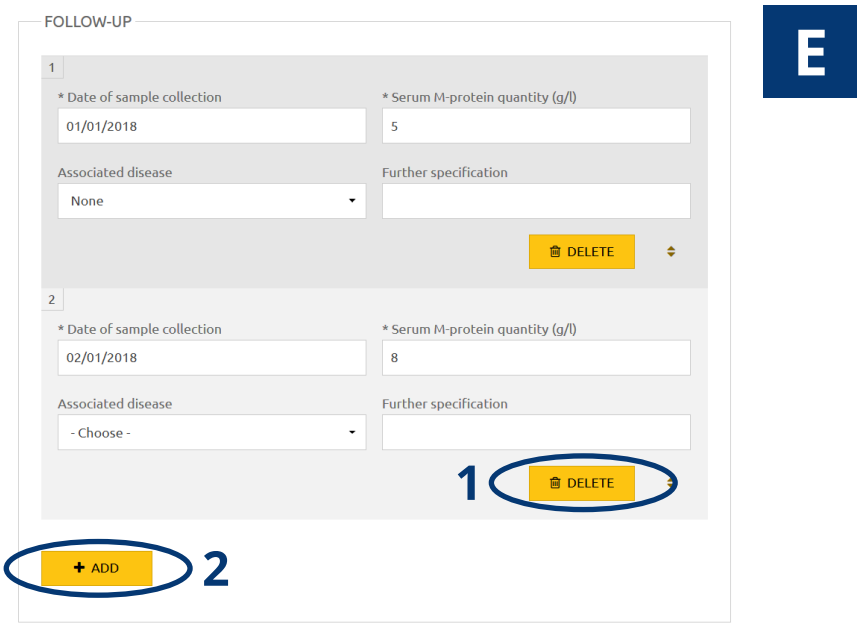


Figure 38 Field types – repeating groups of questions

Specific functions

Mandatory questions are marked by an asterisk (see (*) in Figure 39). Without completing these questions, it is impossible to save the form as complete.

If the entry field is grey, data cannot be entered; the corresponding value is calculated (see Figure 40).

In the case of a mandatory question that is impossible to fill in, there is a slider switch indicating that the information is not known (see B – Figure 41). The question is turned on if the information is known (see A – Figure 41). By clicking on the slider, we can turn the question off or on.

For some questions, there is additional information available via the ⓘ icon next to the question (see **(i)** in Figure 42).

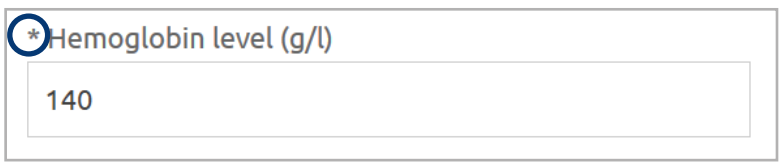


Figure 39 Specific function – mandatory question

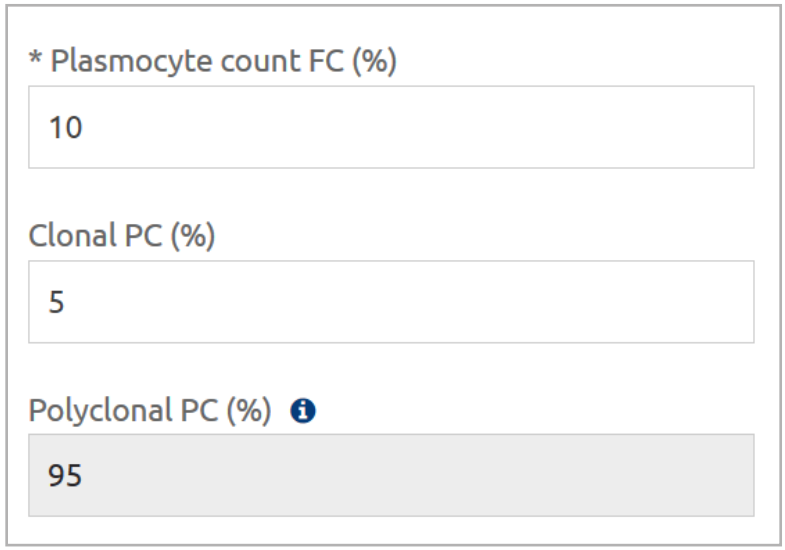


Figure 40 Specific function – calculated question

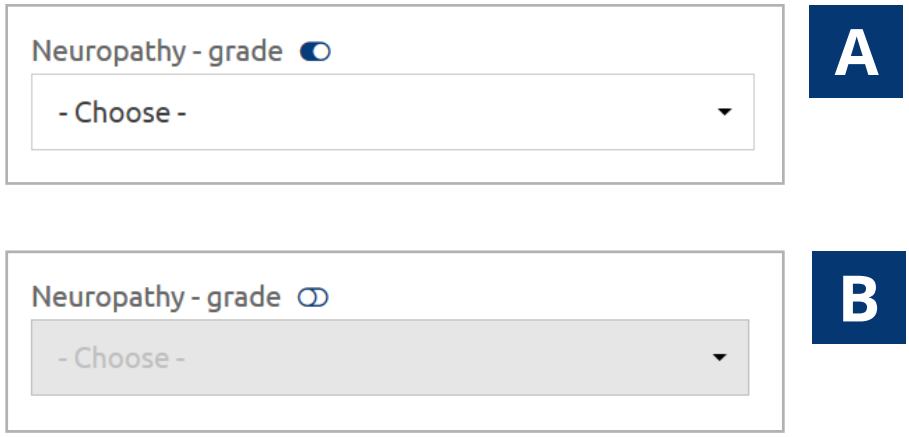


Figure 41 Specific function – the information is not known

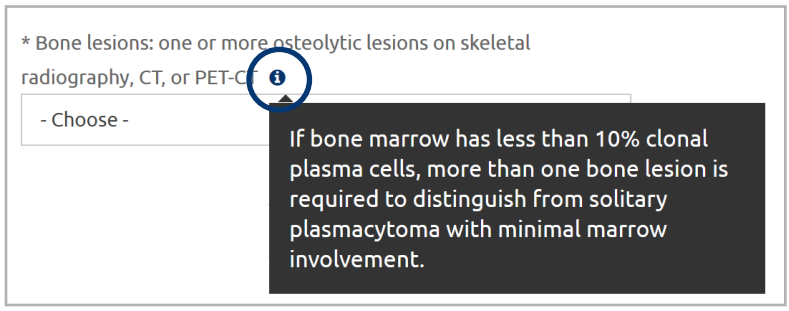


Figure 42 Specific function – additional information

Validation of records

This subchapter describes how the validation rules work and how users are notified if a validation criterion is not met.

A. Mandatory field – the value must be given
B. Special checks – for example, data must be matched to other data
C. Dates – the value must be given in a predefined format (dd/mm/yyyy) and the value must not refer to a future date
D. Numeric values – some numeric values have a minimum and maximum warning limit

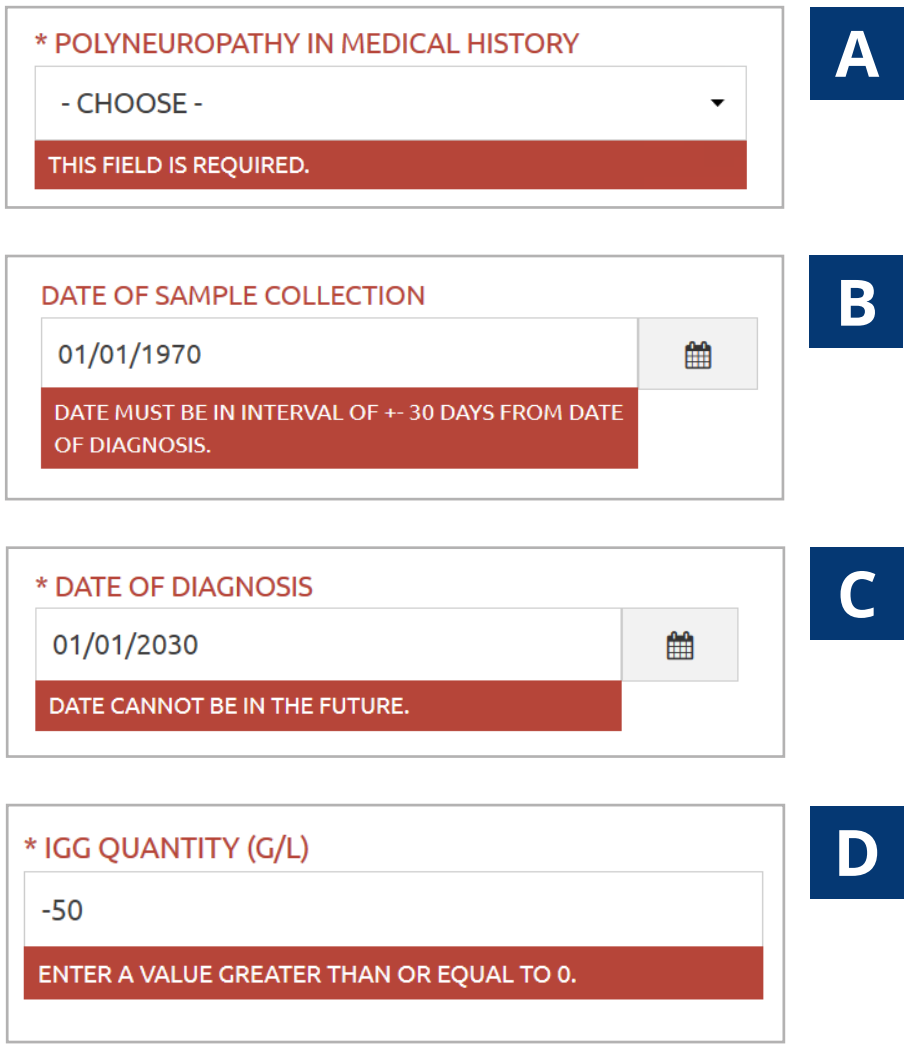


Figure 43 Validation of records

If this validation criterion is not met, the error message is displayed directly for the related question where the error occurred.

3.5 Clinical manual

A clinical manual is prepared for the registry. This manual contains information how to enter individual data and provides the information needed for response evaluation.

The registry database can be accessed from the website of the Czech Myeloma Group (CMG). Comprehensive data entered to the database are structured into several chapters. A range of predefined variables are collected. These include baseline demographic characteristics, laboratory parameters, prognostic markers, clinical stage, symptoms, treatment, toxicity, outcome and death. Dates and results of tests and events are registered to allow for follow-up over time at the level of an individual patient. Continuous collection of follow-up data in the registry is carried out according to a predefined range of variables at the time of significant events: diagnosis, treatment, progression, transplantation or follow-up. The registry covers all variables included in the International Myeloma Working Group (IMWG) criteria for diagnosis and treatment initiation. There is collective agreement on CRF (standardisation of data collection between centres), which is periodically updated.

3.5.1 Clinical manual of the RMG registry

One particularity of the RMG registry is that phases and forms are displayed depending on the selected diagnosis. Depending on the chosen answer of the question "Diagnosis" in the form "Patient personal data", corresponding phases and forms are displayed, which can then be completed for a specific patient. A detailed description of all options is provided in Table 19.

Table 19 Displaying phases and forms depending on the chosen diagnosis

Diagnosis	Phase	Diagnostic	Follow-up	Current status	Secondary amyloidosis	Treatment	Watch and wait
MGUS	MGUS	X	X	X			
	AMYL		X	X	X		
	WM	X				X	X
WM	MGUS	X	X	X			
	WM	X				X	X
MM	MGUS	X	X	X			
	MM	X		X		X	X
	AMYL				X		
AL AMYL	MGUS	X	X	X			
	AMYL	X		X		X	X

3.5.2 Clinical manual of the Registry of monoclonal gammopathies

In the following tables, data collected on patients with multiple myeloma (Table 20, Table 21, Table 23) and monoclonal gammopathy of undetermined significance (MGUS) (Table 24, Table 25, Table 26) are provided as the example of diagnosis-specific forms.

Table 20 Example of form Diagnostics for patients with multiple myeloma

Diagnosis MM – Form Diagnostics Group of questions	Question
Diagnostics	Previous history of MGUS
	SMM diagnosis
	History of solitary plasmacytoma
	Polyneuropathy in medical history
	Date of diagnosis
	WHO ECOG
Laboratory examination	M-protein type (IgG; IgA; IgD; IgE; IgM; Biclonal; Triclonal; Nonsecretory; LC only)
	Serum M-protein quantity (g/l)
	Urine M-protein quantity (mg/24h)
	Urine M-protein quantity (mg/l per 24h)
	FLC quantity (if measured)
	Light chain type
	IgG quantity (g/l)
	IgA quantity (g/l)
	IgM quantity (g/l)
	HLC quantity (if measured)
Bone marrow examination	Bone marrow aspiration cytology/Monoclonal plasmocyte count (%)
	Bone marrow histology/Monoclonal plasmocyte count (%)
Biochemistry	Haemoglobin level (g/l)
	Thrombocyte count (10E9/l)
	Calcium total level (mmol/l)
	Albumin level (g/l)
	Creatinine level (umol/l)
	Beta2 microglobulin (mg/l)
	LDH (ukat/l)
	CRP (mg/l)
Cytogenetics	Date of sample collection
	Conventional (Normal; Changed; Complex)
	FISH: Method used; IGH disruption; t(11;14); t(4;14); t(6;14); t(14;16); del(13)(q14)/monosomy 13; gain 1q21; del(17)(p13); Hyperdiploidy;
Flow cytometry	Plasmocyte count FC (%)
	Clonal PC (%)
	Polyclonal PC (%)

Diagnosis MM – Form Diagnostics Group of questions	Question
Medical Imaging Modalities	Osteolytic lesions - X-ray
	Osteolytic lesions - NMR
	Osteolytic lesions - CT
	Osteolytic lesions - LDCT
	Osteolytic lesions - PET
	Osteolytic lesions - PET/CT
	Osteolytic lesions - MIBI
Extramedullary mass	Extramedullary mass - relation
	Extramedullary mass histology
	Extramedullary mass - lesion count
Stage	Durie-Salmon stage
	Durie-Salmon substage
	ISS classification
	R-ISS classification

Table 21 Example of form Treatment for patients with multiple myeloma

Diagnosis MM – Form Treatment Group of questions	Question
Line of treatment	Line of treatment
SMM (only for 1st line)	Progression SMM to MM (Yes, No)
	Date of progression
Reason for the start of the therapy	Hypercalcaemia: serum calcium >0.25 mmol/L (>1 mg/dL) higher than the upper limit of normal or >2.75 mmol/L (>11 mg/dL)
	Renal insufficiency: creatinine clearance <40 mL per min† or serum creatinine >177 µmol/L (>2 mg/dL)
	Anaemia: haemoglobin value of >20 g/L below the lower limit of normal, or a haemoglobin value <100 g/L
	Bone lesions: one or more osteolytic lesions on skeletal radiography, CT, or PET-CT
	For 1st line: Biomarkers of malignancy: Clonal bone marrow plasma cell percentage* ≥60%; Involved: uninvolved serum free light chain ratio ≥100; >1 focal lesions on MRI studies
	Other: Rapid progression of paraprotein; FLC escape; Hyperviscosity; Extramedullary lesion; Amyloidosis; Recurrent infection; Other

Diagnosis MM – Form Treatment Group of questions	Question
Relapse/progression (for 2nd and higher lines)	Reason for treatment (Relapse/progression; Insufficient response)
	Date of relapse/progression
Performance status before treatment	WHO ECOG
Laboratory examination before treatment	M-protein type (IgG; IgA; IgD; IgE; IgM; Biclonal; Triclonal; Nonsecretory; LC only)
	FLC quantity (if measured)
	Light chain type (Kappa; Lambda; Biclonal)
	Serum M-protein entry level (g/l)
	Urine M-protein entry level (mg/24h)
	Urine M-protein entry level (mg/l per 24h)
Bone marrow examination before treatment	Bone marrow aspiration cytology (Posive, Negative)/Monoclonal plasmocyte count (%)
	Bone marrow histology (Posive, Negative)/ Monoclonal plasmocyte count (%)
Biochemistry before treatment	Haemoglobin level (g/l)
	Thrombocyte count (10E9/l)
	Calcium total level (mmol/l)
	Albumin level (g/l)
	Creatinine level (umol/l)
	Beta2 microglobulin (mg/l)
	LDH (ukat/l)
	CRP (mg/l)
Cytogenetics before treatment	Date of sample collection
	Conventional (Normal; Changed; Complex)
	FISH: Method used; IGH disruption; t(11;14); t(4;14); t(6;14); t(14;16); del(13)(q14)/monosomy 13; gain 1q21; del(17)(p13); Hyperdiploidy;
Flow cytometry before treatment	Yes/No; Sample date
Medical Imaging Modalities before treatment	Osteolytic lesions - X-ray
	Osteolytic lesions - NMR
	Osteolytic lesions - CT
	Osteolytic lesions - LDCT
	Osteolytic lesions - PET
	Osteolytic lesions - PET/CT
	Osteolytic lesions - MIBI

Diagnosis MM – Form Treatment Group of questions	Question
Extramedullary mass before treatment	Extramedullary mass (yes/no)
	Extramedullary mass - relation
	Extramedullary mass histology
	Extramedullary mass - lesion count
Stage before treatment	Durie-Salmon stage
	Durie-Salmon substage
	ISS classification
	R-ISS classification
Treatment	Treatment regimen
	Health insurance company
	Clinical study (Yes/No; Name of clinical study)
	Length of cycle (days)
	Date of treatment initiation
Drugs - overview (depends on drug type)	Dosage
	Number of administrations per cycle
	Number of cycles
	Route of administration
	Dose adjustment (Yes/No)
Dose adjustment	Date of dose adjustment
	New dosage
	Dose interuption (Pernament, Temporary) - if new dosage = 0
	Reason/Grade (Neuropathy; Nausea, vomiting; Anorexia; Diarrhoea; Constipation; Fatigue; Thrombosis/Thrombus/Embolism; Infection; Thrombocytopenia; Neutropenia; Anaemia; Rash; Other)
Drugs - overview of total cumulative doses	Total number of administrations
	Total cumulative dosage (mg)
Laboratory examination after treatment	Serum M-protein level after treatment (g/l)
	Urine M-protein level after treatment (mg/24h)
	Urine M-protein level after treatment (mg/l per 24h)
	Serum M-protein - ratio after treatment/entry (%)
	Urine M-protein (mg/24h) - ratio after treatment/entry (%)
	Urine M-protein (mg/l per 24h) - ratio after treatment/entry (%)

Diagnosis MM – Form Treatment Group of questions	Question
Laboratory examination after treatment	Immunofixation after treatment - serum (Posive, Negative)
	Immunofixation after treatment - urine (Posive, Negative)
	Monoclonal plasmocyte count (%) in bone marrow aspiration after treatment
	4-color plasmocyte cytometry after treatment (Posive, Negative)
	8-color plasmocyte cytometry after treatment (Posive, Negative)
	Kappa/lambda ratio after treatment (Normal, Abnormal)
Toxicity before treatment	Grade of thrombocytopenia before treatment
	Grade of neuropathy before treatment
Toxicity during treatment	Toxicity (Neuropathy; Nausea, vomiting; Anorexia; Diarrhoea; Constipation; Fatigue; Thrombosis/Thrombus/Embolism; Infection; Thrombocytopenia; Neutropenia; Anaemia; Rash; Other)
	Grade/Related
Anticoagulant treatment	Anticoagulant treatment (yes/no)
	Type of anticoagulant treatment
Granulocyte-colony stimulating factors (G-CSF)	Granulocyte-colony stimulating factor (yes/no)
	Dosage 30 IU / 0,5 ml - number of doses
	Dosage 48 IU / 0,5 ml - number of doses
Switch of therapy	Switch of therapy regimen (if yes all of previous questions are collected again)
Radiotherapy	Radiotherapy (yes/no)
	Type of radiotherapy (Analgetic, Therapeutic)
	Total dose [Gy]
Transplantation	Transplantation (yes/no)
	Response before transplantation (sCR; CR; VGPR; PR; MR; SD)
	Date of transplantation
	Special transplantation technique
	Type of conditioning regimen

Diagnosis MM – Form Treatment Group of questions	Question
Treatment withdrawal	Date of treatment withdrawal
	Reason for treatment withdrawal (Treatment response; Treatment response + PBSCT (transplantation); According treatment protocol; Insufficient response; Progression; Toxicity; Exitus; Rash; Other)
	Interruption of treatment (Date, Reason, Grade)
Response to induction treatment	PR after cycle
	Date of partial response (for DOR calculation)
	Maximal response after cycle
	Date of maximal response
	Maximal response to treatment (sCR; CR; VGPR; PR; MR; SD; PD)
	Final response after induction therapy (sCR; CR; VGPR; PR; MR; SD; PD)
Consolidation treatment	Consolidation treatment (yes/no)
	Specify consolidation treatment
	Date of treatment initiation
	Date of treatment withdrawal
	Reason for treatment withdrawal (Treatment response; According treatment protocol; Insufficient response; Progression; Toxicity; Exitus; Rash; Other)
	Number of cycles
	Maximal response (sCR; CR; VGPR; PR; MR; SD; PD)
	Date of maximal response
Consolidation treatment – toxicity	Toxicity (Neuropathy; Nausea, vomiting; Anorexia; Diarrhoea; Constipation; Fatigue; Thrombosis/Thrombus/Embolism; Infection; Thrombocytopenia; Neutropenia; Anaemia; Rash; Other)
	Grade

Diagnosis MM – Form Treatment Group of questions	Question
Maintenance therapy	Maintenance therapy (yes/no)
	Specify maintenance therapy
	Date of treatment initiation
	Date of treatment withdrawal
	Reason for treatment withdrawal (Treatment response; According treatment protocol; Insufficient response; Progression; Toxicity; Exitus; Rash; Other)
	Maximal response (sCR; CR; VGPR; PR; MR; SD; PD)
	Date of maximal response
Maintenance therapy – toxicity	Toxicity (Neuropathy; Nausea, vomiting; Anorexia; Diarrhoea; Constipation; Fatigue; Thrombosis/Thrombus/Embolism; Infection; Thrombocytopenia; Neutropenia; Anaemia; Rash; Other)
	Grade
Response after treatment line	Final response after treatment line (sCR; CR; VGPR; PR; MR; SD; PD)
	MRD evaluation (Method, Date, Result, Sensitivity, Grade)

Table 22 Example of form Current status for patients with multiple myeloma

Diagnosis MM – Form Current status Group of questions	Question
Secondary amyloidosis	MM associated amyloidosis (yes,no)
Current status	Date of the last update
	Patient status (Unknown; Alive; Dead)
	Patient status - alive (Unknown; Treated; Not treated)
	Date of death - if Patient status = death
	Cause of death (Related to diagnosis, Other)

Table 23 Example of form Watch and wait for patients with multiple myeloma

Diagnosis MM – Form Watch and wait Group of questions	Question
After line of treatment	After line
Progression	Date of progression

Table 24 Example of form Diagnostics for patients with MGUS

Diagnosis MGUS – Form Diagnosis Group of questions	Question
Diagnostics	Date of diagnosis
	WHO ECOG
	M-protein type (IgG; IgA; IgD; IgE; IgM; Biclonal; Triclonal; Nonsecretory; LC only)
	Light chain type (Kappa; Lambda; Biclonal)
	Serum M-protein quantity (g/l)
	Urine M-protein quantity (mg/24h)
	Urine M-protein quantity (mg/l per 24h)
	FLC quantity (if measured)
Laboratory examination	IgG quantity (g/l)
	IgA quantity (g/l)
	IgM quantity (g/l)
	IgE quantity (IU/ml)
	HLC quantity (if measured)
Bone marrow examination	Bone marrow aspiration cytology (yes/no; Monoclonal plasmocyte count)
	Bone marrow histology (yes/no; Monoclonal plasmocyte count)
Biochemistry	Haemoglobin level (g/l)
	Thrombocyte count (10E9/l)
	Calcium total level (mmol/l)
	Albumin level (g/l)
	Creatinine level (umol/l)
	Beta2 microglobulin (mg/l)
	LDH (ukat/l)
	CRP (mg/l)

Diagnosis MGUS – Form Diagnosis Group of questions	Question
Cytogenetic - FISH	Cytogenetics (yes/no)
	Date of sample collection
	IGH disruption; t(11;14); t(4;14); t(6;14); t(14;16); del(13)(q14)/monosomy 13; gain 1q21; del(17)(p13); Hyperdiploidy;
Flow cytometry	Plasmocyte count FC (%)
	Clonal PC (%)
	Polyclonal PC (%)
Medical Imaging Modalities	Osteolytic lesions - X-ray
	Osteolytic lesions - NMR
	Osteolytic lesions - CT
	Osteolytic lesions - LDCT
	Osteolytic lesions - PET
	Osteolytic lesions - PET/CT
	Osteolytic lesions - MIBI
Molecular-biology (for M-protein = IgM)	Molecular-biology examination (yes/no)
	MYD88 examination
	Mutation in CXCR4 gene

Table 25 Example of form Current status for patients with MGUS

Diagnosis MGUS – Form Current Status Group of questions	Question
MGUS development	Progression MGUS to (No progression; MM; WM; Lymphoma; AL amyloidosis; CLL; Plasmacytoma; Other)
	Further specification
	Date of MGUS progression
Current status	Date of the last update
	Patient status (Unknown; Alive; Dead)
	Date of death if Patient status = death
	Cause of death (Related to diagnosis, Other)

Table 26 Example of form Follow up for patients with MGUS

Diagnosis MGUS – Form Follow up Group of questions	Question
Follow-up	Date of sample collection
	Serum M-protein quantity (g/l)
	Associated disease (None; Lymphoproliferative disorder; Other haematological disease; Connective tissue disorder; Neurological disease; Dermatological disease; Endocrine disorder; Liver disease; Immunosupression; Other)

3.6 Training

As described in the sections above, it is quite complicated to prepare a complex and functional registry. Many different aspects have to be considered and the result is very specific for each particular registry. Even if you are working with an experienced investigator and well-instructed staff, you should always consider proper training on how to proceed for this concrete project, what is required, recommended and forbidden, and especially how to make the entire project work more efficiently. It is important to know that not only initial training during the project's setup period should be considered, but also training for new users added during the project; moreover, training after an update of the database structure is very important in order to guarantee consistent work with the database system throughout the study's duration for all database users.

