Scientific elaboration of selected aspects of the generation and analysis of routine practice data for benefit assessments of drugs according to §35a of the German Social Code Book V (SGB V)[1]

EXTRACT

Project: A25-13 Version: 1.0 Status: 28 Oct 2025 DOI: 10.60584/A25-13_en

[1] Translation of Chapters 1 to 6 of the rapid report *Wissenschaftliche Ausarbeitung ausgewählter Aspekte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V* (Version 1.0; Status: 28. October 2025). Please note: This translation is provided as a service by IQWiG to English-language readers. However, solely the German original text is absolutely authoritative and legally binding.

Publishing details

Publisher

Institute for Quality and Efficiency in Health Care

Topic

Scientific elaboration of selected aspects of the generation and analysis of routine practice data for benefit assessments of drugs according to §35a of the German Social Code Book V (SGB V)

Commissioning agency

Federal Joint Committee

Commission awarded on

28 January 2025

Internal Project No.

A25-13

DOI-URL

https://doi.org/10.60584/A25-13_en

Address of publisher

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Siegburger Str. 237
50679 Köln
Germany

Phone: +49 221 35685-0
Fax: +49 221 35685-1
E-mail: berichte@iqwig.de
Internet: www.iqwig.de

Recommended citation

Institute for Quality and Efficiency in Health Care. Scientific elaboration of selected aspects of the generation and analysis of routine practice data for benefit assessments of drugs according to §35a of the German Social Code Book V (SGB V); Extract [online]. 2025 [Accessed: DD.MM.YYYY]. URL: https://doi.org/10.60584/A25-13_en.

Keywords

Product Surveillance – Postmarketing, Registries, Data Interpretation – Statistical, Research Design, Models – Statistical, Concept

This report was prepared in collaboration with external experts.

The responsibility for the contents of the report lies solely with IQWiG.

According to §139 b (3) No. 2 of Social Code Book (SGB) V, Statutory Health Insurance, external experts who are involved in the Institute's research commissions must disclose "all connections to interest groups and contract organizations, particularly in the pharmaceutical and medical devices industries, including details on the type and amount of any remuneration received". The Institute received the completed *Form for disclosure of potential conflicts of interest* from each external expert. The information provided was reviewed by a Committee of the Institute specifically established to assess conflicts of interests. The information on conflicts of interest provided by the external experts and external reviewers is presented in Appendix C of the full report. No conflicts of interest were detected that could endanger professional independence with regard to the work on the present commission.

External experts

- Anne Regierer, German Rheumatism Research Centre Berlin – Epidemiology and Health Services Research Programme, Berlin, Germany

IQWiG thanks the external expert for her collaboration in the project.

IQWiG employees

- Caroline Wöhl
- Lars Beckmann
- Katharina Hirsch
- Anne-Kathrin Petri
- Ulrike Seay
- Guido Skipka
- Volker Vervölgyi

Glossary

Terms that are explained in the glossary are underlined upon their first occurrence in the main text and are cross-referenced to the glossary.

Active-comparator new-user design

An active-comparator new-user (ACNU) design can be selected for the replication of randomized comparative studies based on observational data. This refers to a design in which only patients who have an indication for a change in their treatment and are therefore starting a new treatment (intervention or comparator) are included in the comparison of two treatment groups [1]. In this case, the start of follow-up (t_0) is ideally set as the day the treatment decision is made.

Average treatment effect in the population

The "average treatment effect in the population" (ATE) estimand describes the average treatment effect in the entire population. It examines how the results would differ if the entire population had been treated with either the intervention or the comparator [2].

Average treatment effect on the treated

The "average treatment effect on the treated" (ATT) estimand describes the average treatment effect in the population that received the intervention. It examines how the results in this population would differ if they had not been treated with the intervention [2].

Confounder

A confounder is a variable that influences both the treatment decision and the outcome and can therefore distort a treatment effect. In non-randomized comparative studies, adjusting for confounders is intended to compensate for the lack of structural equivalence between treatment groups [3]. However, in statistical analysis, adjustments can only be made for known and measured confounders; bias due to unknown, unmeasured, or unmeasurable confounders persists (so-called residual confounding).

Convergence

Maximum likelihood methods are used for calculating propensity scores, for multiple imputation of missing values, and for effect estimates. These include, among others, generalized linear models (e.g., multivariate logistic regression) and the Cox proportional hazards model. The algorithms underlying these methods are based on iterative processes and multiple convergence criteria, such as convergence of the parameter estimate or the likelihood estimate. An algorithm is said to converge when, in two consecutive iterations, the changes in the respective estimates are smaller than specified values.

If, after a specified number of iterations, convergence is not achieved for one or more criteria, no parameter estimates are available, or the parameter estimates are potentially highly biased. The reasons why an algorithm does not converge are manifold. These include an insufficient number of observations (sample size) and inappropriate modelling (due to too many covariates) [4].

Coverage probability

Maximum likelihood methods are used to calculate propensity scores, perform multiple imputation for missing values, and estimate effects. Simulation studies are conducted to compare the quality of these methods. In these studies, the first step involves systematically generating simulated datasets (replicates) for various scenarios. In a second step, empirical quality measures are calculated using the replicates. The coverage probability, as a quality measure, describes the probability that the (1 − alpha) confidence interval of an effect estimate contains the true value. This requires specifying the alpha error, e.g., alpha = 0.05. In simulation studies, the empirical coverage probability is determined by the proportion of replicates in which the (1 − alpha) confidence interval contains the specified effect. A good statistical method has an empirical coverage probability of 1 − alpha. Coverage probabilities that are too low are an indication of great bias, while coverage probabilities close to 1 are an indication of a lack of power due to overly wide confidence intervals [5].

Disease risk score

The disease risk score (DRS) – also known as the prognosis score – is defined as the probability that a person will develop a disease. The DRS method can thus be used (as an alternative to propensity score methods) for confounder control, with the aim of approximating structural equivalence between treatment groups [6].

Estimand

In propensity score analyses, the concept of estimands is used to describe the population for which the effect of a treatment is to be estimated. The following estimands are typically distinguished: ATE, ATT, average treatment effect on the untreated, and average treatment effect in the overlap [2]. The selection of a specific propensity score method depends on the estimand of interest [7]. Depending on the chosen propensity score method and the overlap of the propensity score distributions, the estimands may merge seamlessly into one another in a given data situation (see above for the estimands ATE and ATT).

The term must be distinguished from the concept of estimands in the context of clinical trials. In this regard, a paper by the European Medicines Agency (Addendum to the ICH E9 Guideline on Statistical Principles for Clinical Trials [8]) describes 5 strategies that lead to different estimands (including an estimand based on the treatment policy strategy). The specifications

made therein focus on the handling of intercurrent events, i.e., events that occur during the course of the follow-up of a patient (e.g., treatment switching).

Immortal-time bias

Immortal-time bias may be present if, due to the study design, there is a period (waiting time) during which an event (e.g., death) cannot occur [9,10]. A classic example of this is when the start of follow-up (t_0) is set at the time the inclusion criteria are met, but before the time of treatment allocation. As a result, treatment allocation is based on information from the follow-up [11]. This applies, for example, to situations in which patients are allocated to the intervention group only after receiving CAR-T cell therapy (and thus after t_0). In these cases, patients must have survived until the start of CAR-T cell therapy (while receiving bridging therapy).

Inverse probability of treatment weighting

Inverse probability of treatment weighting (IPTW) is a weighting approach based on propensity scores. The weight allocated to each person corresponds to the inverse probability of having received the intervention. Patients from the intervention group are allocated a weight of 1/propensity score, and patients from the control group are allocated a weight of 1/(1 − propensity score). This means that patients from the intervention group with a low propensity score (i.e., with a low probability of receiving the intervention) receive a high weight because they resemble patients from the control group in terms of the observed characteristics (as indicated by their low propensity score). In contrast, patients from the control group with a high propensity score (i.e., with a high probability of receiving the intervention) are weighted more heavily, as they are more similar to patients from the intervention group in terms of observed characteristics [12].

Latency-time bias

Latency-time bias occurs when prior treatment leads to an increased risk of the event of interest, but the event itself does not occur until after the study intervention. An example of this is: Secondary primary tumours can occur several months to years after chemotherapy [13]. Since these events are not causally related to the study intervention, latency windows should be defined. Events occurring within this time window are excluded from the analysis [14].

Matching weights

Matching weights is a weighting approach based on propensity scores. The weight allocated to each person corresponds to the ratio of the lower of the two predicted probabilities (propensity score or 1 − propensity score) to the predicted probability of the treatment actually received [7].

Medication possession ratio

Medication possession ratio (MPR) is a measure that incorporates any prior exposure to one or more comparator drugs used in the therapeutic indication being assessed. MPR matching attempts to balance the treatment groups with respect to prior exposure patterns [15].

Missing at random

Missing at random (MAR) refers to an assumption regarding a missing value. The probability of a value being missing – taking into account a person's observed values – does not depend on the missing value itself. Such missing values lead to biased results if this dependency is not accounted for in the analysis. To account for such dependencies in the analysis, the mechanism of missing values must be clearly identifiable and explainable based on the observed values. Proper consideration in an analysis leads to unbiased estimates [16]. An example of this: Men are more likely than women to fail to complete a questionnaire assessing the severity of depression. Completing the questionnaire depends on gender, but not on the actual severity of the depression itself [17].

Missing completely at random

Missing completely at random (MCAR) refers to an assumption regarding a missing value. The probability of a value being missing is independent of other observed values for a person (e.g., covariates or values at earlier time points) and the unobserved, missing value itself. While such missing values do not lead to biased results, they do reduce the precision of the effect estimates [16]. An example of this: Some of the samples are mishandled in the laboratory, so that they can no longer be analysed [17].

Missing not at random

Missing not at random (MNAR) refers to an assumption regarding a missing value. The probability of a value being missing depends on the missing value itself, and this dependency does not disappear even when the observed values of the individual are taken into account. Such missing values lead to biased results. This bias cannot be corrected because the mechanism of missing values depends on unobserved values. In this case, the magnitude of the bias can only be examined using sensitivity analyses based on unverifiable assumptions [16]. An example of this: The probability of missing values regarding the severity of depression depends on the severity of depression itself [17].

Multiple imputation using chained equations

Multiple imputation is a statistical method for handling missing values in datasets. Several complete datasets are generated in which missing values are replaced under model assumptions (MCAR, MAR, or MNAR). Any variables can be used in this process. Multiple imputation using chained equations (MICE) – also known as fully conditional specification (FCS) – has established itself as a variant of multiple imputation [18].

Overlap

Overlap refers to the area in which the distributions of the propensity scores of the groups being compared intersect. The overlap can be assessed by visually comparing the propensity score distributions (propensity distribution plots) (see Figure 1) [19]. As shown in Figure 1 the area of overlap may be minor (approximately 10% here), even if the range of overlap is quite large (ranging from approximately 0.2 to 0.75 here). Good overlap therefore refers to a sufficiently large overlapping area and not to a wide range of overlap.

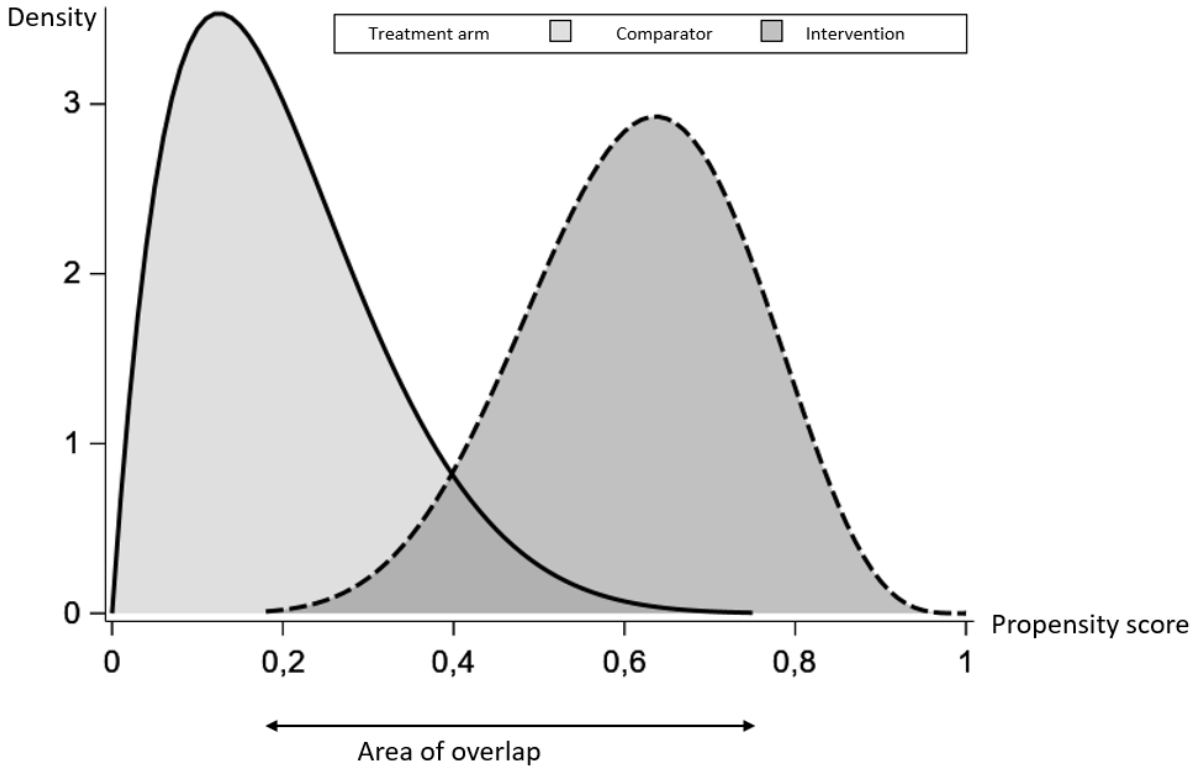


Figure 1: Overlap of the populations (measured by the propensity score)

Overlap weights

Overlap weights is a weighting approach based on propensity scores. The weight allocated to each person corresponds to the probability of having received the opposite treatment [7].

Positivity

Positivity means that, when comparing two patient groups, the treatments being compared (new drug or comparator) represent a therapeutic option for all patients included in the analysis. It follows that, for example, patients with a contraindication to one of the therapies under investigation must not be included in the analysis [20].

Prevalent new-user design

A prevalent new-user (PNU) design can be selected for replicating randomized comparative studies based on observational data. In this design, three different patient populations can be included in the comparison of two treatment groups: (1) patients who begin treatment with the intervention and who were previously treatment-naive (with regard to both the intervention and the comparator) (new users), (2) patients switching from the comparator to the intervention (PNU), and (3) patients who continue their treatment with the comparator [21]. The goal is to reduce potential bias through an appropriate choice of the start of follow-up(t_0) for patients in the control group. For patients in the intervention group (comprising new users and PNU), the selection of t_0 is carried out analogously to the ACNU design (see above).

Propensity score matching

With propensity score matching, patients in the intervention and control groups are paired (i.e., "matched") according to their propensity scores. Matching can be performed using various algorithms, including, for example, nearest-neighbour matching. The allocation can be 1:1 or 1:n, with or without replacement. Depending on the overlap of the treatment groups as measured by the propensity score, this method may be associated with a reduction in the sample size and thus a loss of statistical power (individuals without a matching partner are excluded) [12].

Propensity score methods

Propensity score methods are statistical techniques used in non-randomized comparative studies to achieve balance in terms of observed covariates between treatment groups (structural equivalence). The propensity score is the probability of receiving the intervention, as determined by observed covariates. There are 4 approaches to estimating the treatment effect using propensity scores: weighting, matching, stratification, and regression adjustment [12].

Selection bias

Selection bias arises when patients, follow-up time, or events are excluded from the analysis, and the selection is associated with both the exposure and the outcome. This is the case, among other things, when patients are censored after the start of follow-up (t_0) (e.g., upon treatment switching) or when t_0 is set at the time the inclusion criteria are met but after treatment allocation. In the latter example, only those patients who meet the inclusion criteria exactly at t_0 are included in the analysis [11,22].

Standardized morbidity (or mortality) ratio weighting

Standardized morbidity (or mortality) ratio weighting (SMRW) is a weighting approach based on propensity scores. Patients from the intervention group are allocated a weight of 1, and

patients from the control group are allocated a weight of propensity score/(1 − propensity score). This means that patients from the control group are weighted more heavily the more similar they are to the intervention group in terms of the observed characteristics [7].

Target trial emulation

Target trial emulation refers to the replication of the design of randomized controlled trials based on observational data [23].

Time-lag bias

Time-lag bias arises when comparing treatment groups at different stages of disease progression [24]. This is the case, for example, when the time interval between the diagnosis of a progressive disease and the start of follow-up(t_0) differs between groups.

Executive summary

On 28 January 2025, the Federal Joint Committee (G-BA) commissioned the Institute for Quality and Efficiency in Health Care (IQWiG) to provide a scientific elaboration of methods in the field of routine practice data collection (RPDC), focussing on the assessment of specific topics.

Research question

The objective of this report is to provide a scientific elaboration of methods (generation and analysis of routine practice data for benefit assessments of drugs) in the field of RPDC. The report focuses on the assessment of the following topics, each within the context of non-randomized comparative studies:

- Confounder identification and selection
- Estimation of required sample sizes in the context of insufficient prior information
- Start of follow-up
- Treatment switching
- Patient-reported outcomes (PROs)
- Missing values
- Propensity score analyses in therapeutic indications with small patient populations

This involves an exploratory search for relevant literature on the listed topics and an examination of the deficiencies identified in previous reviews of study protocols and statistical analysis plans (SAPs) (with possible additions to the listed topics).

Existing registries are primarily suitable as potential RPDC data sources. The scientific elaboration therefore also focuses on disease registries as tools for data collection with the aim of generating routine practice data, since other data sources (e.g., routine data from statutory health insurance) do not currently appear suitable for meeting the data (quality) requirements within the framework of benefit assessments.

Methods

Examination of previous reviews of study protocols and statistical analysis plans

The need for changes identified in previous reviews of the study protocols and SAPs submitted by the drug manufacturers were examined. Based on this, additional topics were to be added to the commissioned topics, if necessary. Since all identified needs for changes could be allocated to one of the 7 topics, no additional topics were addressed within the scope of the commission.

Draft proposal for requirements for systematic confounder identification

For the topic of confounder identification and selection, IQWiG working paper GA23-02 (published on 18 June 2025) was used for this commission. Therefore, the methods were not newly developed. The presentation in this report is essentially limited to the derived recommendations for systematic confounder identification in non-randomized comparative studies. For further information, please see working paper GA23-02.

Information retrieval

For the remaining 6 topics, the information retrieval for methodological documents and systematic reviews included an exploratory search in PubMed, a website search of regulatory agencies and other HTA agencies, and a targeted web search. In addition, based on the previously identified methodological documents and systematic reviews, reference lists were screened, and the “Similar Articles” function in PubMed and the “Cited-by” function in OpenAlex were used.

Assessment of selected aspects for generating routine practice data in the context of non-randomized comparative studies

Confounder identification and selection

In IQWiG working paper GA23-02, in accordance with the approach by Pufulete 2022, a systematic confounder identification was conducted for the therapeutic indication of relapsing-remitting multiple sclerosis (RRMS) using an appropriate comparison of two drug therapies with regard to key patient-relevant outcomes. To this end, a working group of external experts was commissioned, which presented its findings in a report. Based on the results of the report, the following recommendations for systematic confounder identification in non-randomized comparative studies (and thus also in RPDC procedures) were derived by IQWiG as part of the work on GA23-02 by means of further investigations:

1) Identification

   Potential confounder variables should be identified through a systematic literature review and interviews with clinical experts.

2) Summarization

   Confounder variables that measure the same construct or overlap in content should be summarized. A clear operationalization must be defined for each of these potential confounders.

3) Assessment of relevance

   The relevance of each potential confounder should be assessed with the involvement of clinical experts. The result can be used for a content-based selection of confounder variables.

The process of systematic confounder identification is resource-intensive, despite the resource-saving options described in working paper GA23-02. However, it is important at this point to distinguish between the resources required prior to RPDC procedures (for the identification and summarization of potential confounders) and the overall resources required (which includes, among other things, the resources required for data collection and documentation). It may be beneficial to invest more resources upfront in order to reduce the overall resources required. In a next step, as part of the preparation of working paper GA25-02, it will be examined whether further measures to reduce the resources required for the systematic identification and recording of confounders are possible. In doing so, potential resource-saving options must be weighed against the associated loss of information and the potential loss of certainty of results.

Estimation of the required sample size in the context of insufficient prior information

The G-BA regularly commissions IQWiG to develop scientific concepts for RPDC procedures and their analysis. The concepts should also describe the requirements for the duration and scope of the respective data collection. The aim is to answer the question of whether the planned RPDC procedure is, in principle, feasible and can be conducted in a meaningful way. For this purpose, no sample size estimation is performed because the necessary information is not available. Rather, the concept development aims to provide an assessment of basic feasibility. Depending on the information available at the time of concept development, the approach taken is either a preliminary sample size estimation or an exploratory analysis of sample size scenarios.

Parameters for the sample size estimation

To determine the necessary sample size within the framework of the preliminary sample size estimation, assumptions must be made for several parameters. These parameters include, among others, the effect size, the significance level (probability of error, also known as Type I error), and the power (1 − Type II error). In sample size planning for randomized controlled trials (RCTs), assumptions regarding the effect size are generally based on information from pilot studies or comparative Phase II studies, which allow for a relatively reliable estimation of this parameter. For RPDC procedures required by the G-BA due to a lack of direct comparative evidence, this information is not available as a starting point for sample size planning. For this reason, sample size estimation as part of the concept development (often based on non-comparative marketing authorization studies) can, at best, serve only as a guide.

In order to derive a conclusion about the benefits or harms of the drug under assessment from a non-randomized comparative study, a sufficiently large observed effect is required, which can be assumed not to be caused solely by systematic bias. This should be implemented statistically in the RPDC procedure in the form of a test for a shifted null hypothesis (H_0). In this context, the confidence interval of the effect observed must lie entirely beyond a specific

threshold. There is currently no scientific standard for how large the threshold for the shifted null hypothesis should be. In rapid report A19-43, IQWiG established a range of 2 to 5 for the relative risk (or 0.2 to 0.5, e.g., for mortality-reducing interventions) relative to the lower (or upper) confidence interval limit as the threshold. However, in light of methodological advancements, it appears appropriate to re-evaluate the established threshold in a subsequent step (outside the scope of the scientific elaboration conducted in this report).

No or insufficient prior information for a preliminary estimate of the RPDC scope

When developing an RPDC concept, a preliminary sample size estimation is always performed whenever it is reasonably feasible. At the time of concept development (usually prior to drug approval), however, the information required to estimate the effect size for a new drug is generally not available (e.g., only preliminary, and in some cases contradictory, results may be available for outcomes presented only selectively in conference abstracts or press releases). Thus, the conditions for a preliminary sample size estimation are not met.

In these cases, an exploratory analysis of sample size scenarios is conducted in order to nonetheless enable conclusions regarding the scope of data collection. This approach essentially follows the principles of preliminary sample size estimation but starts from a different point. In a first step, the estimated number of patients in the relevant therapeutic indication in Germany is determined. In a second step, by specifying the number of patients, the significance level, and the power, the question is answered as to which effect sizes can be detected based on the number of patients generally available for an RPDC procedure in Germany within the relevant therapeutic indication (precision of the estimate) and taking into account the risk of systematic bias due to the non-randomized comparison (shifted null hypothesis).

“Detectable” means that the 95% confidence interval estimated within the RPDC framework, with 80% power for a given number of patients, lies entirely above or entirely below the shifted hypothesis boundary. In addition to the number of patients, if available, the baseline risk – either under treatment with the intervention or with the comparator – for a patient-relevant outcome is used as an anchor to illustrate a spectrum of possible scenarios within the therapeutic indication.

Decision-making for or against a preliminary sample size estimation

No fixed limits are set when deciding for or against a preliminary sample size estimation within the RPDC framework. Whether or not a preliminary sample size estimation is considered reasonably feasible based on the available data constellation must be assessed on a case-by-case basis.

