Systematic confounder identification in the indication of relapsing-remitting multiple sclerosis (RRMS)[1]

EXTRACT

Project: GA23-02 Version: 1.0 Status: 30 Apr 2025 DOI: 10.60584/GA23-02_en

[1] Translation of Chapters 1 to 7 and Appendices A to C of the working paper GA23-02 *Systematische Confounderidentifikation in der Indikation schubförmig remittierende multiple Sklerose (RRMS)* (Version 1.0; Status: 30 April 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

Systematic confounder identification in the indication of relapsing-remitting multiple sclerosis (RRMS)

Commissioning agency

Produced within the context of IQWiG's General Commission

Internal Project No.

GA23-02

DOI-URL

https://doi.org/10.60584/GA23-02_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; Systematic confounder identification in the indication of relapsing-remitting multiple sclerosis (RRMS); Extract [online]. 2025 [Accessed: DD.MM.YYYY]. URL: https://doi.org/10.60584/GA23-02_en.

Keywords

Multiple Sclerosis – Relapsing-Remitting, Propensity Score, Benefit Assessment, Feasibility Studies

This working paper 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 E 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

- Lars G. Hemkens, Research Centre for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University of Basel and University Hospital Basel
- Julian Hirt, RC2NB, University of Basel and University Hospital Basel
- Perrine Janiaud, RC2NB, University of Basel and University Hospital Basel
- Tim Wölfle, RC2NB, University of Basel and University Hospital Basel

IQWiG thanks the external experts for their collaboration in the project.

IQWiG employees

- Jana Göbel
- Lars Beckmann
- Simone Johner
- Claudia Kapp
- Anne-Kathrin Petri
- Daniela Preukschat
- Barbara Spix
- Volker Vervölgyi
- Siw Waffenschmidt

Executive summary

As part of IQWiG's General Commission, the issue of systematic confounder identification in the indication of relapsing-remitting multiple sclerosis (RRMS) was examined.

Background

To adequately control for confounding in comparative non-randomized studies of interventions (NRSI), it is necessary to systematically identify the relevant confounders and consider these in the analysis. The publication by Pufulete 2022 is currently the only concrete proposal for systematic confounder identification that has been published. According to Pufulete 2022, confounder identification is based on 3 methodological components: a systematic literature review, clinician interviews, and a clinician survey.

The aim of this model project is to systematically identify confounders in the indication of RRMS in order to examine the basic feasibility of a systematic approach to confounder identification based on Pufulete 2022. Additionally, the project aims to derive specific recommendations for systematic confounder identification in comparative NRSI, and consequently in routine practice data collection.

Research question

The aims of this investigation are:

- to systematically identify confounders based on Pufulete 2022 for the indication of RRMS in a suitable comparison of 2 drug therapies with regard to key patient-relevant outcomes
- to examine the basic feasibility of confounder identification based on Pufulete 2022
- to derive recommendations for systematic confounder identification in comparative NRSI.

Methods

In this project, external experts were commissioned to systematically identify confounders for the indication of RRMS based on the publication by Pufulete 2022, which describes such identification methods.

The IQWiG project group classified and commented on the report by the external experts ("external report"). Additionally, the external report was used to examine the feasibility of confounder identification based on Pufulete 2022, taking into account the timeframe and resources required.

Based on the external report, further in-house investigations were conducted that went beyond the methodological approach of Pufulete 2022. It was examined whether the potential confounding variables identified in the external report could be further summarized.

Additionally, focused information retrieval was applied to determine whether any other publications exist on systematic confounder identification, aside from the approach proposed by Pufulete 2022.

Finally, specific recommendations for systematic confounder identification in comparative NRSI were derived from Pufulete 2022, the external report on systematic confounder identification based on Pufulete 2022, and the further in-house investigations. In particular, we examined the extent to which special requirements arise for research questions with an expected smaller body of evidence, such as questions relating to routine practice data collection.