3.6.1 Types of training

There are several types of training and the goal of each always involves the organisation of such type of training that will efficiently ensure proper training for all relevant staff with reasonable costs. The paragraphs below represent different categories of training, starting from more detailed (but more expensive) to less proper (but less time-consuming).

Personal on site

It is not surprising that personal training is the most efficient from the point of view of information transfer. Once you have time dedicated to one site, you can answer all questions, explain all the details properly and make sure that the trainees understood everything well. Moreover, the personal knowledge of team members also has a very positive impact on the study process. You can apply this type of training especially in cases with a low number of sites and a more complicated process (extensive CRF, complicated study procedures, strong quality demands etc.) or in projects with a middle or higher budget. Unless there are hundreds of sites in the project, you can understand training on site as an investment that should be repaid in the future during the project's duration.

Personal investigators meeting

This type of training is also very efficient. A very careful preparation is the key aspect of such a training. You should take care to plan the date of the meeting enough in advance (more than 1 month) to be sure that most investigators will have time to participate in the meeting. You should also thoroughly plan the agenda and make sure that you will provide all the important information, such as:

- clinical background (key opinion leader strongly welcomed),
- typical patient,
- details of the protocol, inclusion criteria, exclusion criteria,
- organisation aspects (project duration, number of planned patients, ...),
- specific processes (AE reporting, ...),
- project milestones (analyses, when and what is included),
- nature of data entry, frequency of data entry,
- detailed presentation of eCRF including technical and clinical manual,
- contacts – whom and when to contact (helpdesk, PM, sponsor, medical expert etc.).

In terms of content of the presentation, everything must work with respect to speaking to the personnel directly. As many as possible questions should be answered. It is recommended to have a small discussion after each part in order to keep the audience involved in the meeting. There are several disadvantages when compared to on-site training, such as less time for each investigator, fewer chances to discuss the setup activities for each site and less opportunities for personal contact. On the other hand, advantages include having open discussions between medical experts and investigators, the ability to hear the opinions of colleagues and the collective approach to the project's success. Besides that, the organisation of one meeting should be cheaper than on-site training for tens or hundreds of sites.

Audio/video conference

An audio/video conference is a very popular and efficient way to train many people with very limited costs. You can save time because no travelling is needed and you can train many people at once. Unfortunately, there are several disadvantages of using this kind of training:

- No personal contact
- Limited possibility to ask and answer questions
- Potential technical problems with installation and using the video conference application

We can recommend using this solution especially for:

- Projects with a very easy eCRF
- Special situations like training new staff in an already-running project or training several new functionalities after the database is updated
- Projects where you know the investigators and system users well and they know the system well, too

When using this solution, be sure to prepare the agenda carefully; you should be able to guarantee the duration of meeting agenda. Count on a delay at the beginning because a significant number of users will not be ready at the planned starting time (often due to technical obstacles even if the system is running properly). When presenting for a high number of meeting participants, we recommend keeping everybody on "mute" status to keep the meeting compact and to let the audience discuss at the end of each section only. If you want to be sure you are sticking to the planned schedule, it is recommended to let participants ask questions only in written form and to provide detailed answers later by email (for a large number of participants only).

Video presentation

Video presentation training is also a very efficient and popular method of training, especially for projects with a huge number of investigators, sites and countries participating in the project. It is very cost-effective because you prepare all the materials only once and the rest of the process is very flexible. Participants can attend the training whenever convenient for them, they do not need to install anything and they can read all the information slowly and carefully, as they would prefer. The requirement to have everything perfectly prepared in this type of training is very strong, because the way you prepare it will affect the rate at which users will adopt the system and processes, which will strongly impact the entire study process. It is always recommended to define the period in which the training should be completed and allow the participants to ask questions in a way that will be convenient for both you and them. An optimal option for asking questions is a form linked to the video presentation in order to simplify the process. The advantage is that you can prepare as many video presentations as you need during the study's duration. The only thing you need to think about is to keep everything clear for system users. For example, you can put all the training materials directly into the eCRF (on the dashboard) so that all users can easily come back to it during the duration of the study. Since it is not printed, you can update it anytime if needed.

Text presentation

A text presentation represents a similar type of training as a video presentation, but while the video presentation is more easily understandable, it is more time-consuming in regards to the preparation as well as more expensive to produce.

Finally, it is good to say that we have quite a lot options for training system users, investigators and other staff. We just need to think and choose the right option for our project. An ideal solution is typically combined from the above-mentioned types of training.

Example: We can imagine the initiation of an important and rather complex registry, initially planned for five big sites; additional sites might be added in future. We will start with a group of on-site trainings for all five sites. After one year, 15 more sites are joining, so we will organise an investigator's meeting inviting representatives from all sites; besides the initiation training for the new sites, we will plan a discussion forum focused on sharing the experience from the five initial sites. After some time, we need to update the eCRF due to new methods of treatment, and this update means several modifications in working with the whole system. We will then organise a video conference where we explain the reasons and key points of the update and prepare a video presentation that can be viewed by anybody and anytime to see in further detail what are the new processes and what should be observed. You can see that all types of training could be efficiently realised in one project.

Training on site

Training on site is organised in situations involving more complex projects when more information than a simple training is needed to guarantee sufficient training. The key aspect is to consider all information that should be part of the training:

- Clinical background. It must be clearly described what patients will be included in the project, and the details of their inclusion and exclusion criteria should be discussed as well as the key parts of the study protocol.
- Organisational aspects should be discussed, including details of the contract between the investigator and the sponsor or CRO. The contract should contain information about the project's duration and the planned number of enrolled patients, the level of information entered into the system and the declared responsibility for data quality guaranteed by the investigator.
- The means of working is important, too. It should be clear how often data should be entered into the system, how to work with queries that arise, and compliance with project milestones (e.g. defined data entered into the database before freezing data for interim analyses etc.).
- Data security and technical background should be discussed. The investigator should present on the process of working with data will be undertaken.
- Detailed presentation of eCRF structure including the provision of key documents and technical and clinical manuals. It must be clear that the investigator (system user) understood the system functionalities properly and that he/she is able to enter data into the system properly and in the required quality.
- Specific processes such as the reporting of adverse events should be explained in detail as well.
- Contacts: it must be clear who should be contacted in case of any issue or question (helpdesk, PM, sponsor, medical expert etc.).
- Finally, the investigator(s) should sign the training form confirming that all key parts of the training were done properly and that everything is clear to the investigator(s).

All personnel that will participate in the project should participate in the training, and the site staff should ensure enough time for such a site training in order to be sure that every topic will be properly covered. If the training is well prepared, the site staff has enough time for training and you are clear, open and give significant space for questions, this will lead to a successful training. Last but not least, you should provide relevant contacts to people who will be able to help and leave the impression that all questions are welcome.

4. PHASE 3: HOW TO KEEP THE REGISTRY ALIVE

4.1 Introduction

Having the first patient in the database is a very significant milestone in the project's life cycle. The database is ready now, most of the settings are already done, but there is still a lot of work to do. If the preparation activities were done carefully and the database was set up properly, the majority of the processes could run automatically. Nevertheless, there are still a lot of things to care for. The remaining data management activities are roughly described in the scheme displayed below (Figure 44) and can be divided into three parts:

- The first column (FPI to error messages of coherence checks) represents the communication between investigator entering data and the system accepting data or informing the investigator about any detected inconsistencies. The system alerts the investigator according to checks programmed in the database during the setup phase.
- The second column (data correction to data review) represents a very complex process of data validation. It is one of the crucial activities to obtain the data in the required quality.
- The final column (database lock to statistical analyses) represents the final part of the data management process, preparing the data for statistical analyses and locking the database to prevent any database users from modifying any data.

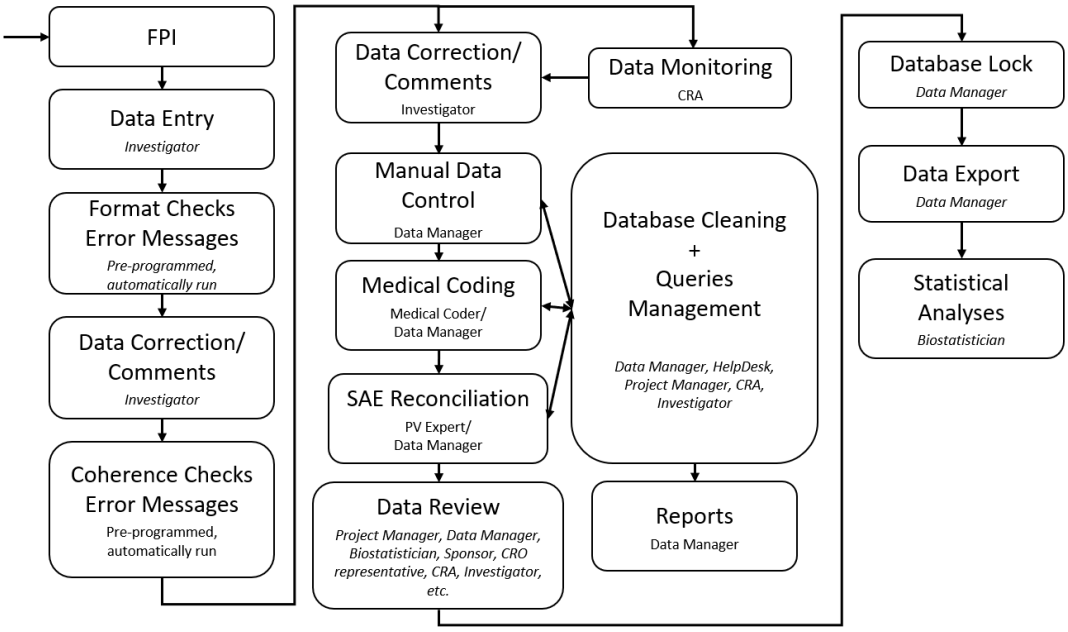


Figure 44 Data management process: FPI to statistical analyses

4.2 Data validation

This chapter describes the continuous process of data validation, including particular steps of data validation. We have already described the process of creating validations (see the chapter Programming of validations). Now, once the project is live, it is obvious that all of the automatic checks planned to be applied are already programmed and should be in progress. This means that work of the database user is continuously checked according the initial plan.

- What is the impact? The primary impact should be better data quality. How is this possible? The investigator enters data into the EDC system – and once the EDC system finds any inconsistency compared to the programmed validations, it alerts the investigator to correct this mistake. However, there can be also negative impact. If the number of validations is too high and their level is too strict, there might be a lot of queries. Some of these queries might be even useless. We are talking about the capturing of real-world data, i.e. data from real practice that, when compared to clinical trials, are not so strict (e.g. regarding the permitted ranges for different parameters). Therefore, we must carefully plan and program validations to keep investigators on friendly terms with the data capture system.
- Is this enough? On the other hand, we might ask a question if all of the programmed checks are enough to guarantee the required data quality. The answer is not so simple. It is good to consider which data are critical, how we can check them and, in combination with the data review, find out if we should do so in practice. If not, we can add a set of manual checks that will help us increase the data quality.
- How to communicate with sites? If the validation finds some inconsistency, it will always result in a query. The simplest possible way to communicate queries to investigators must be found, because they may understand answering queries as an extra job and if there are still new queries it can be very demotivating for investigators, which might have a negative impact on data quality.

All programmed validations are being performed in different steps of the data entry process. During the data processing, in the first step the format of all variables, depending on whether the variable is mandatory (Figure 45), and programmed coherence checks on the form level are checked. As an example of how these types of checks can be presented to the registry user, screenshots from the Clade IS EDC system are presented below.

Figure 45 Example of online validation of mandatory question

The registry user can interact immediately by correcting the inserted value or leave it for a later correction (the pending form status would then be used as it allows invalid data).

The next step in the process of data validation consists of programmed coherence checks that validate parameters across the whole database. Based on the used EDC system and the setup of the concrete project, the results of these checks might lead to the generation of queries in order to correct the data.

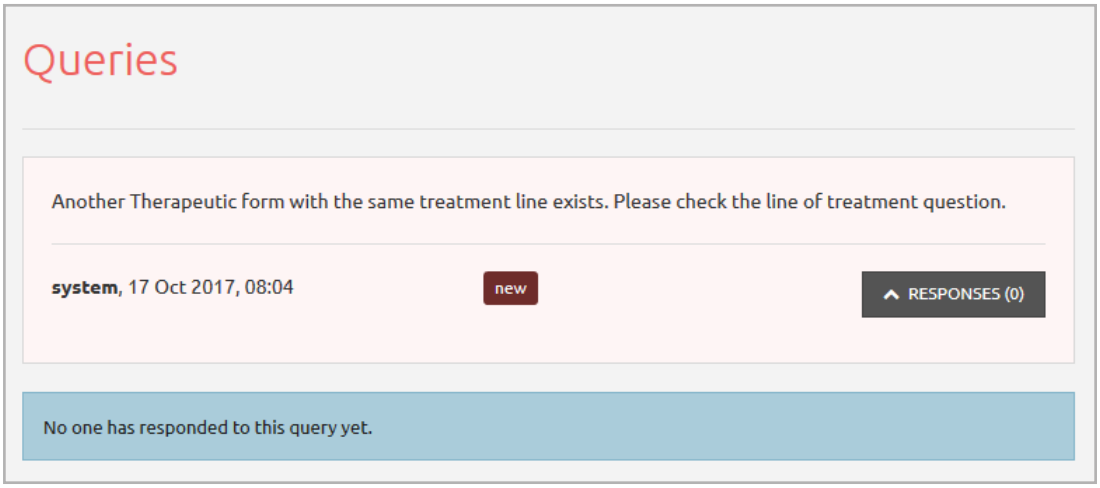


Figure 46 Example of query validating parameters across the whole database

Queries can also be set to allow a user input. The investigator can then respond to those queries and, for example, clarify a rare case with a combination of specific parameters for which those validations should be manually overridden. The project coordinator, monitor or data manager can then reply as well or communicate the case in further detail. The whole communication is attached to that specific query.

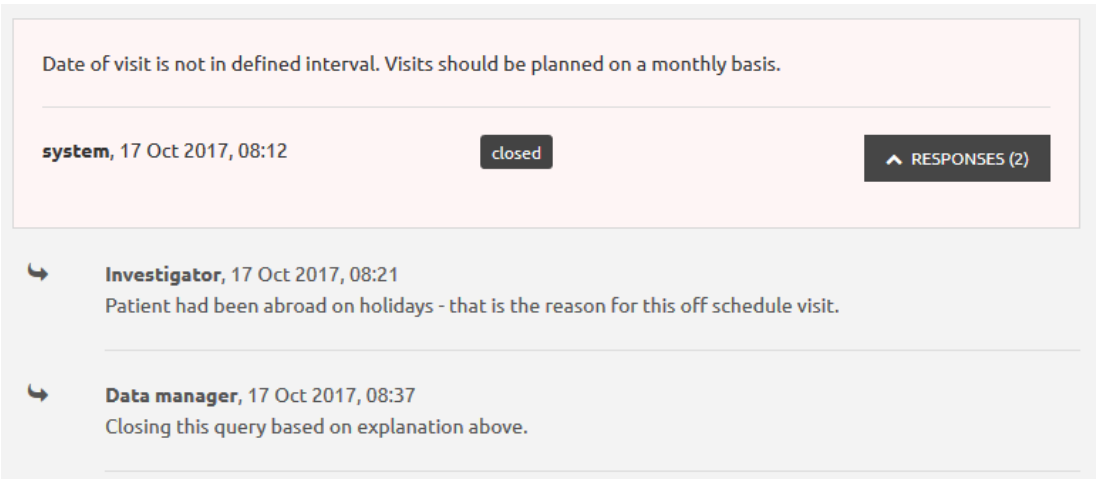


Figure 47 Example of investigator and data manager communication in order to resolve a query

Manual queries

The initial feeling about manual queries might be that they are redundant because of the wide scale of possible automatic checks. The reality is different. There is always something that:

- cannot be checked automatically (for example, the logical consistency of medical terms collected in the text format; the system cannot check these terms automatically, it must be done manually, ideally by a medical expert);

- is more efficient to check manually (checking several values present on different places of the database is too complicated for programming, but simple for a manual check).

Processing of manual queries

Any member of the project team, depending on the set-up database permissions, could create a manual query. Manual queries are typically created by either the data manager, medical expert or pharmacovigilance specialist, i.e. people who are in charge of data review. If the monitoring process is applicable, manual queries might be also be generated by the CRA.

System queries are applicable immediately after the project goes live. The more users are active in entering data, the more automatic queries are generated, so we can say that the amount of new automatic queries mimics the activity of users. For manual queries, this is different. The amount of new manual queries mimics the activity of other team members. Moreover, they cannot start with data control immediately after the project goes live, because there are no data at that time. Alternatively, we can say that the more project is close to its end, the more manual queries appear in the system because some data could be only checked after the data are complete or almost complete. It is important to inform investigators about the planned process of data cleaning in order to let them know that, in addition to the common data entry process, the expected workload will increase before key milestones and before the project's end, because most of the queries are active in these periods.

Manual queries, as well as automatic queries, should be answered by the investigator. The investigator is responsible for the data, so he/she should confirm that the data are entered properly.

Manual queries should be closed by the person who created them. If somebody creates a manual query, he/she should know why the query was created and whether the answer provided by the investigator is relevant or not. If the answer is not relevant, he/she should query the investigator again; only if the query is answered sufficiently, it should be closed.

The output of manual data control can be sent to registry users by email or can be entered into the registry. Some discrepancies can be found during the checks, which offer an incentive to set new limits for the parameters or new validations. After a period of data collection in the registry, there is a need to pool all the generated queries so that registry users can search and work on resolving them in an efficient way. The user has an option to search in all his/her queries or just queries attached to a specific patient. The following images show an example of a query search.

As was the case with patient forms, which have different statuses, queries have statuses as well, informing whether they are open, in progress or resolved. Once a query is created, it is clear that it should be resolved in order to get as clean data as possible.

Since the number of queries is a significant project metric and might be used for reporting the global project status to the project team as well as providing relevant feedback to sites, it is recommended to monitor the number and status of queries in different sites and countries on a regular basis. This can be done optimally via queries reporting.

Results

Show 100 entries

Query ID	Patient ID	Investigator	Form	Message	Status	Answers	Author	Created
7461	RMG-CZ00-0001-test	Pavlíčková Barbora	Diagnostics	The answer filled in the questions Previous history of MGUS and Data not available (on the Diagnosis form in the MM phase) must correspond with creating/not creating of the Diagnosis form in the MGUS phase.	new	0	system	1 Apr 2018, 11:45
7462	RMG-CZ00-0001-test	Pavlíčková Barbora	Treatment	The Date of treatment withdrawal must follow the Date of treatment initiation on the Treatment form (in phase MM).	new	0	system	1 Apr 2018, 11:45

Showing 1 to 2 of 2 entries

Previous 1 Next

EXPORT RECORDS TO XLS

Figure 48 Example of a searchable list of queries for a single patient or for all investigator queries

4.3 Data review

Data review is a very complex process that can be very diverse in different projects. The goal of this process is to review the collected data before a project milestone in order to make a certain decision or to update the established process. This process is very often undervalued because it is usually not strictly defined and required. However, placing the data review process – or at least a part of it – into the clinical project process might have a strong positive impact on the project's success. The data review process offers a unique opportunity to review data globally, not on an individual basis as during the process of query management. Especially for long-term projects, it is recommended to set up the milestones and frequency of particular data review processes in order to enable updating the processes in case some inconsistencies are found. The data review process usually focuses on:

- Preparation of specific data listings containing a list of all/selected data entered into the database
- Medical review of specific predefined consistencies of medical information
- Patient enrolment
- Global overview of subjects and their database statuses (for example enrolled / screened / completed / validated)
- Global overview of forms and their statuses
- Total number of queries in particular sites compared to the number of open queries in order to discover:
  - Systematic mistakes of one (or several) sites. One site has more queries than the others do. New training might be needed.