Regardless of the methodological approach chosen for concept development (preliminary sample size estimation or exploratory analysis of sample size scenarios), the manufacturer of

the drug under assessment should carry out the (final) sample size estimation for the RPDC procedure at the time of the interim analyses based on the preliminary results available at that time. In principle, it is considered appropriate to conduct the final sample size planning while taking into account the results of interim analyses, since the estimation of the necessary sample size in the collection of routine practice data (as opposed to clinical trials) is not performed with a view to active recruitment. Instead, ideally all patients from a registry population who meet the RPDC inclusion criteria are included after giving their consent. This is also done against the backdrop that, depending on the propensity score method, a relevant proportion of the included patients may not be considered in the analysis.

Start of follow-up and treatment switching

Start of follow-up

A challenge in non-randomized studies using routine practice data is the appropriate definition of the start of follow-up (index date [t_0]), since, unlike in RCTs (time of randomization), no fixed starting point exists. Bias in the results (e.g., due to immortal time bias or selection bias) arises when t_0, the assessment of eligibility, and treatment allocation do not align.

In the active-comparator new-user (ACNU) design, the date of the treatment decision (or the best possible approximation thereof) is set as t_0 for both treatment groups. Only those patients who have an indication for switching their treatment are included in the study. The prevalent new-user (PNU) design is suitable for handling situations in which a new treatment is not initiated in either treatment group at the start of follow-up. The goal here is to reduce potential bias resulting from an incorrect choice of the start of follow-up for patients in the control group. The application of this design is contingent upon the use of appropriate methods to account for prior treatment(s) received. The robustness of the results should be verified using sensitivity analyses.

Treatment switching

In comparative studies (with or without randomization), patients have the option to switch from one treatment to another as needed. This process is referred to as “treatment switching”. In the context of the rapid report, the term is limited to switching from the intervention to the comparator or vice versa.

In an RPDC procedure, switching from the intervention to the comparator generally corresponds to switching to a generally accepted treatment that conforms to the medical standard. Switching is thus part of a treatment strategy that corresponds to routine clinical practice. In such a situation, an unbiased estimate of the treatment effect is possible. Switching from the comparator to the intervention is unproblematic if the intervention constitutes an adequate follow-up therapy within the treatment strategy (this applies, for

example, if the intervention has already been approved at an earlier stage for use in a subsequent line of therapy). However, this scenario is usually not to be expected in the RPDC context. In all other cases, switching from the comparator to the intervention can lead to biased results, and it must be demonstrated that the results are transferable to the research question of the RPDC procedure.

Regardless of treatment switching, for benefit assessments of drugs, analyses are primarily required in accordance with the treatment policy strategy (i.e., an estimate of the effect for the entire treatment strategy) and the intention-to-treat (ITT) principle. To investigate the potentially confounding influence of a treatment switch on the ITT analysis using survival time methods, sensitivity analyses can be conducted in which patients who switch from the comparator to the intervention are censored at the time of switching (per-protocol analyses) or in which complex methods are applied to account for treatment switching (such as inverse probability of censoring weighting [IPCW]).

Special considerations for the analysis of routine practice data

Since the G-BA often requires RPDC procedures for drugs used to treat rare diseases (orphan drugs), the number of patients available is inherently small. Depending on the treatments being assessed, there is the additional challenge of recruiting a sufficient number of patients. Against this backdrop, to nevertheless include as many patients as possible receiving a new treatment in the analysis, a pragmatic approach offers the option of allocating patients who switch from the comparator to the intervention to the treatment groups based on their follow-up time under treatment with the comparator (i.e., that patients who switch to the intervention after a shorter period and for whom an adequate follow-up time under the intervention is still expected can be analysed in the intervention group).

Depending on the comparison of interest (new drug vs. comparator), there may be cases in which a treatment option is not immediately available or immediately effective following the treatment decision. It may be necessary to employ a bridging therapy, which should be understood as part of the treatment strategy. In order to handle bridging therapies properly, it is crucial that the date on which the original treatment was decided upon (e.g. the date on which CAR-T cell therapy was approved by the Tumour Board) is set as the start of follow-up(index date [t_0]) in accordance with the target trial concept.

Patient-reported outcomes

To close existing evidence gaps for new drugs, in RPDC procedures, the G-BA routinely requires the recording of outcomes on symptoms and health-related quality of life. These outcomes are typically recorded using PROs. For this purpose, uniform data collection time points must be selected for both treatment groups to generate informative data on the course of the disease during treatment. Data should be collected several times a year at standardized

intervals at defined time points, with narrower intervals at the beginning and wider intervals later on. Appropriate tolerance windows should be predefined for the data collection time points; these windows should not overlap and should allow for a meaningful allocation to a specific data collection time point. In the RPDC context, data are collected in routine clinical practice, so the conditions under which PRO data are collected differ largely from those in a clinical trial. Digital surveys (e.g., via an app and/or a patient portal) enable low-threshold data collection independent of doctor visits. Digital tools are already being successfully used for data collection in registries and in routine clinical practice.

Missing values

Measures to prevent missing values

In studies, missing values generally cannot be completely avoided. In order to still be able to draw sufficiently reliable conclusions about the benefits and harms of a drug – both under conditions of routine clinical practice and with regard to the specific questions of benefit assessments – it is both sensible and necessary to implement measures that keep the proportion of missing values to a minimum. To be able to maintain and provide high-quality data from routine clinical practice in a practical manner and without major obstacles, a permanently available and continuously maintained data infrastructure and documentation system must be established in the registries with adequate personnel and financial resources. Both the establishment and maintenance of this infrastructure could be supported by the drug manufacturers that wish to access the data for conducting registry-based studies. Furthermore, the following factors were identified as promoting factors, among others:

- limiting data collection to those data relevant to answering the research question of interest,
- implementing a suitable monitoring strategy,
- providing regular training on data entry for data collection staff, and
- informing patients about the data stored in the registry (including information on the benefits of collecting patient-reported data).

Methodological handling of missing values in the context of propensity score analyses

Models for estimating propensity scores using logistic regression require complete data for all potentially relevant confounders, meaning that patients with (at least) one missing value are excluded from the analysis. However, if only patients with complete data are included in the analyses (complete-case analyses), this leads to substantial bias in the results if the “missing completely at random” (MCAR) assumption is not met. This is generally the case in studies based on routine practice data. Consequently, statistical methods for handling missing values are necessary in propensity score analyses. In general, multiple imputation methods are considered appropriate. A variety of such methods is available, among which “multiple

imputation by chained equations" (MICE) is recognized as an established method for handling missing values. To assess the robustness of the results calculated in this way, sensitivity analyses using model-based methods may be considered, which are based on the MAR or "missing not at random" (MNAR) assumption.

Propensity score analyses in therapeutic indications with small patient populations

In non-randomized studies, the structural equivalence of the groups to be compared – which is necessary for a fair comparison – is generally not given. Consequently, group differences in potential confounders must be accounted for when estimating effects by adjusting for these confounders. Propensity score methods play a primary role in confounder adjustment within the RPDC context. These methods were originally developed for large datasets (epidemiological questions). In the context of this report, simulation studies on propensity score analyses with small sample sizes were identified. These studies show that propensity score methods are also applicable in therapeutic indications with small patient populations under certain conditions and can lead to interpretable results. However, an important problem is the potential lack of convergence of the models used for effect estimation. The problem of lack of convergence due to an excessive number of confounders can be addressed to a certain extent by gradually removing less important confounders from the model. This approach requires a prespecified ranking of confounders by importance and is carried out while accepting increased uncertainty. Overall, the scenarios examined in the simulation studies only partially reflect the situations that are likely to arise in the ongoing RPDC procedures, as these studies primarily consider different effect measures (such as odds ratios or risk differences) and, in some cases, a small number of potential confounders. At this point in time, due to a lack of evidence, it therefore remains unclear in which cases (and under what conditions) propensity score analyses will yield interpretable results in RPDC procedures.

Conclusion

Benefit assessments of drugs require data for comparison with the standard treatment. Since the approval of orphan drugs is often based on non-comparative data, the RPDC approach was introduced with the aim of closing existing evidence gaps and thus obtaining a better evidence base for benefit assessments. Data collection must be conducted as non-randomized comparative studies. Provided certain quality requirements are met, studies based on registry data can close this evidence gap.

- When planning non-randomized comparative studies based on routine practice data, target trial emulation is a recommended approach to minimize systematic (avoidable) bias. A prerequisite for optimal emulation of a hypothetical RCT using observational data is that the necessary data are available in the registry dataset with the required completeness and depth. High data quality can only be achieved on a broad scale if the generation and utilization of registry data are feasible and resource-efficient. The

establishment and maintenance of a permanently available (operational) data infrastructure is considered beneficial. This could be supported by drug manufacturers that wish to draw on the registries to conduct registry-based studies.

- In non-randomized comparative studies aimed at comparing treatment effects, adequate control for confounders requires the systematic identification of relevant confounders and their consideration in the analysis. Confounder identification following the approach by Pufulete 2022 (via a systematic literature review and clinician involvement) is considered feasible and represents a meaningful approach in principle. Before clinical experts assess the relevance of the confounders, it is recommended to conduct an intensive summarization of the identified confounders. In principle, it may be beneficial to invest more resources in reducing the number of confounders to be recorded prior to an RPDC procedure in order to reduce the overall resources required through resource savings in data collection and analysis.
- In RPDC concepts, it is estimated whether a sufficient number of patients can be enrolled within an acceptable timeframe to enable informative results to be generated for benefit assessments. In general, only uncertain information is available for this estimation. If sufficient information on the intervention and the comparator is available, a preliminary sample size estimation is performed for this purpose. If necessary information is missing for the assumptions underlying a preliminary sample size estimation, an exploratory analysis of sample size scenarios is conducted to demonstrate potential detectable effects. Both approaches follow the same principle (they are based on identical parameters) and differ only in the parameter to be estimated using the remaining parameters.
- For long-term data collection (and patient follow-up) in routine clinical practice, incentives are required to compensate for the resources required for data generation and to motivate both the centres and the patients to collect data as completely as possible.
- To ensure that the resources required for PRO data collection are proportionate to the benefits, digital surveys (e.g., via an app or patient portal) are recommended. This enables low-threshold PRO data collection independent of doctor visits. Digital tools are already being successfully used for data collection in registries and in routine clinical practice and enable PRO data to be made available for research purposes with minimal barriers.
- A challenge in analysing routine practice data without randomization is determining the start of follow-up (index date [t_0]). If a new treatment is initiated in both treatment groups at t_0 (switch indication), the ACNU design can be used in a study. Ideally, t_0 corresponds to the day the treatment decision was made. In a treatment situation

where patients in the control group continue treatment with an established standard treatment, the PNU design represents a suitable alternative.

- Regardless of the choice of design and analysis strategies, the results of the ITT analysis (in accordance with the treatment policy strategy) should generally be presented as the primary results. This requires that all patients (regardless of any treatment switching at some point during follow-up) are analysed according to their original group allocation. The confounding influence of treatment switching during the RPDC procedure can be addressed with sensitivity analyses.
- A commonly used method for accounting for confounders in non-randomized comparative studies based on registries is an analysis using propensity scores. Since models for estimating propensity scores using logistic regression require complete data for all potentially relevant confounders, statistical methods for handling missing values are necessary. The MICE method is recognized as an established method. Propensity score methods are also applicable in small patient populations under certain conditions and can yield interpretable results. The scenarios examined in the identified simulation studies do not fully reflect the situations that are foreseeable in ongoing RPDC procedures. At this point in time, it therefore remains unclear in which cases (and under which conditions) propensity score analyses will yield interpretable results in RPDC procedures.

Table of contents

List of tables

Page

List of figures

Page

List of abbreviations

Abbreviation	Meaning
ACNU	Active-comparator new-user
AHRQ	Agency for Healthcare Research and Quality
AIHTA	Austrian Institute for Health Technology Assessment
ATE	average treatment effect (in the population)
ATMP	Advanced Therapy Medicinal Products
ATT	average treatment effect on the treated
CAR	chimeric antigen receptor
CBDR	Canadian Bleeding Disorders Registry
CDA	Canada's Drug Agency
DHR	Deutsches Hämophilie Register (German Haemophilia Registry)
DKG	Deutsche Krankenhausgesellschaft (German Hospital Association)
DLBCL	diffuse large B-cell lymphoma
DRS	disease risk score
EMA	European Medicines Agency
EMCL	European Mantle Cell Lymphoma
EPAR	European Public Assessment Report
FCS	fully conditional specification
FDA	Food and Drug Administration
G-BA	Gemeinsamer Bundesausschuss (Federal Joint Committee)
GTR	Gene Therapy Registry
HAS	Haute Autorité de Santé (French National Authority for Health)
HIS	hospital information system
HTA	health technology assessment
IMBEI	Institut für Medizinische Biometrie, Epidemiologie und Informatik(Institute for Medical Biometry, Epidemiology, and Informatics)
IPCW	inverse probability of censoring weighting
IPTW	inverse probability of treatment weighting
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (Institute for Quality and Efficiency in Health Care)
ITT	intention-to-treat
MAR	missing at random

Abbreviation	Meaning
MCAR	missing completely at random
MCMC	Markov Chain Monte Carlo
MICE	multiple imputation using chained equations
MNAR	missing not at random
MPR	medication possession ratio
NICE	National Institute for Health and Care Excellence
PICO	population, intervention, comparison, outcome
PNU	prevalent new user
PRO	patient-reported outcome
RCT	randomized controlled trial
RPDC	routine practice data collection
RR	relative risk
RRMS	relapsing-remitting multiple sclerosis
SAP	statistical analysis plan
SCQM	Swiss Clinical Quality Management in Rheumatic Diseases
SGB	Sozialgesetzbuch (Social Code Book)
SMRW	standardized morbidity or mortality ratio weighting
SQR	Swedish Rheumatology Quality Registry
WFH	World Federation of Haemophilia

1 Background

According to §35a (3b), Sentence 1 of the German Social Code Book (SGB) V, the Federal Joint Committee (G-BA) may require the drug manufacturer to conduct routine practice data collection (RPDC) and analysis for benefit assessments of certain drugs [25]. Due to the comparative research questions in benefit assessments, this typically involves a comparative study conducted without randomization, in accordance with the requirements of §35a SGB V (3b).

A resolution requiring an RPDC specifies the requirements for an RPDC and its analysis. If the G-BA requires an RPDC, the drug manufacturer must prepare a study protocol and a statistical analysis plan (SAP). These documents must describe the planning, conduct, and analysis of the study for the purpose of the RPDC. The G-BA reviews whether the approach described in these documents is suitable for meeting the requirements for the RPDC. For this review, it may involve the Institute for Quality and Efficiency in Health Care (IQWiG) [26].

A general elaboration of criteria for the quality of routine practice data and of methodological requirements for the generation of routine practice data and their analysis has already been carried out by IQWiG based on a commission from the G-BA and published as rapid report A19-43 on the IQWiG website [20] (hereinafter referred to as A19-43).

Initial experience is now available both with the review of study documents submitted by the drug manufacturers regarding their suitability in terms of content and methods for conducting the RPDC (including the associated analysis) and with ongoing RPDCs in registries as a data source (registry-based studies). Against this backdrop, there is a need for a further scientific analysis – going beyond A19-43 – of selected aspects regarding the generation of routine practice data and their analysis, in the context of non-randomized comparative studies.

2 Research question

The objective of this report is to provide a scientific elaboration of methods (generation and analysis of routine practice data for benefit assessments of drugs) in the field of RPDC. The report focuses on the assessment of the following topics, each within the context of non-randomized comparative studies:

- Confounder identification and selection (see Section 5.2)
- Estimation of required sample sizes in the context of insufficient prior information (see Section 5.3)
- Start of follow-up (see Section 5.4.1)
- Treatment switching (see Section 5.4.2)
- Patient-reported outcomes (PROs) (see Section 5.5)
- Missing values (see Section 5.6)
- Propensity score analyses in therapeutic indications with small patient populations (see Section 5.7)

This involves an exploratory search for relevant literature on the listed topics and an examination of the deficiencies identified in previous reviews of study protocols and statistical analysis plans (SAPs) (with possible additions to the listed topics).

Existing registries are primarily suitable as potential RPDC data sources. The scientific elaboration therefore also focuses on disease registries as tools for data collection with the aim of generating routine practice data, since other data sources (e.g., routine data from statutory health insurance) do not currently appear suitable for meeting the data (quality) requirements within the framework of benefit assessments.

3 Project timeline

3.1 Project timeline

The G-BA Subcommittee on Medicines decided at its meeting on 28 January 2025, to commission IQWiG to prepare a scientific elaboration of the RPDC methods for the G-BA[2]. This report was to address specific issues arising from previous experience with the RPDC procedure (see Chapter 2).

An external expert was involved in the project.

The preliminary rapid report (Version 1.0) dated 28 July 2025, was published on the IQWiG website on 30 July 2025, and opened for a public hearing procedure. Written comments could be submitted until 19 August 2025. Unclear aspects from the written comments on the preliminary rapid report were discussed on 16 September 2025 in a scientific debate with the commenters. The key arguments from the comments are addressed in Appendix B of the full rapid report.

This rapid report incorporates the changes resulting from the debate.

Following the hearing procedure, IQWiG prepared this rapid report, which was published on the IQWiG website 4 weeks after submission to the G-BA. The comments received on the preliminary rapid report and the minutes of the scientific debate were made available in a separate document (“Documentation of the hearing on the preliminary rapid report ”) on the IQWiG website at the same time as the rapid report.

3.2 Specifications and changes during the course of the project

Final rapid report versus preliminary rapid report

In addition to editorial changes, the following clarifications or changes were made in the rapid report:

- In Section 5.2 it was specified that an evaluation of measures to reduce the resources required for the systematic identification and recording of confounders was to be conducted as part of the preparation of working paper GA25-02 (follow-up commission to GA23-02).
- In Section 5.3.1, the text on the “shifted null hypothesis” parameter was supplemented to state that, in a next step, IQWiG will address a revision of the threshold for the shifted null hypothesis proposed in A19-43.

[2] in accordance with Chapter 1, §4 (2) Sentence 3, Letter a) of the Rules of Procedure

- In Section 5.3.2, the procedural example for epcoritamab was replaced with odronextamab (both for the therapeutic indication of diffuse large B-cell lymphoma [DLBCL]), and a text section titled “Classification of the exploratory analysis of sample size scenarios” was added.
- In Section 5.3.2, Figure 7 was amended, providing a graphical representation of the effect size detectable in an RPDC procedure.
- In the glossary and in Section 5.6.2.1, the definitions of “missing at random” and “missing not at random” were amended.
- In Section 5.6.2.1, the text on the description of suitable sensitivity analyses was amended to note that the inclusion of auxiliary variables carries the risk of biasing study results due to colliders.
- In Section 5.6.2.1, the text on the description of suitable sensitivity analyses was amended to state that sensitivity analyses can be used to estimate the potential influence of residual confounding.
- In Section 5.6.2.2, the threshold for missing values in outcomes with repeated measurements was clarified.
- In Section 5.7, it was clarified that not only propensity score methods are accepted, but that IQWiG assesses whether adequate adjustment for confounders was planned and performed, regardless of the chosen method of causal inference.
- In Section 5.7, a paragraph regarding the order of drawing bootstrap samples and generating multiple imputations was added.
- Appendix B “Evaluation of the hearing on the preliminary rapid report ” was added to the full rapid report.
- Appendix C “Disclosure of relationships of external experts” was added to the full rapid report.

4 Methods

4.1 Examination of previous reviews of study protocols and statistical analysis plans

The changes requested in previous reviews of the study protocols and SAPs submitted by the drug manufacturers were examined. Based on this examination, additional topics were to be added to the commissioned topics, if necessary. Since all identified requested changes could be allocated to one of the 7 topics, no additional topics were addressed within the scope of the commission.

4.2 Proposal for requirements for systematic confounder identification

For the topic of confounder identification and selection, this commission drew upon the working paper GA23-02 [27] published by IQWiG on 18 June 2025. Consequently, no new methods was developed for this rapid report. The presentation in this report is essentially limited to the specific recommendations derived for the systematic identification of confounders in non-randomized comparative studies (and thus also in the RPDC procedure). For further information (e.g., regarding the necessary documentation), please see working paper GA23-02.

4.3 Information retrieval

For the other 6 topics on which this rapid report focuses in accordance with the G-BA's commission, the exploratory information retrieval included:

- a literature search for methodological documents and reviews in PubMed
- a website search for methodological documents from regulatory agencies (European Medicines Agency [EMA], Food and Drug Administration [FDA]), and HTA agencies (National Institute for Health and Care Excellence [NICE], Canada's Drug Agency [CDA], Agency for Healthcare Research and Quality [AHRQ], Haute Autorité de Santé [HAS], and Austrian Institute for Health Technology Assessment [AIHTA])
- a targeted web search
- additional search techniques:
  - screening reference lists of the identified methodological documents and systematic reviews
  - use of the "Similar Articles" function in PubMed and the "Cited-by" function in OpenAlex based on previously identified methodological documents and reviews

The presentation in the rapid report is limited to the specific results.

5 Assessment of selected aspects regarding the generation and analysis of routine practice data in the context of non-randomized comparative studies

5.1 Introductory remarks on the requirements for non-randomized comparative studies for benefit assessments

Analogous to A19-43 [20], routine practice data for the purpose of benefit assessment of drugs are defined as follows for the present report:

- Routine practice data are collected from patient populations for whom there is an indication for the drug under assessment within the scope of its marketing authorization.
- When collecting routine practice data, treatment is provided without specific requirements.

Since the assessment of drugs under SGB V concerns the care of patients in Germany, routine practice data must meet the two aforementioned criteria in such a way that they allow conclusions to be drawn about German healthcare.

According to A19-43, the following fundamental aspects must be observed when collecting and assessing routine practice data:

- An adequate data platform is required for RPDC. Ideally, existing data structures should be used for this purpose. Disease registries are particularly suitable for RPDC for benefit assessments (registry-based studies). Alternatively, study-specific data collection represents an option for conducting RPDC.
- If routine practice non-randomized comparative studies are to be used for benefit assessments, it must be ensured as early as the study planning phase that the study design and the collected data meet the necessary quality standards to generate interpretable results.
  In this context, quality criteria for disease registries and registry-based studies that are largely consistent at both the national and international levels can be derived.
- Major components of study planning include considerations for emulating a hypothetical randomized controlled trial (RCT [target trial]) that addresses the relevant research question, including ensuring the collection of sufficient data for confounder control. The aspects of study planning must be documented in advance in a study protocol (including the SAP).
- A central aspect of the analysis of a non-randomized comparative study is adequate adjustment for confounders to compensate for the influence of structural inequality

among treatment groups. The relevant confounders must be systematically identified and prespecified in the study protocol.

- Due to the inherent uncertainty of results from non-randomized comparative studies caused by potentially unknown confounders, even when high-quality standards are met, a conclusion regarding the benefit or harm of an intervention can only be drawn from the effects observed in a study once a certain effect size is reached. A (positive or negative) conclusion regarding benefit or harm is reached when the confidence interval for the observed effect lies above or below a threshold to be defined (test for a shifted null hypothesis). The specific threshold is determined by the quality of the data in each individual case.

Building on this, the following section describes the key features of planning non-randomized comparative studies based on routine practice data. This forms the starting point for addressing the topics commissioned for this report, which are detailed in Sections 5.2 to 5.7.

For non-randomized comparative studies based on routine practice data that aim to compare the effectiveness of interventions, it is recommended to explicitly emulate the design of randomized comparative studies. The concept of target trial emulation has proven to be a best-practice approach [28,29]. The aim here is to replicate (emulate) the effect of a hypothetical RCT (target trial) addressing the research question using observational data. In this process, both the study protocol and the SAP are harmonized so that they target the same causal effect as closely as possible [30]. In Table 1 the components of study design for replicating a target trial from a non-randomized dataset are illustrated using an example.