Results

Confounder identification based on Pufulete 2022 for the indication of RRMS

Results of confounder identification based on Pufulete 2022 and discussion of the approach of the external experts

The external experts presented a detailed report with a comprehensive search for confounders based on Pufulete 2022.

The external experts used a 3-step systematic approach to confounder identification and identified 160 potential confounders. Of these, 125 were identified via a literature review (search for observational studies and RCTs). A total of 53 potential confounders were identified through interviews with 8 clinicians from Germany and Switzerland. 18 of these were also identified through the search, while the remaining 35 were only identified through the clinician interviews. In the subsequent quantitative survey of 21 participants (7 of whom completed the survey), the identified confounders were assessed based on how strongly they influenced the clinicians' treatment decisions. To this end, the external experts first summarized and reduced the 160 identified confounders to 137 confounders (13 were covered by similar confounders, 2 could not be assessed without the study context, and 8 had not yet been identified at the start of the survey). The confounders were categorized using a scale from "1 - no consideration" to "5 - very strong consideration". 40 of the 137 confounders were rated with an average score greater than 4.

The external experts categorized 148 (93%) of the 160 potential confounders as "likely to be a predictor of treatment", and of those, 128 of 160 (80%) as "likely to be a predictor of the outcomes of interest". 11 (7%) potential confounders were categorized as "unclear", which

the external experts interpreted as "unlikely to be a predictor of either the treatment or the outcomes". The external experts did not categorize the potential confounder "type of MS".

In the final step, the external experts summarized the 160 potential confounders into 136 confounders (i.e. confounders that could be measured or assessed in a similar way).

Overall, the external experts addressed the research question appropriately, in line with the specifications outlined in Pufulete 2022. The following text provides examples of deviations from this approach.

Unlike Pufulete 2022, who applied a saturation criterion for confounder extraction for all systematically identified literature sources, the external experts applied a saturation criterion (15 consecutive studies) only to studies published prior to 2020. By contrast, all relevant studies published since 2020 were extracted. Based on the information in the external report, it cannot be determined whether the additional investment of resources was justified by the additional confounders identified (57/125 [46%] were identified exclusively in studies published in 2020 or later). While the external experts assume that they identified a more complete list, they also describe their approach as very resource-intensive.

Regarding the qualitative interviews, Pufulete 2022 assumed that saturation could be achieved with 12 experts (6 in cardiology and 6 in cardiac surgery). However, due to recruitment difficulties, the external experts ultimately recruited only 8 of the planned 10 clinicians. The external experts stated that the 8 contributions were saturated and sufficiently reflected the factors considered in the treatment decision. Overall, involving 8 clinicians for the qualitative interviews is considered appropriate. It is not possible to assess whether fewer interviews would have also achieved saturation.

In the quantitative survey, participants were not asked to name the 5 most relevant confounders, as in Pufulete 2022, but to rate on a 5-point scale from “1 – no consideration” to “5 – strong consideration” how strongly each individual potential confounder influences the treatment decision. This represents an alternative approach to assessing the relevance of confounders. Although scoring is more resource-intensive than the other possible assessment method, the assessment of the relevance of each individual confounder is preferable overall to the approach chosen by Pufulete 2022. The scores can be used to rank all potential confounders according to their relevance. This ranking can then be used as a decision-making aid (e.g. in cases where the statistical model does not converge, the scores can be used to justify which potential confounder can be eliminated first when adjusting the model). To reduce the effort involved in this confounder assessment, a possible approach is to summarize the list of the potential confounders of interest more intensively (see below).

In the final step, both Pufulete 2022 and the external experts assessed whether the potential

confounder influences both the treatment decision and the treatment result. The purpose of this categorization is to identify "true confounders". However, as assessing the relationship between variables requires strong assumptions and is therefore subject to uncertainty, it is generally not feasible to identify confounders in this manner. Consequently, confounding variables whose influence on the treatment decision and the treatment result is uncertain should not be excluded. Only a literature-based justification can explain the exclusion of a potential confounder.