    - Too many open queries. Investigators are not working properly. Maybe the CRF is too complicated; maybe the investigators have not been trained properly.
    - Feedback to data management. One type of query is more frequent than the others. Maybe something is not clear or is ambiguous. A database update might be needed.
    - Feedback for the DM. Some validations do not produce any query. Maybe it is not well programmed, maybe it is redundant.
- List of specific data inconsistencies, coming from manual queries or messages from investigators
- Fluency of patient enrolment, feedback from sites
- Plan of remaining steps until the nearest milestone (temporary lock, data export, annual analyses etc.)
- Overview and classification of protocol deviations and relation to project populations

The typical process of data review contains a meeting or conference call as the key part of the whole process. Then following steps should be respected:

- Agenda. It is very important to define in advance what will be discussed and ask relevant personnel to prepare materials for this discussion. All pieces of information mentioned in the paragraph above have to be composed of the data in the clinical database by a member of the project team. The agenda is usually prepared by the project manager, i.e. the key person involved in the communication process.
- Participation of all key people. Depending on points planned to be discussed during the data review meeting, it is important to invite the relevant people in order to arrange a constructive discussion leading to appropriate conclusions.
- Minutes. All the decisions should be written, reviewed by all participants, confirmed and optimally signed. Then, it is the perfect base for further steps minimising potential further discrepancies.
- Be sure that the agreed points are performed. Only having the signed document is not a guarantee that everything will work properly. Be sure to manage all consequential steps such as the update of the project database including testing and validation, an update of the technical manual, an update of the clinical manual, new training of sites etc.

The following personnel should participate in the process of data review:

- Project manager, having an overview of the clinical part of the project
- Data manager, having an overview of the database status, number of queries and data quality
- Data analyst, who is able to confirm whether the current project status, structure of the collected data, quality of the collected data etc. is sufficient for the required statistical analyses
- Medical expert, who should confirm that the collected medical terms are consistent across the whole database and that physicians are collecting meaningful data
- Sponsor representative, who should confirm that the project is running according to expectations and in line with the initial documentation and defined processes
- The above-described list of personnel is not ultimate. Any team member whose presence is relevant should be invited, too. As mentioned above, the data review process is important and we should take care to hear all potentially useful comments.

As mentioned above, the process of data review is organised especially before key project timelines such as interim analyses or project revision milestones. Nevertheless, some parts of the data review process can be realised continuously during the project's duration. One of these steps could be the continuous revision of entered data, which could be carried out by the data entry coordinator who can be in permanent contact with investigators and give them regular feedback about the quality of their data. He/she can simultaneously discuss with investigators potential issues with data entry and help them work with the system properly. A key skill of coordinator is the ability to communicate. He/she should be in contact with the database users, but also with the team of service providers in order to inform the helpdesk, data managers or project managers about potential issues or about a situation in the process of data entry, which might help to improve the whole process in a relevant manner.

4.4 Reporting

This chapter describes in more depth the practical aspects of data reporting and displays some practical examples of a few reports.

Within the registry duration, there are several aspects about data reporting that need to be considered. Running a registry is not a short-run activity, and multiple structural and functional changes might occur. All of these may or may not affect reports that have already been set and that are being generated on a regular basis. Any change that is performed should be analysed and the affected report must reflect this change. Data reporting is usually handled by the data management department, so there is only a small chance that a database change (handled by data management as well) of any kind would go unnoticed. In spite of this, it is a good practise to set some kind of regular report testing or at least an overview before sending any report out to its recipient.

An important aspect of data reporting is the feedback from receiver(s) of the report. In general, it is usually the sponsor or client finding out that interim evaluations or analysis are diverging from initial expectations. A few possible outcomes may arise from this. The most common would be the modification of eCRF or patient criteria so that proper and relevant parameters are being collected.

As the volume of entered data grows in time, the informative value of such registry grows as well, and therefore more sites may want to participate in data collection for that registry. It is well known that more does not always translate into better. Specific reports can then reflect certain ways in which data are being collected and the amount of data each site is entering, so these reports may reveal less active or even inactive participants. It is then up to the sponsor or client how to handle this; the discontinuation of such participants would be an option only in extreme cases.

More sites and more sponsors as well may join the registry during its duration. More sponsors may also mean more and different reporting demands, so another setup of the reporting process may occur, as described in the chapter Setup of the reporting process.

As with joining the registry, discontinuation has to be handled as well. Report recipients have to be managed so that no longer involved parties do not receive any kind of regular report from the registry.

Summary report dated on : [redacted]

Site name	Site acronym	Principal investigator	Country	Date of site initiation	No of screened patients						
					All	11/2017	10/2017	9/2017	8/2017	7/2017	6/2017
[redacted]	CZ-001	[redacted]	Czech Republic	26-JUL-2016	33	2	3	2	2	4	0
[redacted]	DE-017	[redacted]	Germany	18-JAN-2017	18	0	0	2	3	1	4
[redacted]	CZ-004	[redacted]	Czech Republic	05-SEP-2016	16	3	0	1	1	1	1
[redacted]	CZ-011	[redacted]	Czech Republic	28-NOV-2016	10	3	0	1	0	1	1
[redacted]	ES-015	[redacted]	Spain	03-JAN-2017	10	2	0	2	0	0	0
[redacted]	ES-012	[redacted]	Spain	13-DEC-2016	9	0	0	0	1	3	3
[redacted]	CZ-006	[redacted]	Czech Republic	22-SEP-2016	8	1	0	0	1	1	0
[redacted]	CZ-002	[redacted]	Czech Republic	12-AUG-2016	8	1	1	1	0	1	0
[redacted]	PT-021	[redacted]	Portugal	20-FEB-2017	8	0	0	0	2	1	1
[redacted]	AR-019	[redacted]	Argentina	09-FEB-2017	7	1	0	1	2	0	1
[redacted]	ES-007	[redacted]	Spain	18-OCT-2016	7	0	0	0	2	0	0
[redacted]	ES-030	[redacted]	Spain	09-MAY-2017	6	3	0	2	0	1	0
[redacted]	PL-013	[redacted]	Poland	20-DEC-2016	5	0	0	0	0	3	0
[redacted]	PL-022	[redacted]	Poland	22-FEB-2017	4	0	0	1	0	2	1
[redacted]	ES-025	[redacted]	Spain	15-MAR-2017	4	1	0	2	0	0	1
[redacted]	SK-008	[redacted]	Slovakia	02-NOV-2016	4	0	1	0	1	1	0

Figure 49 Treatment report

Site name	Sum of all patients on [redacted]	Treated ATM on [redacted]	Enrolled on 10/2017 (now on [redacted])
[redacted]	0	0	0
[redacted]	2	2	0
[redacted]	3	2	0
[redacted]	0	0	0
[redacted]	1	0	0
[redacted]	5	2	0
[redacted]	0	0	0
[redacted]	31	7	4
[redacted]	2	1	0
[redacted]	0	0	0
	44	14	4

Figure 50 Global overview report

[redacted] registry - medication - [redacted]

Updated on: [redacted]

Summary report - all patients in the registry

Medical centre/Hospital	Total number of patients with report of the treatment by the medication	Number of patients who started the biological treatment in 2016				Total number of valid patients	Number of patients with ongoing biological treatment (valid patients)				Number of patients with terminated biological treatment
		First line	Second line	Third and additional line	Total		First line	Second line	Third and additional line	Total	
[redacted]	0	0	0	0	0	0	0	0	0	0	0
[redacted]	0	0	0	0	0	0	0	0	0	0	0
[redacted]	4	1	0	2	3	3	1	0	2	3	0
[redacted]	1	0	0	1	1	1	0	0	1	1	0
[redacted]	4	0	0	2	2	4	0	0	4	4	0
[redacted]	0	0	0	0	0	0	0	0	0	0	0
[redacted]	9	0	0	3	3	5	0	0	5	5	0
[redacted]	1	0	0	1	1	0	0	0	0	0	0

Figure 51 Customised enrolment report

4.5 Import

There are two basic principles of importing data for clinical projects:

One-time data import

The registry is a long-term project, so new sites and users are joining throughout its duration. These new participants may maintain some kind of registry of their own and may be collecting similar data. It may be convenient to import their data into our registry.

Alternatively, big and sophisticated registries may want to join data into one main registry as well. The larger the amount of data that needs to be imported, the greater role the data manager has to play.

The process of data import is not an easy one and many issues have to be considered:

- Mapping process is the main responsibility of the data manager during any data import. This process consists of communication with the data managers responsible for the data that needs to be imported and deciding which parameter from the source database corresponds to which parameter in the target database. This also involves creating rules for the EDC system, which then transform the data from the source database to the target database.
- The decision whether the whole dataset is needed or just a subset would be used.
- We must be sure of whether the dataset is valid and properly cleaned or whether we must perform some additional data validation.
- We must determine if it is required to collect more or different parameters than those already defined in our registry. This may lead to a modification in eCRF (see the chapter Database update), which is not a simple step; it should be therefore decided if those data are worth it.

Data import on a regular basis

Some providers of external data (such as laboratory data, questionnaires or special examinations) may have an option to export data from their devices, so a follow-up import would be an option.

This case should be recognised before the registry launch so an appropriate part of eCRF may be designed for this type of data.

To set up the whole process properly, it is necessary to clarify the following points with data providers:

- List of variables and their data types
- The format of source data
- The frequency of sending data
- Method of providing data (cumulative or partial)
- Other specific requirements

After each data import, it is necessary to perform quality control in order to ensure that all data have been imported completely and correctly. The basic idea of the Quality Control process is to compare the source data with data in the target clinical database. The whole process should be properly documented.

If only a little amount of data should be imported, these can alternatively be entered manually because it might be more efficient, including the quality control step. For this type of process, the data entry specialist would be a great option to handle this.

4.6 Database update

Because running a registry is usually a long-term project, it is quite common to modify the database structure while the project is running.

There can be various reasons for the database update:

- Simplification of the database structure. Even if the database is properly set up and carefully discussed, we might discover during the project's duration that some changes would be welcome for the database users, sponsor or the professional society.
- News in the system of data collection. As each system is developed, after years of the project's duration there might be new functionalities available that can make some parts of the process more efficient. It is good practice to apply them.
- Adding new parameters. This item might be related to new processes in clinical practice, innovations in methods and treatment, a new medicinal product planned to be involved in the registry database etc.
- Removal of some parameters. During the project's duration, we can discover that some items are redundant. Why keep them in the database? Let's simplify the database instead.
- The request for database modification may require the analyst (because of realisation/simplification of the analysis or another practical reason), the data manager (because of improvements in data quality, adding validation, adding limits of parameters etc.), the project manager (because he/she knows the latest news in the surveyed area) and the sponsor (for example, a request of registry users) to be involved.

If the suggestion to modify the database structure comes from the registry users, the proposed modifications have to be approved by both the sponsor and the project manager. Requests to modify the registry are received by the data manager. When the entire project team and the sponsor (possibly represented by the registry users or at least by representatives of the sites) have approved the modifications, then the data manager can start the process of registry modifications. All required modifications to the registry are recorded in the document Request for database modification. A detailed database structure with the required modifications can be added to this document. Before modifying the registry, it is a good idea to download the current database structure with limits of parameters and validations. Sometimes, the database modification also involves data modifications – import, update, removal etc. In this case, it is recommended to download a complete export with all data before modifying the database. Later, we can check the data after and before the modification.

Some modifications require the suspension of the database during the process – deleting questions, changing discrete values etc. In general, we consider it more secure to keep the suspension of the database even for other modifications. The users of the registry are informed about the suspension of the database adequately in advance.

After the database suspension, the data manager can start the process of registry modifications according to the document Request for database modification. He/she begins by modifying the structure and then continues with data modification. All database changes are recorded in the document eCRF tracker, which provides descriptions of the change, the database status before and after the update, the name of the person who updated

the database, the name of the person who approved the change and the reason for the update. Data manager also records the date from when the new version of the database is valid in the eCRF tracker. For data modification, data manager creates a document that contains all commands for the update, insertion or deletion of data. If we modify the validations, we record these changes in the data validation plan document and create a new version of this document. After all database modifications are complete, it is possible to test the database.

Another data manager (i.e. not the data manager who did the modifications) then tests the new version of the database. The tester checks if all required modifications have been done and if no other changes have been made at the same time. The tester also checks the functionality of all parameters, correctness of data types, validations, the setting of skip logic and grammar by creating several test patients and their forms. If more language mutations are available, all of them must be tested. For some modifications, it is recommended to check the setting of the payment export, reports and AE reporting. After testing and correcting any errors, it is possible to launch the new version of the database. After launching, the current database structure (eCRF) is saved and the export is completed; then, the database status before and after the update can be seen.

Some database updates might require updating the data because the new structure affects the placement of the data that might be different from the original. One example could be a new item in the code list where the initial data for this new item were entered into the option "other". These data have to be replaced and carefully checked.

The registry users then typically receive an email about the database re-launch and about the changes made in the registry. If the changes were important, it is appropriate to include these changes in the technical/clinical manual and organise a training for investigators. All of the team members, especially the data analysts, have to be informed about the update that has been just carried out.

4.7 User management

We have already spoken about the setup of database users, their roles and permissions. We have specified roles for countries, sites and users; we know how to manage special requirements.

During the time when the registry is running, there can be requests to add new users or new sites, to change some user's permissions, to add or extend permissions of the existing user, to change their personal details or to deactivate some users or sites from the database due to different reasons.

For all of the above-mentioned steps, the responsible person should initially approve the requests; typically, this tends to be the project manager, who guarantees that only verified persons get access to the database.

The assigned helpdesk officer or data manager can easily create another user or site with the previously created permissions or he/she can easily deactivate users who are no longer participating in the project.

There can be numerous reasons for deactivating a user; mostly commonly, the user is no longer an employee of the participating centre.

The request for these changes can be received through the helpdesk or sent directly to the project manager. All of these changes should be logged in the user list or in some internal system.

When the helpdesk department sends some mandatory information to all database users, they should also check all automatic email responses from the user's e-mails. If they detect some undelivered message due to a non-existing email address or some other reason, they should solve this issue as quickly as possible and the result can be either the deactivation of this particular user or a change to his/her email address.

At the end of the project, there is the phase of locking the registry. User access is then modified at the request of the sponsor. The database can be either completely locked, so all investigators are unable to log in, or the database can be locked to all data changes but still can be accessed, so all data will be only in read-only form for all users. Some users can be excluded from this read-only mode as external data cleaning users, who will have permission to update some of the forms.

4.8 Data export

The purpose of a good registry is to collect high-quality data that contains all important parameters for a specific project – for example, RMG. All of our efforts described in previous chapters leads to exporting data, because a data export is a gate for further processing, mainly data review and statistical analyses.

Export format

Data can be exported into different formats, depending mainly on the purposes of data export:

Export format

- MS Excel format. The typical format used for data review or for additional operative work with the data
- CSV format. The widely-used format used for further processing or for import into another electronic system
- SAS (or another alternative format for statistical processing). Used for importing into statistical software for the purposes of statistical analyses

Format conventions

- Data with a translation of code lists. Optimal for the sponsor, CRA, medical expert etc.
- Data without translations. Optimal for data analysts (numeric codes preferred instead of text translations)
- Format conventions. Simple short names (no spaces, no diacritics, meeting specific conventions) are convenient for a data analyst.
- Format conventions. Long names are convenient for the sponsor, CRA, medical expert etc. for each orientation in the data

- Data exported in the format where each form is on a separate list. Transparent for checking according to the initial CRF, because it has the same structure
- Data exported in the format where repeating question groups are on separate lists
- All data exported in one list. This might be better for the global review as all data are grouped together
- Exported data in different language mutations if applicable
- Data of all form statuses or only chosen ones (complete data, valid forms, signed forms, paid forms etc.)

Use of exports

There can be many processes that require data exports. Moreover, each purpose might have different requirements on the export type, so more export types can exist for one clinical project. Some examples might be:

- Translation of text fields. Relevant data are exported, provided to the person responsible for translation and consequently translated
- Medical coding. Similarly, relevant data are exported, provided to the person responsible for medical coding and consequently coded
- SAE reconciliation. Similarly, relevant data are exported, provided to the person responsible for SAE reconciliation and consequently reconciled
- Data review. Relevant data are exported, usually modified, provided to the personnel who prepares materials for the data review based on these data
- Payments. We might need to export correct data (or valid forms etc.) to name those investigators who should receive a payment based on their actual work activity for the project
- Complete export. There are various purposes for the utilisation of a complete export. The primary one is to serve as a basis for statistical analyses

Export conventions

With any data export used, for the data we obtain, we must be always sure that we exported all of the data that we wanted to export. As described above, there can be more than one export type so we must always validate the export and confirm that the export was done properly and only with relevant data. We have to check, for example:

- Export-focused only on a specific country
- Export-focused only on a specific site
- Only a specific form status (completed, valid, paid etc.) required
- No variable is missing
- No error in formatting
- No error caused by the software (e.g. MS Excel has a limited number of rows and columns)

Complete export

It is possible to check the quantity and quality of the data from users of the registry through a complete export. The data manager downloads the complete export, which can contain all parameters or chosen parameters from the registry. The data manager can then check an export by himself/herself or he/she can send it to an analyst for a more complex analysis. Furthermore, the complete data export can be sent to the analyst for an additional analysis. Different exports can be used for different analyses, so we can define the export according to the requirements of the analyst – different order of parameters, different descriptions of the parameters, selected parameters or a complete export.

The investigators can also use a complete data export for their needs – for an overview of patients, drugs, treatments etc. or for their own analyses or publications. In these cases, the investigator asks the helpdesk for a data export from the registry (this request must be approved by the project manager or sponsor). The helpdesk then sends a request to the data management department and, finally, the data manager will download and define the export according to the requirements of the investigator.

The final database export prepared for the final analyses should be exported in a format preventing any further changes. It should be in read-only mode for everybody in order to ensure that no data will be unjustifiably modified before the statistical analyses are completed.

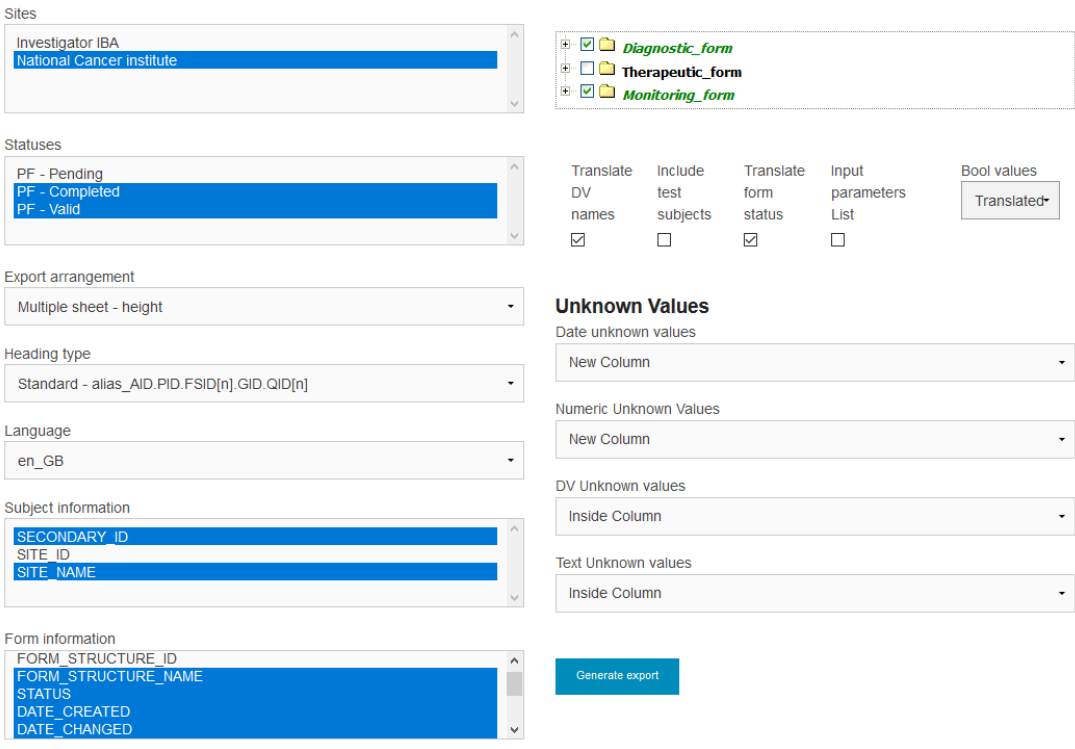


Figure 52 Example of export setting

Data exports can also be used to combine data from different databases (registries) or to connect other data to clinical data in the registry. This is how to combine data from different countries with similar databases focusing on the same area. These combined data are used to jointly analyse or to import data into an international registry. The result is a combination of patients with the same diagnosis from different countries, a larger database and a more representative sample of patients. That is why international companies often request exports from national registries. In addition, patients can request their data for use in other registries or countries for legal purposes; this depends on the legislation process in different countries, but since we collect the data of patients, they should have the right to receive these data.

Before the end of the project, it is good practice that the data manager saves the final complete export – the result of data collection – for further processing, typically for final statistical analysis or payments.

4.9 Database lock

First of all, we need to highlight one of the main differences between a clinical registry and a clinical study. Usually, a registry is observational whereas a study is investigational. Therefore, the registry does not necessarily have to be previewed to be locked one day and the registry's lifetime is not designed to be limited. For a registry, a data export is done periodically (twice a year, annually etc.), and the database lock might consist of the following forms:

- Database freeze. The first step when editing the data is prohibited to all investigators. The only right is that queries can be created, answered and closed. This is used for the phase of the project when we want to stop data entry and focus on data cleaning, typically before the final database lock or before a key milestone. Once the queries are sufficiently resolved, we can either export the data, open the database and continue with the data entry process, or continue with the locking process.
- Database soft lock. If the database is soft-locked, nobody can perform any changes in it. No data can be modified; no queries can be created and answered. This is a natural process after data is frozen before the end of the project. Another option is that we do not want investigators to perform any changes in the database. This option is typically applied during the process of database update.
- Database hard lock. The database is hard-locked only after the end of the project. The data manager cannot work in the database either.
- Complete/partial freeze/lock. The database can be frozen/locked completely, which means it is either fully applied to all sites, or partly applied only to selected sites or countries.

Once the clinical project is finished, the clinical database has to be locked. The database lock is performed according to the following checklist (Table 27) only if these components are relevant for the concrete clinical project.

Table 27: Example of Database lock checklist

☑	All data planned to be collected (CRF data; laboratory data; data from patient questionnaires) are available and entered into the database
☑	All subjects/patients in the database are valid (or in the database status required for the project)
☑	All data validation checks planned to be processed are performed
☑	Translation is performed if applicable
☑	Medical coding is performed if applicable
☑	Medical review of relevant data performed by the client/medical expert
☑	All generated/created queries are resolved
☑	SAE Reconciliation process performed and all issues resolved if applicable
☑	Quality control performed with an acceptable error rate if applicable
☑	Relevant data management documentation is approved and signed
☑	Data review performed and meeting minutes are approved, signed and implemented
☑	Approval for database lock is given by the client

The database lock ensures that there will not be any future data modification neither from the investigator's activity nor from anyone else. When the criteria for a database lock are accomplished, then the data manager can revoke the access by all investigator. As a final step, a complete data export is generated and must be checked so that all data are included in this export. This file consists of all data entered by all investigators into the database. It is provided to the analyst in order to carry out the final statistical analysis for the project.