Table 1: Comparison of an RCT study protocol and a study based on routine practice data using target trial emulation (IQWiG's own illustration based on Hoffmann 2021 [31])

Component	RCT (target trial)	Study with routine practice data (target trial emulation)
Objective	Investigation of the effect of hormone replacement therapy in postmenopausal women on breast cancer risk within 5 years	Identical to the target trial
Inclusion criteria	▪ Women who have been postmenopausal for at least 5 years ▪ No history of cancer ▪ No hormone replacement therapy in the past 2 years	Identical to the target trial
Interventions	**A:** Initiation of hormone replacement therapy **B:** No initiation of hormone replacement therapy	Identical to the target trial
Allocation	Randomization at baseline, no blinding	**Differences:** ▪ Allocation at the start of treatment (start of hormone replacement therapy) or at the time of the treatment decision (no initiation of hormone replacement therapy) ▪ Replication of randomization through comprehensive adjustment for baseline characteristics (all relevant confounders must be included)
Follow-up	**Start:** Treatment allocation **End:** confirmed breast cancer diagnosis, death, lost to follow-up, or 5 years after baseline, whichever occurs first	Identical to the target trial (a particular challenge in observational studies is the start of follow-up in the non-treatment group)
Outcome	Breast cancer diagnosis, confirmed by biopsy, within 5 years of randomization	Identical to the target trial
Effect between interventions ("causal contrast") and analysis	Effect of allocation to hormone replacement therapy vs. no hormone replacement therapy	**Different:** Effect of starting hormone replacement therapy vs. not starting hormone replacement therapy (adjustment for relevant confounders is mandatory)
RCT: randomized controlled trial		

To replicate randomized comparative studies, the dataset intended for the non-randomized comparative study must contain all necessary information in high quality [32].

Using the concept of target trial emulation at the study planning stage helps identify and avoid sources of systematic bias and emulation discrepancies (lack of sufficient alignment between the target trial and the emulation) [11,33]. Common sources of bias include a structural violation of positivity, structural inequality between groups (bias due to confounding), and an

incorrect choice of the start time of follow-up (leading, among other things, to immortal-time bias and/or selection bias).

Ensuring positivity

In an RCT, each participant is randomly allocated to a treatment strategy through randomization. Based on the target trial approach, a prerequisite for routine practice non-randomized comparative studies is that, for all patients included in the analysis, the treatments being compared (both the new drug and the comparator) represent a potential treatment option (positivity) [34,35]. This means, for example, that patients with contraindications to one of the treatments of interest must not be included in the analysis. This must be ensured by reviewing patient eligibility based on the prespecified inclusion criteria (of the target trial and the emulation) [34].

Adequate adjustment for confounders

In non-randomized comparative studies, the lack of structural equivalence between treatment groups should be compensated for by means of adequate adjustment for confounders (bias due to confounding). Bias due to confounding arises from a violation of the principles of randomization during sampling, i.e., during the allocation of patients to treatment groups. Particularly in group comparisons, bias due to confounding can lead to systematic differences between the groups. If, as a result, important confounders are unevenly distributed across the groups, the results of a comparison are generally no longer interpretable [34]. A variable is defined as a confounder if it influences both the treatment decision and the outcome and can thereby distort a treatment effect (such as age or disease severity) [3]. When comparing groups, randomization is the best method for avoiding bias due to confounding, since the resulting groups do not differ systematically in either known or unknown confounders [36].

In non-randomized comparative studies, when determining treatment effects, the systematic identification of potentially relevant confounders and their consideration in the analysis are prerequisites for adequate confounder control. The goal of this confounder control is to obtain an estimate of the causal effect of interest that is as unbiased as possible, despite the lack of randomization. To achieve this, it is necessary to identify all relevant confounders (including important interactions) in advance using a systematic approach, to document them comprehensively, and to account for them appropriately in the model during data analysis [20].

Correct choice of the start of follow-up (index date [t_0])

A key challenge when emulating a target trial is adequately determining the start of follow-up (index date [t_0]). Unlike RCTs (time of randomization), non-randomized routine practice studies do not have a fixed starting point for follow-up [23]. To address this problem, the conditions shown in Figure 2 must be met at t_0 (in both the target trial and the non-randomized study).

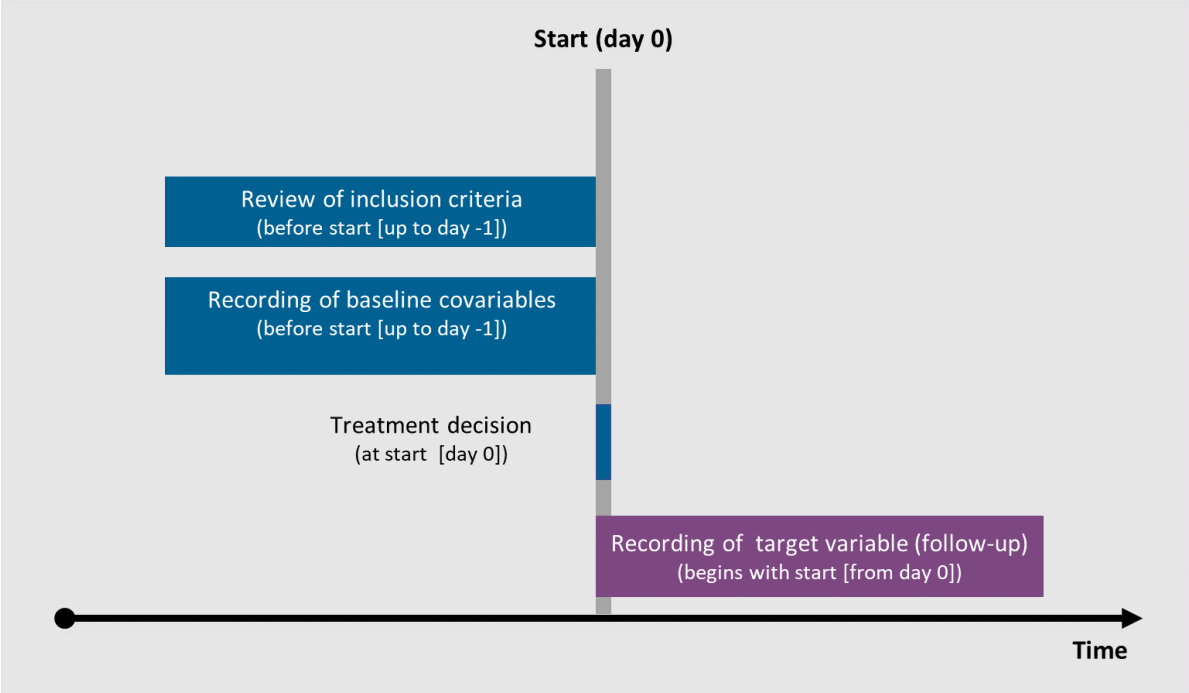


Figure 2: Temporal alignment of key elements of the study design at the start of follow-up (t_0) (IQWiG's own illustration, based on Braitmaier 2022 [37])

The inclusion and exclusion criteria must be met at the time of treatment allocation [11,28]. Ideally, the day of the treatment decision should be selected as t_0 in both groups. This requires information on treatment intention (in an RCT, this is derived from the allocation sequence) [33,38]. The follow-up period to record the outcomes begins with the treatment allocation t_0.

5.2 Identification and selection of confounders

In non-randomized comparative studies, the proper identification of and adjustment for potential confounders plays a central role. According to Pufulete 2022 [39], confounder identification comprises 3 methodological components: (1) a systematic literature review, (2) interviews with clinicians, and (3) a survey among clinicians.

In IQWiG working paper GA23-02 [27], a systematic confounder identification was conducted in accordance with Pufulete 2022 for the therapeutic indication of relapsing-remitting multiple sclerosis (RRMS) using a suitable comparison of two drug therapies with regard to key patient-relevant outcomes. The goal was to assess the basic feasibility of the confounder identification procedure proposed by Pufulete 2022. To this end, a working group of external experts was commissioned, which presented its findings in a report. Based on this report, concrete recommendations regarding systematic confounder identification in non-randomized comparative studies (and thus also in RPDC) were derived by IQWiG as part of the work on GA23-02 through further investigations.

Although the report by the external experts (as well as the work by Pufulete 2022) refers to a therapeutic indication with a large body of studies (RRMS), the situation of rare(r) diseases was also taken into account when deriving the recommendations, as this represents the typical situation, e.g., within the RPDC context. Where specific requirements apply in this regard, these are explicitly described below.

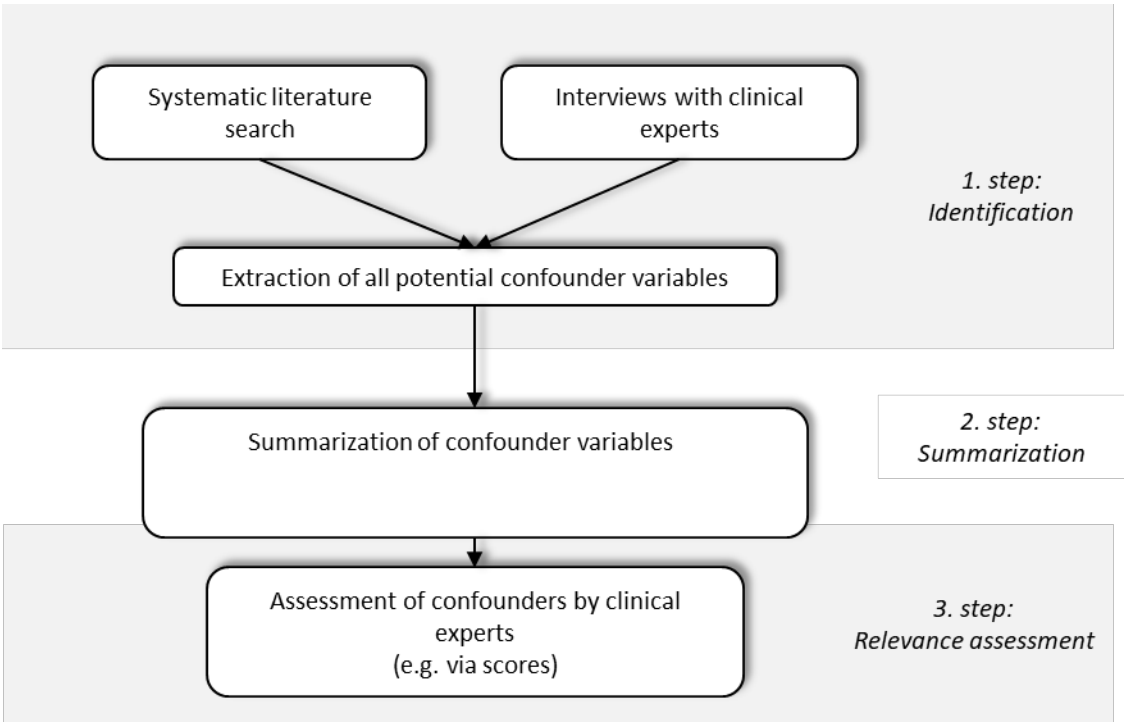


Figure 3: Proposed 3-step procedure for systematic confounder identification

Step 1: Identification through systematic literature review and interviews with clinical experts

Based on Pufulete 2022, the first step of confounder identification comprises two components: (1) a systematic literature search and (2) interviews with clinical experts. The systematic literature search should include primary publications on RCTs and cohort studies. For research questions regarding therapeutic indications with a foreseeably limited body of studies (such as rare diseases), in the absence of suitable RCTs and cohort studies, one may rely on single-arm studies (including before-and-after comparisons) and the patient characteristics listed in tabular form in the respective primary publications. Similarly, in these cases, there is the option of, for example, including guidelines in the literature search and screening them for prognostic or risk factors and for factors that determine the treatment decision. The inclusion and exclusion criteria that studies must meet to be included in the study pool for confounder extraction should be closely aligned with the research question of interest. For comparative observational studies, the extraction should account for all factors used for adjustment in the statistical analyses, regardless of statistical significance. For RCTs

or (depending on the data available) single-arm studies, the patient characteristics reported in tables (“Table 1”) should be extracted.

For data extraction from primary publications, the use of an appropriate saturation criterion is recommended to save resources, particularly when a large number of publications are identified.

The second component for confounder identification involves conducting interviews with clinical experts. Decision criteria for determining the number of interviews to be conducted may include, for example, the therapeutic indication, the expected number and quality of studies, or variability in disease progression. It can be assumed that in those therapeutic indications where there may be a smaller number of publications available, the interviews will be given greater weight. In principle, it should be transparently documented how many clinical experts from which specialties were involved. The experts involved must be named, and any potential conflicts of interest must be disclosed.

Step 2: Summarization

In the second step, after extracting all identified potential confounder variables from the systematic literature review and the qualitative interviews with clinical experts, it is recommended – where appropriate – to summarize confounder variables with overlapping content. For each of the identified confounder variables, it must be verified whether it falls within the scope of the research question of interest. If a variable is excluded from the list of identified potential confounder variables, a literature-based justification must be provided. In general, transparent and comprehensible documentation should be ensured when summarizing potential confounder variables into potential confounders. For all identified potential confounders, a suitable operationalization must be defined a priori. Although potential confounders that are not measurable and/or cannot be operationalized cannot be included in the analysis, they should nevertheless be identified and taken into account when interpreting the results.

Step 3: Assessment of the relevance of potential confounders by clinical experts

In the third step, clinical experts conduct a quantitative assessment of the potential confounders that have already been summarized to evaluate their relevance. An assessment using a scoring system is recommended. The result of this assessment step can be used in the data analysis for a content-based variable selection in cases where the statistical model does not converge. In such cases, for example, the potential confounder with the lowest score is the first to be excluded from the model.

In summary, the procedure for confounder identification developed in working paper GA23-02 is considered feasible. However, the present investigations also underscore how resource-intensive confounder identification is for non-randomized comparative studies. It is important

at this point, however, to distinguish between (1) the resources required prior to RPDC (for the identification and summarization of potential confounders) and (2) the total resources required for the RPDC (which includes, among other things, the resources required for data collection and documentation). It may be beneficial to invest more resources upfront in order to reduce the overall resources required for the RPDC. In a next step, as part of the preparation of working paper GA25-02 [40], it is examined whether further measures to reduce the resources involved in the systematic identification and recording of confounders are possible in order to increase the feasibility of the RPDC. In doing so, possible options for reducing resources must be weighed against the associated loss of information and potential loss of certainty of results.

5.3 Estimating the required sample size given insufficient prior information

The G-BA regularly commissions IQWiG to develop scientific concepts for RPDCs and their analysis. These concepts should also describe the requirements regarding the duration and scope of the respective data collection. The aim is to determine whether the planned RPDC is, in principle, feasible and can be carried out in a meaningful manner. Based on the available findings regarding the drug under assessment and the comparator, as well as the number of patients available in the respective therapeutic indication, an estimate is made as to whether a sufficient number of patients or events (i.e., the required sample size or events) can be included or observed within an acceptable timeframe to enable the RPDC to generate informative results for a benefit assessment in principle [26,34].

It should be noted that, for this purpose, no sample size planning (as for a clinical trial) can be carried out during concept development, as the necessary prior information is not available. Sample size planning for an RCT is generally based on information from pilot studies or comparative Phase II studies, which allow for a relatively reliable determination of the required sample size. However, a key criterion for requesting an RPDC is precisely that no comparative data are available in the relevant therapeutic indication. Therefore, the concept development aims rather to provide an assessment of the basic feasibility of an RPDC. Depending on the information available at the time of concept development, the approach taken involves either a preliminary sample size estimation or an exploratory analysis of sample size scenarios.

The preliminary sample size estimation is explained in more detail as one of the approaches in the following Section 5.3.1 (including a presentation of the fundamentals [parameters] of a sample size estimation). A description of the exploratory analysis of sample size scenarios is provided in Section 5.3.2. The classification of the two approaches is finally discussed in Section 5.3.3.

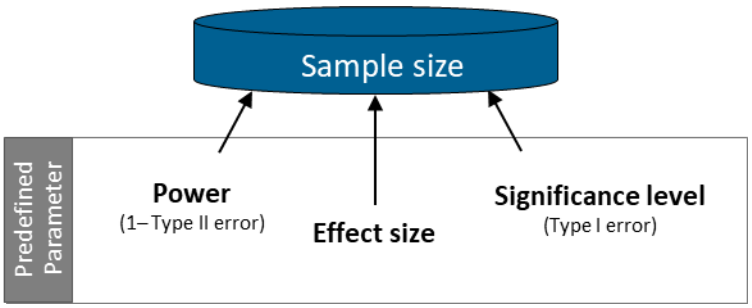
5.3.1 Fundamentals of sample size estimation

The goal of every study is to answer a scientific research question. Within the framework of evidence-based medicine, the scientific research question is guided by the so-called PICO framework (Population, Intervention, Comparison, Outcome). The study should thus enable conclusions to be drawn regarding the comparison – e.g., of the benefit – of an intervention with a comparator for a specific patient population. However, within the scope of a study, in general the entire patient population of interest is not analysed; instead, a sample that is as representative as possible is included, and the results are subsequently extrapolated to the entire patient population.

A sample size calculation is performed to estimate how large this sample should be in order to obtain informative results from a study. It serves to determine the necessary sample size.

Parameters for sample size estimation

To determine the required sample size, assumptions must be made regarding various parameters. These are described below for the effect measure "relative risk" (RR). However, similar considerations apply to other effect measures (such as the hazard ratio or the standardized mean difference). The parameters include the expected effect size in RPDC, the significance level (probability of error, also known as Type I error), and the power (1 − Type II error) [41] (see Figure 4). Here, the expected effect size in RPDC corresponds to the true treatment effect plus a systematic bias caused by the non-randomized design (e.g., due to unknown confounders).



Depending on the effect measure, additional parameters are included in the sample size estimation, such as the baseline risk for the relative risk.

Figure 4: Required sample size (N) when the parameters expected effect size, significance level (Type I error), and power (1 − Type II error) are specified

By specifying the power, the significance level, the risk in the control group, the expected effect in RPDC (RR_1) and the distribution ratio of patients between the treatment groups (intervention to comparator), the estimated required sample size is determined to infer a statistically significant difference between the intervention and the comparator, taking into account the shifted null hypothesis.

One key criterion for initiating a comparative study is insufficient evidence for a new drug compared to the standard treatment. RPDC therefore provides comparative data for the first time. Against this backdrop, it may make sense in the RPDC context not to conduct a comparative study solely with the aim of demonstrating a benefit of the new drug and to base the sample size estimation on that goal. Instead, it is conceivable to base the estimation of the required sample size on the fact that a selected propensity score method, while ensuring adequate adjustment for confounders, yields usable results in the form of an effect estimate with an associated 95% confidence interval (see also Section 5.7), regardless of whether a benefit of the drug under assessment can be inferred when taking the shifted null hypothesis into account. Under this premise, every result of RPDC represents a gain in knowledge.

Power and significance level

The assumptions for power (at least 80%) and the significance level (2.5% [one-sided test]) are generally based on standard specifications.

Expected effect size

To answer with sufficient certainty the question of what magnitude the expected effect in the RPDC (RR_1) or the risk in the control group might be, prior information is required from existing evidence on observed event rates both under treatment with the intervention and under treatment with the comparators for a (patient-)relevant outcome with appropriate operationalization in the patient population of interest.

In general, only (in some cases preliminary) data from non-comparative (single-arm) approval studies are available for the drug for which an RPDC concept is being developed [42-44]. For the comparators, study data from various sources (e.g., from publications or from manufacturer dossiers) are used. Consequently, the information for the intervention and the comparators is drawn from various studies; furthermore, the available information is usually not very detailed, particularly regarding the description of the population of interest and/or the intervention. Overall, it becomes clear that only very uncertain prior information is available for sample size estimation within the RPDC framework. For this reason, sample size estimation within the RPDC framework can, at best, serve as a guide.

Distribution ratio between treatment groups

Another parameter that must be considered when estimating the required sample size is the distribution ratio of patients between the treatment groups to be compared, i.e., establishing an assumption about how patients are likely to be distributed across the treatment groups.

In an RCT, the distribution ratio is predetermined by the randomization ratio specified in the study protocol (e.g., in the form of an intervention-to-comparator ratio of 1:1 or 2:1). In the RPDC context, where there is no active allocation to a treatment group due to the lack of

randomization, it is necessary to estimate how the market share of the drug under assessment will develop in the future (after market access). In practice, this depends on many different factors. For example, if the new drug is associated with the promise of curing an otherwise incurable disease or if the availability of effective alternative therapies is limited, patient preference for this drug is expected to be very high. In extreme cases, this can even lead to a situation where hardly any patients can be found for the control group of the RPDC. Conversely, there may be situations in which patients' conditions are well-managed with the currently available treatment options, so that the new drug – especially if it involves an entirely new therapeutic principle – gains acceptance only slowly and with delay.

It is difficult to estimate how the market share for a specific drug will develop at the time the RPDC concept is developed, which usually occurs prior to approval. These comments make it clear that the assumptions for a realistic distribution ratio in an RPDC procedure are very uncertain. Therefore, during the concept development phase, various distribution ratios are typically assumed to illustrate their impact on the required sample size. It should be noted that the required sample size increases only slightly for distribution ratios beyond 5:1 or 1:5 [45-47], so that the distribution ratios presented in the RPDC concepts generally range between 5:1 and 1:5.

Shifted null hypothesis

Another key aspect of sample size estimation in RPDC concerns uncertainty in the results due to the non-randomized study design. It is undisputed that in RPDC, even with the most careful analysis and compliance with extensive quality requirements, relevant uncertainties remain due to the lack of randomization, as unknown confounders, among other factors, may play a role [48]. To derive a conclusion regarding the benefit or harm of the drug under assessment from a non-randomized comparative study, a sufficiently large observed effect is required, which can be assumed not to be caused solely by systematic bias. In RPDC, this is to be implemented statistically in the form of a test for a shifted null hypothesis (H_0). A (positive or negative) conclusion regarding the benefit or harm of the drug under assessment is reached when the confidence interval for the observed effect lies above or below a threshold to be defined [49]. Various alternative methods, such as quantitative bias analysis, are proposed in the literature [50]. These essentially follow a similar approach and share the common goal of enabling conclusions about the certainty of the results from a non-randomized comparative study.

There is currently no scientific standard for how high the threshold for the shifted null hypothesis should be. Since the generation of routine practice data within the RPDC framework requires compliance with considerable quality requirements as a prerequisite for assessing effects, A19-43 [20] initially stipulated that this threshold should be clearly below the value for a "dramatic effect" (RR of 5 to 10 [51]), e.g., in a range of 2 to 5 for the RR (or 0.2

to 0.5, e.g., for mortality-reducing interventions) relative to the lower (or upper) confidence interval limit.

Since the publication of A19-43 (in May 2020), there has been a growing awareness among experts that high-quality data are indispensable for obtaining informative results from observational studies [52]. A methodological advancement is the systematic identification of confounders according to the procedure described by Pufulete 2022 [39]. Experience to date with reviewing the study documents submitted by the drug manufacturers regarding their suitability in terms of content and methods for RPDC conduct shows that it is possible to successfully plan non-randomized comparative studies based on routine practice data that address existing sources of systematic bias as effectively as possible.

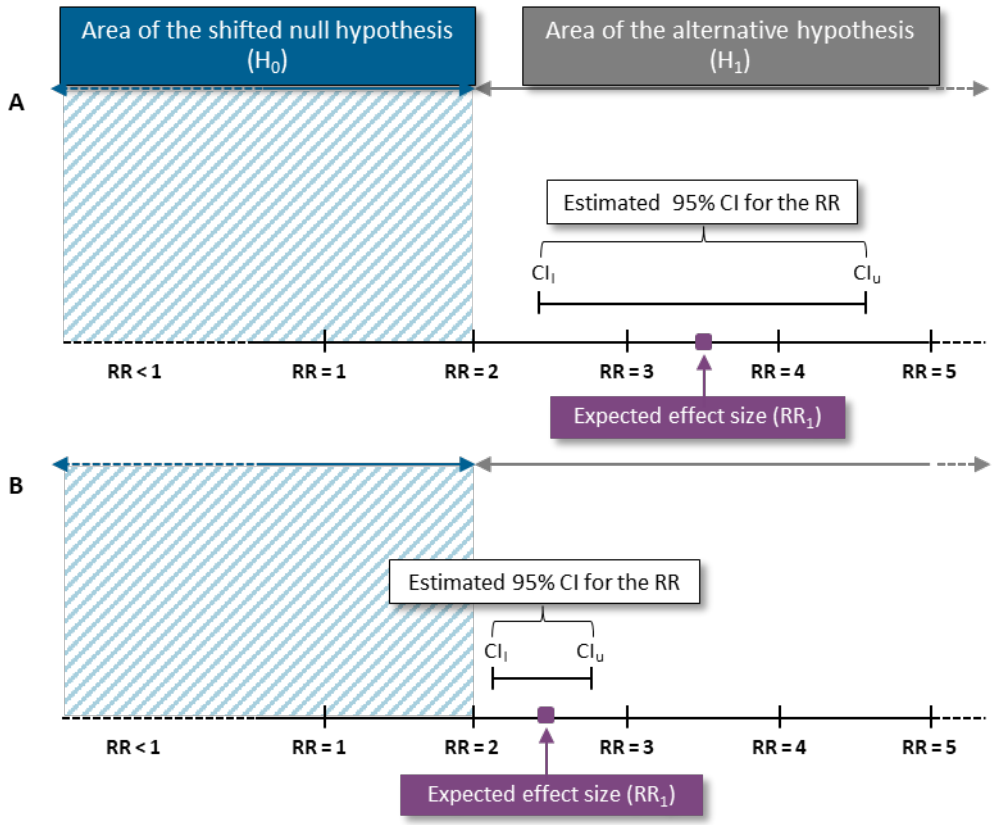
In summary, it is undisputed that relevant uncertainties remain in a non-randomized comparative study, which must be accounted for through appropriate measures – such as the application of the shifted null hypothesis (H_0) in the RPDC. Nevertheless, in light of methodological advancements, it seems appropriate to undertake, as a next step (outside the scope of this report), a review of the threshold specified in A19-43.