Feasibility of confounder identification based on Pufulete 2022

The review of the external report showed that the approach essentially met the specifications outlined in Pufulete 2022. Therefore, the implementation of confounder identification based on the Pufulete 2022 approach is considered feasible in principle.

The external experts required 62 person-days (8 working hours each) to identify the 160 confounders. The systematic literature review was the most resource-intensive component of confounder identification, accounting for an estimated 33 person-days. In this context, it should be noted that RRMS was selected on the basis that the overall body of evidence available is assumed to be sufficient. In indications where the expected body of evidence is smaller, substantially fewer resources will be required for this component, as there will be fewer hits to screen and less data to extract in the literature review. Additionally, the approach adopted by the external experts not to select a saturation criterion for confounder identification in publications after 2020 is highly sensitive. Therefore, resources could be saved by applying a saturation criterion to all publications in this step.

In-house investigations

Potential confounders for the indication of RRMS

External experts identified 160 potential confounding variables and summarized confounders that could be assessed or measured in a similar way into 136 potential confounders. Summarizing the confounders identified through literature reviews and clinician interviews was not part of the approach of Pufulete 2022; however, the authors discuss this as a logical next step. Currently, however, there are no recommendations for this step.

We examined the extent to which it is possible to summarize the identified potential confounding variables more intensively, without having to accept a relevant loss of information regarding the identified potential confounders. The present systematic identification of confounders should inform a comparative NRSI of dimethyl fumarate and glatiramer acetate in patients aged 18 years and over with RRMS. In the following text, it is assumed that the CONFIRM RCT will be replicated as a model study using target trial emulation. Thus, the CONFIRM RCT's specific inclusion and exclusion criteria were used to

exclude irrelevant variables. Based on the research question, 28 out of 160 variables were excluded.

The remaining 132 potential confounding variables were summarized to reduce the number of duplicates and overlaps. For example, the following potential confounding variables were assigned to the confounder (health-related) “Quality of Life”: EQ-5D, SF-36 scores, the Multiple Sclerosis Impact Scale, the Multiple Sclerosis International Quality of Life Questionnaire, and the Sexual Satisfaction Scale Score. In total, the 132 potential confounding variables were summarized into 28 potential confounders.

26 of the 28 potential confounders were identified from the literature, and 18 were identified from clinician interviews. The potential confounders "patient wish" and "type of professional activity" were identified only through interviews, not the literature.

Operationalization of potential confounders

Finally, suitable operationalizations of the potential confounders were proposed. Where possible, a continuous operationalization was favoured as this provides the greatest amount of information. If this was not possible, a categorical operationalization was selected instead. Summarizations of manifestations (e.g., dichotomizations) should only be considered if statistical models do not converge. These summarizations must be pre-specified.

Due to their complexity and multi-layered nature, it was not possible to adequately operationalize the 2 confounders "patient wish" and "tolerance of previous therapy and expectations of the treatment". Therefore, these 2 relevant confounders cannot be included in the analysis. This must be considered when interpreting the results.

Further suggestions for systematic confounder identification in the literature

Apart from Pufulete 2022, no other publications on systematic approaches to confounder identification were found.

Proposal for a requirement for systematic confounder identification

When evaluating intervention effects based on a non-randomized comparison, a set of potential confounders should be identified beforehand as part of the study planning process. Based on the results of a confounder identification by external experts and the further in-house investigations, we derived specific recommendations for systematic confounder identification. These recommendations represent a proposal for specific requirements for systematic confounder identification.

Although both the external report and the publication by Pufulete 2022 refer to an indication with an extensive body of evidence, the situation of rare(r) diseases was also considered when deriving the recommendations, as this represents the usual situation for areas such as routine

practice data collection. Should special requirements arise in this regard, they will be explicitly stated below.