5. PHASE 4: HOW TO ANALYSE MM DATA

5.1 Introduction

A clinical registry is an important source of data regarding clinical features of patients with the same disease or condition. Figure 53 shows the most important steps of a **statistical workflow**, which describes how to translate the information hidden in the data into clinically meaningful outputs. Data quality control has to be the first step in any research; successful "cleaning" of registry data improves the quality of the outputs. Depending on the objectives of given research, statistical analysis is performed in the next step. During the hypothesis generating process, the sample size (number of patients who fulfil inclusion criteria for the analysis) should be considered; advanced multivariate methods can be performed in cases when the sample size is large enough (i.e. more than 10 patients per one variable); in cases with a small sample size, research is restricted to a descriptive summary of data. Finally, the results can be presented in many different ways depending on the type of statistical methods used and the objectives in general.

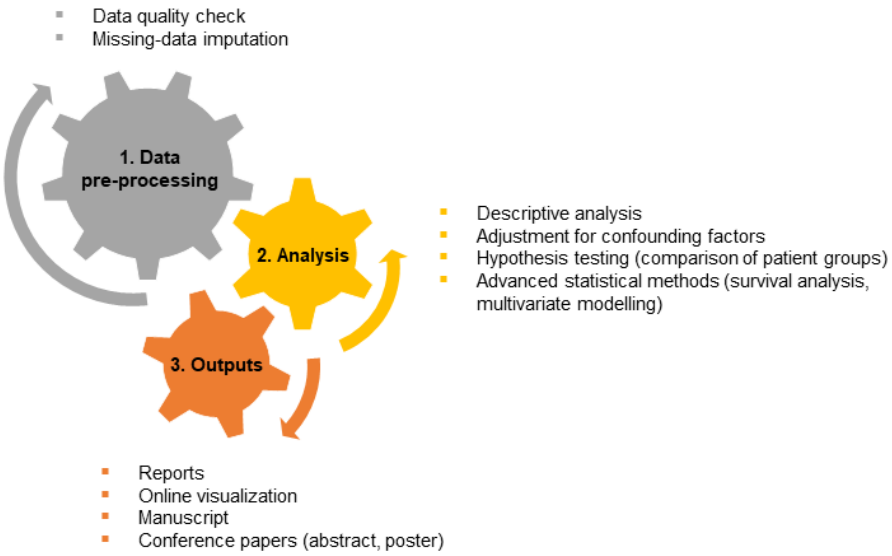


Figure 53 Statistical workflow

Correct decision making processes during the analysis together with the users' ability to interpret the results crucially influence the utility and applicability of the registry data [64]. This chapter explains how to assess clinical data of patients diagnosed with multiple myeloma (MM).

5.2 Randomisation in clinical registries?

One of the most frequent objectives is to determine the effectiveness and safety of therapy in a real world setting. Compared to randomised clinical trials (CTs), clinical registries reflect real clinical practice, where the treatment of patients is not strictly regulated by the protocol instructions; compliance is lower but the target population is wider – also in-

cluding elderly patients, patients with comorbidities and other groups, which are usually excluded from clinical trials. The effectiveness gathered from the clinical registry can be, undoubtedly, lower than the efficacy in CTs. However, the results from registries provide **real-world evidence** and therefore have a high external validity, i.e. are more generalisable to a wider target population.

Randomisation is not carried out in clinical registries. Data collected in the registries reflects real clinical practice where the treatment selection, care and its evaluation fall within the physician's decision. The attending physician may change anything during the treatment according to current clinical outcomes and his/her expertise. In such non randomised studies, it is important to consider the adjustment of treatment effect for a measured set of confounders using complex statistical methods (for an overview, see [65]). The easiest and the most commonly used methodology is to add the potential confounders as predictors in a multivariable statistical model. statistical matching or weighting techniques based on **propensity score** [66] might be another option. The propensity score describes the probability of a person to be assigned into a specific treatment group based on measured covariates (selected confounders). The procedure then matches/weights participants from two treatment conditions using this score so that the final samples are well balanced in this score and consequently also in considered covariates. This method actually performs the so-called **"post randomisation"**. The **matching-adjusted indirect comparison** (MAIC) [67] is a similar method, which also ensures the similar distribution of baseline characteristics to be the same or almost the same as the compared dataset; in this case, however, the reference dataset is not directly available as patient level data but only as aggregated data – clinical trials, other studies or experiments are most commonly used.

The main **confounders in MM data**, which should be interpreted if available, include: sex, age, duration of follow up, stage of disease (International Staging System, ISS [68]; revised ISS, R ISS [69]), patient's performance status (Performance Status developed by Eastern Cooperative Oncology Group, ECOG PS [70]), comorbidities, previous therapy (novel drugs, transplantation), reaction to the previous therapy (relapsed and/or refractory [71]) and the number of previous treatment lines.

5.3 Data pre processing

Validation processes in electronic data capture (EDC) system should significantly reduce the collection of erroneous data. As an example, the system does not allow to fill the date of treatment withdrawal which would preceding date of treatment initiation; as another example, the system can automatically calculate staging systems using laboratory values. The character of every item and its limits (range of numerical variables) in the registry should be defined before the registry establishment, in this way, the system itself restricts the collection of extreme or incorrectly recorded values. In case of MM treatment, a standardised collection of data regarding the specification of induction therapy is crucial, as there are many options (regimens) depending on the patient's condition and an important regulatory system in the respective country might apply.

After importing data into a specialised statistical software, quality of these data has to be checked. The analysis has to be based on predefined objectives (hypotheses). Usually not all patients are selected for analysis: some hypotheses could be related to a specific subset of patients who fulfil the **inclusion** or those who do not fulfil the exclusion criteria (e.g. include patients aged up to 70 and exclude patients with therapies conducted within

clinical trials). **Representativeness** of the selected sample compared to the whole population of patients with MM could be disturbed. The final interpretation of results has to consider the selection criteria; i.e. if only patients with ISS stage III were selected for the analysis, the results cannot be generalised to the whole population of patients with MM.

The selection criteria in a standard analysis of MM treatment usually specify the treatment regimen and lines of therapy (e.g. lenalidomide ¬+ dexamethasone in 2nd–4th line of therapy); as for all other characteristics (e.g. age, stage, performance status), the goal is to be representative of the whole population which may receive the given therapy. Finally, the subset analysis can describe the outcomes in specific groups of patients (e.g. ISS I II III). Nevertheless, many hypotheses can be generated above MM data, e.g. focused on patients with extramedullary disease, high risk cytogenetics or not yet mentioned progression from the asymptomatic stage of disease.

Descriptive summary of data is the first step of analysis. This phase itself reveals most troubles regarding the quality of data. Incorrectly recorded values (e.g. continuous parameters out of the defined range) or results directly reported as NA (not available) or "cannot be measured" used to be recoded as missing. In general, **there are three types of missing data** [72]:

1) Missing completely at random (MCAR): missing data are not related to any other patient's characteristics.
2) Missing not completely at random (MNCAR): missing data are associated with information not observed in data.
3) Missing at random (MAR): missing data are related to patient's characteristics observed in data.

How to handle missing values depends mainly on the character and the amount of missing data and objectives of the respective study. If the objective is the univariate analysis of individual variables, researchers usually report the number of missing values and do not use any imputation technique to complete gaps in the data. In case of multivariate analysis (assessing simultaneous effect of multiple variable), most statistical softwares automatically perform a complete case analysis, i.e. they only select patients with all the characteristics available. Nevertheless, if missing data is of type MAR or MCAR (which is assumed for the majority of missing registry data), **complete case analysis** causes a selection bias: the analysed sample is no longer a random sample from the population. There are many options how to handle the missing values; for a review concerned on sophisticated imputation techniques, see [72]. **Multiple imputation** belongs to the most recommended techniques outperforming other methods (complete case analysis, missing indicator method, overall mean imputation). Using this technique, missing values are predicted in multiple steps by other known patient's characteristics. Multiple imputation can be performed relatively simply, e.g. by using the freeware software R (www.r-project.org) or the package MICE (Multivariate Imputation by Chained Equation) [73]. In conclusion, missing data is a challenge for most of clinical research; imputation techniques should be considered mainly if multivariate analysis is included in the statistical analysis plan.

5.4 Descriptive analysis

Usually, the first step in any analysis is to summarise or to display gathered data in order to get some information about the analysed population of patients. Even the main outcomes of therapy can be described by descriptive statistics. Descriptive analysis is therefore often the first and sometimes also the final step, and researchers do not have to proceed with more advanced statistical methods.

This chapter will show several ways of describing patients' demographic and clinical characteristics depending on the type of data (i.e. if the data are categorical or continuous). In the analyses of MM patients, the following variables are often among the most reported ones:

- **Patient's characteristics:** sex, age, duration of follow up, performance status, comorbidities.
- **Clinical features:** ISS, R ISS, biochemistry (e.g. beta2 microglobulin, creatinine, haemoglobin), percentage of plasma cells in the bone marrow, type of monoclonal immunoglobulin, cytogenetic risk (changes in the genome of plasma cells), extramedullary disease, bone lesions.
- **Characteristics of therapy:** line of therapy, induction regimen, transplantation, reason for treatment termination, therapy in previous treatment lines.
- **Outcomes:** maximal and final response to therapy, time to event endpoints (e.g. the overall survival, as discussed in chapter 5.6).
- **Safety:** grade of adverse events.

5.4.1 Summary of categorical data

An exploratory analysis of qualitative (i.e. categorical) data (e.g. sex, stage of disease) is performed in two complementary ways – description by frequency tables and visualisation by bar plots or pie charts.

Frequency tables describe the representation of each category in a data set by displaying absolute frequencies, usually accompanied by relative frequencies (i.e. percentages) as well (see Figure 54). If the frequency of some categories is low, we should consider merging adjacent (or similar) categories together. Apart from frequency tables, categorical data can be also summarised by the modus, which is represented by the variable's category with the highest frequency. Furthermore, in case of categories that can be ordered (i.e. ordinal data), the median value can also be calculated.

Bar plots and pie **charts** are undoubtedly the most frequently used graphs for the visualisation of qualitative data. A bar plot is suitable for displaying both absolute and relative frequencies (the height of the bars corresponds to category frequencies). A pie chart is appropriate for the depiction of relative frequencies because the entire area of a pie chart represents 100% [74]. Examples of a bar plot and a pie chart are shown in Figure 54.

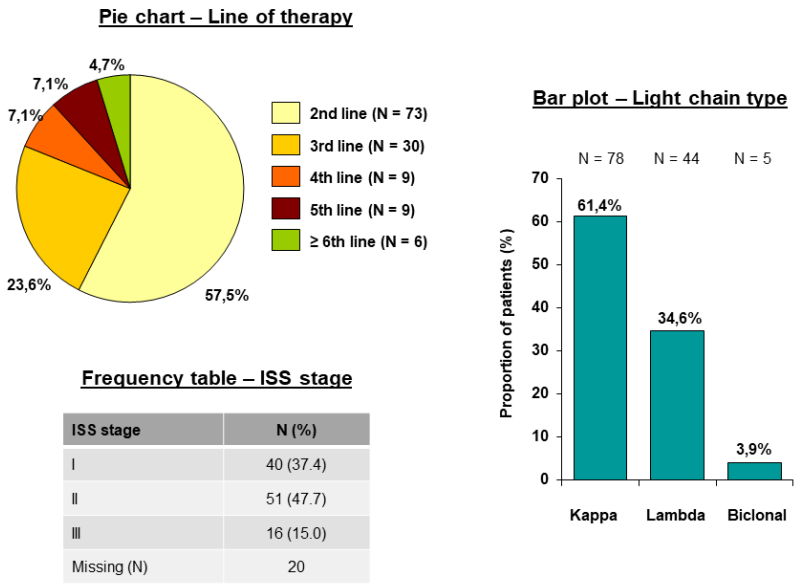


Figure 54 Visualisation of categorical data

Very often, it is useful to display two categorical variables on one graph to see their relationship. This can be done by drawing clustered or stacked bar plots (see Figure 55). The statistical evaluation of the relationship is described in the chapter Association between clinical parameters.

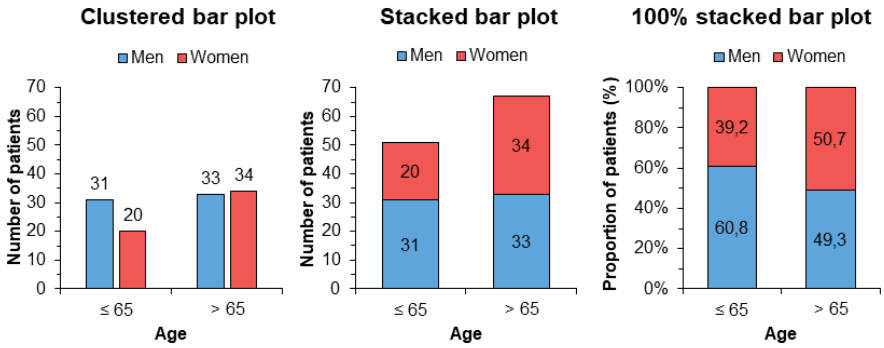


Figure 55 Visualisation of categorical data – bar charts for two variables

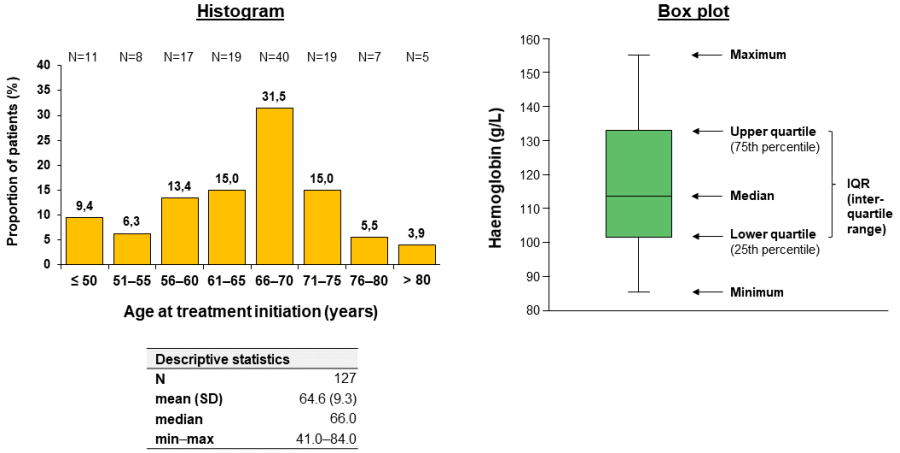
5.4.2 Summary of continuous data

Not only qualitative parameters, but also quantitative (both discrete and continuous) parameters often represent an important part of collected data; age, creatinine level or treatment duration can be given as typical examples. Summary statistics can be divided into measures of central tendency (e.g. mean or median) and **measures of dispersion** (e.g. standard deviation or variance, percentiles, range):

- **Mean** defines the centre value a...round which other values fluctuate. If data contains atypical values (i.e. outliers), the mean is often considerably biased.
- **Median** represents the value that separates data into two halves – the higher half and the lower half. Unlike the mean, the median is not affected by atypical values.
- **Variance** is defined as the average of squared deviation from the mean. Similar to the mean, it can be greatly influenced by outliers.
- **Standard deviation (SD)** is represented by the square root of the variance. The favourable feature of the standard deviation is that it is measured in the same units as the mean.
- **Standard error of mean (SE)** is defined as the standard deviation of the sample mean and characterises the accuracy of the mean. This statistic is calculated as the standard deviation divided by a number of patients; it is therefore significantly affected by the sample size.
- **Range** presents another measure of dispersion and is defined as the interval between the **minimum** and **maximum**. As well as variance and standard deviation, it is affected by atypical values.
- **Quantiles** are measures of dispersion, which are robust against outliers. For example, the 0.25 quantile (i.e. the 25th percentile, called lower quartile) divides the data set into the first quarter and the remaining three quarters.
- **Interquartile range (IQR)** is defined as the difference between the upper (i.e. 0.75 quantile) and lower quartiles (i.e. 0.25 quantile) and therefore covers 50% of observed values. IQR is sometimes reported as an interval between the lower and the upper quartile.

To sum up, if data are **symmetric**, then both the mean and the median are appropriate characteristics of central tendency and result in similar values. On the other hand, if data are asymmetric or if they contain outliers, then only the median is a good estimate of the frequency centre of the data because it is not affected by atypical values (unlike the mean). Thus, in practice, if the data do not follow a normal – or at least symmetrical – probability distribution, it is common to calculate the medians and percentiles (5th and 95th) rather than the means and standard deviations. In some cases, it may help to **transform** the data, so that we can obtain data with the required probability distribution (e.g. normalising log-normal data using the logarithmic transformation).

Descriptive statistics alone are usually sufficient, but it is often convenient to add some graphs which can show more than just numbers. For the visualisation of continuous or quantitative data, we can draw **histograms, box plots or scatter plots** [64]. Examples of a box plot and a histogram together with a table containing descriptive statistics are shown in Figure 56. Both summary statistics (e.g. range) and graphs (e.g. boxplot) can reveal outliers, extreme values or incorrect values. A scatter plot allows to study the relationship of two continuous variables. For more details about the correlation of two variables, see the chapter Association between clinical parameters.



Descriptive statistics	
N	127
mean (SD)	64.6 (9.3)
median	66.0
min–max	41.0–84.0

Figure 56 Visualisation of continuous data

5.5 Hypotheses testing

Summary description and the visualisation of data form an integral part of almost every statistical analysis. To compare certain groups of patients (e.g. according to different types of treatment), we can draw graphs (e.g. bar plots) or calculate descriptive statistics (e.g. means, medians, counts) for each group (for more details, see the chapters Summary of categorical data and Summary of continuous data). However, human eyes alone are often not sufficient to discover significant differences among compared subgroups and formal testing is therefore inevitable. **P-values** of statistical tests help us to make conclusions about the tested null hypotheses (e.g. there is no difference in mean value of haemoglobin in two groups of patients). P values describe the probability of getting the same or larger difference while the null hypothesis is true. A low p-value (less than the significance level usually set as 0.05) thus indicates rejection of the null hypothesis; in our example, we could conclude that there is a significant difference in haemoglobin level between the two groups. In case of high p values, the inference is that we failed to reject the null hypothesis. The p value is influenced to a large extent by the sample size and variability of data (with higher sample size and lower variability, the test provides lower p values even if the means in both groups did not change). The statistical report should therefore always include p values describing the **statistical significance** together with the summary statistics (mean, percentage) describing the **clinical significance** [74].

This chapter deals with the statistical comparisons of patient subgroups according to values of categorical parameters (such as performance status or ISS stage) or continuous parameters (such as age or creatinine level). It also includes ways to measure the association between two variables (such as clinical parameters). The last part of this chapter is focused on statistical measures of the accuracy of diagnostic tests. The **choice of a statistical test** depends on the data type and study settings. The decision-making process during the selection of appropriate statistical methods explained in this chapter is shown in Figure 57. A real example of statistical comparison of patients from RMG in two groups is shown in [75]; the manuscript focuses on the effectiveness and adverse events of intravenous and subcutaneous administration of bortezomib.

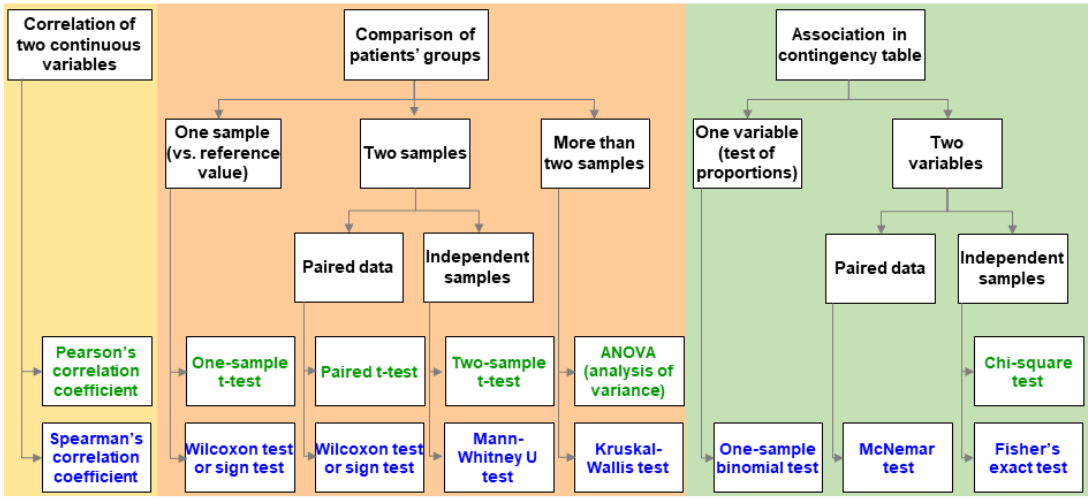


Figure 57 Decision-making process in the selection of conventional statistical tests

5.5.1 Comparison of patients' groups in continuous variable

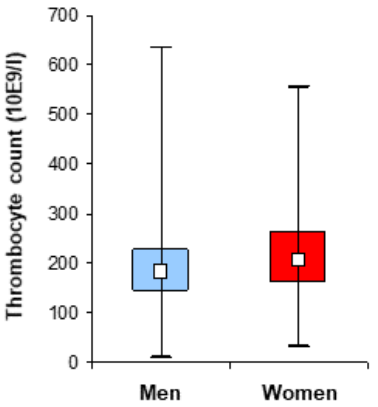
Continuous variables can be compared across different subgroups via the parametric t-test (in case of one or two samples; for a summary with practical examples, see [76]) and the analysis of variance (generalisation of t test for more than two samples) or via non-parametric alternatives: Mann Whitney U test [77] and Kruskal Wallis test [78].

Parametric tests should be only used if the data follow a **normal distribution** (in a simplified way, if there are no outliers and the data are symmetrical) and have the **same variance across all subgroups** of patients. The normality assumption can be checked with normality tests, such as the Shapiro Wilk or Kolmogorov Smirnov test [79], [80]. An alternative option is to use a graphical means of verification, such as histograms, boxplots or normal-quantile plots. The equality of variance in subgroups can be tested formally using tests such as the F-test, the Levene's test, the Bartlett's test or the Brown Forsythe test [81][82][83]. Plotting quantile-quantile plots, histograms or boxplots represents a graphical way to check the equality of variances. The violation of equal variances is not crucial if patient subgroups have the same sample sizes. The independent samples **t-test** is used for testing the null hypothesis that the means of two subgroups are equal. In case of inequality of variances or different sample sizes, **Welch's t-test** can be used instead (also known as unequal variance t-test) [84]. Testing the equality of means in more than two subgroups can be performed with the **analysis of variance (ANOVA)**. Testing through ANOVA tells us whether an overall difference exists among groups, but it does not say which particular groups differ. **Post-hoc tests** (e.g. Tukey's method or Scheffé's method) enable us to discover which subgroups pairs differ.

Non-parametric tests represent an alternative to parametric tests in cases of violation of assumptions (such as skewed data or presence of atypical values in data) or a small sample size. The advantage of non-parametric tests is that they do not assume a certain shape of the probability distribution and that they are not affected by outliers (they work with signs or ranks of values). On the other hand, they are said to be less powerful than the parametric tests [64], [85], [86]. The **Mann-Whitney U test** [77] is a non-para-

metric counterpart of the two sample t-test. Unlike the t-test, the Mann Whitney U test is focused on the comparison of the probability distribution of data across tested subgroups, not the means. **The Kruskal-Wallis test**, a non parametric alternative to ANOVA, is used for the evaluation of differences in probability distributions between more than two subgroups [78]. Both the t-test and ANOVA are quite robust against the violation of the equality of variances and of normality. However, if the data are substantially skewed, the sample size is small or outliers are present, then it is necessary to use non-parametric tests [64], [85].