To demonstrate the effects of the shifted null hypothesis on the sample size estimation, the threshold specified in A19-43 is used in the following.

The starting point is the planning of a study to test the shifted hypotheses

$$H_0: RR \leq RR_0 \text{ vs. } H_1: RR > RR_0$$

based on the RR with the threshold $RR_0 = 2$ (or $H_0 : RR \geq RR_0$ vs. $H_1 : RR < RR_0$ with the threshold $RR_0 = 0.5$, e.g., for mortality-reducing interventions). A sample size estimation is then used to determine the minimum number of patients or events (i.e., required sample size) that must be included and observed in order to detect an expected effect size (RR_1) in an RPDC. This is achieved using a statistical test with shifted hypothesis boundaries (the thresholds described above) at a given significance level (Type I error) and power (1 − Type II error) [53]. As schematically illustrated in Figure 5, the following scenarios, among others, may arise:



H_0: null hypothesis; H_1: alternative hypothesis; CI: confidence interval; CI_l: lower limit of the confidence interval; CI_u: upper limit of the confidence interval; RR: relative risk; RR_0: threshold (hypothesis boundary) for the lower limit of the two-sided 95% confidence interval for the relative risk; RR_1: expected relative risk

Figure 5: Scenarios for preliminary sample size estimation with a shifted null hypothesis ($RR_0 = 2$)

- If the expected effect size (RR_1) in RPDC clearly exceeds the threshold of 2, a relatively small sample size or few events (and consequently minor precision [wide confidence interval]) is sufficient for the 95% confidence interval for the RR to lie entirely above the shifted hypothesis boundary (Scenario A).
- If the expected effect size (RR_1) only just exceeds the threshold of 2, a larger sample size or many events (and consequently high precision [narrow confidence interval]) is required for the 95% confidence interval for the RR to lie entirely above the shifted hypothesis boundary (Scenario B).

5.3.2 No or insufficient prior information for preliminary estimates of the RPDC scope

For drugs for which IQWiG is commissioned to develop a scientific concept for RPDC and to conduct analyses, the information required to estimate the effect size and baseline risk is generally not available at the time the concept is developed, meaning that the prerequisites for a preliminary sample size estimation are not met (see the following box).

Example RPDC concept in which a preliminary sample size estimation was not reasonably feasible due to insufficient prior information

On 1 February 2024, the G-BA commissioned IQWiG to develop a concept for an RPDC on odronextamab for the treatment of adults with relapsed or refractory diffuse large B-cell lymphoma (DLBCL) following at least two prior systemic therapies (RPDC concept A24-18 [54]).

As part of the concept development, two patient populations were considered in accordance with the appropriate comparator therapy specified by the G-BA:

- Adults with relapsed or refractory DLBCL following at least two prior systemic therapies for whom chimeric antigen receptor (CAR)-T cell therapy or stem cell transplantation is an option (Question 1)
- Adults with relapsed or refractory DLBCL following at least two prior systemic therapies for whom CAR T-cell therapy and stem cell transplantation are not an option (Question 2)

Since the primary treatment goal in the therapeutic indication is to prolong overall survival, the plan during concept development was to perform a preliminary sample size estimation based on the outcome of overall survival for the comparison of odronextamab with the comparators designated by the G-BA in each case.

For odronextamab, at the time the concept was developed (completed on 3 June 2024) – prior to the drug's approval in the European Union (22 August 2024) – only initial preliminary results from the ongoing pivotal Phase 2 study ELM-2 were available in the form of a press release [55]. For the comparators in the above-mentioned questions regarding odronextamab, either final results from studies published in medical journals or a benefit assessment dossier in accordance with §35a SGB V were available. However, data on overall survival were not available for either odronextamab or the comparators specified by the G-BA for the two research questions. However, due to the differing prognoses of the patient groups in the two research questions (curative vs. palliative intent), only question-specific sample size estimations allow for a meaningful assessment of the RPDC scope. For this reason, a preliminary sample size estimation was not included in the RPDC concept A24-18.

An overview of the concepts developed to date in RPDC procedures and the associated addenda shows the following constellations in particular:

- The new drug under assessment is often not yet approved at the time of concept development. Since publications on the main results of the (pivotal) study are generally not yet available in these cases, only preliminary, sometimes contradictory, results are available for outcomes presented only selectively in conference abstracts or press releases.
- The information on existing studies is not sufficiently detailed to assess whether patients were treated with the comparators in accordance with the G-BA's requirements for the appropriate comparator therapy (e.g., with regard to prior treatments).
- Results from single arms of various studies must usually be used for the intervention and the comparators. A review of the similarity of the studies (e.g., with regard to disease severity) is not possible due to the lack of publicly available specific data relevant for such a review.
- Differences between the studies available for the intervention and the comparators in the operationalization of a patient-relevant outcome, on the basis of which a preliminary sample size estimation is sought, lead to methodological uncertainties (e.g., regarding the operationalization of bleeding events in the therapeutic indication of haemophilia).
- When determining the appropriate comparator therapy (the comparators), the G-BA may subdivide the patient population (e.g., based on suitability for a particular form of therapy). Results differentiated by research question generally cannot be derived for the intervention and the comparators from publicly available documents.

Based on the described data situations, it is evident that, in the context of developing RPDC concepts, there are in many cases content and methodological problems with the approach of a preliminary sample size estimation. The necessary information for the assumptions underlying a preliminary sample size estimation is usually missing. Consequently, preliminary sample size estimations have increasingly been omitted in recent RPDC procedures, as they could not be meaningfully conducted.

Exploratory analysis of sample size scenarios

In order to be able to make statements regarding the scope of data collection even in cases where a preliminary sample size estimation is not meaningfully possible, an exploratory analysis of sample size scenarios is conducted. This essentially follows the principle of the preliminary sample size estimation. However, the starting point for assessing the basic feasibility of an RPDC is the number of patients potentially available in the relevant therapeutic indication. This is because, in cases where a sufficient estimate of the expected effect is not possible, the number of patients is generally the most reliable estimate available.

To provide an exploratory analysis of sample size scenarios, the estimated number of patients in the relevant therapeutic indication in Germany is first determined. This is based primarily – where available – on previous G-BA decisions regarding early benefit assessments pursuant to §35a SGB V in the relevant therapeutic indication. Alternatively, information on patient numbers from available IQWiG reports is used to estimate patient numbers, with which the Institute may be commissioned by the G-BA as part of the necessity evaluation for an RPDC [26]. If neither G-BA decisions nor estimates of patient numbers are available, the number of patients is estimated by IQWiG as part of the concept development. Unlike in the preliminary sample size estimation, the number of patients available in the therapeutic indication is taken as the basis for determining the detectable effect sizes. The procedure is explained in more detail below.

Due to the uncertainty of the estimate, information on patient numbers is usually presented as a range. This range is used to define a spectrum of different patient numbers that are incorporated into further calculations. It is generally not possible to estimate how many patients in the therapeutic indication are actually available for RPDC, as this is influenced by various factors (including the willingness of centres and patients to participate).

In a second step, by specifying the patient numbers, the significance level (2.5% [one-sided test]), and the power (at least 80%), the question is then answered as to which expected effect sizes (RR_1) can be detected based on the number of patients generally available for RPDC in Germany within the relevant therapeutic indication (precision of the estimate) and taking into account the risk of systematic bias due to the non-randomized comparison (shifted null hypothesis) (see Figure 6).

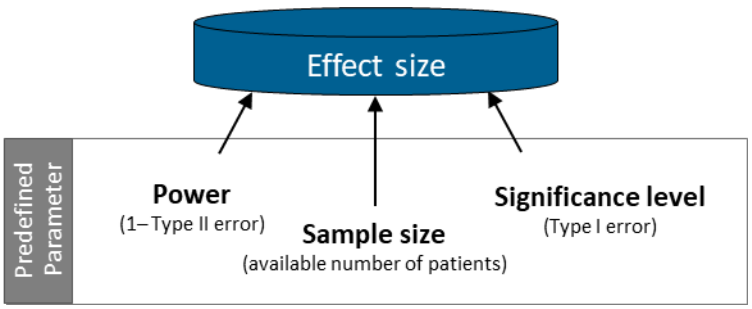
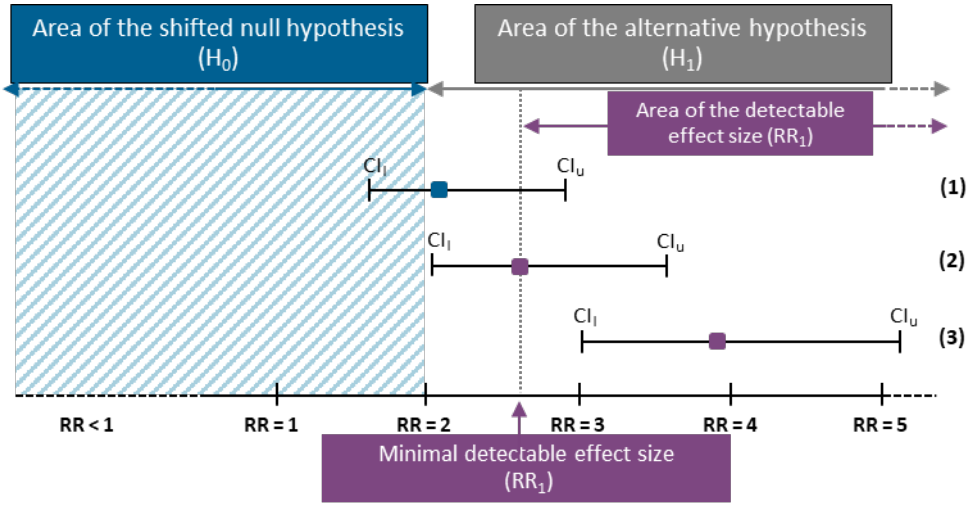


Figure 6: Detectable effect size when specifying the parameters sample size (N), significance level (Type I error), and power (1 − Type II error)

“Detectable” means that the 95% confidence interval estimated within the RPDC framework, with 80% power for a given number of patients, lies completely above ($RR_0 = 2$) or below ($RR_0 = 0.5$) the shifted hypothesis boundary. This is illustrated by the following Figure 7.



H_0: null hypothesis; H_1: alternative hypothesis; CI_l: lower limit of the confidence interval; CI_u: upper limit of the confidence interval; RR: relative risk; RR_0: threshold (hypothesis boundary) for the lower limit of the two-sided 95% confidence interval for the relative risk; RR_1: detectable relative risk

Figure 7: Detectable effect (effect measure: relative risk [RR_1]) with shifted null hypothesis (RR_0= 2)

The point estimates of effects (2) and (3) correspond to the minimum detectable effect size or lie beyond it. This means that effects of this magnitude observed in the study are statistically significant with a power of 80% given a certain number of patients, whereas an effect estimate of the magnitude of effect (1) is not.

In addition to the estimated number of patients, and provided that relevant data are available, either the baseline risk under treatment with the intervention or that under treatment with the comparators for a patient-relevant outcome is used as an anchor to illustrate a range of possible scenarios within the therapeutic indication. Whether the baseline risk in the intervention group or that in the control group is used depends on which of the treatments has information available for the patient population of interest (see the following box).

Example of an addendum to an RPDC concept in which an exploratory analysis of various sample size scenarios was conducted

Addendum A24-120 [56] to the aforementioned RPDC concept A24-18 on odronextamab is, in accordance with the G-BA's follow-up commission, limited to patients for whom CAR-T cell therapy and stem cell transplantation are not an option (Question 2). For the patient population in Question 2, there was insufficient information available at the time the follow-up commission was processed to provide a preliminary sample size estimation (neither for odronextamab nor for the comparators specified by the G-BA).

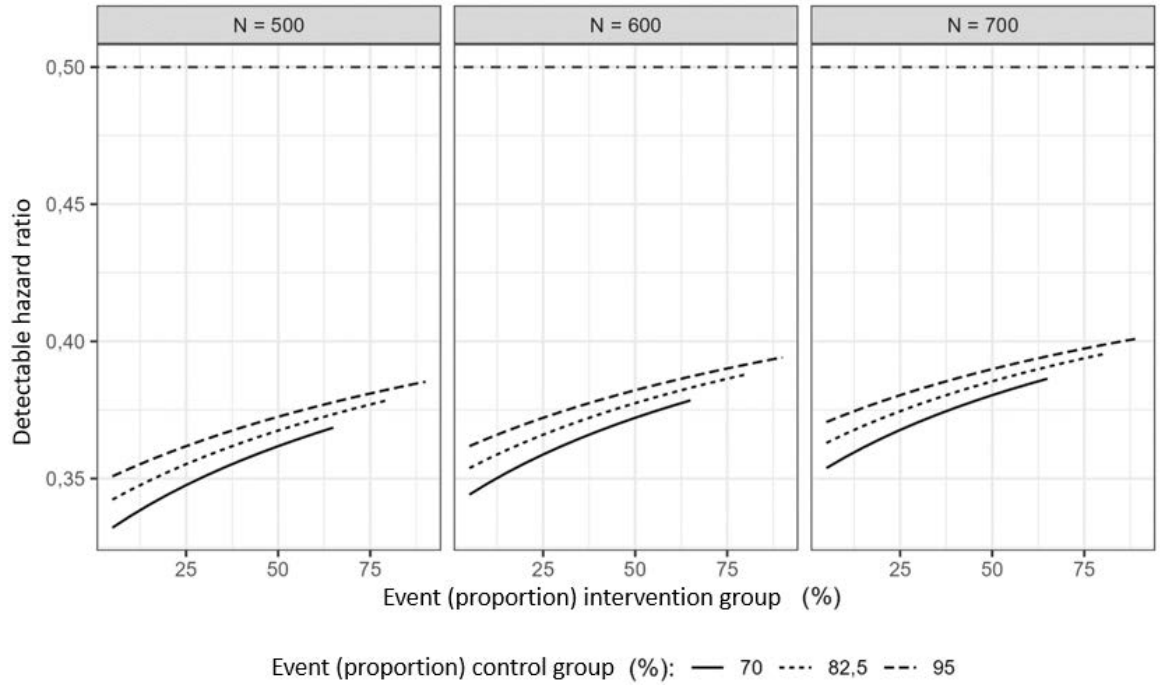
As was already the case during concept development, it was unclear:

- what proportion of patients in each study is ineligible for CAR-T cell therapy and stem cell transplantation,
- whether the respective proportions under odronextamab and the comparators are comparable, and
- whether there are differences between the studies on odronextamab and the comparators in the subpopulation of patients for whom CAR-T cell therapy and stem cell transplantation are not an option, in terms of the number of prior lines of therapy and thus the severity and progression of the disease.

To address the question of whether RPDC for odronextamab is fundamentally feasible, an exploratory analysis of various sample size scenarios was conducted. The effects of an outcome analysed using event time analyses are presented (effect measure: hazard ratio) – in this case, the outcome of overall survival. These effects can be detected with a power of 80% given the estimated patient numbers in the therapeutic indication (N = 500, N = 600, and N = 700) and a 1:1 allocation ratio (intervention to comparators).

To determine the effect detectable in RPDC (HR_1), the available evidence for the control group regarding the outcome of overall survival was used as an anchor. Based on the reviewed data, proportions of deceased patients (hereinafter referred to as the “event rate”) of 70%, 82.5%, and 95% were assumed for the control group (see “Classification of the exploratory analysis of the sample size scenarios” below). For the intervention group, the resulting event rates of 5% to 65%, 5% to 80%, and 5% to 90% are presented.

The following Figure 8 shows the effect detectable in RPDC (with a power of 80%) with a range of approximately HR_1 = 0.35 to HR_1 = 0.40 in favour of odronextamab compared to the comparators, using specified sample sizes as well as event rates in both treatment groups.



36-month follow-up period; one-sided test with a shifted null hypothesis (H_0: HR ≥ 0.5, indicated by the horizontal line at HR = 0.5) and a significance level of 2.5% (HR: hazard ratio; N: number of cases)

Figure 8: Detectable effect (effect measure: hazard ratio [HR1]) as a function of sample size and event rates in both treatment groups (intervention-to-comparator ratio of 1:1)

Classification of the exploratory analysis of the sample size scenarios

To classify the exploratory analysis of the sample size scenarios, the results for the outcome of overall survival were described from the available studies on the intervention and the comparators (including the comparators additionally designated by the G-BA for the follow-up commission).

Data on the comparators

Regarding the comparators for Question 2 submitted by the G-BA for concept development, published results on tafasitamab + lenalidomide [57] and the dossier on the benefit assessment of polatuzumab vedotin [58] were available for the relevant therapeutic indication.

For the 3 comparators glofitamab, epcoritamab, and loncastuximab tesirin, which are to be additionally considered according to the G-BA's commission, data on the outcome of overall survival were available in the benefit assessment dossiers (glofitamab [59], epcoritamab [60], and loncastuximab tesirin [61]).

After extrapolating the available data for all aforementioned comparators to the 36-month period relevant for data collection [53], event rates ranging from 72% to 91% were obtained.

Based on this, event rates between 70% and 95% were considered in the exploratory analysis of sample size scenarios for the control group.

Data on odronextamab

For odronextamab, results from the two studies ELM-2 and ELM-1 were available in the European Public Assessment Report (EPAR) [62]. After extrapolating the data to the 36-month period relevant for data collection [53], event rates ranging from 83% to 96% were observed with odronextamab. This range corresponds in part to the event rates used as event rates in the intervention group for the exploratory analysis of possible sample size scenarios (see Figure 8).

Uncertainties in the interpretability of the study results

Due to the uncertainties described above, which considerably limit the interpretability of the results, the study results to date do not allow for a sufficient assessment of the effects that would be necessary for a preliminary sample size estimation when comparing odronextamab and the comparators. However, the data can serve as reference points for the event rates in the control group in the presented exploratory analysis of sample size scenarios.

5.3.3 Decision-making for or against a preliminary sample size estimation

In summary, it is clear from the above descriptions that the approaches of preliminary sample size estimation and the preliminary analysis of sample size scenarios essentially follow the same principles and differ only in their starting point. Both approaches thus reflect different variations of the same methods, one of which is applied depending on the available prior information. This is illustrated by the following Figure 9.

Figure 9: Approaches to assessing the feasibility of RPDC: (A) preliminary sample size estimation and (B) exploratory analysis of sample size scenarios

In Figure 9, the content of Figure 4 and Figure 6 is compared. This comparison makes it clear that the parameters under consideration are identical and that only the input and output parameters (effect size and sample size) are swapped.

When developing an RPDC concept, a preliminary sample size estimation is always performed whenever it is reasonably feasible. For this, the following conditions must be met:

- The patient populations on which the evidence for the intervention and the comparator is based are sufficiently similar and each corresponds to a sufficient extent to the population of interest as defined by the research question.
- The intervention and the comparator are each applied in the studies in a manner sufficiently consistent with the research question.
- The operationalizations of the outcome analysed are sufficiently comparable between the evidence for the intervention and that for the comparator.
- The available information is sufficient to assess the aforementioned conditions.

When deciding for or against preliminary sample size estimation within the RPDC framework, IQWiG does not set any fixed limits. Whether a preliminary sample size estimation is considered feasible and appropriate given a specific data set, or whether an exploratory analysis of sample size scenarios is conducted, must be evaluated on a case-by-case basis. In principle, however, it is important to note that even a preliminary sample size estimation can, at best, provide only a rough estimate of the magnitude of the required sample size due to the uncertainties described.

Ensuring sufficient recruitment of the study population

It is the responsibility of the drug manufacturer to recruit a sufficiently large number of patients to demonstrate, with sufficient certainty within the RPDC framework, an existing difference between the intervention and the comparators.

Particularly for rare diseases, it may be sensible and necessary to collect routine practice data through international collaboration in order to recruit and observe a sufficient number of patients within an acceptable timeframe to generate informative data for benefit assessments pursuant to §35a SGB V [20]. In addition to data collection in an existing (generally suitable) European or international (disease-specific) registry, data collection in a national registry can also be supported by integrating additional (international) registries, supplemented as necessary by study-specific collection of data not available in the registry.

In this context, the data collected in the respective registry must be suitable for benefit assessments of drugs. This requires that

- the data meets the RPDC requirements in terms of scope and quality, and the analysis of the data is conducted in accordance with these requirements, and
- healthcare provision in the countries where the data are collected is sufficiently similar to that in Germany, or the findings derived from the respective registry are transferable to the situation in Germany (German healthcare context).

Final sample size estimation during the RPDC procedure

Regardless of the methodological approach chosen for concept development (preliminary sample size estimation or exploratory analysis of sample size scenarios), the drug manufacturer of the drug to be evaluated should carry out the (final) sample size estimation for the RPDC at the time of the interim analyses based on the preliminary results available at that time [63]. This approach is considered fundamentally appropriate, as the estimation of the necessary sample size in RPDC – in contrast to clinical trials – is not performed with a view to active recruitment. Instead, as many patients as possible from a registry population who meet the RPDC inclusion criteria are included in the study after giving their consent. This is also done against the backdrop that, depending on the selected propensity score method, a relevant proportion of patients may not be included in the analysis. It is not possible to estimate in advance how many patients will not be included in the analysis population.

5.4 Start of follow-up and treatment switching

5.4.1 Start of follow-up

As already explained in Section 5.1, choosing the correct starting point for the follow-up (index date [t_0]) is one of the key challenges in target-trial emulation. If t_0 is chosen incorrectly for a comparison, this can lead to erroneous results and distort the conclusions drawn from them [64]. The reason for this is what is known as time-related forms of bias. These include immortal-time bias, time-lag bias as well as latency-time bias. Risks of time-related bias arise when the elements of the study design associated with t_0 differ between treatment groups [64]. Accordingly, it is crucial that the start of follow-up (t_0), the assessment of eligibility, and treatment allocation align [11]. To address this, various approaches exist, which are described below.

Active-comparator new-user design

The study design considered the current standard in pharmacological research is the active-comparator new-user (ACNU) design [1]. In the ACNU design, only patients who have an indication for treatment switching (such as the initiation of a new line of therapy) are included in a study. This approach is associated with a reduction in bias due to confounding and selection bias. Another advantage of the design is that the time points for assessing eligibility, treatment allocation, and the start of follow-up can be aligned, and the day of the treatment decision can be defined as t_0 for both treatment groups. In cases where the treatment decision is not documented, the best possible approximation should be used. What constitutes the best possible approximation depends on the specific treatment situation. If treatment is typically initiated shortly after the treatment decision, the date of treatment initiation may be considered an adequate approximation. If mandatory preparatory steps are required before treatment begins (such as the preparation of CAR-T cell therapy), a time point should be selected that is as close as possible to the time of the treatment decision. The study protocol should specify how

to handle situations where the "t_0" differs from the time of the treatment decision. Potential discrepancies should be transparently presented in the results.

Prevalent new-user design

The prevalent new-user design (PNU) design is suitable for situations in which patients either remain on an established standard treatment (e.g., for haemophilia A and B) or can switch to a new drug. In this design, both treatment-naive patients and patients already treated with the comparator are included. Thus, 3 different patient populations can be included in the comparison of 2 treatment groups:

1) Patients who begin treatment with the intervention and were previously treatment-naive (with respect to both the intervention and the comparator) (new user)
2) Patients switching from the comparator to the intervention (prevalent new users) as well as
3) Patients who continue their treatment with the comparator.

Since it is assumed that prior treatments are associated with the respective outcome of interest, the PNU design can lead to time-related bias [65]. Thus, within the PNU design, both the choice of t_0 and exposure to prior treatment(s) (e.g., in terms of treatment duration and/or number of prescriptions) must essentially be considered as confounders [66].