The proposal for systematic confounder identification is divided into 3 main steps.

Step 1: Identification through a systematic literature review and clinician interviews

According to Pufulete 2022, the first step of identification consists of 2 components: a systematic literature review and clinician interviews. Primary publications on RCTs and cohort studies should be considered for systematic information retrieval. Guidelines or single-arm studies may be considered subsequently. The inclusion and exclusion criteria for the study pool for confounder extraction should closely align with the research question. In comparative observational studies, extraction should include all adjustment variables, regardless of statistical significance. For randomized or non-comparative studies, all patient characteristics should be extracted. When extracting data from primary publications, it is advisable that a suitable saturation criterion is applied. Clinician interviews should be conducted as a second component. The clinicians should be named and any potential conflicts of interest disclosed.

Step 2: Summarization

After all potential confounding variables have been extracted, the second step involves summarizing confounding variables with overlapping content. For each confounding variable identified, it should be checked whether it is within the scope of the research question. The exclusion of a potential confounder based on content considerations must be justified based on the literature. Transparent and comprehensible documentation should be ensured. An operationalization should be defined for all identified potential confounders. Although potential confounders that cannot be measured or operationalized cannot be included in the analysis, they should be named and considered when interpreting the results.

Step 3: Assessment of the relevance of potential confounders by clinicians

In the third step, the clinicians quantitatively assess the relevance of the potential confounders that have already been summarized. It is advisable that they use a scoring system.

Documentation of confounder identification

In order to understand confounder identification, the approach and the results must be documented in detail and transparently. This documentation should be published as part of the study protocol for the planned comparative NRSI.

Conclusion

The external experts applied the approach suggested by Pufulete 2022 to identify a total of 160 potential confounding variables. These were summarized into 28 potential confounders

in further in-house investigations. This demonstrated that confounder identification based on Pufulete 2022 was feasible in principle, for example in the context of routine data collection.

The following recommendations for systematic confounder identification in 3 steps were derived from the experiences gained during the investigations:

1) Confounder identification

Potential confounding variables should be identified through a systematic literature review and clinician interviews.

2) Confounder summarization

Confounders that measure the same construct or overlap in terms of content should be summarized. A clear operationalization must be defined for each potential confounder.

3) Confounder assessment

The relevance of each potential confounder should be assessed in a clinician survey. The result can inform a content-based selection of variables.

The second step in particular builds on the methods proposed by Pufulete 2022 and aims to save resources without losing relevant information.

The systematic approach to confounder identification is resource-intensive, despite the described resource-saving options. The next step is to examine other possible ways to save resources in both the process of confounder identification and the associated recording of potential confounders. Any resource-saving options must be weighed against the associated loss of information and potential reduction in the certainty of results.

In order to minimize confounding bias in a comparative NRSI, systematic confounder identification is required. However, there will always be uncertainty as to whether all relevant confounders have actually been identified and can be considered in the analysis. Both unknown and unmeasurable confounders remain an unresolved problem, meaning that in comparative NRSI, the added benefit of a treatment can only be inferred if the effects are sufficiently large. The only sensible alternative is to conduct an RCT, such as a registry-based RCT, which does not require confounder identification and control.

Table of contents

List of tables

Page

List of figures

Page

List of abbreviations

Abbreviation	Meaning
AE	adverse event
DAG	directed acyclic graph
EDSS	Expanded Disability Status Scale
G-BA	Gemeinsamer Bundesausschuss (Federal Joint Committee)
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (Institute for Quality and Efficiency in Health Care)
MRI	magnetic resonance imaging
MS	multiple sclerosis
NRSI	non-randomized studies of interventions
PICO	population, intervention, comparison, outcome
RCT	randomized controlled trial
RRMS	relapsing-remitting multiple sclerosis
SAE	serious adverse event
SGB	Sozialgesetzbuch (Social Code Book)

Glossary

For further reading on the definitions described below, see, for example, Feeney 2024 [1], Williams 2018 [2], Catalogue of bias collaboration [3], and Sendor 2022 [4].