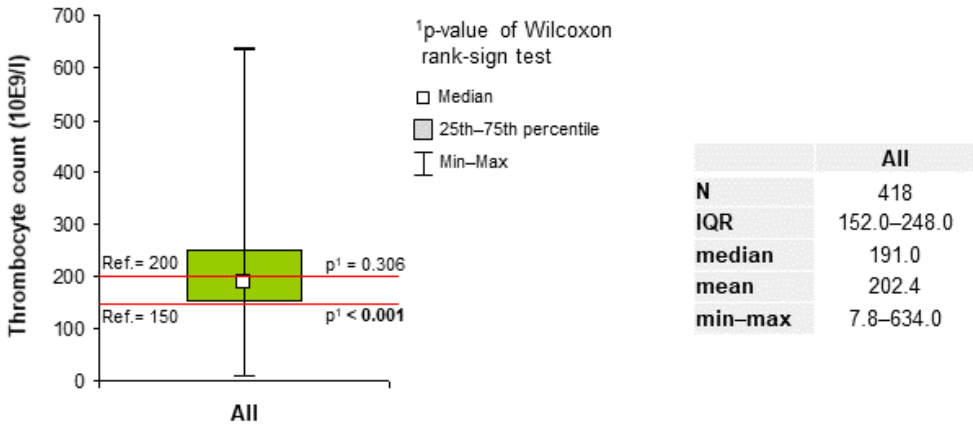
A comparison of men and women in terms of thrombocyte count is shown in Figure 58. The Mann Whitney U test revealed that there is a statistically significant difference ($p < 0.05$) between the two groups. Clinical significance was proved as well: the median of thrombocyte count is approximately 20 units higher in women than in men. The box plot shows that the distribution in respective groups is right-skewed (with less frequent high values): the boxes (corresponding to IQR) are not in the middle of the way between minimum and maximum, which points on an asymmetry of distribution. The value of mean is higher than the value of median, as explained in the chapter Summary of continuous data; mean is greatly influenced by outliers and asymmetry of distribution; therefore, it can be concluded that the mean does not describe the centre of distribution in this case (unlike the median).



	Men	Women
N	222	196
IQR	145–227	161–265.5
median	181.5	205.5
mean	191.7	214.5
min–max	7.8–634	31–555

Figure 58 Example of visualisation and statistical comparison of two samples in continuous parameters using a non-parametric test

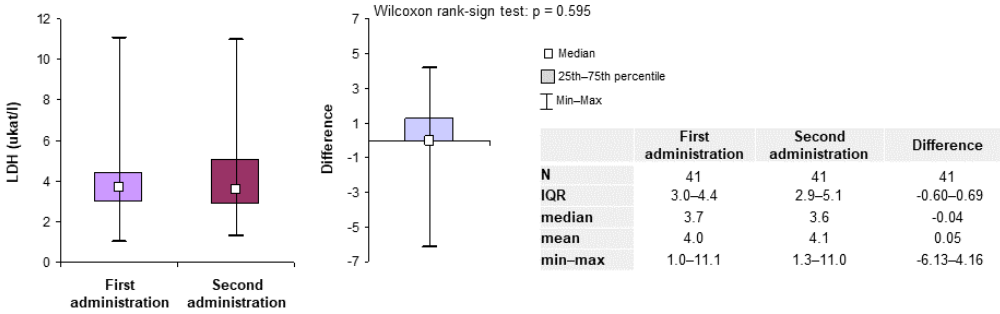
Apart from testing the differences of parameter values across different subgroups, one can evaluate if the mean or median value of the parameter of interest is smaller, larger or equal to a given reference value (e.g. testing the haemoglobin in MM patients against the reference value of haemoglobin level in a healthy population). This can be done using the one-sample test: either a parametric t-test or a non-parametric Wilcoxon signed-rank test [77]. In contrast to the one-sample t test, which compares the mean to a reference value, the Wilcoxon test works with ranks and compares the median to a reference value. Figure 59 shows the result of Wilcoxon test comparing the median of thrombocyte count with reference value 150 x 109/l and 200 x 109/l respectively. The median of thrombocyte count is significantly different from 150 ($p < 0.05$) but not significantly different from 200 ($p > 0.05$). As expected, the p-value decreases with a larger effect size (due to a difference in the observed median and value and the reference value).



	All
N	418
IQR	152.0–248.0
median	191.0
mean	202.4
min–max	7.8–634.0

Figure 59 Example of visualisation and statistical comparison of one sample to reference value (ref.)

Sometimes, it is also needed to test two related samples (or repeated measurements on a single sample). For example, we want to compare the values of the parameters of interest (such as the quantity of monoclonal immunoglobulin) before initiation and after termination of the therapy. Testing one sample over time can be performed with the parametric **paired t-test** or the non-parametric **Wilcoxon signed-rank test**. The paired test actually corresponds to the one sample test with the difference between the paired values as an input. The assumption (normal distribution) therefore has to be controlled on the difference of the paired values. One of real examples from RMG is shown in Figure **60** [87]. In this study, the effectiveness of the first and second administration of lenalidomide from one treatment centre was compared. Descriptive analysis included paired comparison of baseline characteristics. Figure **60** shows the difference in LDH (lactate dehydrogenase) at the first and second administration; the Wilcoxon test proved that there was no significant change in the level of LDH.



	First administration	Second administration	Difference
N	41	41	41
IQR	3.0–4.4	2.9–5.1	-0.60–0.69
median	3.7	3.6	-0.04
mean	4.0	4.1	0.05
min–max	1.0–11.1	1.3–11.0	-6.13–4.16

Figure 60 Example of visualisation and statistical comparison of paired samples

5.5.2 Association between clinical parameters

The magnitude of the association (correlation) between two continuous variables can be described using the parametric **Pearson's correlation coefficients** or the non-parametric **Spearman's correlation coefficient**. A correlation coefficient ranges from 1 (negative correlation) through 0 (no correlation) to +1 (positive correlation) [88]. The correlation coefficient (r) used to be reported together with the p value and the test compares the value of observed coefficient with 0. Association can be shown via a scatter plot (see

Figure 61). In the left graph, we can see a positive correlation, i.e. that with the increasing values of the beta2 microglobulin, the values of the creatinine increase as well. The right graph shows an opposite trend; as the values of haemoglobin go up, the values of the beta2 microgobulin tend to go down, which is a negative correlation. The trend of association could be derived only using the correlation coefficient without the need of plotting the scatterplot, but it is highly recommended to check the distribution as the coefficient could reach high value only thanks to extreme values (in case of parametric approach).

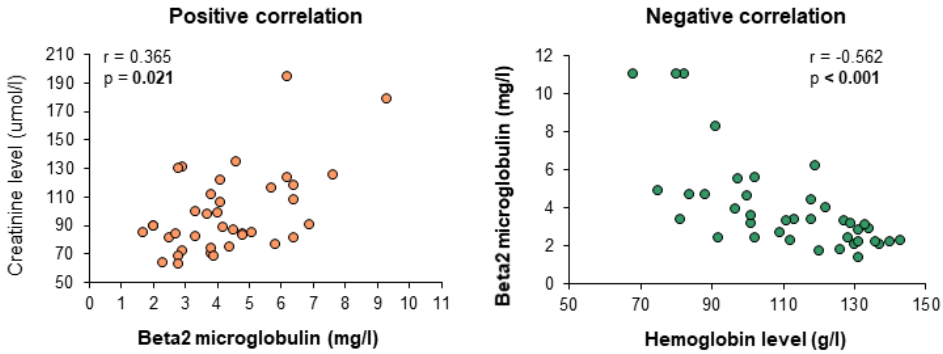


Figure 61 Example of visualisation and statistical assessment of correlation between two continuous variables

The association of categorical (i.e. qualitative) variables can be assessed using the Pearson's chi-squared test or the non-parametric alternative Fisher's exact test [89]. These tests are used when we want to decide whether the proportions of a variable's categories are different depending on the value of the other categorical variable. Let us imagine that we want to study whether the proportion of men and women is the same across patients with and without extramedullary disease (EMM and MM). Counts of patients in categories of two variables can be displayed using a **contingency table** (see Table 28). The assumption of the **Pearson's chi-squared test** is that at least 80% of cells in the contingency table have an expected count (assuming independence) of at least 5 and 100% of cells of at least 2. If this assumption is violated, the **Fisher's exact test** should be used instead. Table 28 shows that the analysed groups (EMM and MM) significantly differ in sex and age. The column percentages (to make the groups comparable, sum of percentages in the target groups should be 100) revealed that there is a higher proportion of male and young patients in the EMM group.

Table 28 Pearson's chi-square test as a test of independence in the contingency table

Baseline characteristics[1]	**EMM (N = 309)**	**MM (N = 2 092)**	**p**[2]
Sex			
Women	112 (36.2%)	1 014 (48.5%)	**< 0.001**
Men	197 (63.8%)	1 078 (51.5%)	
Age (years)			
≤ 65	172 (55.7%)	992 (48.3%)	**0.011**
66–75	99 (32.0%)	685 (33.4%)	
> 75	38 (12.3%)	375 (18.3%)	

[1] Described using N (column %)

[2] p-value of the chi-square test

If we wish to compare the counts in categories within related samples (e.g. two different risk stratification algorithms), we can apply the **McNemar's test** [90] (see an example in Table 29). The McNemar's test points to a statistically significant difference in the classification of patients into stages of ISS and R ISS. Using percentages, we can conclude that the classification differs in 27% of patients.

Table 29 McNemar's test in the evaluation of paired categorical data (blue = reclassification from a higher ISS to a lower R ISS stage; red = reclassification from a lower ISS to a higher R ISS stage)

Staging systems[1]	**Stage 1**	**ISS Stage 2**	**Stage 3**	**p**[2]
R-ISS				
Stage 1	97 (17.5%)	0 (0.0%)	0 (0.0%)	
Stage 2	57 (10.3%)	157 (28.3%)	95 (17.1%)	**< 0.001**
Stage 3	0 (0.0%)	0 (0.0%)	149 (26.8%)	

[1] Classification of patients described using N (total %)

[2] p-value of the McNemar's test

The relationship of two or more categorical variables (e.g. types of treatment, types of response) can be illustrated using the contingency tables. Statistical tests only say if there is a statistically significant relationship between the two categorical variables or not. However, these tests do not quantify this association. The **risk ratio** and **odds ratio** are measures of effect for assessing the strength of the relationship between two qualitative outcomes. They are usually applied to evaluate the association between a given exposure (e.g. presence/absence of a prognostic factor) and a specific outcome (e.g. presence/absence of a complete response) [91]. For example, we can be interested whether patients obtaining a new treatment have bigger odds for a complete remission than patients obtaining a conventional treatment. We can also study whether patients who are administered a novel drug have a higher risk of adverse events than patients who are administered an old drug.

Relative risk (also risk ratio, RR) can be calculated as the ratio of two probabilities – the probability of the presence of a specific outcome (e.g. adverse events) in a group with a given exposure (e.g. a new treatment combination) and without that exposure (e.g. a standard treatment combination). The **odds ratio** (OR) does not compare the probabilities in two groups: it evaluates the comparison of the odds instead. Odds are defined as the ratio of two probabilities – the probability of the occurrence of an outcome of interest (e.g. partial remission) and the probability of the non-occurrence of that outcome [92]. For a better understanding of the difference between the relative risk and the odds ratio, see Figure 62 with computational formulas. If the value of the RR is 1, then there is no difference in the risks between two groups. Values higher (or lower) than 1 mean that the outcome is more (or less) likely to occur in the group with the examined factor. Similarly, if the OR is equal to 1, there is no difference in the odds of the outcome between the compared groups. If the OR is higher (or lower) than one, then the exposure is associated with higher (or lower) odds of the event. Risk ratios and odds ratios provide similar results if the probability of an event is generally small. Nonetheless, if the probability of occurrence is high, then both measures result in different values. The relative risk should

not be used in cases of retrospective studies; odds ratios should be applied instead [92], [93]. A MM registry should be designed as a prospective study; therefore, both RR and OR could be used in the analysis. It is always more than appropriate to add confidence intervals (CI) to both statistics.

		Partial remission		
		Yes	No	Total
Treatment combination	New	a	b	a + b
	Standard	c	d	c + d
	Total	a + c	b + d	a + b+ c + d

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \qquad OR = \frac{\frac{a/(a+b)}{b/(a+b)}}{\frac{c/(c+d)}{d/(c+d)}} = \frac{\frac{a}{b}}{\frac{c}{d}}$$

Figure 62 Calculation of relative risk (RR) and odds ratio (OR)

Odds ratios are closely connected with the **logistic regression**. Regression coefficients in the logistic regression represent the log odds of the outcome (e.g. response to treatment) per unit increase in the predictor (e.g. age). In other words, the exponential function of the regression coefficient corresponds to the odds ratio associated with a one-unit increase in a predictor [93]. Table 30 shows the ORs from the assessment of association between baseline characteristics and the EMM status. Based on the results, we can conclude that the population of men has a 1.66 times (i.e. 66%) larger chance of having EMM when compared to the population of women. When evaluating the age, there is an opposite direction in OR (< 1), which reflects a lower chance of having EMM in older age. The odds ratios are often visualised via forest plots, i.e. using the same type of graph as hazard ratios in survival analysis (see the chapter Cox proportional hazard model). A p-value lower than the significance level reflects a statistically significantly different value of OR from 1 (where 1 means no association).

Table 30 Odds ratio as a help in interpretation of contingency table

Baseline characteristics[1]	EMM (N = 309)	MM (N = 2,092)	OR (95% CI)	p
Sex				
Women	112 (36.2%)	1,014 (48.5%)	reference	**< 0.001**
Men	197 (63.8%)	1,078 (51.5%)	1.66 (1.29–2.12)	
Age (years)				
≤ 65	172 (55.7%)	992 (48.3%)	reference	
66–75	99 (32.0%)	685 (33.4%)	0.87 (0.67–1.13)	0.294
> 75	38 (12.3%)	375 (18.3%)	0.59 (0.41–0.87)	**0.007**

[1] Described using N (column %)

5.5.3 Evaluation of diagnostic tests

Diagnostic tests (e.g. measurement of the quantity of monoclonal immunoglobulin in the serum) are frequently used in medicine in order to confirm the presence of disease in individuals suspected of having that condition or to discriminate between subgroups of patients. **Receiver operating characteristic (ROC)** analysis helps to evaluate the diagnostic ability of such medical tests, as well as classifiers or models. If the result of

a diagnostic test is dichotomous (e.g. positive or negative), then **sensitivity** and **specificity** are conventionally used as measures of the efficiency of the test. Sensitivity (true positive rate) is defined as the conditional probability that individuals with a given disease are correctly identified by the test as having that disease. In other words, it measures the ability of the test to recognise individuals with a given disease. Conversely, specificity (true negative rate) represents the percentage of healthy people who are correctly identified as not having the disease. Thus, specificity measures the ability of the test to recognise healthy individuals. Other useful indicators of discrimination ability in clinical practice include positive and negative predictive values. The **positive predictive value (PPV)** represents the probability that a person with a positive test actually has a given disease. Similarly, the **negative predictive value (NPV)** is defined as the probability of being healthy after obtaining negative test results. These two measures are influenced by the prior prevalence of disease [94]. The derivation of aforementioned characteristics is illustrated in Figure 63. Sensitivity, specificity, PPV and NPV can be reported as probabilities (with range 0–1) or expressed as percentages.

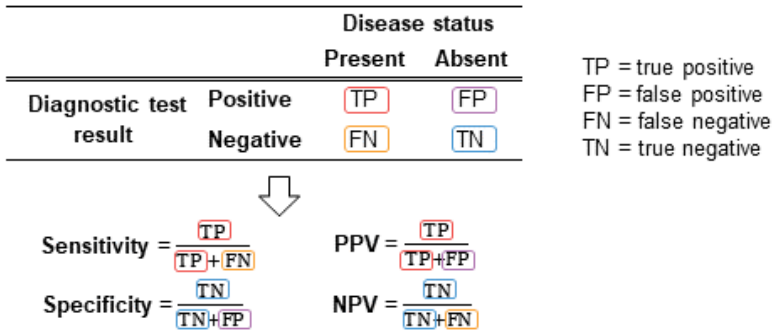


Figure 63 Comparison of diagnostic test results and real disease status through sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV)

However, the outcome of diagnostic tests does not have to be a binary result (positive/negative); it can also be a continuous/quantitative variable (e.g. albumin level). In that case, the previously mentioned indicators cannot be used because both sensitivity and specificity can be calculated across all possible values of the variable. Unless the cut-off value for the variable discriminating between healthy individuals and those with a given disease has already been found in previous research, we do not know which cut-off value of the continuous variable is the best. The ROC analysis represents an effective way to overcome this problem because it can help to find out whether the continuous variable is good for discriminating between healthy individuals and those with a given disease or not. Moreover, the ROC curve also enables us to identify the best cut-off point in terms of discriminating abilities. The ROC curve can be created by plotting sensitivity on the y-axis and one minus specificity on the x-axis (see Figure 64). The area **under the curve (AUC)** represents an effective summary measure of the accuracy of the diagnostic power of the test. It takes on values from 0 to 1, where 1 means the best discriminating ability and 0.5 denotes no discriminating ability. Accordingly, one can compare different tests through AUC. Figure 64B illustrates different shapes of the ROC curve with excellent, no and reversed discriminating ability. In the case of "no ability to discriminate between groups" (the grey diagonal line), the result of diagnostic test is basically no better than chance level (i.e. random). The closer to the upper left-hand corner the ROC curve lies, the better diagnostic ability it has [94].

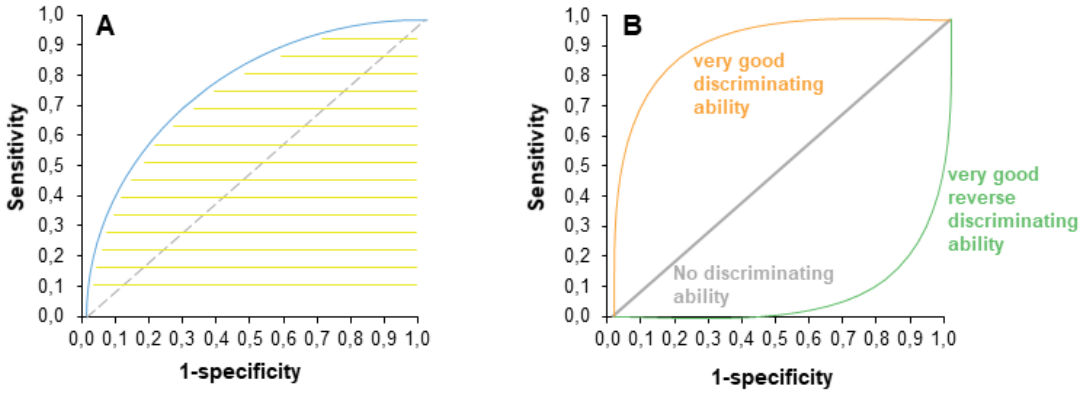


Figure 64 Example of ROC curve (blue line) with area under the curve (AUC; yellow) (A) and ROC curves with different discriminating ability (B)

There are several criteria to determine the best cut-off values. One of them is, for example, the **Youden index**, which searches for the maximum distance of the ROC curve from the diagonal line. That basically means that the sum of sensitivity and specificity is maximised [95]. Another commonly used method works with the squared distance between the upper left corner (x-axis=0, y-axis=1) and any point on the ROC curve and tries to minimise that distance [94].

ROC analysis clearly has a place when working with MM data. In case of MM patients, it is very helpful to categorise the patients into different groups according to their stage of disease, prognosis or values of prognostic factors, as treatment should be well suited for each patient. Therefore, the stratification of patients into various groups with different demands for treatment is much needed. For example, the ROC analysis was used in the evaluation of bortezomib base therapy in the first line not followed by autologous stem cell transplantation [96]. The ROC analysis pointed to a significant association of concentration of monoclonal immunoglobulin (M Ig) with treatment response PR (partial response) or better. Using the Youden index, the cut off 2.3 g/dl was identified (the cut off was associated with sensitivity and specificity value around 60%). The logistic regression finally revealed that patients with low levels of M Ig (< 2.3 g/dl) had 1.7 times higher chance of reaching PR or better when compared to patients with high levels of M Ig. The ROC analysis is therefore used very often in order to simplify the interpretation of regression modelling: the coefficients in the regression equation explain the change of outcome in association with unit increase of continuous variables, while the comparison of risk vs. reference group in case of binary variable (using the cut off from ROC) is much easier to interpret and more useful in practice.

In case of MM data, the endpoints used to be time dependent variables (e.g. the overall survival or the progression-free survival described in the chapter Survival analysis). In case of such endpoints, conventional ROC is not sufficient because it requires binary outcomes (e.g. the reaching of at least PR, as mentioned in the previous example). The prediction performance of the continuous variable can change over time; with increasing time from the assessment of the marker value, the predictive ability usually decreases. A special approach called **time-dependent ROC** should be adopted in these cases [97]. This technique assesses the predictive value of a continuous variable in predefined times of a patient's follow up (for a review, see [98]). In one study [99], time dependent ROC was performed to compare the survival prediction between two stratification algorithms; the one with a higher value of AUC was reported as the better one (i.e. having a better predictive ability).

5.6 Survival analysis

The main objective of the **survival analysis** is to analyse the expected duration of time until one or more events happen (e.g. death, progression or relapse). Time to event endpoints are the most important outcomes in the sense of **long-term assessment of treatment effectiveness**. Traditional statistical methods (such as t-test) cannot be used for modelling due to the presence of incomplete data (i.e. the event does not occur in all patients at the time of evaluation). Therefore, appropriate analytical techniques must be considered instead. This chapter provides a basic overview of statistical methods in survival analysis. The principles of the methods, their usage and several examples will be presented.

5.6.1 Kaplan-Meier method

As previously mentioned, the incompleteness of survival data is an important feature that requires the application of non-standard statistical methods. An observed event can be, for instance, death, progression/relapse, first treatment, next treatment, partial response and many others. The main goal is to find out how often and when such events occur since the beginning of the follow-up. Here are two examples of **scientific questions**: "Does the new treatment prolong the patients' survival time?" or "What proportion of patients will have disease progression at a certain time after treatment initiation?". The term "survival time" is used even if the observed event is not death.

Another crucial part of survival analysis is the definition from which time point the occurrence of events will be followed. The observation period can start, for example, at the time when a patient enters into the registry, at the time of diagnosis or at the time of treatment initiation. In MM data, survival analysis is mostly focused on the evaluation of the effect of therapy; therefore, the date of treatment initiation is usually the starting point. Even in this case (i.e. observation started at the date of treatment initiation), we can compare the survival according to the characteristics assessed in previous time periods (e.g. ISS at diagnosis, refractory status).