Various options for selecting the start time of follow-up (index date [t_0])

As a naive approach, the time of inclusion in the study can be set as t_0 for both treatment groups. However, since this approach leads to a biased estimate of the treatment effect due to immortal-time bias [64], this approach is generally not recommended.

Another option is to define the day the treatment decision is made as t_0 for the intervention group, whereas the time of enrolment in the study is used for the control group [64,67]. This approach can lead to time-lag bias and latency-time bias, particularly when the incidence of the event of interest decreases over time, causing early events in the control group to carry greater weight in the comparison [64].

One possible solution is to conduct sensitivity analyses in which, for example, patients with events occurring early in the study are excluded from the analysis [64]. A similar recommendation can be found in addendum A23-99 to the RPDC concept A22-83 regarding etranacogene dezaparvovec for the therapeutic indication of haemophilia B (review of the study protocol and SAP [Version 1.0] [68]). Since potential time-related bias could not be ruled out, addendum A23-99 described the necessity of sensitivity analyses (in addition to the choice of t_0 proposed above). These analyses should adjust for the time patients spent under treatment with the comparator (see below for the procedure according to Suissa 2017 [21]).

Furthermore, it is conceivable to define the day of the treatment decision as t_0 for the intervention group, while for the control group, a random specification based on visit times (e.g., hospital visits) following inclusion in the study could be used [64]. Since it is unclear to what extent such an approach adjusts for time-related bias, this is generally not recommended.

A PNU approach that accounts for the time patients spent under prior treatment with the comparator was proposed by Suissa in 2017 [21]. The procedure is carried out step-by-step as follows:

- **Step 1:** A baseline population is formed that meets the study's inclusion and exclusion criteria. For all patients in this baseline population, the assumption of positivity applies.
- **Step 2:** An exposure set is formed for each patient who switched from the comparator to the intervention. For this purpose, the respective exposure to the comparator (e.g., time on comparator therapy or number of prescriptions since the start of comparator therapy) is used for each patient at the time of switching. The set then includes all patients in the control group with similar exposure to the comparator at that time [21,69].
- **Step 3:** In the PNU design, the propensity score is defined as the probability that a patient will switch from the comparator to the intervention. A time-dependent propensity score is calculated. The calculation can be performed using a time-dependent Cox proportional hazards model or an equivalent conditional logistic regression model, in each case taking all patients into account. (The literature describes an additional approach in which the time-dependent propensity score model is adjusted separately for treatment-naive and pretreated patients [70]).
- **Step 4:** Pairing is performed within an exposure set (each treatment switcher is matched with the control group patient who is most similar to the switcher in terms of the time-dependent propensity score).
- **Steps 1 to 4** result in a (matched) analysis population. The t_0 for each pair is the day of the decision to begin treatment with the intervention (e.g., initial prescription of the intervention). To avoid selection bias, Suissa 2017 [21] recommend a chronological and systematic allocation without replacement using nearest-neighbour matching in a 1:1 ratio. Depending on the data situation (e.g., a small number of patients continuing treatment with the comparator), other matching methods and allocation schemes (1:n matching, with replacement) may be selected [70].

The approach proposed by Suissa in 2017 can be modified in various ways. If patients who switch (immediately or after a delay) from the comparator to the intervention differ in terms of their prior treatment(s), methods should be chosen that take into account additional

characteristics of these prior treatment(s) through matching, stratification, or adjustment for confounders [66]. This can be done equally during the definition and formation of the exposure set [70], but potentially leads to smaller and thus more restrictive exposure sets.

A further modification was made by Yang 2022 [15]. To balance the two treatment groups with respect to all potential prior treatments, the approach by Suissa 2017 is extended by one selection step in the form of calculating medication possession ratios (MPRs) followed by matching. In the data example used by Yang 2022, however, this approach led to a substantial reduction in sample size, particularly in the control group, which can cause problems in therapeutic indications with already small patient populations.

The method according to Suissa 2017 (including the aforementioned modifications) entails various requirements for data collection. When forming exposure sets, the following verification steps must be observed [21]:

1) Within an exposure set, positivity must be verified and present for all patients.
2) All patients in the exposure set must meet the study's inclusion and exclusion criteria. For patients who switch from the comparator to the intervention during the RPDC procedure, the inclusion and exclusion criteria are re-evaluated at the time of switching. Patients who no longer fully meet the inclusion and exclusion criteria at that time may still potentially be included in the analysis population as matching partners of the control group.
3) For all patients who are included in an exposure set from the control group, a new baseline assessment must be conducted within a previously defined assessment window.

The aforementioned verification steps involve both a complex assessment of the patients' actual eligibility and a high data collection burden. This must be taken into account when planning an RPDC.

In addition, methods without time-dependent propensity scores are being discussed. Webster-Clark 2022 [65] examine different methods in a simulation study using very large datasets. These include a method using standardized morbidity (or mortality) ratio weighting (SMRW) and a method using disease risk scores (DRS). Compared to the method described by Suissa 2017 [21], these methods involve fewer resources for data collection. Both methods yield precise estimates with minimal computational resources. However, since the simulations are based on very large sample sizes, it is unclear whether the methods mentioned will be suitable for use within the RPDC framework, as relatively small patient populations are expected here.

Another approach involves creating clones of the patients [64]. In this method, two clones (copies) are created for each patient; one is allocated to the intervention group and the other to the control group. One of the clones is then censored if it no longer meets the definition of the allocated group. Since this artificial censoring leads to selection bias (informative censoring), the "inverse probability of censoring weighting" (IPCW) method is additionally used in this approach [71] (for an explanation of IPCW, see GA14-04 [72]).

Classification and recommendation

To reduce time-related bias in the analysis of routine practice data (and thus also data from RPDC), the appropriate selection of t_0 is of great importance. Depending on the therapeutic indication, determining t_0 poses a particular challenge. If patients in both groups begin treatment with a new treatment (or treatment line), the ACNU design can be used in a study. In this case, t_0 ideally corresponds to the day the treatment decision was made.

In a treatment situation where patients can either continue treatment with an established standard of care or switch to a new drug, the ACNU design may reach its limits. In such cases, the PNU design offers a suitable alternative. With this design, there is no gold standard for determining t_0. The approach according to Suissa 2017 [21] using matching (systematic, without replacement), according to Yang 2022 [15] using SMR and DRS (described in Webster-Clark 2022 [65]), and the formation of clones [64] are discussed as approaches to reducing time-related bias. The aforementioned methods represent a way to approximate the structural equivalence of the treatment groups required for the comparison of interest [64]. When using a PNU design, it should be ensured that prior treatment(s) are adequately accounted for through the selection of appropriate methods. A prerequisite is that the completeness and accuracy of the data on prior treatments in the respective registry are ensured. To verify the robustness of the respective effect estimate, further methods described above should be used to conduct sensitivity analyses.

5.4.2 Treatment switching

Allocation to treatment groups

In comparative studies (with or without randomization), patients have the option to switch from one treatment to another as needed. This process is referred to as "treatment switching." It may occur for several reasons, such as because a patient does not respond to the study treatment administered at the start of follow-up or experiences disease progression despite the treatment [72].

The term "treatment switching" can refer to switching from the comparator to the intervention, switching from the intervention to the comparator, or switching from the study treatment to a treatment outside the protocol-specified study treatment [73,74]. This rapid report exclusively considers switching from the intervention to the comparator (Scenario A)

and switching from the comparator to the intervention (Scenario B). Depending on the scenario, different consequences typically arise:

- **Scenario A:** In RPDC, switching from the intervention to the comparator generally corresponds to switching to a widely accepted treatment that conforms to the medical standard. Switching is thus part of a treatment strategy consistent with normal clinical practice. In such a situation, despite switching occurring during the course of follow-up, an unbiased estimate of the treatment effect is possible [74].
- **Scenario B:** Switching from the comparator to the intervention is unproblematic if the intervention constitutes an adequate follow-up therapy within the treatment strategy (this applies, for example, if the intervention has already been approved at an earlier stage for use in a subsequent line of therapy). However, this scenario is unlikely to occur in the RPDC context. In all other cases, switching from the comparator to the intervention can lead to biased results, and it must be demonstrated that the results are transferable to the research question of the RPDC procedure.

Regardless of treatment switching, for benefit assessments of drugs, analyses are primarily required in accordance with the treatment policy strategy (i.e., an estimate of the effect for the entire treatment strategy) and the intention-to-treat (ITT) principle [8,34]. To investigate a potentially confounding influence of switching on the ITT analysis when using survival time methods, sensitivity analyses can be conducted in which patients who switch from the control group to the intervention group are censored at the time of switching (per-protocol analyses). However, in the presence of confounding factors that influence both survival and the mechanism of treatment switching, this approach leads to informative censoring (i.e., censoring is not independent of the analysed outcome and the treatment group, but depends on patient/treatment characteristics, e.g., end of follow-up after progression). Since survival time analyses (e.g., the Cox proportional hazards model) require non-informative censoring, the approach of censoring patients who switch treatments at the time of switching leads to potentially biased effect estimates [72,74].

Furthermore, to account for treatment switching, more complex methods for estimating the treatment effect based on parametric models of the censoring mechanism are also used (e.g., IPCW). However, the application of these more complex methods is subject to a number of prerequisites and assumptions that, in practice, are generally not verifiable [72]. For this reason, results from such analyses should be presented as a supplement to the ITT analysis. Regardless of which methods are planned to account for treatment switching in a target trial, corresponding statistical procedures based on routine practice data are only appropriate if the registry dataset contains the necessary data (particularly regarding potential confounders) [75].

Special situations in the analysis of routine practice data

The G-BA frequently requires RPDCs for drugs for rare diseases (orphan drugs). This inherently results in small patient numbers. In addition, there are other special circumstances that make it difficult to recruit a sufficient number of patients for RPDC procedures for new drugs, such as minor utilization of gene therapies for the therapeutic indication of haemophilia (due, among other things, to delayed development of suitable reimbursement models [76]).

The following describes a pragmatic approach that can be adopted to include as many patients as possible receiving the treatment of interest (in terms of the feasibility of an RPDC) in the analysis. Furthermore, the handling of bridging therapies (i.e., therapies administered after treatment allocation but before the intervention of interest) is explained.

Allocation to treatment groups depending on the follow-up period under treatment with the comparator

As already explained above, the analysis of study results in the context of benefit assessments is generally conducted as an ITT analysis. This requires that all patients, regardless of any treatment switching during follow-up, be analysed according to their original group allocation. To counteract the problem of sometimes very small patient numbers in the intervention group (e.g., due to recruitment problems), it is conceivable, when conducting comparative studies to generate routine practice data, to allocate patients to treatment groups based on the duration of their follow-up period while receiving treatment with the comparator (corresponding to the time from the start of follow-up [t_0] until treatment switching).

The therapeutic indication of haemophilia can serve as an example to illustrate this approach, as only a few patients have been treated with gene therapy to date [76]. According to the information in the study documents of the RPDC regarding etranacogene dezaparvovec for haemophilia B, the following is planned [77,78]:

- Patients who have already been followed for > 2 years while on treatment with the comparator (out of a planned 3-year follow-up period), so that informative data for the control group are already available, and only then switch to treatment with etranacogene dezaparvovec, remain (in accordance with the ITT principle) in the control group, are followed until the end of the study, and are analysed in this group. To investigate the effect of etranacogene dezaparvovec in the control group, sensitivity analyses are conducted in which patients who switch to etranacogene dezaparvovec in the control group and are not allocated to the intervention group in the primary analysis are censored at the time of switching.
- Patients who switch to treatment with etranacogene dezaparvovec after a shorter period (≤ 2 years) and for whom an adequate follow-up period under treatment with etranacogene dezaparvovec is still expected, on the other hand, are to be analysed in

the intervention group. For these patients, the time of treatment switching is used as the start of follow-up (index date [t_0]), and baseline characteristics are documented again. The follow-up period for these patients under treatment with the comparator is not included in the analysis.

Handling of bridge therapies

In cases where a treatment option is not immediately available or immediately effective following the treatment decision, the use of a bridging therapy may be indicated for disease control. In previous RPDC procedures, bridging therapies have played a role in the following scenarios:

- **B-cell neoplasms:** Use of antitumour therapy between the extraction of T-cells via leukapheresis to bridge the waiting period until CAR-T cells are available [79]
- **Spinal muscular atrophy:** Pretreatment with nusinersen or risdiplam until the planned gene therapy is administered (with a prespecified clear limit defining the duration up to which bridging therapy is considered as such) [80,81]
- **Haemophilia:** Haemostatic support with exogenous human factor VIII / factor IX following infusion of the planned gene therapy to ensure an adequate supply of factor VIII / factor IX in the first weeks after treatment [77]

Bridging therapy should be viewed as part of the treatment strategy [82]. A key factor in the appropriate management of bridge therapies is that the date of the decision regarding the originally planned treatment (e.g., the date the decision to proceed with CAR-T cell therapy is made by the Tumour Board) is designated as t_0 in accordance with the concept of target trial emulation. This means that follow-up begins as of t_0 and that all events of interest occurring during treatment with the bridging therapy are also included in the analyses in accordance with the ITT principle. The reason for the use of a bridging therapy and the type of bridging therapy must be documented.

5.5 Patient-reported outcomes

The benefit of a drug, as defined by the Ordinance for the Benefit Assessment of Medicinal Products[3], is the patient-relevant therapeutic effect, including improvements in health status (symptoms) or quality of life [83]. To close existing gaps in the evidence for a new drug, including within RPDCs, the G-BA therefore routinely requires the recording of outcomes related to symptoms and health-related quality of life. These outcomes are typically assessed using PROs.

[3] “Arzneimittel-Nutzenbewertungsverordnung”

Requirements for PRO assessment instruments

For PRO assessment, instruments that have been developed and validated in accordance with established standards should be used [84]. Where possible, disease-specific instruments for collecting data on symptoms or health-related quality of life should be preferred over generic instruments.

Major challenges in collecting PRO data in routine clinical practice

Since PRO data are collected in RPDC for the purpose of benefit assessments, they must meet certain quality requirements to yield interpretable results. It should be noted that data in the RPDC context are collected in routine clinical practice. Consequently, the conditions for data collection differ significantly from those in a clinical trial. For example, patient visits occur less regularly, and patient care may take place at different levels of care (e.g., primary care and care in specialized centres). The following section outlines ways to successfully collect PRO data as efficiently as possible under these conditions. Information on approaches to avoid missing values, particularly in PRO data collection, with the aim of ensuring high data quality, can be found in Section 5.6.1.

Previous experience with PRO data collection as part of RPDC

When developing an RPDC concept, it is routinely assessed whether the relevant registry has the capability to collect PRO data. To date, experiences have shown a mixed picture. In addition to registries in which PRO data are not collected and there are no plans to do so, there are those in which PRO data are included in the regular dataset or are collected via additional modules. In some cases, solutions were found specifically for RPDCs required by the G-BA to enable the collection of PRO data. Examples of this include:

- In the RPDC procedure for brexucabtagen autoleucel for relapsed or refractory mantle cell lymphoma, PRO data collection is not directly integrated into the (primary) European Mantle Cell Lymphoma (EMCL) Registry, but rather an external institute has been commissioned to collect the data. For this purpose, a trust office and central PRO unit (based at the Institute for Medical Biometry, Epidemiology, and Information [IMBEI] at Mainz University Medical Centre) has been established [85,86]. This approach reduces the administrative burden on the registry, and the collection of PRO data is no longer necessarily tied to a visit at the treatment centre. Since the questionnaires in the RPDC are sent and received by mail (with reminders by mail and, if necessary, by phone in the event of a missing response), the potential associated with digital PRO data collection cannot be fully realized (e.g., the questionnaires completed by patients must first be entered into a database by the PRO Unit) [86]. Furthermore, the treating centres are not informed of the results of the respective PRO survey [85], even though the positive impact of using PRO data in routine care is well proven in studies (see below).

- In the German Haemophilia Registry (DHR), PRO data are not routinely collected in the dataset. In the RPDC procedure for valoctocogene roxaparvovec (indication: haemophilia A) and etranacogene dezaparvovec (indication: haemophilia B), exclusively for this procedure, the collection of PRO data was established in an external module [87,88]. The function for completing the patient questionnaires is activated following online training provided by the DHR office. Patients are provided with a (patient-specific) generated link or QR code on the day they consent to the RPDC to complete the questionnaire. The same link or QR code is used for each questionnaire (data collection every 3 months) throughout the entire duration of the RPDC. Centres can view their patients' completed PROs in the DHR database (with status [e.g., "due" or "survey interrupted"]). If, in exceptional cases, a questionnaire in a follow-up survey is completed on paper rather than electronically, the centre can enter the data from the paper questionnaire into the database [88].

PRO data collection using digital applications

A trend toward the digital collection of PRO data is emerging in the generation of routine practice data (in registries and pilot projects within routine practice care) [89,90]. In routine clinical practice, digital data collection via an app and/or a patient portal enables a low-threshold survey that helps structure the doctor-patient dialogue, independent of in-person visits [91,92]. Another advantage of digital data collection is seen in the reduction of the potential burden on patients associated with PRO measurement [91] (e.g., completed paper questionnaires do not need to be mailed). Illustrative examples of successful digital applications in registries can be found in Appendix A.

Data collection time points and tolerance windows

To generate informative results in terms of data on the course of the disease under treatment, uniform data collection time points are necessary for both treatment groups. This is particularly important when therapies that largely differ in their application are compared (e.g., CAR-T cell therapy vs. chemotherapy). Data collection should take place several times a year at standardized intervals and at defined time points [91,93]. In this context, a frequency of data collection adjusted over the course of the study should be selected (with narrower intervals at the beginning and wider intervals later on) to reduce the burden caused by frequent data collection [91]. A low-threshold yet standardized collection of PRO data within the RPDC framework can be achieved through the use of digital tools.

For the data collection time points of PROs, appropriate tolerance windows should be pre-specified in the RPDC study documents; these windows should not overlap and should allow for a meaningful allocation to a specific data collection time point. It can be assumed that returns of the PRO questionnaires are generally received closer to a specified data collection time point. For this reason, it may be advisable to select asymmetric tolerance windows that

include a shorter time window before the planned data collection time point (e.g., 2 weeks) and a longer post-data collection window (e.g., 4 to 6 weeks) (project A24-38, RPDC on etranacogene dezaparvovec [haemophilia B]: 3rd addendum to commission A22-83 [94]).

Usability of digitally collected PRO data in routine clinical practice (also outside of RPDC)

Digital tools can play a central role in clinical settings (in contrast to paper-based surveys sent by mail or email) through the digital transmission of PROs and their development over time. Reasons for this include:

- Retrieving real-time data enables symptom tracking of patients. When clinically relevant thresholds are exceeded, an alarm is triggered at the treating centre (early warning system). As a result, treatment adjustments can be initiated early, or patients can be advised by phone or invited to an appointment [90,95].
- Insights from existing PRO data (e.g., retrieved from daily updated dashboards) can be used in patient counselling and for shared decision-making [90,95,96].
- For specific oncological entities, it has been demonstrated that medical care that integrates patient-reported data throughout the course of treatment is superior to care without such information [97-103]. Among these, two studies examined the effect of incorporating digitally transmitted PRO data into routine care (in the form of symptom monitoring) on overall survival. For this outcome, a statistically significant difference was observed in each case, favouring the use of PROs for symptom monitoring over standard care [100,103].

The integration of digitally recorded PRO data into electronic case or patient records, including adequate graphical presentation for rapid data capture, is cited as a key factor for more frequent use of the data in daily clinical practice, as this can minimize the preparatory workload for the treating staff [90].

Easy-to-use, efficient digital solutions help minimize the very high personnel and logistical costs in treatment centres – which pose a major challenge, for example, if paper-based surveys are used – through the automated sending of invitations to complete questionnaires (with reminders for missing responses) and the automated collection, entry, and analysis of PRO data [86,104].

At the patient level, successful collection of PRO data requires that the process be as low-burden and as integrated into treatment as possible (see Section 5.6.1). To ensure that the resources invested are proportionate to the benefits, the use of electronic PROs is recommended whenever possible [91,92,105-107]. Other data collection methods (such as by telephone or on paper) are discussed only as a supplement for patients who would otherwise be difficult to reach [92,105,106,108,109].

An overall review of the literature screened indicates that the collection of PRO data is fundamentally feasible and meaningful in routine clinical practice. Digital collection of PRO data offers the possibility of making these data available in a practical manner and without major obstacles for studies based on routine practice data (and thus also for RPDC).

5.6 Missing values

5.6.1 Measures to prevent missing values

Missing values in scientific studies are associated with a loss of information and, if these missing values do not result from a random mechanism, with the risk of bias in the study results [16]. If necessary, statistical methods (imputation strategies) are required to adequately handle missing values. The validity of all methods decreases with an increasing proportion of missing values and depends on the plausibility of the underlying assumptions regarding why a missing value occurs. However, these assumptions are generally not verifiable for the selection of appropriate methods for handling missing values. Missing values generally cannot be completely avoided in studies. However, in order to draw sufficiently reliable conclusions about the benefits and harms of the drugs under assessment – both under routine care conditions and with regard to the specific research questions of the benefit assessment – measures to keep the proportion of missing values to a minimum are therefore both sensible and necessary.

The causes of missing values in data collection from a registry can be categorized thematically into the following 3 areas:

- registry structure,
- motivation of the centres (medical staff) to collect and transmit data, and
- willingness of patients to make their data available to the registry.

Feedback from registry operators in A19-43

A19-43 listed factors that are particularly conducive and particularly hindering to the establishment and operation of registries. In the interviews conducted with registry operators to identify these factors, the following approaches were mentioned that suggest a way to minimize incomplete data [20]:

- Establishment of electronic interfaces for automated data transfer from (external) data sources, e.g., from hospital information systems (HIS), practice management systems, or death registries (standardization of IT infrastructure)
- Ensuring semantic interoperability through the use of uniform data standards (harmonization of coding systems)
- Establishment of standardized communication channels between registry operators and participating centres (centre support)

- Regular training for data collection staff; if necessary, deployment of specially trained specialists (professional support for on-site data collection)
- Establishment of a feedback system to participating centres (usability of facility-specific data from the registry, e.g., to illustrate temporal trends or for benchmarking)
- Provision of data for research purposes and involvement of the centres (medical staff) in scientific publications

Additional measures identified

Building on the discussions in A19-43, further approaches should be identified to minimize the proportion of missing values in RPDC as much as possible. To this end, the results of previous reviews of study protocols and SAPs for RPDCs required by the G-BA, as well as the literature searched for in the context of this report, were examined. Additionally, comments by external experts regarding measures that have proven effective in registry practice were incorporated. The following approaches offer potential for increasing the completeness of the data sets:

Registry structure

- High data quality is associated with a large amount of resources required for data entry. Due to the already limited resources available for healthcare, high data quality can only be achieved on a broad scale if data collection in a registry – which must be managed by the centres alongside their daily clinical work – can be carried out as efficiently as possible. To achieve this, it is necessary to limit data collection to those data that align with the registry's objectives and are relevant to answering the research questions of a registry study. An easy-to-use, user-friendly software solution can support data entry and data transmission (software ergonomics) [110,111].
- In order to maintain and provide high-quality health data from routine clinical practice in a practical manner and without major obstacles, a permanently available and continuously maintained data infrastructure and documentation must be established in the registries with adequate personnel and financial resources. The establishment and maintenance of this infrastructure could be supported by the drug manufacturers that wish to access the data for conducting registry-based studies.
- A systematic, risk-based monitoring strategy, in which centralized monitoring (including query management) is supplemented by monitoring measures at centres with anomalies (targeted on-site monitoring), is considered an appropriate measure for improving the completeness of data sets [112,113]. Given the significant resource requirements associated with monitoring measures, initial positive experiences have been reported with the use of an automated monitoring system (IT-supported review [missing data analyses] with a query system) [114].

Maintaining the motivation of participating centres and patients

In addition to the information provided in A19-43, the following factors were identified as conducive to maintaining the motivation of the centres and patients participating in the registry for as long as possible:

Centres

- One potential incentive for ensuring the long-term usability of registry data is the visibility of high-quality care to a broader public, which arises from the transparency of the results achieved. Facility-specific data from the registry can be used for awards (certificates) or the certification of participating centres and (at the aggregate level) for public reporting (to communicate with the professional community) [90,95,111,115].
- Promoting network formation within a registry (registry community) and the exchange of knowledge to further develop expertise (e.g., in the form of regular meetings to share best practices) can contribute to increased willingness among medical staff to participate [96,111,115].
- Participating centres can be compensated for the invested resources, e.g., based on the quality of the collected data [111,115,116].
- One approach to raising awareness of data collection among medical staff involves promoting the integration of data into routine clinical practice and decision-making through the retrieval of real-time data (e.g., via an IT platform or the electronic health record [case file]) [117-123]. Dedicated teams, targeted training programmes, and graphical presentation of the data for rapid capture and minimization of resources required (through automated provision of the analysis) can support the transition toward use in routine clinical practice [90,95]. The immediate visibility of the benefits of the collected data provides an incentive to make it available in full and in high quality (see also Section 5.5).