Confounder

A confounder is a variable that influences both the treatment decision and the outcome (see Figure 1) and can therefore distort a treatment effect. In comparative non-randomized studies of interventions (NRSI), adjustment for confounders is used to compensate for the lack of structural equality between the treatment groups.

Unmeasurable confounders

In this report, unmeasurable confounders are defined as confounders that are known but cannot be operationalized and therefore cannot be recorded, such as patient wish or physician's decision. Furthermore, even with careful and systematic confounder identification, completely unknown confounders may still exist.

These confounders must be distinguished from confounders that can be clearly operationalized but are not available in the data source (see also residual confounding).

Residual confounding / unmeasured confounding

The terms residual confounding and unmeasured confounding are used in the literature to describe 3 situations in which analyses can lead to biased results.

Firstly, no values are available for a known confounder, or the confounder cannot be recorded and operationalized. Secondly, although values are available for a known confounder, these are incorrectly recorded. In the third situation, a confounder is unknown. Residual confounding is therefore understood as the totality of confounders that are not measured and therefore not taken into account in the statistical analysis.

Confounding by indication

Confounding by indication occurs when a variable (e.g., a symptom or prognostic factor) that influences treatment decisions also influences the results of the outcomes of interest. A classic example of this is when a disease (or the severity of a disease) is a risk factor for the outcome of interest. If the disease is treated with the study intervention, the treated patients have a higher risk of a poor treatment result for the outcome of interest than the untreated patients simply because their disease is more severe to begin with.

Directed acyclic graph (DAG)

DAGs are a graphical tool that provide a way to visually represent and better understand the (assumed) causal relationships and dependencies between variables.

Collider

A collider is influenced by both the intervention and the outcome (see Figure 1). Adjustment for a collider should be avoided, as this can create a spurious correlation between the intervention and the outcome.

Mediator

A mediator lies directly between the intervention and the outcome. Adjustment for a mediator should also be avoided.

Target trial emulation

Target trial emulation refers to the replication of the design of a randomized controlled trial (RCT) based on observational data.