As mentioned before, the event of interest does not have to occur in all patients during the follow-up period; in this case, therefore, **censoring** is needed. The patients in which the event of interest has not occurred during the follow up period have to be censored in the survival analysis, i.e. considered as without that event at the time of last evaluation. This can be due to a patient's drop-out, loss to follow-up during the study or the end of the study (i.e. the closure of the database). These observations are called **right-censored** and are very common in real-world data. Patients with censoring still have their survival time (i.e. censoring time) included in the analysis and should not be removed from the analysis, as they are highly valuable. The stretch of time without experiencing the event of interest is itself very informative. In addition to 'right censoring', we can also come across the so-called '**interval censoring**', which means that we do not know the exact time of event occurrence; we only know that it happened between two known time points (e.g. visits), i.e. in an interval of time. Yet another type is '**left censoring**', which means that the event of interest has already occurred before the time of enrolment in the study but we do not know when [100]–[102]. For a better understanding of right censoring, see Figure 65.

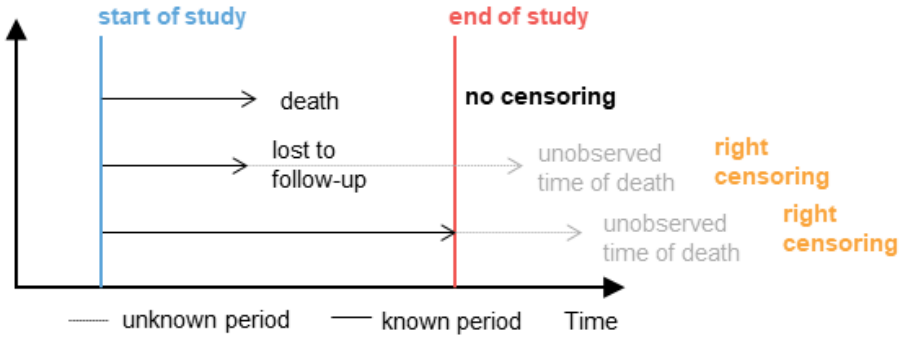


Figure 65 Illustration of right censoring

When analysing survival data, there are several important quantities to be defined. One of the crucial quantities is the survival function which describes the probability of survival (i.e. the non-occurrence of the event of interest) at given times. The survival function can be illustrated through a non ascending survival curve [102]. The real probability distribution of the survival data is usually not known. Therefore, non-parametric estimators represent frequently used means of estimating that do not require a specification of the distribution. The non-parametric Kaplan-Meier estimator is the most commonly used estimation technique of the survival function [103]. The Kaplan-Meier method is based on the probability of survival until a given time point, conditional to the probability of being alive in previous intervals. The assumptions of Kaplan-Meier analysis include the knowledge of exact time points at which events occur and the independence of censoring and event rate [104]. In MM data, the most frequently used intervals assessed using Kaplan Meier methodology are [71], [105]:

- **Overall survival (OS):** duration from the start of treatment (or date of diagnosis) to death from any cause.
- **Progression-free survival (PFS):** duration from the start of treatment to date of relapse/progression or death from any cause (whichever occurred first).
- **Time to progression (TTP):** duration from the start of treatment to date of relapse/progression or death related to MM (whichever occurred first).
- **Duration of response (DOR):** duration from the first date of documented PR to date of relapse/progression or death related to MM (whichever occurred first).
- **Time to next therapy (TNT):** duration from the start of treatment to start date of next therapy (time interval between two consecutive lines).
- **Time to treatment failure (TTF):** duration from the start of treatment to treatment discontinuation from any reason (progression, toxicity, etc.).
- **Event-free survival (EFS):** definition of EFS often corresponds to PFS, but EFS may include additional events, such as serious toxicity.

If the event of interest is not observed, patients are censored at the time of last evaluation (date of last contact with the patient). The interval thus corresponds to the duration of the patients' follow up. Definitions of treatment intervals should always correspond to the up to date guidelines of International Myeloma Working Group (IMWG) [71]. Examples of endpoints are illustrated on the timeline of an imaginary patient in Figure 66.

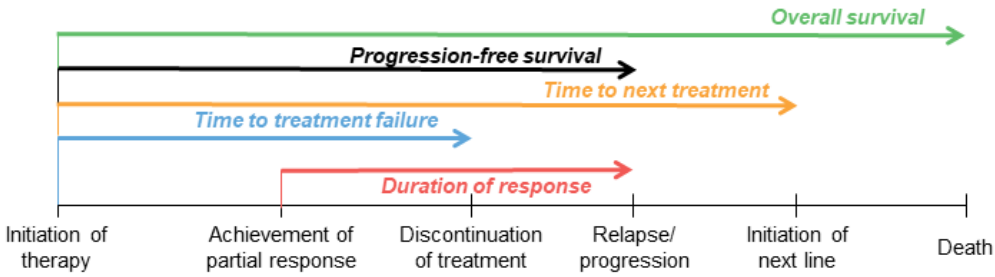


Figure 66 Timeline illustrating the set of various endpoints in a survival analysis

An example of the Kaplan-Meier curve is shown in Figure 67. The vertical axis of the curve represents the estimated probability of survival and takes on values between 0 and 1. Sometimes, it is also expressed as values from 0 to 100 (i.e. as the percentage of patients without the event of interest). The horizontal axis displays the time from starting the observation. The Kaplan-Meier curve begins at 1 on the vertical axis, representing the entire study population. When the events of interest occur, the line drops and the height of the step is dependent on the proportion of patients with that event at a given time point. Confidence intervals can be drawn around the curve to show the precision of estimated probabilities. Censored observations are very often indicated by notches or crosses in the Kaplan-Meier curve.

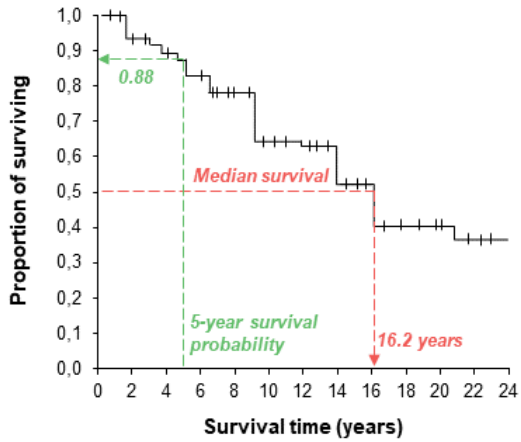


Figure 67 Example of Kaplan-Meier curve

The Kaplan-Meier curve is often accompanied by **median survival time**, which corresponds to the amount of time after which 50% of patients reach an event of interest. The estimated median survival time can be found by looking at the intersection of the survival curve and the horizontal line corresponding to the probability of survival equal to 0.5 (see the red dashed line in Figure 67). Sometimes, it may happen that median survival time is not reached, which occurs when the survival curve does not drop below 0.5. The Kaplan-Meier curve also enables us to obtain the **x-year survival rate** (i.e. the proportion of patients that survived for x years). See the green dashed line in Figure 67, which shows how to find the 5-year survival. It is also useful to provide, together with the Kaplan-Meier plot, the **number of patients** at risk at various time points in a table underneath the graph. Patients are 'at risk' if they are still alive and the event has not yet occurred at that time. It is important to keep in mind that survival curves are estimated less precisely if

only few subjects are at risk, usually near the end of a follow-up. Apart from summary of the survival time, it is also helpful to summarise the follow-up time through the median, since dissimilar follow-ups in compared groups could lead to bias [100]–[102].

Very often it is desirable to compare two or more Kaplan-Meier curves that can correspond, for example, to subgroups of patients with different treatment options. This can be done using the **log-rank test** [106]. The null hypothesis is that there is no difference in survival curves across different subgroups. The resulting p-value is usually compared with the level of significance $\alpha = 0.05$ (less frequently 0.001 or 0.1). If $p < 0.05$, the difference in survival between the subgroups of patients is considered as statistically significant; otherwise it is statistically insignificant [101].

Kaplan Meier curves represents the most frequent outcome in the analysis of MM data; accordingly, numerous studies focused on survival analysis of MM patients are available. An overall summary of survival of MM patients recorded in the RMG is reported in [107]. Figure 68A describes the overall survival of 4,738 patients who were diagnosed with MM. The median overall survival from the time of diagnosis was 5.7 years (95% CI: 5.4–6.0). From the table underneath the graph, we can find out how many patients were at risk every 5 years. During the time period, the number of patients at risk decreases; for example, the 10 year-long follow up was only reached in 358 patients, i.e. 92% of patients had either already died or had been followed for less than 10 years without the occurrence of event (censored). Together with the number of patients at risk, the probability of survival (i.e. proportion of surviving patients) was reported, again with a downward trend over time. At 10 years of follow up, the proportion of surviving patients was 32%. The censored data are not shown in the figure (it is an option of researcher if he/she wants to show the marks for censoring). Figure 68B describes the overall survival in respective lines of therapy. As expected, the OS decreases with higher line of therapy: the median OS was 60.5, 34.3, 22.6 and 13.8 months in 1st, 2nd, 3rd and ≥ 4th line of therapy, respectively. As the analysis was strictly descriptive, no statistical comparison was performed.

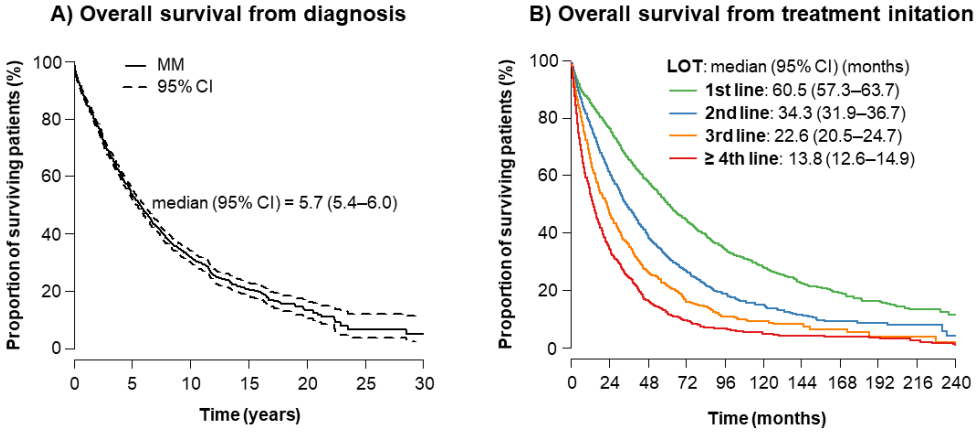


Figure 68 Overall survival from MM diagnosis (A) and from initiation of respective line of therapy (B)[107]

If the database includes data on asymptotic stages of MM, it is often in the researchers' interest to assess the time from diagnosis of asymptomatic disease to treatment requiring MM, as shown in several previous studies [107]–[110]. Figure 69 shows that the Kaplan Meier curve does not necessarily have to start at 1 (i.e. at 100%). It can be

plotted in a reverse way if the goal is to show that the number of events increases. In case of OS, an increasing duration of follow up is associated with a decreasing probability of survival. In case of evaluation of time to progression, an increasing duration of follow up is associated with an increasing risk of progression, and therefore the curve is plotted in a reverse way. An analysis of patients diagnosed with MGUS (monoclonal gammopathy of undetermined significance) revealed that the risk of progression in 5, 10 and 15 years of follow up was 7.6%, 16.5% and 26.5%, respectively [108].

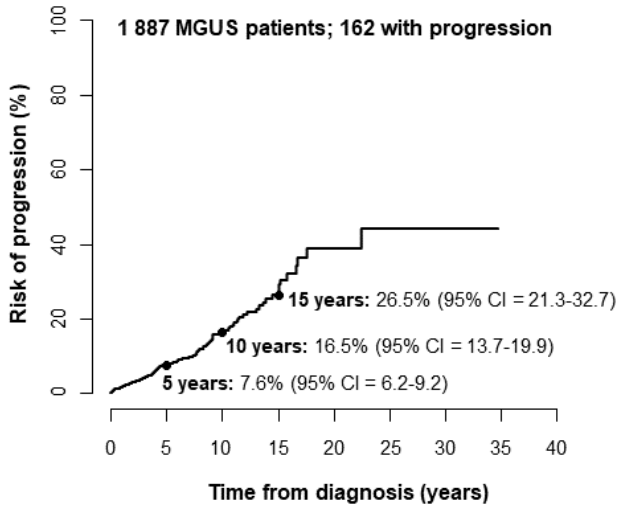


Figure 69 Time to progression from MGUS diagnosis [108]

A marked heterogeneity in survival is observed in MM patients, depending on many prognostic factors [111]. The Kaplan Meier methodology is therefore often used to verify the OS or other intervals per risk groups. An explanatory analysis of OS and PFS in patient subgroups (according age, ISS, etc.) can be found in [112]. As another example, a study [113] was focused on the assessment of performance of staging systems in MM patients recorded in the RMG. Figure 70 displays the OS from diagnosis according to the IMWG risk score [111] and the R ISS stratification [69], respectively. The Kaplan-Meier curves were compared using two sided log rank test. The figure clearly shows that the OS was significantly worse in higher stages for both risk stratification algorithms ($p < 0.05$).

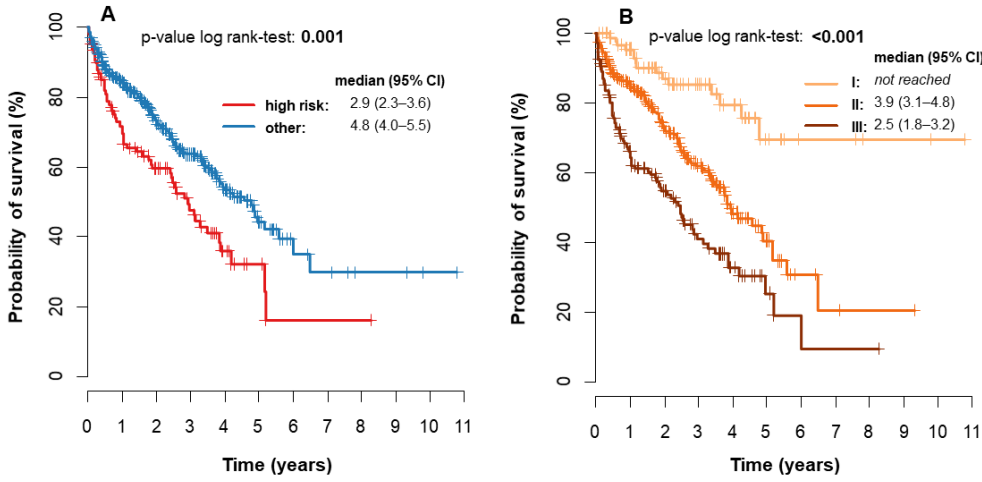


Figure 70 Overall survival from diagnosis according to the IMWG risk score (A) and the R-ISS risk stratification (B)

Registry data, reflecting real world evidence (as discussed in chapter Randomisation in clinical registries?), can serve as a comparator to the trial-based therapy. In study [114], data regarding daratumumab monotherapy conducted within clinical trials were compared to data of heavily pretreated multiple myeloma patients recorded in the RMG. The median observed PFS for daratumumab and the RMG cohort was 4.0 and 5.8 months, respectively. As another example, the study [96] was focused on the assessment of bortezomib based therapy in patients who initiated the first line of therapy and who were not eligible for transplantation. Survival analysis revealed that VMP (bortezomib + melphalan + prednisone) regimen outperforms the other regimens based on the evaluation of OS, PFS, TTP and DOR.

5.6.2 Cox proportional hazard model

The hazard function is another fundamental quantity in survival analysis. The hazard function does not represent probability; instead, it is an indicator of the risk of experiencing the event of interest at a given point in time [102]. The **Cox proportional hazards** model enables us to investigate the relationship of predictors (e.g. patient or treatment characteristics) and survival time through the hazard function [115], i.e. it can quantify the effect of predictors on patient survival. The advantage of the Cox model as opposed to parametric survival regression models is that it does not require the definition of a particular survival distribution (it is semi-parametric) [101]. The interpretation of the Cox model is done using the **hazards ratios (HRs)**. A hazard ratio is defined as the ratio of the predicted hazard function under two different values of a predictor variable (e.g. treated vs. untreated patients) while controlling for the effects of other covariates in the model (e.g. age, sex or thrombocyte count). If the hazard ratio is greater than 1, the occurrence of a given event is more likely. A hazard ratio less than 1 means an event is less likely to occur [102]. This allows us to distinguish between **risk factors** (increasing risk of event) and **protective factors** (decreasing risk of event) according to values of the hazard ratios.

Hazard ratios of predictors can be graphically represented via **forest plots** (see Figure 71). The figure is one of the outcomes of a real analysis from the RMG, focusing on patients treated with pomalidomide in relapse of MM; it was presented in an abstract in the EHA (European Hematology Association) conference in 2017 [116]. Association of predictors with the overall survival was assessed, HR > 1 points to higher risk of death for risk category compared to the reference in categorical variables and risk associated with unit increase of continuous variables. Specifically, for example patients with ISS III have almost three times (exactly 2.9 times) higher risk of death compared to patients with ISS I. Similarly, patients with extramedullary myeloma have almost three times higher risk of death compared to the patients without extramedullary mass. In terms of continuous variables, unit increase of beta2 miroglobulin was associated with 1.1 times higher risk of death. By contrast, a unit increase in thrombocyte count (platelet count) or haemoglobin level was associated with a lower risk of death (HR < 1). Using the **p-values**, we can interpret if the HR is significantly different from 1 (1 means no association with a given endpoint). In case of continuous variables, significant HRs close to 1 were observed; this is the consequence of the fact that model assesses a small change (the unit increase). If we used the time dependent ROC (as discussed in the chapter Evaluation of diagnostic tests) to recode the parameter into binary variables (e.g. haemoglobin < 110 g/l and ≥ 110 g/l) and use this as a predictor in the model, HR would reach higher values (compared to the continuous approach).

Forest plot are presented in a different way if the goal is to compare two treatment regimens, see for example [117]. The rows in the plot describe the selection of patients (e.g. women) and HR describes the effect of therapy in this subset.

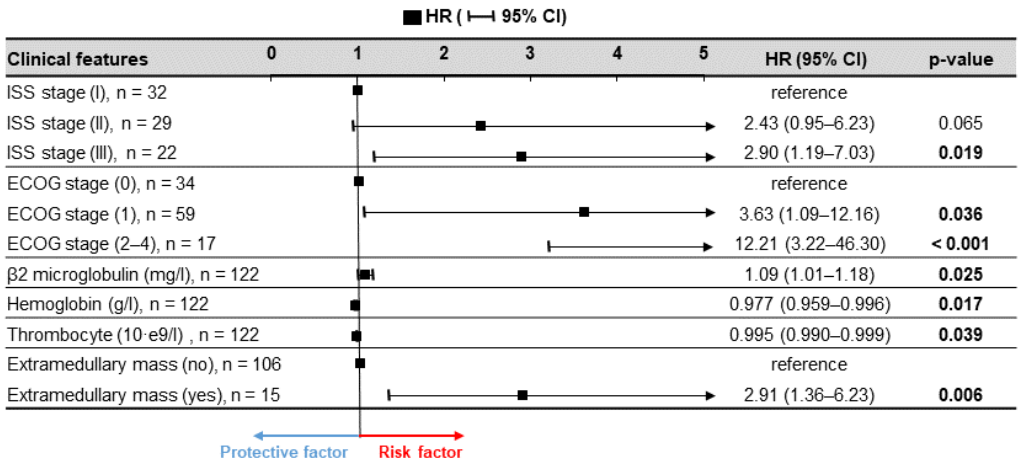


Figure 71 Example of forest plot - association of risk factors with OS in patients with pomalidomide in relapse of MM [116]

The Cox proportional hazards model assumes **proportional hazards** for different values of predictors. In other words, the effects of covariates should be constant over time. Therefore, it is important to check the assumption for all predictors tested. There are several graphical methods to do this. The simplest one is to look at whether survival curves move apart gradually and without intersections (see Figure 72). A small crossing at earlier points in time is usually fine. If the researcher wishes to examine the results more thoroughly, he/she can plot a cumulative hazards function or Schoenfeld residuals vs. time to identify patterns [103], [119].

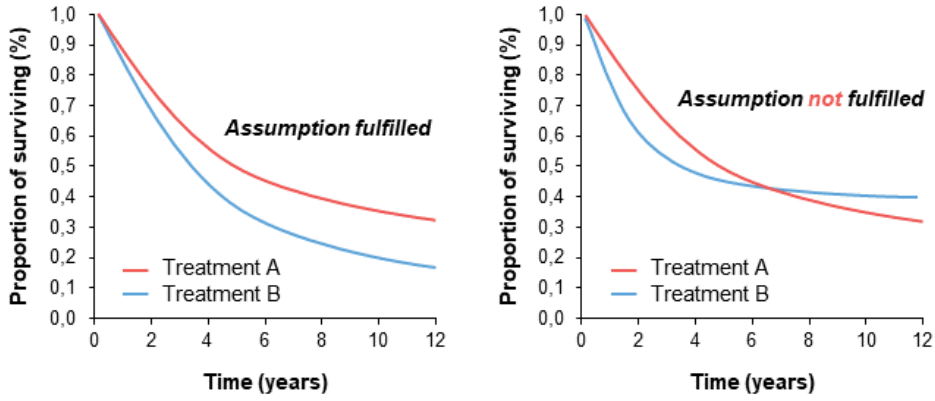


Figure 72 Checking the assumption of proportional hazards using Kaplan-Meier curves

The **univariate** Cox proportional hazards model can be used as a complement to the log-rank test. The log-rank test gives us information on whether there is a difference between given subgroups in terms of survival; the univariate Cox regression model quantifies the difference in the form of unadjusted hazard ratios (it quantifies the magnitude of association with endpoint similarly as OR in a contingency table does). Statistically significant variables in univariate analyses (e.g. when $p < 0.05$ or $p < 0.10$) are usually included in a **multivariate** Cox analysis as well, to obtain adjusted hazard ratios. This enables us to study prognostic factors independently. Using results from the univariate analysis, the CMG (Czech Myeloma Group) model was developed to classify patients into risk groups with different risks of progression from MGUS to MM [108]. Table 31 presents results of

both, univariate and multivariate analysis; even if adjusted by another covariate (from CMG model), all predictors significantly ($p < 0.1$) increase the risk of progression. Adjusting for confounding factors (as mentioned in the chapter Randomisation in clinical registries?) could be another option; for an example, see study [114].

Table 31 Univariate and multivariate analysis of predictors from CMG model associated with progression from MGUS to treatment requiring MM [108]

Risk model parameters	Univariate		Multivariate	
	HR (95% CI)	p	HR (95% CI)	p
CMG model				
Monoclonal protein in serum: ≥ 1.5 g/dL	3.65 (2.29–5.82)	< 0.001	2.83 (1.75-4.56)	< 0.001
Kappa/lambda ratio: < 0.26 or > 1.65	3.20 (1.85–5.52)	< 0.001	2.68 (1.53-4.69)	0.001
Bone marrow infiltration - cytology: > 5 %	2.57 (1.50–4.41)	0.001	1.60 (0.93-2.75)	0.092
Haemoglobin: < 12 g/dL	1.99 (1.15–3.46)	0.015	1.92 (1.11-3.32)	0.019
Immunoparesis: Any Ig lower	2.73 (1.72–4.34)	< 0.001	1.78 (1.09-2.88)	0.020

Immunoparesis: IgA < 0.61 g/l, IgG < 7.67 g/l, IgM < 0.5 g/l

5.7 Types of outputs from the registry

Data collected in the registry can be processed in several different ways, depending on the collected parameters and objectives of the registry (see the chapter Purpose and goals of the registry). However, the main goal of all data analyses should be more or less the same – to contribute to improvements in the quality of patient care. Only thanks to the collection of data from multiple treatment centres, the analysis of registry data reflects the representative results regarding the effectiveness of therapy and its adverse events. Typical types of outputs that can be obtained from the registry include periodical reports (monthly/quarterly/annual), articles or publications, conference papers and online charts.