Patients

- Incorporating patients' interests into the organization of a registry through patient representatives ensures that the collected data are relevant to patients and that participation in the registry is perceived as meaningful (participatory approach) [115,124,125].
- Patients must be informed about the benefits of data collection (including patient-reported data) for the quality of their care and for their health. An appropriate selection of instruments for collecting PRO data (based on relevance to patients), a selection of PRO instruments that are as short as possible, direct feedback on digital data entries, and the use of data for immediate treatment guidance or symptom management are

frequently cited success factors (in the sense of added value for individual patient care) [17,90,126,127].

- Presenting and communicating registry results in a way that is understandable to the general public, tailoring patient information to the target audience, and ongoing press and public relations efforts (e.g., through an informative website or newsletter) have a positive impact on willingness to participate [110,111,124,126].
- By creating a specific access point (e.g., via an IT platform [dashboard] or via an app on a smartphone/tablet), the data stored in the registry can directly be made accessible to patients in a personalized, patient-friendly, and explanatory format (individual progress data or data compared with a peer group) to encourage their engagement [90,117,119,121,125,126,128-131].
- Using a platform that offers the opportunity to access additional information about the condition or self-management and to connect with others can provide patients with an additional motivational incentive (access to information and opportunities for exchange) [119,121,126,128,132].
- In addition to non-material incentives (such as the provision of registry data), patients find incentives in the form of financial rewards (gift vouchers) or material gifts motivating [115,128,133].

Recommendations for avoiding missing values in the RPDC context

Not all of the above-mentioned measures to prevent missing values can be implemented by (individual) drug manufacturers under the current RPDC framework conditions. This includes, for example, the establishment of electronic interfaces for automated data transfer from (external) data sources. However, ensuring the implementation of some of the measures mentioned falls within the responsibility of the respective drug manufacturer. In the RPDC context, the implementation of the following measures, among others, is considered feasible:

- Limiting data collection to those data relevant to answering the research question of the RPDC procedure
- Implementing a suitable monitoring strategy, if necessary using an automated monitoring system
- Compensating the participating centres for the invested resources (e.g., based on the quality of the collected data)
- Providing (regular) training for teams at treatment centres (with regard to data collection and the active application of PRO results in clinical practice)
- Informing patients about the benefits of data collection (including patient-reported data)

- Selecting appropriate PRO instruments (including considerations regarding questionnaire length and the relevance of the questions)
- Informing patients directly about the data stored about them in the registry, in a personalized, patient-friendly, and explanatory format
- Use incentives in the form of financial rewards, such as gift vouchers, or material gifts for patients

5.6.2 Methodological handling of missing values in the context of propensity score analyses

5.6.2.1 Identification and discussion of suitable statistical methods

Propensity score methods are complex. Starting with the choice of the estimand (average treatment effect in the population (ATE) or average treatment effect on the treated (ATT)), it is necessary, on the one hand, to appropriately identify and select confounders (see Section 5.2). On the other hand, a decision must be made on both a method for handling missing values and a specific propensity score method (matching, weighting [e.g., with inverse probability of treatment weighting, IPTW], stratification, or regression adjustment).

In clinical research, the problem of missing values arises regularly, for example with outcomes or with other variables such as baseline characteristics (see Table 2).

Table 2: Pattern of observed values for N individuals at the outcome and in confounders L1 through L7

Subjects	Treatment	Outcome	L1	L2	L3	L4	L5	L6	L7	Percentage of missing values per person
1	Intervention	●		●		●	●	●		
2	Intervention	●	●	●	●	●	●	●	●	
3	Comparator		●				●			
4	Intervention	●	●		●	●		●		
5	Comparator	●	●		●	●		●		
…										
N	Comparator	●		●	●	●		●		
Proportion of missing values per outcome or confounder										
● Observed value										

From Table 2, it becomes clear that different perspectives can be taken when considering missing values. In addition to the proportion of variables with missing values per patient, one

can also consider the proportion of patients with missing values per variable or the proportion of the total data volume that ideally should have been collected. To outline the pattern of missing values in a dataset, 3 mechanisms are distinguished:

- Missing completely at random (MCAR): The probability of a value being missing is independent of other observed values for a person (e.g., covariates or values at earlier time points) and the unobserved, missing value itself. Such missing values generally do not lead to biased results. However, the precision of the effect estimates is reduced [16]. The MCAR assumption is the strongest of the 3 assumptions, under which relatively simple methods of data analysis can be applied.
- Missing at random (MAR): Taking a person's observed values into account, the probability of a value being missing does not depend on the missing value itself. Such missing values lead to biased results if this dependency is not accounted for in the analysis [16]. If information on other values for the individual (e.g., covariates or values from earlier time points) is available, this information can be used to handle missing values.
- Missing not at random (MNAR): The probability of a value being missing depends on the missing value itself, and this dependency does not disappear even when the person's observed values are taken into account. Such missing values lead to biased results. This bias cannot be corrected either, since the mechanism of missing values depends on unobserved values [16]. To handle missing values under the MNAR assumption, extensive distribution assumptions regarding the relationship between the missing value and other observed values are necessary.

In general, it is not possible to rule out the MAR and MNAR assumptions using statistical analyses. However, ignoring MAR and MNAR leads to biased results.

Models for estimating propensity scores using logistic regression require complete data for all potentially relevant confounders, meaning that patients with (at least) one missing value are excluded from the analysis. However, if only patients with complete data are included in the analyses – so-called complete-case analyses – this can lead to major bias if the MCAR assumption is not met [134,135]. This is generally to be expected in observational studies. For this reason, statistical methods for handling missing values in confounders within the context of propensity score analyses are described and classified below.

Exploratory information retrieval identified a series of simulation studies [134-144]. These studies examined various methods for handling missing values in the context of propensity score methods. Components of the simulations include the sample size (typically $N \geq 1000$), the nature of the outcomes (binary or continuous), the number (2 to 10) and nature (binary or continuous) of the confounders, as well as the correlation among them. Additionally, the

number of confounders with missing values, the proportion of missing values, and the underlying mechanism for the missing values (MCAR, MAR, or MNAR) vary. Models are examined both under the null hypothesis and with treatment effects. To compare the various methods, bias and coverage probability are typically considered, whereas the balance of confounders, overlap, and convergence of the underlying algorithms are presented descriptively.

In summary, a review of the literature confirms that multiple imputation methods have become the standard approach for handling missing data. These methods are considered appropriate for addressing the problem of missing values when applying propensity score methods. In particular, the "multiple imputation by chained equations" (MICE) method – also known as fully conditional specification (FCS) – is proposed [134-140,142,145,146]. In addition to the MICE method, there are a number of other variants of multiple imputation (model-based, regression-based, predictive mean matching, or Bayesian) that can be applied in a study.

The classic MICE approach involves the following steps:

1) For each variable included in the model, a regression model is specified, e.g., a linear regression model for a continuous variable.
2) Missing values are imputed M times by sampling the model parameters for missing data. In addition to using the observed values, the algorithm uses the imputed values of one variable to impute other missing values in other variables (chain).
3) M complete data sets are generated, and the propensity score calculation is performed separately for each data set.

The application of MICE was examined in all simulation studies considered for this report and, in a large number of the situations investigated, demonstrates better properties with regard to bias and coverage probability than alternative methods. The specific features of applying MICE in propensity score methods are discussed below.

Application of MICE

"Across" or "within" approach?

A propensity score method consists of two steps. First, a propensity score is calculated for each patient. In a second step, the treatment effect is estimated while adjusting for the propensity score (e.g., through weighting methods). If there are missing values that require multiple imputation, the question arises as to how multiple imputation is combined with these two steps. Two main approaches are proposed for this: the "within" approach and the "across" approach [134,135,137-139,141,142,146-149]. Both approaches impute missing data and estimate individual propensity scores in each imputed data set. Subsequently, in the

"within" approach, the treatment effect is estimated in each data set, and the estimates are pooled using the Rubin rule. In the "across" approach, however, the average of the individual propensity scores estimated in each complete data set is calculated, and the effect is estimated once based on the previously averaged propensity scores. Both approaches have been examined in several simulation studies. The results from earlier studies did not allow for a clear conclusion as to which of the two approaches should be used in practice. More recent studies, however, conclude that the "within" approach is preferable, as it is associated with comparatively minor bias and a higher coverage probability [135,138,149]. These advantages of the "within" approach, however, come at the cost of greater computational resources. In addition to this, assessing balance and overlap is challenging. While the final, averaged propensity scores can be used to assess balance and overlap in the "across" approach, this should be done for each imputed data set with subsequent propensity score calculation in the "within" approach [134,135]. In the event that there is insufficient balance or overlap in an imputation step, the reviewed literature describes no procedure for either the "across" approach or the "within" approach.

Influence of the number of imputations

In the simulation studies examined, a number of imputed data records ranging from 5 to 20 was used. Overall, the influence of the number of imputations is not considered to be high. Due to the data situations expected in RPDCs, with a substantially lower sample size and a higher number of confounders, a substantially higher number of imputations should be used in the RPDC context (e.g., 50). When the Bayesian extension of MICE is applied using an iterative Markov-Chain Monte Carlo (MCMC) algorithm, it should be performed with an appropriate number of iterations. In a case study, 40 iterations are proposed for the MCMC [145]. This number is appropriate.

Inclusion of the outcome in the MICE method

There is ongoing debate in the literature regarding whether, in addition to confounders, the analysed outcome should also be included in the modelling of the MICE method [137,150]. The inclusion of the outcome is supported not only by recent simulation results [135,138] but also by theoretical considerations [151].

Operationalization of confounders

There may be data situations in which the algorithm for imputation and/or the algorithm for calculating propensity scores is not appropriate. These can be situations in which sufficient balance between the two treatment groups is not achieved or the algorithm does not converge. In these cases, changing the operationalization of the confounders is an option. For continuous confounders, such as age, this can be done via transformations (log function or cubic spline function). For categorical variables, categories can be grouped together. The use of interaction terms is also possible. Eiset and Frydenberg 2022 [145] provide an illustrative

example of this. The goal of their analysis was to achieve balance for all 5 included confounders using the criterion of a standardized mean difference < 0.10. To achieve this goal, different operationalizations were selected for the underlying models, depending on the confounder.

Further approaches to handling missing values

In addition to the MICE method, there are not only other variants of multiple imputation but also methodological approaches that do not rely on multiple imputation, e.g. missing-pattern approach [152], missing-indicator method [147], general boosting modelling, and random forest-based missing imputation methods [139]. Although these methods also show some promising results in simulation studies, there is currently insufficient evidence to question the use of MICE.

Description of appropriate sensitivity analyses

As described above, MICE has established itself as the method of choice for handling missing values for confounders in propensity score methods. The advantage of this method is that it requires only the MAR assumption to be met and that only a few parameters, such as the underlying regression models and the number of imputations, are necessary for specification. Suitable sensitivity analyses include methods that are, in principle, capable of calling the results of the primary analysis into question, insofar as it cannot be ruled out that the assumptions underlying the primary analysis are fully met. Since MICE requires only the MAR assumption, other model-based methods based on the MAR or MNAR assumption are suitable for this purpose [136,139,147]. Lee 2021 [146], Ling 2020 [140] as well as Carpenter and Smuk 2021 [153] provide practical information on this. The relevant component here is generally the MAR assumption [152,153]. In addition, so-called auxiliary variables can be included [140,145], provided they are recorded in the registry. These variables are not confounders in the strict sense, but are associated with confounders, and their inclusion in the multiple imputation model can ultimately improve the imputation of missing values for the confounders [145,153]. When selecting auxiliary variables, however, it is important to note that there is a risk of collider bias [154].

In further sensitivity analyses, the operationalization of potential confounders can be varied, e.g., by transforming continuous variables [145] or by using calibration methods [143]. An instructive example of a propensity score analysis combined with multiple imputation to handle missing values can be found in the publication by Eiset 2022 [145]. This publication highlights the complexity of the procedure and discusses the individual assumptions underlying the model on which the imputations are based. Based on the points discussed by Eiset 2022, appropriate sensitivity analyses can be identified and planned. In addition, sensitivity analyses can be used to estimate the potential influence of residual confounding, e.g., through quantitative bias analysis (which includes the E-value, which measures how

strongly an unaccounted-for confounder must be associated with treatment and the analysed outcome to explain the effect) [50,155,156] or by using negative controls [157]. However, the specific assumptions underlying these methods must be adequately described in the study documents.

Complete-case analyses, which are based exclusively on patients for whom complete data are available, are not suitable as sensitivity analyses. A complete-case analysis is based on the MCAR assumption, which is a considerably stricter assumption than the MAR assumption. Simulation studies have repeatedly shown that performing a complete-case analysis leads to a substantial increase in bias compared to other methods [137,138,140,146-148].

5.6.2.2 Interpretability of analyses with missing and/or imputed values

This section outlines the maximum proportions of missing and/or imputed values for which analyses remain interpretable. The discussion pertains to both the proportion of missing values per patient and the proportion of missing values per outcome/confounder.

Thresholds for missing values for outcomes

A relevant lack of values for outcomes generally means that the affected patients cannot be included in the effect estimation. If no values are available for > 30% of the study population or the difference between the treatment groups is > 15 percentage points, the results for the corresponding outcome are generally not considered [34]. If the proportion is less than 30%, the consequences regarding the certainty of results are discussed on an outcome-specific basis. For outcomes with repeated measurements, e.g., PROs, both the value at the start of the study and at least one observed value at a later time point should be available for at least 70% of the patients. In addition to the aforementioned condition, for analyses at a specific analysis time point, values for at least 50% of the patients should be available. These thresholds serve as guidance when it cannot be assumed with sufficient certainty that the values are missing at random. Exceptions arise, for example, when data from patients at specific centres are missing for a questionnaire within a study and the centres are not considered effect modifiers, or in the presence of large treatment effects.

Thresholds for missing values for confounders

The literature provides only limited guidance on deriving specific thresholds for the proportion of missing values for confounders beyond which results can no longer be meaningfully interpreted. None of the identified studies offered specific recommendations regarding the proportion of missing values for confounders that would affect the interpretation of effect estimates. The informative value of the results from the simulation studies presented is limited with regard to the data situations expected in RPDC. The studies examine large sample sizes with ≥ 1,000 patients while simultaneously considering a limited number of confounders (mostly in the range of 2 to 5). In the RPDC context, however, small sample sizes and

considerably more than 5 confounders are to be expected. It should be noted that the number of patients as well as the number of relevant confounders (and their measurement level) influence the feasibility of multiple imputation, i.e., both the convergence of the underlying algorithm and the quality of the multiple imputation, and thus ultimately also the effect estimation performed using the respective propensity score method. Furthermore, the simulation studies did not sufficiently examine either the balance or the overlap with regard to a specific procedure [144,145,148].

The proportions of missing values vary greatly across simulation studies. The overall proportions of missing values range from 10% to 80%, while the proportions for individual confounders range from 10% to 60%. In a simulation study by de Vries and Groenwold (2017) [137], the “across” approach shows major bias ranging from 20% to 80% for the total proportion of missing values (across all confounders), whereas the “within” approach is associated with acceptable bias in the range of 20% to 40%. The IPTW method shows higher bias compared to adjustment using propensity scores or propensity score matching. In Granger 2019 [138], all propensity score methods demonstrate good performance in terms of bias across a wide range of missing values (10% to 75%). However, the results of the aforementioned simulation studies are only partially comparable, as the studies differ not only in the statistical methods used but also in the simulated scenarios.

Overall, these simulation studies yield the following thresholds as a guideline:

A complete-case analysis is appropriate provided that fewer than 5% of patients are excluded from the analysis and the total proportion of missing values does not exceed 10%. If the proportion of missing values is higher, the results may be considerably biased due to the excluded observations [140]. In cases where the proportion of missing values per confounder is less than 10% and the overall proportion of missing values is < 20%, multiple imputation may be considered appropriate. Higher proportions of missing values per confounder and/or overall may lead to lower certainty of results when multiple imputation is applied. In such cases, adequate sensitivity analyses should be performed to assess the robustness of the results of the primary analysis. In situations where the overall proportion of missing values is >50%, the data are considered unsuitable for benefit assessments.

It is not appropriate to exclude relevant confounders due to excessively high proportions of missing values. In these situations, an added benefit can only be inferred in the case of very large effects, i.e., a dramatic effect.

5.6.2.3 Requirements for presentation

The complex question – dealing with missing values in propensity score – gives rise to specific requirements for reporting. Imputation of missing values can lead to biased effect estimates, even if the algorithm itself is successful, since the imputations are made under assumptions

that generally cannot be verified. This must be taken into account when interpreting the results for the effect estimates.

Requirements for the presentation of the methods in the study protocol and statistical analysis plan

- Methodological details regarding the implementation of multiple imputation
  - Justification for the choice of method for handling missing values
  - Description of the model for the primary analyses, including operationalizations of the confounders
  - Number of imputed data sets
  - Number of iterations for the MCMC algorithm of a Bayesian approach and/or number of iterations for a bootstrap procedure (if applied)
- Methodological details regarding sensitivity analyses
- Description of how balance and overlap were handled in the imputations
- Description of criteria for interpreting results with regard to certainty of results

Reporting requirements for benefit assessments

The reporting should include the following aspects:

- Methodological details regarding the implementation of multiple imputation, in particular the approaches and parameters pre-specified in the study protocol and/or SAP
- Descriptive presentation of missing values prior to multiple imputation
- Descriptive presentation of the characteristics after multiple imputation
- Description of the convergence of the underlying algorithms for multiple imputation and propensity score calculation, as well as justifications for the operationalization of the included confounders [145]
- Justifications for methodological aspects that could not be specified in advance in the study protocol and/or SAP [145]
- Appropriate descriptions of the approach regarding balance and overlap, as well as their presentation
- Complete presentation of the planned sensitivity analyses
- Discussion of the certainty of results for the effect estimates, taking into account possible bias due to missing values

5.7 Propensity score analyses in therapeutic indications with small patient populations

In non-randomized studies, the structural equivalence of the groups to be compared – which is necessary for a fair comparison – is generally not present. Therefore, group differences in potential confounders must be accounted for when estimating effects by adjusting for these confounders [20]. Various causal methods (such as multifactorial regression models, instrumental variables, and propensity scores) are available for data analysis with adequate adjustment for confounders.

A sufficiently large number of patients is required to ensure adequate adjustment for confounders. Although there are no recommendations in the literature regarding how many patients are needed for the analysis, it can generally be stated that the required number of patients increases as the number of confounders rises. The analysis of non-randomized comparative studies therefore poses a particular challenge when dealing with small patient populations. This applies generally to all statistical methods for confounder control. In the following sections, given their central importance, reference is made exclusively to propensity score methods for confounder adjustment. Notwithstanding the above, IQWiG will, in every review of study documentation and in every assessment, evaluate in an unbiased and open-ended manner whether adequate confounder adjustment was planned and carried out, regardless of the causal inference method chosen by the drug manufacturer.

In the RPDC context, the propensity score method plays a particularly important role in confounder adjustment [49], among other reasons because there are approaches here to test and present the central assumptions of causal inference (i.e., sufficient positivity, overlap, and balance) [158]. Propensity score methods were originally developed for large datasets, some of which included several thousand patients (for epidemiological questions) [159]. However, they are increasingly being used to analyse non-randomized comparative studies based on smaller sample sizes. Since the number of patients available in the RPDC context is usually small, the question arises as to what extent propensity score methods can also provide informative results with small sample sizes in the range of 100 to 500 patients, and what special considerations must be taken into account when dealing with small sample sizes.

In recent years, a number of studies on propensity score methods for small sample sizes have been published, in which the influence of various parameters was examined using simulation studies, some of which were very extensive. These parameters include:

- the total sample size and the distribution ratio of patients between the groups,
- the effect measure (such as odds ratio or mean difference),
- the type of propensity score method (matching, weighting, stratification, or adjustment),
- the number of covariates and their nature (binary or continuous), and

- the model for the relationships between covariates, treatment, and analysed outcomes.

Due to the large number of possible configurations arising from the high-dimensional parameter space, even large simulation studies can only cover a limited subset of scenarios. This applies in particular to the strength of the covariates' influence on treatment and the outcome.

Simulation studies on propensity score analyses with small sample sizes

Through exploratory information retrieval, simulation studies on propensity score analyses with small sample sizes were identified. These are described below:

In a large simulation study by Pirracchio (2012) [160], the statistical properties of the two most common propensity score methods – matching and weighting using IPTW – were compared for the effect measure "odds ratio". The study considered sample sizes ranging from 40 to 1000 and up to 4 covariates with varying effects on the treatment and the analysed outcome. For all simulations, the authors assumed a baseline risk of 50%. The study shows that Type I error is well controlled even with small sample sizes, provided that all confounders are accounted for in the model. The relative bias (i.e., the percentage deviation of the mean effect estimate from the simulations from the assumed effect in the simulation model) is less than 10% even with sample sizes of 40 patients, which is an acceptable level. In some cases, propensity score weighting using IPTW performs slightly better than propensity score matching.

A simulation study by Friedrich (2020) [161] examined various propensity score methods for small sample sizes (N = 40, N = 100, and N = 1000) across different scenarios involving up to 9 covariates with varying effects on the treatment and the analysed outcome. The study compared the effect measures "odds ratio" and "risk difference". The results show that models using the odds ratio are, in some cases, more biased than those using the risk difference. Furthermore, the study indicates that matching procedures using both effect measures often lead to convergence problems, particularly with very small sample sizes (N = 40). The smaller the selected sample size, the higher the proportion of scenarios in which the models do not converge for the effect estimates. The authors of this simulation study therefore recommend using the risk difference as the effect measure for small sample sizes. Of the adjustment methods considered in their study, they recommend doubly robust estimators for quintiles and advise against propensity score weighting using IPTW due to a low coverage probability.

A low coverage probability for propensity score weighting using IPTW is also evident in a simulation study by Austin 2022 [162]. It should be noted that this study considered only sample sizes of at least 250. For various scenarios, each with 5 continuous and 5 dichotomous covariates, the coverage probability was compared for, among others, the 3 weighting

methods IPTW, matching weights, and overlap weights. The effect measures selected were mean difference, risk difference, and relative risk. Particular attention was paid to the question of the extent to which a variance estimate using bootstrap is superior to the asymptotic variance estimate. While the matching weights and overlap weights weighting methods each show a good coverage probability when using an asymptotic variance estimate (with no improvement when using bootstrap), the coverage probability for the IPTW weighting method can be considerably improved using bootstrap, so that it approaches that achieved with matching weights and overlap weights. This is particularly true when the ratio of sample sizes between the groups varies greatly. When using bootstrap methods, the question of the order of drawing bootstrap samples and generating multiple imputations arises in the handling of missing values. In a simulation study by Schomaker (2018) [163], the approach of first drawing bootstrap samples and then imputing each sample multiple times proves to be the most reliable, but also the most computationally intensive, method. According to the results, the approach of first generating multiple imputations and then drawing bootstrap samples from each imputation also appears to yield valid results, provided that the proportion of missing values is not too high.

In a simulation study by Wilkinson 2022 [164], 5 scenarios with varying overlap of propensity scores for 1 dichotomous covariate and 4 continuous covariates, as well as sample sizes starting at 100 for the effect measure “odds ratio”, were analysed. Additionally, the distribution ratio between the groups varied from 1:20 (highly unbalanced) to 1:1 (balanced). In addition to direct adjustment using logistic regression, the 3 propensity score methods – matching, weighting using IPTW, and adjustment – were compared. For all methods, it was found that the coverage probability decreases as the imbalance of patients between groups increases, with the decrease being more pronounced for matching compared to weighting and adjustment. Furthermore, as imbalance between groups increases and overlap in propensity scores decreases, effect estimation may become impossible in some cases, as the models used for estimating the effect fail to converge. While convergence is very high for weighting and adjustment, the convergence rate for matching methods can drop to 50% to 70% [164], particularly with moderate overlap, even when group sizes are balanced.