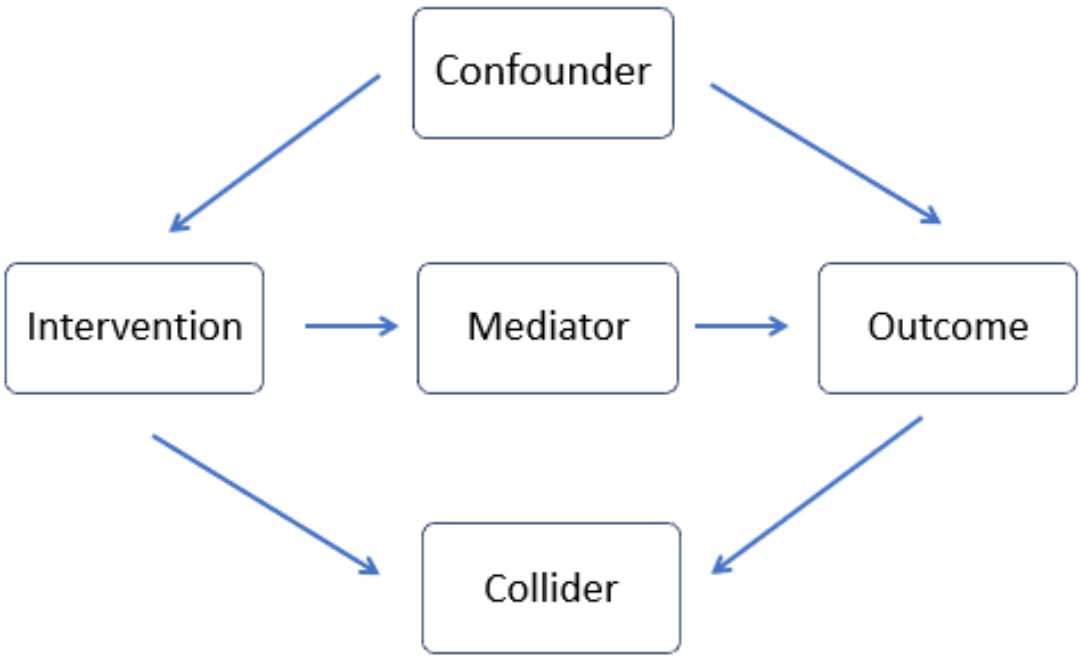

Figure 1: Graphical representation (directed acyclic graph, DAG) of the dependencies of confounders, mediators and colliders with the intervention and the outcome of interest

1 Background

General relevance of systematic confounder identification

When new drugs are authorized, gaps in the evidence often remain, for example regarding comparisons of different treatment options. These gaps can be closed by registry-based randomized controlled trials (RCTs), for example. As an alternative, non-randomized studies using target trial emulation [5-9], including careful correction for confounders (e.g., using propensity score methods), are being discussed. In comparative non-randomized studies of interventions (NRSI), it is necessary to systematically identify the relevant confounders and consider these in the analysis for adequate control for confounding. Since routine practice data collections pursuant to §35a (3b) of Social Code Book V (SGB V) are usually NRSI, adequate adjustment for confounders also plays a central role for these studies. All relevant confounders (including important interactions) must be identified in advance using a systematic approach. These confounders must be recorded in full and taken into account appropriately in the model [10]. A specific situation arises in the context of routine practice data collection, which is usually performed for indications involving small patient numbers and limited evidence.

Approach of Pufulete 2022

The 2022 publication by Pufulete [11] describes how confounders can be identified using 3 methodological components: a systematic literature review, clinician interviews, and a clinician survey. Firstly, 2 researchers independently extract data on potential confounders and co-interventions using a saturation criterion. Secondly, further potential confounders are identified through clinician interviews and categorized into different groups. Thirdly, the 5 most important potential confounders are identified using the survey. Finally, the identified confounders are classified according to their presumed influence on the treatment decision and/or the treatment result, and the presumed direction of bias is described (see Table 1 in Pufulete 2022). Although the literature calls for a systematic approach to confounder identification [12-14], Pufulete 2022 is the only publication to provide a concrete proposal for doing so in clinical and epidemiological studies. Accordingly, IQWiG mentions it as an example in its General Methods [15] and in its projects on routine practice data collection [16,17].

Rationale for the project

Various pharmaceutical companies have already carried out systematic confounder identification in the planning of studies on routine practice data collection (e.g., [18,19]). However, none of the submitted studies fully met the requirements for this (see, for example, [20-23]). In this context, the procedure proposed by Pufulete 2022 for systematic confounder identification is often criticized for being very resource-intensive.

The aim of this model project is to systematically identify confounders in the indication of relapsing-remitting multiple sclerosis (RRMS) in order to examine the basic feasibility of a systematic approach to confounder identification based on Pufulete 2022. RRMS is an indication with an expected large body of evidence, corresponding to the coronary heart disease indication investigated by Pufulete 2022.

Additionally, the project aims to derive specific recommendations for systematic confounder identification in comparative NRSI, and consequently in routine practice data collection. This working paper explicitly addresses the special features of indications with an expected smaller body of evidence.

Furthermore, the confounders identified in this working paper for the indication of RRMS are to inform a follow-up project, in which treatment effects would be derived based on a non-randomized comparison. This is to be carried out in the context of replicating a specific RCT, the CONFIRM RCT [24], which compared dimethyl fumarate and glatiramer acetate in adult patients with RRMS, and is to be used as a model study for target trial emulation.