Periodical reports may serve as a useful tool to obtain an overview about the size of the MM population, demographic and clinical characteristics of patients, types and effectiveness of treatment and adverse events. The reports are often in the form of slideshows (e.g. PPTX or PDF documents) which usually contain a summary of all data in the database altogether. However, it can be also useful to prepare periodical reports for each centre separately; these can serve as a means of data control (e.g. the centres can check if the numbers of patients in which a certain type of therapy was started during the last month correspond to true numbers, confirming that they did not forget to enter the data into the database etc.).

A detailed analysis (usually focused only on a specific treatment regimen, line of therapy etc.) is the most frequent output of MM data. MM represents a very complex disease and a high number of hypotheses can therefore be generated. The structure of a final report is therefore tailor made, depending on the original objectives. Results can be presented via **articles**, **publications** and **conference papers (abstract or poster**). Considering the fact that the results are presented to the broader public, it represents the most valuable form of outputs from the registry.

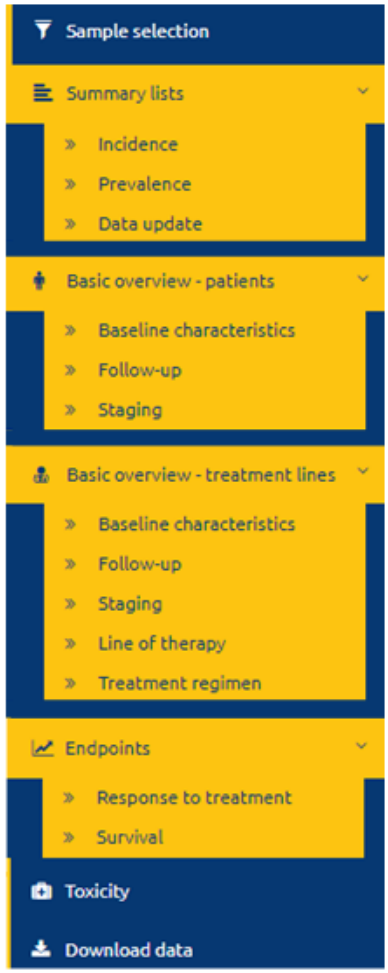


Figure 73 Parameter options for visualisation

Besides, another very helpful tool is represented by **online visualisations**, which can interactively draw graphs and tables according to the selected patients' groups and parameters. Obtaining necessary information about the patient population of interest (e.g. the size of the population or types of treatment) can be particularly interesting for clinicians or researchers before the initiation of analyses or studies. Additionally, leaders of individual treatment centres use the visualisation for a quick check of collected data. Visualisations are divided into a public part and an authenticated part, revealing only relevant information to a given user. The public part contains basic information about the number of patients diagnosed with MM and their basic characteristics. Figure 73 shows the options offered by the authenticated part of online visualisation of Czech data on MM patients. In the first step, the user needs to select groups of patients to be included in the analysis (example of selection criteria are presented in Figure 74). Figure 75 shows some of many possible online outputs that can be obtained by the online visualisation.

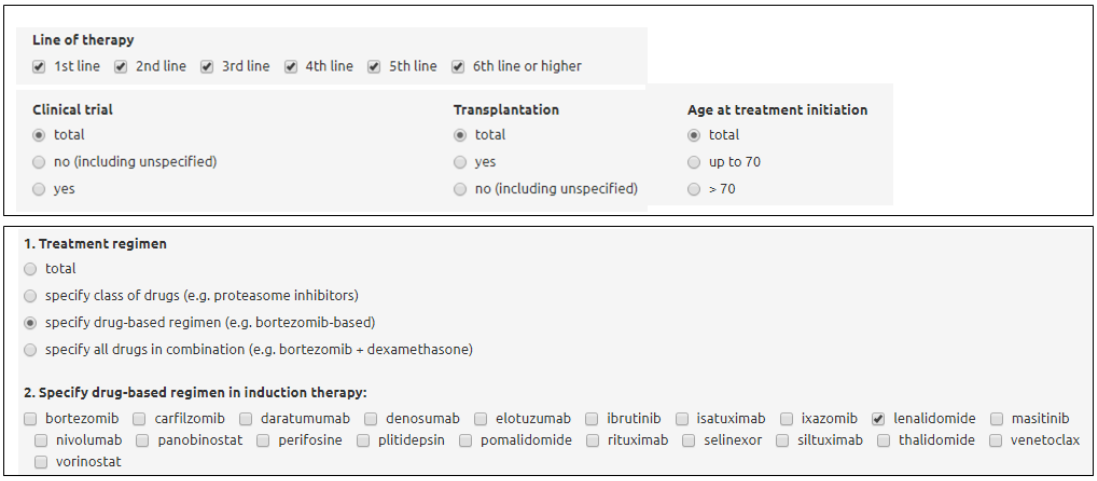


Figure 74 Selection of data in an online visualisation

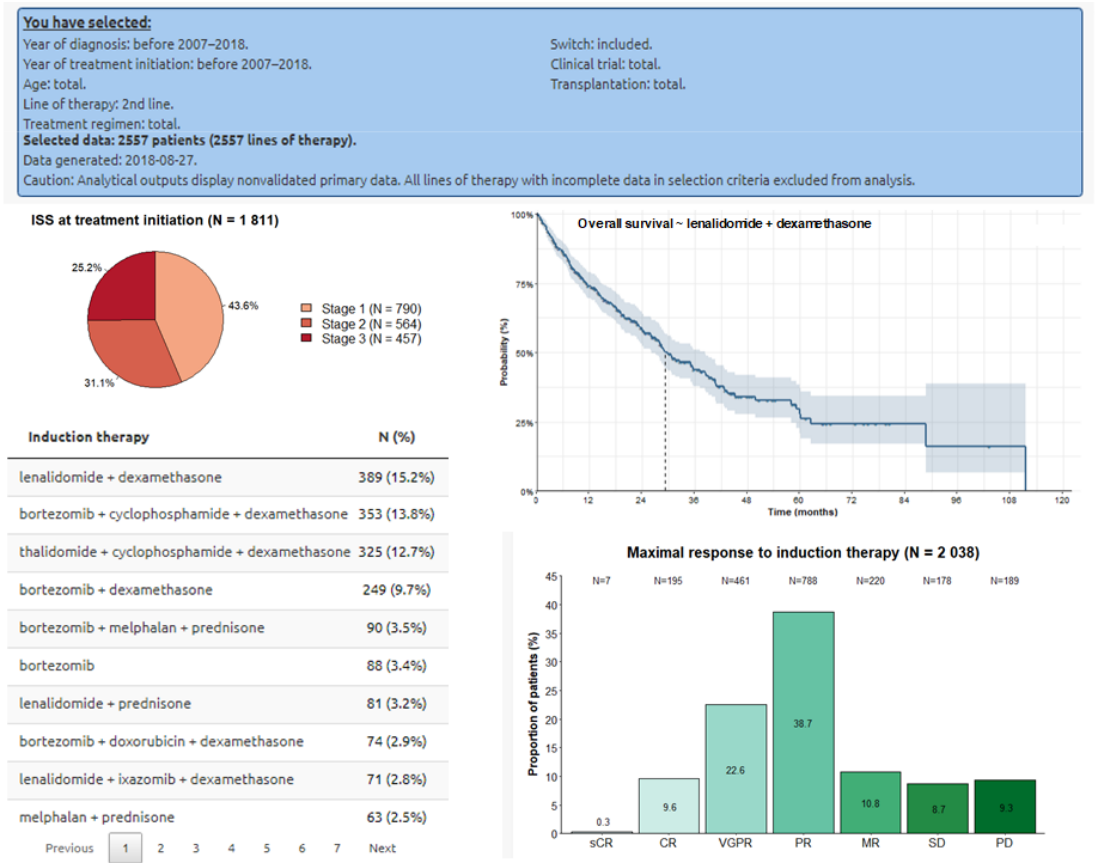


Figure 75 Example of outputs from an online visualisation

6. PHASE 5: WHAT COULD BE THE IMPACT OF THE RESULTS TO CLINICAL PRACTICE

6.1 Impact of the results to clinical practice

The impact of the results of MM data on a patient registry depends on the design, data scope, execution and quality of the registry itself – and the interpretation of results should reflect this. Different stakeholders benefit from the registry information in different ways. The major impacts for a clinician include real-world evidence of patient outcomes, clinical practice, safety, and clinical, comparative, and cost-effectiveness, and can serve qualified decision-making [33].

Describing the natural history of the disease

MM data may help to evaluate the natural history of the disease [33], its characteristics, changing management and outcomes with and/or without specific treatment. The natural history of some diseases is often not well described, may vary across different groups or geographic regions and may change after the introduction of new therapies.

Determining effectiveness

Disparities have been demonstrated between the results of clinical trials (efficacy) and the results in actual clinical practice (effectiveness) [33].

Efficacy in a clinical trial determines whether an intervention produces the expected result under ideal circumstances for a well-defined population and these results may not be generalisable to other populations or subgroups of interest (the underrepresentation of older patients in cancer clinical trials is a typical example [119]).

Effectiveness measures the degree of beneficial effects under "real-world" clinical settings. Registries may help to determine clinical effectiveness and/or cost-effectiveness in real-world clinical practice.

Registries may be useful for tracking effectiveness outcomes for a longer period, which is typically not feasible with clinical trials, and may help to answer new questions (e.g. secondary malignancies caused by MM treatment). Furthermore, if the registries collect cost data and effectiveness data they can be used in modelling and assessing cost-effectiveness.

Cost-effectiveness is a means to describe the comparative value of a healthcare product or service in terms of its ability to achieve a desired outcome for a given unit of resources [120].

There is increasing interest and investment in registries for comparative effectiveness research (CER) across a number of stakeholders. Patient registries are considered a core component of CER data.

Assessing safety or harm

Patient registries can measure and monitor safety and harm and serve as an active surveillance system for the occurrence of unexpected or harmful events [33].

Measuring quality of care

Patient registries may be useful in measuring the quality of care. They can help to discover differences between providers or patient populations, compare treatments outcomes, identify disparities in access to care and show opportunities for quality improvement.

Providing epidemiological data

If all patients from a specific region (city, country) are included, the registry can provide valuable epidemiological data such as incidence and prevalence. With the advent of new therapies, the overall survival for MM patients is longer. These data may help for health care planning and negotiation with payers.

The interpretation of results should include the validation of key results from the registry by an external data source, e.g. treatment consumption (resp. ratio of treatments) can be correlated with the amount of purchased drugs (or their ratio); treatment outcomes can be validated by results from other studies – both non-interventional and clinical trials. This method demonstrates the external validity of the results, i.e. that the distribution of patients in the registry corresponds to the reference (target) population that the researcher wishes to make claims about.

Example from practice

The comparison of the results of lenalidomide treatment in patients with refractory / relapsing MM is a good example of the importance of data collection from practice. According to the recommendation for the use of lenalidomide in this indication, the drug should be administered until the next disease relapse / progression. Until 2014, the treatment with lenalidomide in the Czech Republic and Slovakia was limited by specific “stop rules” of state regulators and insurance companies (namely, the maximal duration of therapy 10-12 cycles or the cumulative dose reaching 4200 mg). In our analysis, we compared the results achieved in real-world setting with results of clinical trials on similar patient populations. A comparison of these data clearly showed that the use of lenalidomide within clinical trials (= until disease progression) is associated with better outcomes for both progression-free survival (PFS) and overall survival (OS) than the treatment of a comparable group of patients in routine practice Figure 76 [121]. The results of our analysis subsequently contributed to the abolition of the above-mentioned “stop rules” for lenalidomide treatment in the Czech Republic and Slovakia.

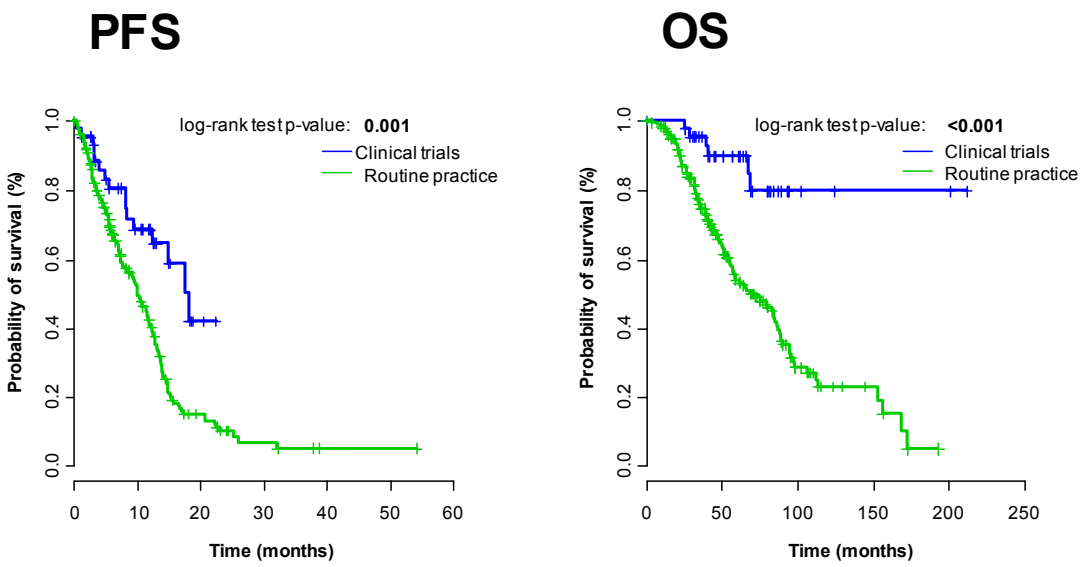


Figure 76 Comparison of PFS and OS from diagnosis for lenalidomide treatment that were achieved in real-world setting (green dots) with results of clinical trials on similar patient populations (blue dots).

7. LIST OF ABBREVIATIONS

AE	Adverse event
AL	amyloid light-chain
ANOVA	Analysis of variance
AUC	Area under the curve
CAPA	Corrective and preventive action
CAT	Central administration tool
CER	Comparative effectiveness research
CFR	Code of Federal Regulations
CI	Confidence intervals
CLADE-IS	Clinical Data Warehousing Information System
CMG	Czech Myeloma Group
CR	Complete remission
CRAB	Hypercalcaemia, renal failure, anaemia, bone lesions
CRF	Case Report Form
CRO	Contract Research Organisation
CRP	C-Reactive Protein
CT	Clinical trial
CT	Computed tomography
CTMS	Clinical trial management systems
DB	Database
DM	Data management
DMP	Data management plan
DNA	Deoxyribonucleic acid
DPO	Data protection officer
DSS	Durie-Salmon Staging
DVP	Data validation plan
EAV	Entity–Attribute–Value
EBM	Evidence-based medicine
EC	Ethics Committee
ECOG	Eastern Cooperative Oncology Group
ECOG PS	Performance status developed by Eastern Cooperative Oncology Group
eCRF	Electronic case report form
EDC	Electronic Data Capture
EFS	Event-free survival
EHA	European Hematology Association
EMA	European Medicines Agency
EMM	Extramedullary myeloma

ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
EU	European Union
FDA	Food and Drug Administration
FISH	Fluorescent in situ hybridisation
FLC	Free light chain
FMV	Fair market value
FN	False negative
FP	False positive
FPI	First patient in
GCP	Good cclinical practice
GDPR	General Data Protection Regulation
GPP	Good pharmacoepidemiology practice
HIPAA	Health Insurance Portability and Accountability Act
HLC	Heavy/light chain pair
HR	Hazard ratio
HTA	Health Technology Assessment
HW	Hardware
ICD	International Classification of Diseases
ICF	Informed consent form
ICT	Information and communication technology
IEC	International Electrotechnical Committee
ICH	The International Council for Harmonisation
IMiDs	Immunomodulatory imide drugs
IMWG	International Myeloma Working Group
IQR	Interquartile range
IRB	Institutional Review Board
ISMS	Information security management system
ISO	International Organization for Standardization
ISPE	International Society for Pharmacoepidemiology
ISS	International Staging System
IT	Information technology
ITSM	Information technology service management
IU	International Unit
LDCT	Low-dose computed tomography
LDH	Lactate dehydrogenase
LEC	Local Ethics Committee
MAR	Missing at random
MCAR	Missing completely at random
MEC	Multicentre Ethics Committee

MGUS	Monoclonal gammopathy of undetermined significance
MIBI	99Technitium sestamibi
MICE	Multivariate imputation by chained equation
M-Ig	Monoclonal immunoglobulin
MM	Multiple myeloma
MNCAR	Missing not completely at random
MRD	Minimal residual disease
MRI	Magnetic resonance imaging
N	Number
NA	Not available
NIS	Non-interventional study
NMR	Nuclear magnetic resonance
NPV	Negative predictive value
OR	Odds ratio
OS	Overall survival
PAES	Post--authorisation effectiveness study
PASS	Post-authorisation safety study
PBSCT	peripheral blood stem cell transplantation
PC	Plasma cells
PD	Progressive disease
PDCA	Plan, Do, Check, Act cycle
PET	Positron Emission Tomography
PFS	Progression-free survival
PM	Project manager
PMP	Project management plan
PPV	Positive predictive value
PR	Partial response
PROMs	Patient-reported outcome measures
PV	Pharmacovigilance
QA	Quality assurance
QMP	Quality management plan
QMS	Quality management system
QoL	Quality of life
RA	Regulatory Authority
RCTs	Randomised controlled trials
RISS	Revised International Staging System
R-ISS	Revised-international staging system
RMG	Registry of Monoclonal Gammopathies
ROC	Receiver operating characteristic

RR	Relative risk (risk ratio)
RWD	Real-world data
RWE	Real-world evidence
SaaS	Software as a service
SAE	Severe adverse event
SAP	Statistical analysis plan
SAS	Statistical analysis software
sCR	Stringent complete response
SCT	Stem cell transplantation
SD	Stable disease
SD	Standard deviation
SE	Standard error of mean
SFLC	Serum Free Light Chains
SMM	Smoldering MM
SOP	Standard operating procedures
SQL	Structured Query Language
SSL	Secure sockets layer
SW	Software
TFL	Tables, figures and listings
TN	True negative
TP	True positive
VGPR	Very good partial response
WHO	World Health Organization
WI	Working instructions
WM	Waldenström macroglobulinaemia

8. REFERENCES

1. A. Palumbo and K. Anderson, 'Multiple myeloma.', N. Engl. J. Med., vol. 364, no. 11, pp. 1046–1060, Mar. 2011.

2. C. Rollig, S. Knop, and M. Bornhauser, 'Multiple myeloma.', Lancet Lond. Engl., vol. 385, no. 9983, pp. 2197–2208, May 2015.

3. O. Landgren and B. M. Weiss, 'Patterns of monoclonal gammopathy of undetermined significance and multiple myeloma in various ethnic/racial groups: support for genetic factors in pathogenesis.', Leukemia, vol. 23, no. 10, pp. 1691–1697, Oct. 2009.

4. J.-L. Harousseau and M. Dreyling, 'Multiple myeloma: ESMO clinical recommendations for diagnosis, treatment and follow-up.', Ann. Oncol. Off. J. Eur. Soc. Med. Oncol., vol. 19 Suppl 2, pp. ii55-57, May 2008.

5. 'Ústav zdravotnických informací a statistiky ČR. Národní zdravotnický informační systém (NZIS), Národní onkologický registr (NOR) [online]', Aug-2018. [Online]. Available: http://www.uzis.cz/registry-nzis/nor.

6. O. Landgren et al., 'Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study.', Blood, vol. 113, no. 22, pp. 5412–5417, May 2009.

7. R. A. Kyle et al., 'A long-term study of prognosis in monoclonal gammopathy of undetermined significance.', N. Engl. J. Med., vol. 346, no. 8, pp. 564–569, Feb. 2002.

8. S. V. Rajkumar et al., 'International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma.', Lancet Oncol., vol. 15, no. 12, pp. e538-548, Nov. 2014.

9. B. M. Weiss, J. Abadie, P. Verma, R. S. Howard, and W. M. Kuehl, 'A monoclonal gammopathy precedes multiple myeloma in most patients.', Blood, vol. 113, no. 22, pp. 5418–5422, May 2009.

10. S. V. Rajkumar, G. Merlini, and J. F. San Miguel, 'Haematological cancer: Redefining myeloma.', Nat. Rev. Clin. Oncol., vol. 9, no. 9, pp. 494–496, Sep. 2012.

11. R. A. Kyle et al., 'Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma.', N. Engl. J. Med., vol. 356, no. 25, pp. 2582–2590, Jun. 2007.

12. 'Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group.', Br. J. Haematol., vol. 121, no. 5, pp. 749–757, Jun. 2003.

13. E. Kastritis et al., 'Extensive bone marrow infiltration and abnormal free light chain ratio identifies patients with asymptomatic myeloma at high risk for progression to symptomatic disease.', Leukemia, vol. 27, no. 4, pp. 947–953, Apr. 2013.

14. S. J. Russell and S. V. Rajkumar, 'Multiple myeloma and the road to personalised medicine.', Lancet Oncol., vol. 12, no. 7, pp. 617–619, Jul. 2011.

15. B. G. Durie and S. E. Salmon, 'A clinical staging system for multiple myeloma. Correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival.', Cancer, vol. 36, no. 3, pp. 842–854, Sep. 1975.

16. P. R. Greipp et al., 'International staging system for multiple myeloma.', J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol., vol. 23, no. 15, pp. 3412–3420, May 2005.

17. P. N. Hari et al., 'Is the International Staging System superior to the Durie-Salmon staging system? A comparison in multiple myeloma patients undergoing autologous transplant.', Leukemia, vol. 23, no. 8, pp. 1528–1534, Aug. 2009.

18. S. K. Kumar et al., 'Management of newly diagnosed symptomatic multiple myeloma: updated Mayo Stratification of Myeloma and Risk-Adapted Therapy (mSMART) consensus guidelines.', Mayo Clin. Proc., vol. 84, no. 12, pp. 1095–1110, Dec. 2009.

19. S. Kumar et al., 'Trisomies in multiple myeloma: impact on survival in patients with high-risk cytogenetics.', Blood, vol. 119, no. 9, pp. 2100–2105, Mar. 2012.

20. A. Palumbo et al., 'Revised International Staging System for Multiple Myeloma: A Report From International Myeloma Working Group.', J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol., vol. 33, no. 26, pp. 2863–2869, Sep. 2015.

21. J. G. Lohr et al., 'Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy.', Cancer Cell, vol. 25, no. 1, pp. 91–101, Jan. 2014.

22. N. Bolli et al., 'Heterogeneity of genomic evolution and mutational profiles in multiple myeloma.', Nat. Commun., vol. 5, p. 2997, 2014.

23. S. K. Kumar et al., 'Continued improvement in survival in multiple myeloma: changes in early mortality and outcomes in older patients.', Leukemia, vol. 28, no. 5, pp. 1122–1128, May 2014.