Interpretation of the simulation study results and recommendations

The scenarios and parameter specifications of the 4 simulation studies described above differ from one another, in some cases substantially, and no single study provides a comprehensive picture.

Furthermore, the scenarios examined in the studies do not fully reflect the situations that have arisen in previous RPDC procedures. Although the 4 simulation studies examined the propensity score methods (matching and weighting) frequently used in RPDCs, the number of relevant confounders in the ongoing RPDC procedures is in some cases higher (5 to 22

confounders [77,79-81,165]) than that underlying the aforementioned simulation studies (4 to 10 confounders [10 confounders for a sample size of N ≥ 250]). Moreover, the effect measures “odds ratio” and “risk difference” were primarily examined. In the early benefit assessment of drugs (and thus also in RPDC), however, the RR, the hazard ratio, or the (standardized) mean difference are the standard effect measures. Only the study by Austin 2022 [162] used the two effect measures “mean difference” and “RR” used. For these effect measures, this study shows no major differences compared to the odds ratio or the risk difference. Whether this holds true in general cannot be assessed based on the identified simulation studies. Overall, it is therefore not possible to assess the extent to which the results of the simulation studies can be applied to the RPDC context.

In summary, the aspects mentioned above make it difficult both to compare the simulation results with one another and to derive general recommendations for RPDC.

The following section discusses various influencing factors that must be taken into account when using a propensity score method.

Influence of the number of confounders

To minimize bias caused by confounding, it is essential in propensity score methods to identify all potentially relevant confounders and account for them in the analysis (see Section 5.2). This is supported by the results of simulation studies by Pirracchio 2012 [160] and Friedrich 2020 [161], in which models that did not include all confounders yielded considerably worse results, particularly with regard to bias and coverage probability. However, the studies also show that when using matching procedures with small sample sizes, the number of confounders that can be included in the models is limited, as otherwise this may lead to a lack of convergence of the models for the effect estimates and, consequently, to the inability to calculate effect estimates. In addition, continuous confounders should be included in the model as continuous variables for the estimation of propensity scores and should not be dichotomized. The reviewed literature does not indicate how many patients per confounder to be included must be enrolled in a study to address the convergence problem.

The problem of lack of convergence due to an excessive number of confounders can be addressed to a certain extent by prespecifying an order of the confounders based on their importance (for an example, see Section 5.2, Step 3). If the models do not converge when all originally identified confounders are included, less important confounders must be gradually removed from the model. It should be noted that the uncertainty of the results increases as more confounders are omitted. This increased uncertainty must be taken into account when interpreting the results, e.g., by further shifting the null hypothesis. However, this approach is of limited practical use. If too many confounders are excluded in this manner, sufficient confounder adjustment can no longer be assumed, and the results are interpretable only in

the case of very large effects – i.e., dramatic effects. The data situation must be assessed on a case-by-case basis.

Selection of the propensity score method for small sample sizes

With regard to the selection of an appropriate propensity score method for small sample sizes, studies show that propensity score methods based on weighting or adjustment have advantages over matching methods, particularly with regard to convergence issues. When using the common IPTW weighting method, the coverage probability is too low in the case of small sample sizes and low baseline risks; however, this can be compensated for by variance estimation using bootstrapping. For the risk difference, the doubly robust estimator for quintiles offers advantages over IPTW.

The identified simulation studies cover only a small fraction of possible scenarios. For this reason, no clear recommendation for the use of propensity scores for small sample sizes that is applicable to all situations can be derived from the existing literature. In principle, the studies show that the application of propensity score methods is possible for small patient populations. At present, however, no simulation studies are available that adequately reflect the conditions under which an RPDC is conducted, as these studies predominantly examine different effect measures and, in some cases, only a small number of potential confounders. Consequently, there is currently no evidence to assess under which conditions propensity score analyses in RPDC yield interpretable results.

Summary

Propensity score methods were originally developed for very large datasets. The identified simulation studies show that these methods are also applicable in therapeutic indications with small patient populations under certain conditions and can lead to interpretable results. A relevant problem, however, is the risk of a lack of convergence in the models used for effect estimation. This problem can be addressed through various measures, though in some cases this entails accepting increased uncertainty. The scenarios examined in the simulation studies do not fully reflect the situations that are foreseeable in ongoing RPDC procedures. At this point in time, it therefore remains unclear in which cases (and under which conditions) propensity score analyses will yield interpretable results in RPDC procedures.

6 Conclusion

Benefit assessments of drugs require data for comparison with the standard treatment. Since the approval of orphan drugs is often based on non-comparative data, the RPDC approach was introduced with the aim of closing existing evidence gaps and thus obtaining a better evidence base for benefit assessments. Data collection must be conducted as non-randomized comparative studies. Provided certain quality requirements are met, studies based on registry data can close this evidence gap.

- When planning non-randomized comparative studies based on routine practice data, target trial emulation is a recommended approach to minimize systematic (avoidable) bias. A prerequisite for optimal emulation of a hypothetical RCT using observational data is that the necessary data are available in the registry dataset with the required completeness and depth. High data quality can only be achieved on a broad scale if the generation and utilization of registry data are feasible and resource-efficient. The establishment and maintenance of a permanently available (operational) data infrastructure is considered beneficial. This could be supported by drug manufacturers that wish to draw on the registries to conduct registry-based studies.
- In non-randomized comparative studies aimed at comparing treatment effects, adequate control for confounders requires the systematic identification of relevant confounders and their consideration in the analysis. Confounder identification following the approach by Pufulete 2022 (via a systematic literature review and clinician involvement) is considered feasible and represents a meaningful approach in principle. Before clinical experts assess the relevance of the confounders, it is recommended to conduct an intensive summarization of the identified confounders. In principle, it may be beneficial to invest more resources in reducing the number of confounders to be recorded prior to an RPDC procedure in order to reduce the overall resources required through resource savings in data collection and analysis.
- In RPDC concepts, it is estimated whether a sufficient number of patients can be enrolled within an acceptable timeframe to enable informative results to be generated for benefit assessments. In general, only uncertain information is available for this estimation. If sufficient information on the intervention and the comparator is available, a preliminary sample size estimation is performed for this purpose. If necessary information is missing for the assumptions underlying a preliminary sample size estimation, an exploratory analysis of sample size scenarios is conducted to demonstrate potential detectable effects. Both approaches follow the same principle (they are based on identical parameters) and differ only in the parameter to be estimated using the remaining parameters.

- For long-term data collection (and patient follow-up) in routine clinical practice, incentives are required to compensate for the resources required for data generation and to motivate both the centres and the patients to collect data as completely as possible.
- To ensure that the resources required for PRO data collection are proportionate to the benefits, digital surveys (e.g., via an app or patient portal) are recommended. This enables low-threshold PRO data collection independent of doctor visits. Digital tools are already being successfully used for data collection in registries and in routine clinical practice and enable PRO data to be made available for research purposes with minimal barriers.
- A challenge in analysing routine practice data without randomization is determining the start of follow-up (index date [t_0]). If a new treatment is initiated in both treatment groups at t_0 (switch indication), the ACNU design can be used in a study. Ideally, t_0 corresponds to the day the treatment decision was made. In a treatment situation where patients in the control group continue treatment with an established standard treatment, the PNU design represents a suitable alternative.
- Regardless of the choice of design and analysis strategies, the results of the ITT analysis (in accordance with the treatment policy strategy) should generally be presented as the primary results. This requires that all patients (regardless of any treatment switching at some point during follow-up) are analysed according to their original group allocation. The confounding influence of treatment switching during the RPDC procedure can be addressed with sensitivity analyses.
- A commonly used method for accounting for confounders in non-randomized comparative studies based on registries is an analysis using propensity scores. Since models for estimating propensity scores using logistic regression require complete data for all potentially relevant confounders, statistical methods for handling missing values are necessary. The MICE method is recognized as an established method. Propensity score methods are also applicable in small patient populations under certain conditions and can yield interpretable results. The scenarios examined in the identified simulation studies do not fully reflect the situations that are foreseeable in ongoing RPDC procedures. At this point in time, it therefore remains unclear in which cases (and under which conditions) propensity score analyses will yield interpretable results in RPDC procedures.

References for English extract

Please see rapid report for full reference list.

1. Stürmer T, Wang T, Golightly YM et al. Methodological considerations when analysing and interpreting real-world data. Rheumatology 2020; 59(1): 14-25. https://doi.org/10.1093/rheumatology/kez320.

2. Greifer N, Stuart EA. Choosing the Causal Estimand for Propensity Score Analysis of Observational Studies [online]. 2021 [Accessed: 18.06.2025]. URL: https://arxiv.org/pdf/2106.10577.

3. Rhodes AE, Lin E, Streiner DL. Confronting the Confounders: The Meaning, Detection, and Treatment of Confounders in Research. Can J Psychiatry 1999; 44(2): 175-179. https://doi.org/10.1177/070674379904400209.

4. Puhr R, Heinze G, Nold M et al. Firth's logistic regression with rare events: accurate effect estimates and predictions? Stat Med 2017; 36(14): 2302-2317. https://doi.org/10.1002/sim.7273.

5. Hartung J, Elpert B, Klösener KH. Statistik; Lehr- und Handbuch der angewandten Statistik. München: Oldenbourg Wissenschaftsverlag; 2009.

6. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. Am J Epidemiol 2011; 174(5): 613-620. https://doi.org/10.1093/aje/kwr143.

7. Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. BMJ 2019; 367: l5657. https://doi.org/10.1136/bmj.l5657.

8. European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials [online]. 2020 [Accessed: 18.03.2025]. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf.

9. Lévesque LE, Hanley JA, Kezouh A et al. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 2010; 340: b5087. https://doi.org/10.1136/bmj.b5087.

10. Suissa S. Immortal Time Bias in Pharmacoepidemiology. Am J Epidemiol 2008; 167(4): 492-499. https://doi.org/10.1093/aje/kwm324.

11. Hernán MA, Sauer BC, Hernández-Díaz S et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 2016; 79: 70-75. https://doi.org/10.1016/j.jclinepi.2016.04.014.

12. Kuss O, Blettner M, Börgermann J. Propensity Score: an Alternative Method of Analyzing Treatment Effects. Dtsch Arztebl Int 2016; 113: 597-603. https://doi.org/10.3238/arztebl.2016.0597.

13. Pottegård A, Friis S, Stürmer T et al. Considerations for pharmacoepidemiological studies of drug-cancer associations. Basic Clin Pharmacol Toxicol 2018; 122(5): 451-459. https://doi.org/10.1111/bcpt.12946.

14. Suissa S, Dell'Aniello S. Time-related biases in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2020; 29(9): 1101-1110. https://doi.org/10.1002/pds.5083.

15. Yang CY, Kuo S, Lai ECC et al. Three-step matching algorithm to enhance between-group comparability and minimize confounding in comparative effectiveness studies. Nature 2022; 12: 214. https://doi.org/10.1038/s41598-021-04014-z.

16. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Hoboken: Wiley; 2020.

17. Mack C, Su Z, Westreich D. Managing Missing Data in Patient Registries; White Paper; Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition [online]. 2018 [Accessed: 27.06.2025]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/wysiwyg/missing-data-registries-guide-3rd-ed-addendum-white-paper.pdf.

18. Azur MJ, Stuart EA, Frangakis C et al. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011; 20(1): 40-49. https://doi.org/10.1002/mpr.329.

19. Hashimoto Y, Yasunaga H. Theory and practice of propensity score analysis. Ann Clin Epidemiol 2022; 4(4): 101-109. https://doi.org/10.37737/ace.22013.

20. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzepte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V; Rapid Report [online]. 2020 [Accessed: 07.03.2025]. URL: https://www.iqwig.de/download/a19-43_versorgungsnahe-daten-zum-zwecke-der-nutzenbewertung_rapid-report_v1-1.pdf.

21. Suissa S, Moodie EE, Dell'Aniello S. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. Pharmacoepidemiol Drug Saf 2017; 26(4): 459-468. https://doi.org/10.1002/pds.4107.

22. Sterne JAC, Hernán MA, McAleenan A et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.5; Chapter 25: Assessing risk of bias in a non-randomized study [last updated October 2019] [online]. 2024 [Accessed: 13.03.2025]. URL: https://training.cochrane.org/handbook/current/chapter-25.

23. Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. JAMA 2022; 328(24): 2446-2447. https://doi.org/10.1001/jama.2022.21383.

24. Suissa S, Azoulay L. Metformin and the risk of cancer; Time-relared biases in observational studies. Diabetes Care 2012; 35(12): 266-273. https://doi.org/10.2337/dc12-0788.

25. SGB V Handbuch: Sozialgesetzbuch V; Krankenversicherung. Altötting: KKF-Verlag; 2020.

26. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses [online]. URL: https://www.g-ba.de/richtlinien/42/.

27. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Systematische Confounderidentifikation in der Indikation schubförmige remittierende multiple Sklerose (RRMS); Arbeitspapier [online]. 2025 [Accessed: 27.06.2025]. URL: https://doi.org/10.60584/GA23-02.

28. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016; 183(8): 758-764. https://doi.org/10.1093/aje/kwv254.

29. Matthews AA, Danaei G, Islam N et al. Target trial emulation: applying principles of randomised trials to observational studies. BMJ 2022; 378: e071108. https://doi.org/10.1136/bmj-2022-071108.

30. Lodi S, Phillips A, Lundgren J et al. Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples With Apples. Am J Epidemiol 2019; 188(8): 1569-1577. https://doi.org/10.1093/aje/kwz100.

31. Hoffmann F, Kaiser T, Apfelbacher C et al. Versorgungsnahe Daten zur Evaluation von Interventionseffekten: Teil 2 des Manuals. Gesundheitswesen 2021; 83: 470-480. https://doi.org/10.1055/a-1484-7235.

32. Desai RJ, Wang SV, Kattinakere Sreedhara S et al. Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center. BMJ 2023; 384: e076460. https://doi.org/10.1136/bmj-2023-076460.

33. Mathes T. Anforderungen an die Daten für die Target-Trial-Emulation: Eine Diskussion unter Betrachtung von Patientenregistern. GMS Med Inform Biom Epidemiol 2024; 20: Doc03. https://doi.org/10.3205/mibe000259.

34. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden; Version 7.0 [online]. 2023 [Accessed: 07.03.2025]. URL: https://www.iqwig.de/methoden/allgemeine-methoden_version-7-0.pdf.

35. Zhu Y, Hubbard RA, Chubak J et al. Core Concepts in Pharmacoepidemiology: Violations of the Positivity Assumption in the Causal Analysis of Observational Data: Consequences and Statistical Approaches. Pharmacoepidemiol Drug Saf 2021; 30(11): 1471-1485. https://doi.org/10.1002/pds.5338.

36. Higgins JPT, Savović J, Page MJ et al. Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. p. 205-228.

37. Braitmaier M, Didelez V. Emulierung von "target trials" mit Real-world-Daten; Ein allgemeines Prinzip, um den Herausforderungen von Beobachtungsdaten zu begegnen. Prävention und Gesundheitsförderung 2022. https://doi.org/10.1007/s11553-022-00967-9.

38. Brookhart MA. Counterpoint: The Treatment Decision Design. Am J Epidemiol 2015; 182(10): 840-845. https://doi.org/10.1093/aje/kwv214.

39. Pufulete M, Mahadevan K, Johnson TW et al. Confounders and co-interventions identified in non-randomized studies of interventions. J Clin Epidemiol 2022; 148: 115-123. https://doi.org/10.1016/j.jclinepi.2022.03.018.

40. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Evaluation von Maßnahmen zur Aufwandsreduktion bei der systematischen Identifikation und Erhebung von Confoundern [online]. 2025 [Accessed: 04.09.2025]. URL: https://www.iqwig.de/projekte/ga25-02.html.

41. Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat Med 1990; 9(12): 1447-1454. https://doi.org/10.1002/sim.4780091208.

42. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzept für eine anwendungsbegleitende Datenerhebung – Brexucabtagen autoleucel; Rapid Report [online]. 2022 [Accessed: 07.03.2025]. URL: https://www.iqwig.de/download/a21-130_anwendungsbegleitende-datenerhebung-brexucabtagen-autoleucel_rapid-report_v1-0.pdf.

43. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzept für eine anwendungsbegleitende Datenerhebung – Valoctocogen Roxaparvovec; Rapid Report [online]. 2022 [Accessed: 07.03.2025]. URL: https://www.iqwig.de/download/a22-20_anwendungsbegleitende-datenerhebung-valoctocogen-roxaparvovec_rapid-report_v1-0.pdf.

44. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Etranacogen Dezaparvovec (Hämophilie B) – Bewertung gemäß § 35a SGB V; AbD-Konzept [online]. 2023 [Accessed: 07.03.2025]. URL: https://www.iqwig.de/download/a22-83_etranacogen-dezaparvovec-haemophilie-b_abd-konzept_v1-1.pdf.

45. Walter SD. Matched case-control studies with a variable number of controls per case. Appl Statist 1980; 29(2): 172-179. https://doi.org/10.2307/2986303.

46. Breslow NE, Lubin JH, Marek P et al. Multiplicative models and cohort analysis. J Am Stat Assoc 1983; 78(381): 1-12. https://doi.org/10.1080/01621459.1983.10477915.

47. Taylor JM. Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. Stat Med 1986; 5(1): 29-36. https://doi.org/10.1002/sim.4780050106.

48. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annu Rev Public Health 2000; 21: 121-145. https://doi.org/10.1146/annurev.publhealth.21.1.121.

49. Member State Coordination Group on Health Technology Assessment. Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons [online]. 2024 [Accessed: 04.07.2025]. URL: https://health.ec.europa.eu/publications/practical-guideline-quantitative-evidence-synthesis-direct-and-indirect-comparisons_en.

50. Brown PJ, Hunnicutt JN, Sanni Ali M et al. Quantifying possible bias in clinical and epidemiological studies with quantitative bias analysis: common approaches and limitations. BMJ 2024; 385: e076365. https://doi.org/10.1136/bmj-2023-076365.

51. Glasziou P, Chalmers I, Rawlins M et al. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007; 334(7589): 349-351. https://doi.org/10.1136/bmj.39070.527986.68.

52. International Society for Pharmacoeconomics and Outcomes Research. Towards a Vision for HEOR: Opportunities for Enhancement and Evolution; An ISPOR Health Science Policy Council White Paper [online]. 2024 [Accessed: 26.09.2025]. URL: https://www.ispor.org/docs/default-source/councils/health-science-policy-council/towards-a-vision-for-heor---opportunities-for-enhancement-and-evolution-an-ispor-health-science-policy-council-white-papere53f0015-8dd9-4a79-90a5-e8efe89f03e4.pdf?sfvrsn=fdbcb152_1.

53. Chow SC, Wang H, Shao J. Sample Size Calculations in Clinical Research. Boca Raton: Taylor & Francis; 2003.

54. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Odronextamab (DLBCL); Bewertung gemäß § 35a SGB V; AbD-Konzept [online]. 2024 [Accessed: 05.08.2025]. URL: https://doi.org/10.60584/A24-18.

55. Kahl KL. Single-Agent Odronextamab Efficacy, Safety Upheld in Relapsed/Refractory DLBCL [online]. 2023 [Accessed: 07.03.2025]. URL: https://www.cancernetwork.com/view/single-agent-odronextamab-efficacy-safety-upheld-in-relapsed-refractory-dlbcl.

56. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Odronextamab (DLBCL); 1. Addendum zum Projekt A24-18 (AbD-Konzept) [online]. 2025 [Accessed: 05.08.2025]. URL: https://doi.org/10.60584/A24-120.

57. Duell J, Abrisqueta P, Andre M et al. Tafasitamab for patients with relapsed or refractory diffuse large B-cell lymphoma: final 5-year efficacy and safety findings in the phase II L-MIND study. Haematologica 2024; 109(2): 553-566. https://doi.org/10.3324/haematol.2023.283480.

58. Roche Pharma. Polatuzumab Vedotin (POLIVY); Dossier zur Nutzenbewertung gemäß § 35a SGB V [online]. 2023 [Accessed: 18.07.2025]. URL: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/1041/#dossier.

59. Roche Pharma. Glofitamab (Columvi); Dossier zur Nutzenbewertung gemäß § 35a SGB V [online]. 2023 [Accessed: 18.07.2025]. URL: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/978/#dossier.

60. AbbVie Deutschland. Epcoritamab (Tepkinly); Dossier zur Nutzenbewertung gemäß § 35a SGB V [online]. 2023 [Accessed: 18.07.2025]. URL: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/1004/#dossier.

61. Swedish Orphan Biovitrum. Loncastuximab tesirin (Zynlonta); Dossier zur Nutzenbewertung gemäß § 35a SGB V [online]. 2023 [Accessed: 18.07.2025]. URL: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/955/#dossier.

62. European Medicines Agency. Ordspono; Assessment report [online]. 2024 [Accessed: 26.09.2025]. URL: https://www.ema.europa.eu/en/documents/assessment-report/ordspono-epar-public-assessment-report_en.pdf.

63. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anwendungsbegleitende Datenerhebung zu Risdiplam: Prüfung des Studienprotokolls (Version 3.0) und des statistischen Analyseplans (Version 3.0); 3. Addendum zum Projekt A21-131 [online]. 2024 [Accessed: 07.03.2025]. URL: https://doi.org/10.60584/A24-57.

64. Wakabayashi R, Hirano T, Laurent T et al. Impact of "time zero" of Follow-Up Settings in a Comparative Effectiveness Study Using Real-World Data with a Non-user Comparator: Comparison of Six Different Settings. Drugs Real World Outcomes 2023; 10(1): 107-117. https://doi.org/10.1007/s40801-022-00343-1.

65. Webster-Clark M, Mavros P, Garry EM et al. Alternative analytic and matching approaches for the prevalent new-user design: A simulation study. Pharmacoepidemiol Drug Saf 2022; 31(7): 796-803. https://doi.org/10.1002/pds.5446.

66. Webster-Clark M, Ross RK, Lund JL. Initiator Types and the Causal Question of the Prevalent New-User Design: A Simulation Study. Am J Epidemiol 2021; 190(7): 1341-1348. https://doi.org/10.1093/aje/kwaa283.

67. Her QL, Rouette J, Young JC et al. Core Concepts in Pharmacoepidemiology: New-User Designs. Pharmacoepidemiol Drug Saf 2024; 33: e70048. https://doi.org/10.1002/pds.70048.

68. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anwendungsbegleitende Datenerhebung zu Etranacogen Dezaparvovec: Prüfung des Studienprotokolls (Version 1.0) und des statistischen Analyseplans (Version 1.0); 2. Addendum zum Projekt A22-83 [online]. 2023 [Accessed: 10.06.2025]. URL: https://doi.org/10.60584/A23-99.

69. Wintzell V, Svanström H, Pasternak B. Selection of Comparator Group in Observational Drug Safety Studies; Alternatives to the Active Comparator New User Design. Epidemiology 2022; 33(5): 707-714. https://doi.org/10.1097/EDE.0000000000001521.

70. Tazare J, Gibbons DC, Bokern M et al. Prevalent new user designs: A literature review of current implementation practice. Pharmacoepidemiol Drug Saf 2023; 32(11): 1252-1260. https://doi.org/10.1002/pds.5656.

71. Willems S, Schat A, van Noorden MS et al. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. Stat Methods Med Res 2018; 27(2): 323-335. https://doi.org/10.1177/0962280216628900.

72. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Treatment Switching in onkologischen Studien; Arbeitspapier [online]. 2018 [Accessed: 07.03.2025]. URL: https://www.iqwig.de/download/ga14-04_treatment-switching-in-onkologischen-studien_arbeitspapier_v1-0.pdf.

73. Latimer NR, Abrams KR, Lambert PC et al. Adjusting survival time estimates to account for treatment switching in randomized controlled trials – an economic evaluation context: methods, limitations, and recommendations. Med Decis Making 2014; 34(3): 387-402. https://doi.org/10.1177/0272989X13520192.

74. Gorrod HB, Latimer NR, Abrams KR. NICE DSU Technical Support Document 24: Adjusting survival time estimates in the presence of treatment switching: an update to TSD 16; Report by the Decision Support Unit [online]. 2024 [Accessed: 18.03.2025]. URL: https://www.sheffield.ac.uk/media/65536/download?attachment.

75. Aslop J, Latimer N, Olson M et al. Estimating comparative effectiveness when patients are switching treatments: a real-world challenge. Value & Outcomes Spotlight 2020: 31-33.