2 Research question

The aims of this investigation are:

- to systematically identify confounders based on Pufulete 2022 for the indication of RRMS in a suitable comparison of 2 drug therapies with regard to key patient-relevant outcomes
- to examine the basic feasibility of confounder identification based on Pufulete 2022
- to derive recommendations for systematic confounder identification in comparative NRSI.

3 Course of the project

As part of IQWIG's General Commission, the issue of systematic confounder identification in the indication of RRMS was examined. The project began on 5 May 2023.

External experts were involved in the project and were commissioned to carry out systematic confounder identification for the indication of RRMS, based on Pufulete 2022. Based on the report from these experts, we assessed the basic feasibility of the methodological approach proposed by Pufulete 2022 and also conducted further in-house investigations that went beyond this approach, ultimately resulting in overarching recommendations for systematic confounder identification.

A working paper was prepared based on an internal project outline. This was submitted to the Federal Joint Committee (G-BA) and published on the IQWiG website 4 weeks later.

4 Methods

4.1 Confounder identification for the indication of RRMS based on Pufulete 2022

4.1.1 Confounder identification by external experts

In this project, external experts were commissioned to systematically identify confounders for the indication of RRMS based on Pufulete 2022 [11], which describes the methods to be used for confounder identification.

The external group, who have relevant expertise in non-randomized studies (non-RCTs) and confounders, initially completed the commission independently and in full. In this working paper, IQWiG evaluates the methodological approach and results of the external experts. To this end, the methodological approach is discussed in comparison with Pufulete 2022.

Selected research question

The identified confounders are to be used to emulate the CONFIRM RCT [24] based on a non-randomized comparison of dimethyl fumarate and glatiramer acetate in patients aged 18 years and over with RRMS. The CONFIRM RCT was primarily selected because it is well suited to the planned target trial emulation (e.g., due to the comparison with an active control). In addition, it is assumed that the overall body of evidence is sufficient for identifying confounders for the underlying research question. The central, patient-relevant outcomes for which the treatment effects were investigated in the non-randomized comparison, with corresponding adjustments for confounders, were the annual relapse rate, confirmed disability progression (as measured by the Expanded Disability Status Scale, EDSS), health-related quality of life, and adverse events (AEs), particularly serious adverse events (SAEs). Section 5.3 discusses the transferability of the feasibility of confounder identification for the present research question to other research questions.

4.1.2 Classification and comment on the external report

In Section 5.1.2 of this working paper, the IQWiG project group classifies and comments on the report by the external experts (“external report”). This report is included as Appendix D to the original working paper. The internal review of the report's methodological approach focused on whether the approach was implemented according to the specifications in Pufulete 2022 and on whether any adjustments (e.g., measures to save resources) were reasonable and justifiable.

The working paper discusses possible optimization procedures and variation options for the approach based on the 2 examples of application: confounder identification by Pufulete 2022 in the indication of coronary heart disease and by external experts in the indication of RRMS. For example, it considers whether any of the components of Pufulete 2022 – systematic

literature review, clinician interviews, and clinician survey - are dispensable when the resources required are considered.

4.1.3 Examination of the feasibility of confounder identification based on Pufulete 2022

If the external group provides a report that includes the identified confounders, and if IQWiG's review shows that the approach described in Pufulete 2022 was adequately followed, then this approach should be considered feasible in principle. It was also assessed whether the timeframe outlined in the commission of the external report was met.

4.2 In-house investigations

Further investigations were carried out based on the external report that go beyond the methodological approach of Pufulete 2022. These are described below.

4.2.1 Summarization of the identified potential confounders

We examined whether the potential confounding variables identified in the external report could be summarized. To this end, 2 researchers independently checked which confounding variables measured the same construct or overlapped in terms of content. They documented which individual confounding variables were summarized. Then, suitable operationalizations were defined for the summarized confounders.