24. O. Landgren and S. Giralt, 'MRD-driven treatment paradigm for newly diagnosed transplant eligible multiple myeloma patients.', Bone Marrow Transplant., vol. 51, no. 7, pp. 913–914, Jul. 2016.

25. S. Mailankody et al., 'Minimal residual disease in multiple myeloma: bringing the bench to the bedside.', Nat. Rev. Clin. Oncol., vol. 12, no. 5, pp. 286–295, May 2015.

26. V. Magarotto et al., 'Triplet vs doublet lenalidomide-containing regimens for the treatment of elderly patients with newly diagnosed multiple myeloma.', Blood, vol. 127, no. 9, pp. 1102–1108, Mar. 2016.

27. S. Y. Kristinsson, W. F. Anderson, and O. Landgren, 'Improved long-term survival in multiple myeloma up to the age of 80 years.', Leukemia, vol. 28, no. 6, pp. 1346–1348, Jun. 2014.

28. S. Kumar et al., 'International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma.', Lancet Oncol., vol. 17, no. 8, pp. e328–e346, Aug. 2016.

29. O. Landgren, S. Devlin, M. Boulad, and S. Mailankody, 'Role of MRD status in relation to clinical outcomes in newly diagnosed multiple myeloma patients: a meta-analysis.', Bone Marrow Transplant., vol. 51, no. 12, pp. 1565–1568, Dec. 2016.

30. M. Roshal et al., 'MRD detection in multiple myeloma: comparison between MSKCC 10-color single-tube and EuroFlow 8-color 2-tube methods.', Blood Adv., vol. 1, no. 12, pp. 728–732, May 2017.

31. Y. Bai, A. Orfao, and C. S. Chim, 'Molecular detection of minimal residual disease in multiple myeloma.', Br. J. Haematol., vol. 181, no. 1, pp. 11–26, Apr. 2018.

32. M. Cavo et al., 'Role of (18)F-FDG PET/CT in the diagnosis and management of multiple myeloma and other plasma cell disorders: a consensus statement by the International Myeloma Working Group.', Lancet Oncol., vol. 18, no. 4, pp. e206–e217, Apr. 2017.

33. R. E. Gliklich, N. A. Dreyer, and M. B. Leavy, Eds., Registries for Evaluating Patient Outcomes: A User's Guide, 3rd ed. Rockville (MD): Agency for Healthcare Research and Quality (US), 2014.

34. International Society for Pharmacoepidemiology, 'Data Privacy, Medical Record Confidentiality, and Research in the Interest of Public', INTERNATIONAL SOCIETY FOR PHARMACOEPIDEMIOLOGY (ISPE), 1998. [Online]. Available: https://www.pharmacoepi.org/resources/privacy.cfm. [Accessed: 22-Oct-2017].

35. K. Breckenridge et al., 'How to routinely collect data on patient-reported outcome and experience measures in renal registries in Europe: an expert consensus meeting', Nephrol. Dial. Transplant., vol. 30, no. 10, pp. 1605–1614, Oct. 2015.

36. Kevin Lee, 'How to keep the project on budget in the clinical trial study'. PharmaSUG 2012, 2012.

37. Project Management Institute, Guide to the Project Management Body of Knowledge (PMBOK Guide), vol. 2013. Ingram International Inc.

38. K. Theobald, M. Capan, M. Herbold, S. Schinzel, and F. Hundt, 'Quality assurance in non-interventional studies', GMS Ger. Med. Sci., vol. 7, Nov. 2009.

39. 'Resources - Policies', INTERNATIONAL SOCIETY FOR PHARMACOEPIDEMIOLOGY (ISPE), 03-Oct-2017. [Online]. Available: https://www.pharmacoepi.org/resources/guidelines_08027.cfm. [Accessed: 03-Oct-2017].

40. 'The Importance of Risk Management in Clinical Trials', Pharmaceutical Training – Astra Nova, 22-Nov-2016. [Online]. Available: https://crotraining.co.uk/the-importance-of-risk-management-in-clinical-trials/. [Accessed: 06-Oct-2017].

41. Emily Haglund, 'Risk Management in Clinical Research: Process and Application'. 2015.

42. Valérie JOURNOT, 'Guideline on risk management for clinical research Version 1.0'. 2015.

43. M. Zaletel, M. Kralj, M. Magajne, and P. Doupi, 'Methodological Guidelines and Recommendations for Efficient and Rationale Governance of Patient RegistriesMetka Zalatel', Eur. J. Public Health, vol. 25, no. suppl_3, Oct. 2015.

44. Rodwell C., and Aymé S., '2014 Report on the State of the Art of Rare Disease Activities in Europe', 2014.

45. S. de Groot et al., 'Balancing the Optimal and the Feasible: A Practical Guide for Setting Up Patient Registries for the Collection of Real-World Data for Health Care Decision Making Based on Dutch Experiences', Value Health, vol. 20, no. 4, pp. 627–636, Apr. 2017.

46. V. Maisnar et al., 'RMG - Registr Monoklonalnich Gamapatii', Onkologie, vol. 5, no. 3, pp. 138–140, 2011.

47. L. Brozova et al., 'Czech Registry of Monoclonal Gammopathies - Technical Solution, Data Collection and Visualisation.', Klin. Onkol. Cas. Ceske Slov. Onkol. Spolecnosti, vol. 30, no. Supplementum2, pp. 43–50, Summer 2017.

48. B. Harper, 'Meshing EDC with CTMS', BioITWorld, 2007. [Online]. Available: http://www.bio-itworld.com/issues/2007/feb/cover-story/. [Accessed: 31-Oct-2017].

49. Gazali, S. Kaur, and I. Singh, 'Artificial intelligence based clinical data management systems: A review', Inform. Med. Unlocked, vol. 9, no. Supplement C, pp. 219–229, Jan. 2017.

50. FDA, 'Title 21 CFR Part 11', Code of Federal Regulations. [Online]. Available: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=11.

51. J. D. Franklin, A. Guidry, and J. F. Brinkley, 'A partnership approach for Electronic Data Capture in small-scale clinical trials', J. Biomed. Inform., vol. 44 Suppl 1, pp. S103-108, Dec. 2011.

52. R. E. Kibbelaar et al., 'Bridging the gap between the randomised clinical trial world and the real world by combination of population-based registry and electronic health record data: A case study in haemato-oncology', Eur. J. Cancer Oxf. Engl. 1990, vol. 86, pp. 178–185, Oct. 2017.

53. ISO/IEC, 'ISO/IEC 27000 family - Information security management systems', International Organization for Standardization. .

54. J. Dutton, 'How ISO 27001 can help to achieve GDPR compliance', IT Governance, Aug-2017. [Online]. Available: https://www.itgovernance.co.uk/blog/how-iso-27001-can-help-to-achieve-gdpr-compliance/.

55. B.Ewatt, 'World Federation of Hemophilia: Guide to developing a national patient registry'. World Federation of Hemophilia, 2005.

56. ['REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL'. THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION, 27-Apr-2016.

57. '« Disease Registries in Europe », Orphanet Report Series, Rare Diseases collection, January 2011'. 2011.

58. 'Patient Registries Workshop, 28 October 2016'. European Medicines Agency, 2017.

59. Y. R. Rubinstein et al., 'Creating a Global Rare Disease Patient Registry linked to a Rare Diseases Biorepository Database: Rare Disease - HUB (RD – HUB)', Contemp. Clin. Trials, vol. 31, no. 5, pp. 394–404, Sep. 2010.

60. 'EN ISO 9001:2015 Quality management system'. European Committee for Standardization, Sep-2015.

61. International Council for Harmonization ot Technical Requirements for Pharmaceuticals Use, 'Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2). Version 4.' 09-Nov-2016.

62. EMEA, 'Statistical Principles for Clinical Trials (E9), in The International Council for Harmonisation (ICH) of Technical Requirements for Pharmaceuticals for Human Use'. 1998.

63. K. F. Schulz, D. G. Altman, and D. Moher, 'CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials.', BMC Med., vol. 8, p. 18, Mar. 2010.

64. D. M. Windish and M. Diener-West, 'REVIEW: A Clinician-Educator's Roadmap to Choosing and Interpreting Statistical Tests', J. Gen. Intern. Med., vol. 21, no. 6, pp. 656–660, 2006.

65. E. A. Stuart, 'Matching Methods for Causal Inference: A Review and a Look Forward', Stat. Sci., vol. 25, no. 1, pp. 1–21, 2010.

66. P. R. Rosenbaum and D. B. Rubin, 'The central role of the propensity score in observational studies for causal effects', Biometrika, vol. 70, no. 1, pp. 41–55, 1983.

67. J. E. Signorovitch et al., 'Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research', Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res., vol. 15, no. 6, pp. 940–947, 2012.

68. P. R. Greipp et al., 'International staging system for multiple myeloma', J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol., vol. 23, no. 15, pp. 3412–3420, 2005.

69. A. Palumbo et al., 'Revised International Staging System for Multiple Myeloma: A Report From International Myeloma Working Group', J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol., vol. 33, no. 26, pp. 2863–2869, 2015.

70. M. M. Oken et al., 'Toxicity and response criteria of the Eastern Cooperative Oncology Group', Am. J. Clin. Oncol., vol. 5, no. 6, pp. 649–655, 1982.

71. S. V. Rajkumar et al., 'Consensus recommendations for the uniform reporting of clinical trials: report of the International Myeloma Workshop Consensus Panel 1', Blood, vol. 117, no. 18, pp. 4691–4695, 2011.

72. A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, 'Review: a gentle introduction to imputation of missing values', J. Clin. Epidemiol., vol. 59, no. 10, pp. 1087–1091, 2006.

73. S. Van Buuren, 'mice: Multivariate Imputation by Chained Equations in R', Journal of Statistical Software, vol. 45, no. 3, 2011.

74. T. Pavlík and L. Dušek, Biostatistika. Brno: Akademické nakladatelství CERM, s.r.o., 2012.

75. J. Minarik et al., 'Subcutaneous Bortezomib in Multiple Myeloma Patients Induces Similar Therapeutic Response Rates as Intravenous Application But It Does Not Reduce the Incidence of Peripheral Neuropathy', PLoS ONE, vol. 10, no. 4, 2015.

76. T. K. Kim, 'T test as a parametric statistic', Korean J. Anesthesiol., vol. 68, no. 6, pp. 540–546, 2015.

77. F. Wilcoxon, 'Individual Comparisons by Ranking Methods', Biom. Bull., vol. 1, no. 6, pp. 80–83, 1945.

78. W. H. Kruskal and W. A. Wallis, 'Use of Ranks in One-Criterion Variance Analysis', J. Am. Stat. Assoc., vol. 47, no. 260, pp. 583–621, 1952.

79. S. S. Shapiro and M. B. Wilk, 'An Analysis of Variance Test for Normality (Complete Samples)', Biometrika, vol. 52, no. 3/4, pp. 591–611, 1965.

80. F. J. Massey, 'The Kolmogorov-Smirnov Test for Goodness of Fit', J. Am. Stat. Assoc., vol. 46, no. 253, pp. 68–78, 1951.

81. H. Levene, 'Robust tests for equality of variances', in Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I., Stanford University Press, 1960, pp. 278–292.

82. G. W. Snedecor and W. G. Cochran, Statistical Methods, Eight. Ames, Iowa: Iowa State University Press, 1989.

83. M. B. Brown and A. B. Forsythe, 'Robust Tests for the Equality of Variances', J. Am. Stat. Assoc., vol. 69, no. 346, pp. 364–367, 1974.

84. B. L. Welch, 'The generalisation of student's problems when several different population variances are involved', Biometrika, vol. 34, no. 1–2, pp. 28–35, 1947.

85. J. W. Song, A. Haas, and K. C. Chung, 'Applications of Statistical Tests in Hand Surgery', J. Hand Surg., vol. 34, no. 10, pp. 1872–1881, 2009.

86. F. S. Nahm, 'Nonparametric statistical tests for the continuous data: the basic concept and the practical use', Korean J. Anesthesiol., vol. 69, no. 1, pp. 8–14, 2016.

87. M. Stork et al., 'Retreatment with lenalidomide is an effective option in heavily pretreated refractory multiple myeloma patients', Neoplasma, vol. 65, no. 4, pp. 585–591, 2018.

88. M. Mukaka, 'A guide to appropriate use of Correlation coefficient in medical research', Malawi Med. J. J. Med. Assoc. Malawi, vol. 24, no. 3, pp. 69–71, 2012.

89. M. L. McHugh, 'The Chi-square test of independence', Biochem. Medica, vol. 23, no. 2, pp. 143–149, 2013.

90. Q. McNemar, 'Note on the sampling error of the difference between correlated proportions or percentages', Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.

91. G. Tripepi, C. Torino, G. D'Arrigo, D. Bolignano, F. Provenzano, and C. Zoccali, 'An overview of standard statistical methods for assessing exposure-outcome link in survival analysis (Part I): basic concepts', Aging Clin. Exp. Res., vol. 24, no. 2, pp. 109–112, 2012.

92. J. Stare and D. Maucort-Boulch, 'Odds Ratio, Hazard Ratio and Relative Risk', Metodološki Zv., vol. 13, no. 1, pp. 59–67, 2016.

93. M. Szumilas, 'Explaining Odds Ratios', J. Can. Acad. Child Adolesc. Psychiatry, vol. 19, no. 3, pp. 227–229, 2010.

94. K. Hajian-Tilaki, 'Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation', Casp. J. Intern. Med., vol. 4, no. 2, pp. 627–635, 2013.

95. W. J. Youden, 'Index for rating diagnostic tests', Cancer, vol. 3, no. 1, pp. 32–35, 1950.

96. V. Sandecka et al., 'Bortezomib-Based Induction Treatment for Patients with Multiple Myeloma Ineligible for Autologous Stem Cell Transplantation: A Retrospective Multicenter Study in Czech Republic', Blood, vol. 130, no. Suppl 1, pp. 5435–5435, 2017.

97. P. J. Heagerty, T. Lumley, and M. S. Pepe, 'Time-dependent ROC curves for censored survival data and a diagnostic marker', Biometrics, vol. 56, no. 2, pp. 337–344, 2000.

98. A. N. Kamarudin, T. Cox, and R. Kolamunnage-Dona, 'Time-dependent ROC curve analysis in medical research: current methods and applications', BMC Med. Res. Methodol., vol. 17, no. 1, p. 53, Apr. 2017.

99. M. Zhou et al., 'Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma', J. Exp. Clin. Cancer Res., vol. 34, p. 102, 2015.

100. S. May and B. McKnight, 'Graphics and statistics for cardiology: survival analysis', Heart Br. Card. Soc., vol. 103, no. 5, pp. 335–340, 2017.

101. J. Kim, 'Survival analysis', Pediatr. Rev. Am. Acad. Pediatr., vol. 33, no. 4, pp. 172–174, 2012.

102. B. George, S. Seals, and I. Aban, 'Survival analysis and regression models', J. Nucl. Cardiol. Off. Publ. Am. Soc. Nucl. Cardiol., vol. 21, no. 4, pp. 686–694, 2014.

103. E. L. Kaplan and P. Meier, 'Nonparametric Estimation from Incomplete Observations', J. Am. Stat. Assoc., vol. 53, no. 282, pp. 457–481, 1958.

104. S. A. ElHafeez et al., 'An overview on standard statistical methods for assessing exposure-outcome link in survival analysis (Part II): the Kaplan-Meier analysis and the Cox regression method', Aging Clin. Exp. Res., vol. 24, no. 3, pp. 203–206, 2012.

105. R. Pazdur, 'Endpoints for assessing drug activity in clinical trials', The Oncologist, vol. 13 Suppl 2, pp. 19–21, 2008.

106. N. Mantel, 'Evaluation of survival data and two new rank order statistics arising in its consideration', Cancer Chemother. Rep., vol. 50, no. 3, pp. 163–170, 1966.

107. L. Brozova et al., 'Asymptomatic and Treatment-requiring Multiple Myeloma - Data from the Czech Registry of Monoclonal Gammopathies', Klin. Onkol. Cas. Ceske Slov. Onkol. Spolecnosti, vol. 30, no. Supplementum2, pp. 51–59, 2017.

108. V. Sandecká et al., 'A first Czech analysis of 1887 cases with monoclonal gammopathy of undetermined significance', Eur. J. Haematol., vol. 99, no. 1, pp. 80–90, 2017.

109. R. A. Kyle et al., 'Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma', N. Engl. J. Med., vol. 356, no. 25, pp. 2582–2590, 2007.

110. R. A. Kyle et al., 'A long-term study of prognosis in monoclonal gammopathy of undetermined significance', N. Engl. J. Med., vol. 346, no. 8, pp. 564–569, 2002.

111. W. J. Chng et al., 'IMWG consensus on risk stratification in multiple myeloma', Leukemia, vol. 28, no. 2, pp. 269–277, 2014.

112. R. Hájek et al., 'Real-world Outcomes of Multiple Myeloma: Retrospective Analysis of the Czech Registry of Monoclonal Gammopathies', Clin. Lymphoma Myeloma Leuk., vol. 18, no. 6, pp. 219–240, 2018.

113. J. Radocha et al., 'Validation of multiple myeloma risk stratification indices in routine clinical practice: Analysis of data from the Czech Myeloma Group Registry of Monoclonal Gammopathies', Cancer Med., vol. 7, no. 8, pp. 4132–4145, 2018.

114. T. Jelínek et al., 'Adjusted comparison of daratumumab monotherapy versus real-world historical control data from the Czech Republic in heavily pretreated and highly refractory multiple myeloma patients', Curr. Med. Res. Opin., vol. 34, no. 5, pp. 775–783, 2018.

115. D. R. Cox, 'Regression Models and Life-Tables', J. R. Stat. Soc. Ser. B Methodol., vol. 34, no. 2, pp. 187–220, 1972.

116. L. Pour et al., 'Pomalidomide is more effective in real clinical practice than in randomized trial - an observational study of the Czech Myeloma Group', in Session topic: 14. Myeloma and other monoclonal gammopathies - Clinical, Madrid, 2017, vol. 22nd Congress of EHA.

117. J. S. Miguel et al., 'Pomalidomide plus low-dose dexamethasone versus high-dose dexamethasone alone for patients with relapsed and refractory multiple myeloma (MM-003): a randomised, open-label, phase 3 trial', Lancet Oncol., vol. 14, no. 11, pp. 1055–1066, 2013.

118. K. R. Hess, 'Graphical methods for assessing violations of the proportional hazards assumption in Cox regression', Stat. Med., vol. 14, no. 15, pp. 1707–1723, 1995.

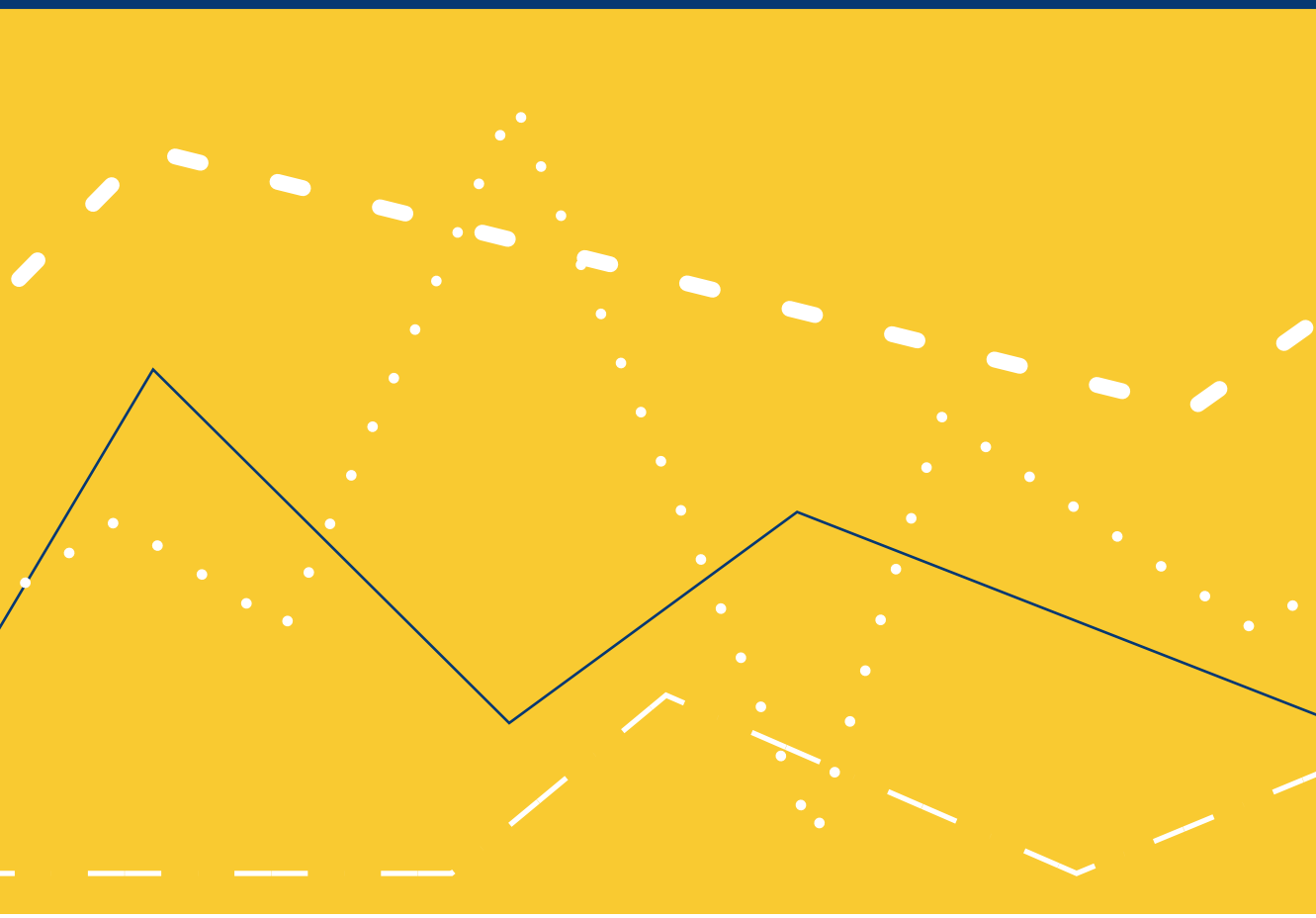
119. L. F. Hutchins, J. M. Unger, J. J. Crowley, C. A. J. Coltman, and K. S. Albain, 'Underrepresentation of patients 65 years of age or older in cancer-treatment trials.', N. Engl. J. Med., vol. 341, no. 27, pp. 2061–2067, Dec. 1999.

120. H.-G. Eichler, S. X. Kong, W. C. Gerth, P. Mavros, and B. Jonsson, 'Use of cost-effectiveness analysis in health-care resource allocation decision-making: how are cost-effectiveness thresholds expected to emerge?', Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res., vol. 7, no. 5, pp. 518–528, Oct. 2004.

121. V. Maisnar, Z. Stefanikova, and I. Spicka, 'Lenalidomide and dexamethasone in treatment of patients with relapsed and refractory multiple myeloma - Analysis of data from the Czech Myeloma Group Registry of Monoclonal Gammopathies.', manuscript was recently sent for publication to Neoplasma.


janssen
Oncology


iba
Institute
of Biostatistics
and Analyses



janssen Oncology

iba Institute of Biostatistics and Analyses