76. World Federation of Hemophilia, International Society on Thrombosis and Haemostasis, European Association for Haemophilia and Allied Disorders. Critical Juncture in the Development of Hemophilia Therapies; A Statement and Call-to-Action from the International Society on Thrombosis and Haemostasis (ISTH), European Association for Haemophilia and Allied Disorders (EAHAD), World Federation of Hemophilia (WFH) for the importance of continued research into and development of effective and safe hemophilia treatments [online]. 2025 [Accessed: 11.06.2025]. URL: https://wfh.org/wp-content/uploads/2025/05/WFH-Statement-at-WHO-25th-EML-Expert-Committe-Meeting.pdf.

77. CSL Behring. Routine Practice Data Collection and Evaluation of etranacogene dezaparvovec (Hemgenix) and prophylactic factor IX (FIX) replacement in severe and moderately severe haemophilia B without a history of FIX inhibitors; a prospective, non-interventional study mandated by GBA; Study Protocol; Version 3.0 [online]. 2024 [Accessed: 02.06.2025]. URL: https://www.g-ba.de/downloads/40-268-10667/2024-07-18_AM-RL-XII_Etranacogen-Dezaparvovec_2022-AbD-005_Studienunterlagen.pdf.

78. CSL Behring. Routine Practice Data Collection and Evaluation of etranacogene dezaparvovec (Hemgenix) and prophylactic factor IX (FIX) replacement in severe and moderately severe haemophilia B without a history of FIX inhibitors; a prospective, non-interventional study mandated by GBA; Statistical Analysis Plan (SAP); Version 3.0 [online]. 2024 [Accessed: 02.06.2025]. URL: https://www.g-ba.de/downloads/40-268-10667/2024-07-18_AM-RL-XII_Etranacogen-Dezaparvovec_2022-AbD-005_Studienunterlagen.pdf.

79. Gilead Sciences. Real world effectiveness and safety of brexucabtagene autoleucel versus patient-individual therapy in relapsed/refractory mantle cell lymphoma: A European Mantle Cell Lymphoma Network (EMCL) registry study mandated by the G-BA; Study Project Plan; Project Plan Number: RW-X19-2206; Version 4.0 [online]. 2025 [Accessed: 20.06.2025]. URL: https://www.g-ba.de/downloads/40-268-11585/2025-06-18_AM-RL-XII_BrexCel_2021-AbD-008_Feststellung_Studienunterlagen.pdf.

80. Novartis. Routine data collection and evaluations of onasemnogene abeparvovec in Germany; Study Protocol; Protocol Number: COAV101A1DE01; Version 4.01 [online]. 2024 [Accessed: 23.06.2025]. URL: https://www.g-ba.de/downloads/40-268-10961/2024-06-06_AM-RL-XII_Onasemnogen-Abeparvovec_2020-AbD-001_Ueberpruefung-SP-SAP_Studienunterlagen.pdf.

81. Roche Pharma. Evaluation of a real world data collection for the reassessment of the additional benefit or evrysdi (risdiplam); Protocol; Protocol Number: ML44661; Version 3.0 [online]. 2024 [Accessed: 23.06.2025]. URL: https://www.g-ba.de/downloads/40-268-10838/2024-09-19_AM-RL-XII_Risdiplam_AbD-004_Feststellung_Studienunterlagen.pdf.

82. Di Staso R, Casadei B, Locke FL et al. Is CAR T a drug or a therapeutic pathway? Intention to treat versus per protocol analysis of real world studies of CAR-T cell therapy in relapsed refractory diffuse large B cell lymphoma. Blood Cancer J 2024; 14: 197. https://doi.org/10.1038/s41408-024-01183-8.

83. Bundesministerium für Gesundheit. Verordnung über die Nutzenbewertung von Arzneimitteln nach § 35a Absatz 1 SGB V für Erstattungsvereinbarungen nach § 130b SGB V (Arzneimittel-Nutzenbewertungsverordnung - AM-NutzenV) [online]. 2023 [Accessed: 09.10.2023]. URL: https://www.gesetze-im-internet.de/am-nutzenv/AM-NutzenV.pdf.

84. de Vet H, Terwee CB, Mokkink LB et al. Measurement in Medicine; A Practical Guide. Cambridge: Cambridge University Press; 2011.

85. Gilead Sciences. Real world effectiveness and safety of brexucabtagene autoleucel versus patient-individual therapy in relapsed/refractory mantle cell lymphoma: A European Mantle Cell Lymphoma Network (EMCL) registry study mandated by the G-BA; Study Project Plan; Project Plan Number: RW-X19-2206; Version 3.0 [online]. 2023 [Accessed: 10.06.2025]. URL: https://www.g-ba.de/downloads/40-268-9958/2023-11-16_AM-RL-XII_BrexCel_2021-AbD-008_finale-Ueberpruefung_Studienunterlagen.pdf.

86. Singer S, Bayer O, Schranz M et al. "Patient-reported outcomes" in medizinischen Registern; Erfahrungen aus einer Anwendungsbegleitenden Datenerhebung. Onkologie 2024; 30: 304-311. https://doi.org/10.1007/s00761-024-01494-2.

87. Deutsches Hämophilieregister. Gesamtdatensatz DHR 2.0 [online]. 2024 [Accessed: 23.06.2025]. URL: https://www.pei.de/SharedDocs/Downloads/DE/regulation/meldung/dhr-deutsches-haemophilieregister/dhr-20-datensatz.pdf?__blob=publicationFile&v=20.

88. Deutsches Hämophilieregister. Handbuch; Version 2.7 [online]. 2025 [Accessed: 23.06.2025]. URL: https://www.pei.de/SharedDocs/Downloads/DE/regulation/meldung/dhr-deutsches-haemophilieregister/dhr-20-handbuch.pdf?__blob=publicationFile&v=15.

89. Wang K, Eftang CN, Jakobsen RB et al. Review of response rates over time in registry-based studies using patient-reported outcome measures. BMJ Open 2020; 10(8): e030808. https://doi.org/10.1136/bmjopen-2019-030808.

90. Amelung V, Arnold M, Altendorf M et al. Patient-Reported Outcomes (PROs) in der Routineversorgung bei Krebserkrankungen; Anwendungsbeispiele aus ausgewählten Ländern [online]. 2024 [Accessed: 30.05.2025]. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Studie_ST-G_QS_PROM__Krebserkrankung.pdf.

91. Food and Drug Administration. Core Patient-Reported Outcomes in Cancer Clinical Trials; Guidance for Industry [online]. 2024 [Accessed: 10.06.2025]. URL: https://www.fda.gov/media/149994/download.

92. Di Maio M, Basch E, Denis F et al. The role of patient-reported outcome measures in the continuum of cancer clinical care: ESMO Clinical Practice Guideline. Ann Oncol 2022; 33(9): 878-892. https://doi.org/10.1016/j.annonc.2022.04.007.

93. Nielsen LK, King M, Möller S et al. Strategies to improve patient-reported outcome completion rates in longitudinal studies. Qual Life Res 2020; 29(2): 335-346. https://doi.org/10.1007/s11136-019-02304-8.

94. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anwendungsbegleitende Datenerhebung zu Etranacogen Dezaparvovec: Prüfung des Studienprotokolls (Version 3.0) und des statistischen Analyseplans (Version 3.0); 3. Addendum zum Projekt A22-83 [online]. 2024 [Accessed: 07.03.2025]. URL: https://doi.org/10.60584/A24-38.

95. Fürchtenicht A, Wehling H, Grote Westrick M et al. Patient-Reported Outcomes; Wie die Patientenperspektive die Versorgung transformieren wird [online]. 2023 [Accessed: 30.05.2025]. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Patient-Reported_Outcomes_Bertelsmann_Stiftung_2023.pdf.

96. Steinbeck V, Ernst SC, Pross C. Patient-Reported Outcome Measures (PROMs): ein internationaler Vergleich; Herausforderungen und Erfolgsstrategien für die Umsetzung von PROMs in Deutschland [online]. 2021 [Accessed: 30.05.2025]. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/BSt_PROMs-Implementierung_final.pdf.

97. Lindberg-Scharf P, Steinger B, Koller M et al. Long-term improvement of quality of life in patients with breast cancer: supporting patient-physician communication by an electronic tool for inpatient and outpatient care. Support Care Cancer 2021; 29(12): 7865-7875. https://doi.org/10.1007/s00520-021-06270-1.

98. Klinkhammer-Schalke M, Steinger B, Koller M et al. Diagnosing deficits in quality of life and providing tailored therapeutic options: Results of a randomised trial in 220 patients with colorectal cancer. Eur J Cancer 2020; 130: 102-113. https://doi.org/10.1016/j.ejca.2020.01.025.

99. Absolom K, Warrington L, Hudson E et al. Phase III Randomized Controlled Trial of eRAPID eHealth Intervention Drug Chemotherapy. J Clin Oncol 2021; 39: 734-747. https://doi.org/10.1200/jco.20.02015.

100. Basch E, Deal AM, Dueck AC et al. Overall Survival Results of a Trial Assessing Patient-Reported Outcomes for Symptom Monitoring During Routine Cancer Treatment. JAMA 2017; 318(2): 197-198. https://doi.org/10.1001/jama.2017.7156.

101. Basch E, Schrag D, Henson S et al. Effect of Electronic Symptom Monitoring on Patient Reported Outcomes Among Patients With Metastatic Cancer; A Randomized Clinical Trial. JAMA 2022; 327(24): 2413-2422. https://doi.org/10.1001/jama.2022.9265.

102. Basch E, Schrag D, Jansen J et al. Symptom monitoring with electronic patient-reported outcomes during cancer treatment: final results of the PRO-TECT cluster-randomized trial. Nat Med 2025; 31(4): 1225-1232. https://doi.org/10.1038/s41591-025-03507-y.

103. Denis F, Basch E, Septans AL et al. Two-Year Survival Comparing Web-Based Symptom Monitoring vs Routine Surveillance Following Treatment for Lung Cancer. JAMA 2019; 321(3): 306-307. https://doi.org/10.1001/jama.2018.18085.

104. Kowalski C, Wesselmann S, van Oorschot B et al. „Patient-reported outcomes“ in der onkologischen Versorgung – aktuelle Anwendungsfelder und Initiativen der Deutschen Krebsgesellschaft. Onkologie 2024; 30: 411-420. https://doi.org/10.1007/s00761-024-01508-z.

105. Aiyegbusi OL, Cruz Rivera S, Roydhouse J et al. Recommendations to address respondent burden associated with patient-reported outcome assessment. Nat Med 2024; 30(3): 650-659. https://doi.org/10.1038/s41591-024-02827-9.

106. Winter LM, Sztankay M, Giesinger JM et al. Manual for the use of EORTC measures in daily clinical practice [online]. 2016 [Accessed: 10.06.2025]. URL: https://www.eortc.org/app/uploads/sites/2/2018/02/EORTC_QLQ_Clinical_Practice_User_Manual-1.0.pdf.

107. Calvert MJ, O'Connor DJ, Basch EM. Harnessing the patient voice in real-world evidence: the essential role of patient-reported outcomes. Nat Rev Drug Discov 2019; 18(10): 731-732. https://doi.org/10.1038/ d41573-019-00088-7.

108. Al-Antary N, Tam S, Alzouhayli S et al. Interventions influencing patient-reported outcomes (PROs) response rates in cancer: a scoping review. J Cancer Surviv 2025. https://doi.org/10.1007/s11764-025-01801-9.

109. Meirte J, Hellemans N, Anthonissen M et al. Benefits and Disadvantages of Electronic Patient-reported Outcome Measures: Systematic Review. JMIR Perioper Med 2020. https://doi.org/10.2196/15588.

110. Korngut L, MacKean G, Casselman L et al. Perspectives on neurological patient registries: a literature review and focus group study. BMC Med Res Methodol 2013; 13: 135. https://doi.org/10.1186/1471-2288-13-135.

111. Niemeyer A, Semler SC, Veit C et al. Gutachten zur Weiterentwicklung medizinischer Register zur Verbesserung der Dateneinspeisung und -anschlussfähigkeit [online]. 2021 [Accessed: 30.05.2025]. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Gesundheit/Berichte/REG-GUT-2021_Registergutachten_BQS-TMF-Gutachtenteam_2021-10-29.pdf.

112. Food and Drug Administration. A Risk-Based Approach to Monitoring of Clinical Investigations; Questions and Answers; Guidance for Industry [online]. 2023 [Accessed: 30.05.2025]. URL: https://www.fda.gov/media/121479/download.

113. European Medicines Agency. Guideline on registry-based studies [online]. 2021 [Accessed: 30.05.2025]. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en.pdf.

114. Lindner L, Weiß A, Reich A et al. Implementing an automated monitoring process in a digital, longitudinal observational cohort study. Arthritis Res Ther 2021; 23(1): 181. https://doi.org/10.1186/s13075-021-02563-2.

115. Gliklich RE, Leavy MB, Dreyer NA. Registries for Evaluating Patient Outcomes: A User's Guide; Fourth Edition [online]. 2020 [Accessed: 27.06.2025]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/registries-evaluating-patient-outcomes-4th-edition.pdf.

116. Hageman IC, van Rooij I, de Blaauw I et al. A systematic overview of rare disease patient registries: challenges in design, quality management, and maintenance. Orphanet J Rare Dis 2023; 18: 106. https://doi.org/10.1186/s13023-023-02719-0.

117. Nährlich L, Burkhart M, Registry Working Group of the German CF Registry. Success of the German Cystic Fibrosis Registry. Pharmacoepidemiol Drug Saf 2025; 34(1): e70076. https://doi.org/10.1002/pds.70076.

118. Lissbrant IF, Erikkson MH, Lambe M et al. Set-up and preliminary results from the Patient-overview Prostate Cancer. Longitudinal registration of treatment of advanced prostate cancer in the National Prostate Cancer Register of Sweden. Scand J Urol 2020; 54(3): 227-234. https://doi.org/10.1080/21681805.2020.1756402.

119. Universitätsklinikum Heidelberg. INTEGRATE ATMP; Innovative Versorgungsstrukturen für neue Therapien; Informationsbroschüre [online]. 2023 [Accessed: 05.06.2025]. URL: https://integrate-atmp.de/integrate-broschuere-2023-final.pdf.

120. Hansen S, Hildberg O, Suppli Ulrik C et al. The Danish severe asthma register: an electronic platform for severe asthma management and research. Eur Clin Respir J 2020; 8(1): 1842117. https://doi.org/10.1080/20018525.2020.1842117.

121. Oliver BJ, Nelson EC, Kerrigan CL. Turning Feed-forward and Feedback Processes on Patient-reported Data into Intelligent Action and Informed Decision-making: Case Studies and Principles. Med Care 2019: S31-S37. https://doi.org/10.1097/MLR.0000000000001088.

122. Lehmann J, Wintner LM, Sztankay M et al. Patient-reported outcomes and psycho-oncological screening in hematology: a practical example of routine electronic monitoring. Magazine of European Medical Oncology 2020; 13: 285-293. https://doi.org/10.1007/s12254-020-00628-7.

123. SCQM Foundation. SCQM-Datenbank [online]. [Accessed: 23.06.2025]. URL: https://www.scqm.ch/de/medizinisches-fachpersonal/scqm-datenbank/.

124. Klein TL, Bender J, Bolton S et al. A rare partnership: patient community and industry collaboration to shape the impact of real-world evidence on the rare disease ecosystem. Orphanet J Rare Dis 2024. https://doi.org/10.1186/s13023-024-03262-2.

125. Stubbs E, Exley J, Wittenberg R et al. How to establish and sustain a disease registry: insights from a qualitative study of six disease registries in the UK. BMC Med Inform Decis Mak 2024; 24(1): 361. https://doi.org/10.1186/s12911-024-02775-x.

126. Gliklich RE, Dreyer NA, Leavy MB et al. Registries for Evaluating Patient Outcomes: A User's Guide; Addendum – 21st Century Patient Registries [online]. 2018 [Accessed: 27.06.2025]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/wysiwyg/registries-guide-3rd-ed-addendum-research-2018-revised.pdf.

127. Nelson EC, Dixon-Woods M, Batalden PB et al. Patient focused registies can improve health, care, and science. BMJ 2016; 354: i3319. https://doi.org/10.1136/bmj.i3319.

128. Lee SB, Zak A, Iversen MD et al. Participation in Clinical Research Registries: A Focus Group Study Examining Views From Patients With Arthritis and Other Chronic Illnesses. Arthritis Care Res 2016; 68(7): 974-980. https://doi.org/10.1002/acr.22767.

129. European Medicines Agency. Patient Registries Workshop, 28 October 2016; Observations and recommendations arising from the workshop [online]. 2017 [Accessed: 13.06.2025]. URL: https://www.ema.europa.eu/en/documents/report/report-patient-registries-workshop_en.pdf.

130. World Federation of Hemophilia. WFH Gene Therapy Registry; User guide for people with hemophilia [online]. 2022 [Accessed: 23.06.2025]. URL: https://www1.wfh.org/publications/files/pdf-2214.pdf.

131. SCQM Foundation. mySCQM Webapplikation [online]. [Accessed: 23.06.2025]. URL: https://www.scqm.ch/de/betroffene/myscqm-ihre-webapplikation/.

132. Osara Y, Coakley K, Devarajan A et al. Development of newborn screening connect (NBS connect): a self-reported patient registry and its role in improvement of care for patients with inherited metabolic disorders. Orphanet J Rare Dis 2017; 12(1): 132. https://doi.org/10.1186/s13023-017-0684-3.

133. Deutsches Rheuma-Forschungszentrum. ActiMON; Aktivitätsmonitoring bei Jugendlichen und jungen Erwachsenen mit Rheuma; Flyer [online]. [Accessed: 05.06.2025]. URL: https://www.drfz.de/wp-content/uploads/2025/02/Flyer-Actimon_compressed3.pdf.

134. Leite WL, Aydın B, Cetin-Berber DD. Imputation of Missing Covariate Data Prior to Propensity Score Analysis: A Tutorial and Evaluation of the Robustness of Practical Approaches. Eval Rev 2021. https://doi.org/10.1177/0193841x211020245.

135. Leyrat C, Seaman SR, White IR et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Res 2019; 28(1): 3-19. https://doi.org/10.1177/0962280217713032.

136. Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. Eur J Epidemiol 2019; 34(1): 23-36. https://doi.org/10.1007/s10654-018-0447-z.

137. de Vries B, Groenwold R. A comparison of two approaches to implementing propensity score methods following multiple imputation. Epidemiology, Biostatistics, and Public Health 2017; 14(4). https://doi.org/10.2427/12630.

138. Granger E, Sergeant JC, Lunt M. Avoiding pitfalls when combining multiple imputation and propensity scores. Stat Med 2019; 38(26): 5120-5132. https://doi.org/10.1002/sim.8355.

139. Lee Y, Leite W. A Comparison of Random Forest-Based Missing Imputation Methods for Covariates in Propensity Score Analysis. PsyArXiv Preprints 2023. https://doi.org/10.31234/osf.io/a47w6.

140. Ling A, Montez-Rath M, Mathur M et al. How to Apply Multiple Imputation in Propensity Score Matching with Partially Observed Confounders: A Simulation Study and Practical Recommendations. J Mod Appl Stat Methods 2020; 19(1). https://doi.org/10.22237/jmasm/1608552120.

141. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Stat Methods Med Res 2016; 25(1): 188-204. https://doi.org/10.1177/0962280212445945.

142. Ségalas C, Leyrat C, Carpenter JR et al. Propensity score matching after multiple imputation when a confounder has missing data. Stat Med 2023; 42(7): 1082-1095. https://doi.org/10.1002/sim.9658.

143. Vader DT, Mamtani R, Li Y et al. Inverse Probability of Treatment Weighting and Confounder Missingness in Electronic Health Record-based Analyses: A Comparison of Approaches Using Plasmode Simulation. Epidemiology 2023; 34(4): 520-530. https://doi.org/10.1097/ede.0000000000001618.

144. Yücel S, Ünal I. Balance diagnostics in propensity score analysis following multiple imputation: A new method. Pharm Stat 2024; 23(5): 763-777. https://doi.org/10.1002/pst.2389.

145. Eiset AH, Frydenberg M. Considerations for Using Multiple Imputation in Propensity Score-Weighted Analysis - A Tutorial with Applied Example. Clin Epidemiol 2022; 14: 835-847. https://doi.org/10.2147/CLEP.S354733.

146. Lee KJ, Tilling KM, Cornish RP et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. J Clin Epidemiol 2021; 134: 79-88. https://doi.org/10.1016/j.jclinepi.2021.01.008.

147. Bottigliengo D, Lorenzoni G, Ocagli H et al. Propensity Score Analysis with Partially Observed Baseline Covariates: A Practical Comparison of Methods for Handling Missing Data. Int J Environ Res Public Health 2021; 18(13): 6694. https://doi.org/10.3390/ijerph18136694.

148. Leyrat C, Seaman SS, White IR et al. Propensity score analysis with partially observed confounders: How should multiple imputation be used? arXiv (Cornell University) 2016. https://doi.org/10.48550/arxiv.1608.05606.

149. Nguyen T, Stuart E. Multiple imputation for propensity score analysis with covariates missing at random: some clarity on “within” and “across” methods. Am J Epidemiol 2024; 193(10): 1470-1476. https://doi.org/10.1093/aje/kwae105.

150. Moons KGM, Donders RART, Stijnen T et al. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006; 59(10): 1092-1101. https://doi.org/10.1016/j.jclinepi.2006.01.009.

151. Carpenter JR, Kenward MG, Bartlett JW et al. Multiple Imputation and its Application. Chichester: Wiley; 2023.

152. Blake H, Leyrat C, Mansfield KE et al. Propensity scores using missingness pattern information: a practical guide. Stat Med 2020; 39(11): 1641-1657. https://doi.org/10.1002/sim.8503.

153. Carpenter JR, Smuk M. Missing data: A statistical framework for practice. Biom J 2021; 63(5): 915-947. https://doi.org/10.1002/bimj.202000196.

154. Curnow E, Tilling K, Heron JE et al. Multiple imputation of missing data under missing at random: including a collider as an auxiliary variable in the imputation model can induce bias. Front Epidemiol 2023; 3: 1237447. https://doi.org/10.3389/fepid.2023.1237447.

155. Ioannidis JPA, Tan YJ, Blum MR. Limitations and Misinterpretations of E-Values for Sensitivity Analyses of Observational Studies. Ann Intern Med 2019; 170(2): 108-111. https://doi.org/10.7326/M18-2159.

156. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Ann Intern Med 2017; 167(4): 268-274. https://doi.org/10.7326/M16-2607.

157. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 2010; 21(3): 383-388. https://doi.org/10.1097/EDE.0b013e3181d61eeb.

158. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res 2011; 46(3): 399-424. https://doi.org/10.1080/00273171.2011.568786.

159. Stürmer T, Joshi M, Glynn RJ et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 2006; 59(5): 437-447. https://doi.org/10.1016/j.jclinepi.2005.07.004.

160. Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. BMC Med Res Methodol 2012; 12: 70. https://doi.org/10.1186/1471-2288-12-70.

161. Friedrich S, Friede T. Causal inference methods for small non-randomized studies: Methods and and application in COVID-19. Contemp Clin Trials 2020; 99: 106213. https://doi.org/10.1016/j.cct.2020.106213.

162. Austin PC. Bootstrap vs asymptotic variance estimation when using propensity score weighting with continuous and binary outcomes. Stat Med 2022; 41(22): 4426-4443. https://doi.org/10.1002/sim.9519.

163. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med 2018; 37(14): 2252-2266. https://doi.org/10.1002/sim.7654.

164. Wilkinson JD, Mamas MA, Kontopantelis E. Logistic regression frequently outperformed propensity score methods, especially for large datasets: a simulation study. J Clin Epidemiol 2022; 152: 176-184. https://doi.org/10.1016/j.jclinepi.2022.09.009.

165. BioMarin. Comparative Effectiveness of Roctavian to Standard of Care Hemostatic Therapies in Germany Among People with Severe Hemophilia A: A Prospective Non-Interventional Study Utilizing Data Collected in the German Hemophilia Register; Study Protocol; Protocol Number: 270-603; Version 3.0 [online]. 2024 [Accessed: 09.07.2025]. URL: https://www.g-ba.de/downloads/40-268-10669/2024-07-18_AM-RL-XII_Valoctocogen-Roxaparvovec_2020-AbD-002_Studienunterlagen.pdf.

The full rapid report (German version) is published under

https://www.iqwig.de/en/projects/a25-13.html.