All steps were agreed upon by the project group. Any discrepancies were resolved through discussion. Additionally, a workshop was held with the external experts to discuss and revise the list of potential confounders and their operationalizations, drawing on the expertise of both the external experts and the internal project group.

4.2.2 Focused information retrieval for publications on the systematic approach to confounder identification

Focused information retrieval was used to determine whether there were any additional publications on systematic confounder identification beyond the approach proposed by Pufulete 2022.

Publications were included if they described a systematic approach to confounder identification, with the aim of defining a confounder set a priori, similar to the approach of Pufulete 2022. Publications employing data-supported or data-based identification approaches (e.g., statistical methods based on electronic patient record data [25,26]) were excluded. There was no restriction on the publication period. Publications had to be written in German or English.

The following information sources and search techniques were used for focused information retrieval: MEDLINE and the “Similar Articles” function in PubMed for known publications [2,11,27,28].

Firstly, the identified hits were assessed based on their titles and, if available, their abstracts, to determine their potential relevance to the inclusion criteria (see above). Secondly, the full texts of potentially relevant documents were assessed for relevance. One researcher carried out both selection steps.

4.3 Derivation of requirements for systematic confounder identification (recommendations)

Based on Pufulete 2022 and the external report on confounder identification following Pufulete 2022, and further in-house investigations, we derived specific recommendations for systematic confounder identification in comparative NRSI. In particular, we examined the extent to which special requirements arise for research questions with an expected smaller body of evidence, such as questions regarding routine practice data collection.

5 Results

5.1 Confounder identification for the indication of RRMS based on Pufulete 2022

5.1.1 Results of confounder identification based on Pufulete 2022

The external report aimed to systematically and independently identify potential confounders to consider when designing and conducting a non-randomized comparison of dimethyl fumarate and glatiramer acetate in adults with RRMS, focussing on patient-relevant outcomes. The results of the external report are summarized below. The methodological approach is then discussed in comparison with Pufulete 2022 [11]. The full external report is presented in Appendix D of the German original version.

The external experts used a 3-step systematic approach to confounder identification and identified 160 potential confounders. Of these, 125 were identified via a literature review (search for observational studies and RCTs). A total of 53 potential confounders were identified through interviews with 8 clinicians from Germany and Switzerland. 18 of these were also identified through the search, while the remaining 35 were only identified through the clinician interviews. In the subsequent quantitative survey, which was completed by 7 of the 21 participants, 137 of the 160 confounders were assessed based on how strongly they influenced the clinicians' treatment decisions. The confounders were categorized using a scale from "1 - no consideration" to "5 - very strong consideration". Of the 137 confounders, 40 were rated with an average score greater than 4.

The external experts categorized 148 (93%) of the 160 potential confounders as "likely to be a predictor of treatment," and of those, 128 of the 160 (80%) as "likely to be a predictor of the outcomes of interest". 11 (7%) potential confounders were categorized as "unclear", which the external experts interpreted as "unlikely to be a predictor of either the treatment or the outcomes". The external experts did not categorize the potential confounder "type of MS" because only RRMS was considered.

In the final step, the external experts summarized the 160 potential confounders into 136 confounders (i.e. confounders that could be measured or assessed in a similar way).

5.1.2 Discussion of the approach of the external experts

The external experts presented a detailed report on their comprehensive search for confounders, based on Pufulete 2022 (see Appendix D of the original working paper for more information). The report documented the overarching results of the individual steps in detail. A detailed record of the data source for each confounder, including verbatim citations, is available in a separate Excel file (see Appendix F of the external report [dataset] [29]).

Overall, the external experts addressed the issue appropriately, adhering to the specifications outlined in Pufulete 2022. The following section details any deviations from this approach, any adjustments made (e.g., to save resources), and possible optimization procedures and variation options.

Systematic literature review

Inclusion and exclusion criteria and selection

In addition to RCTs and cohort studies, the external experts explicitly considered non-RCTs in their literature review (according to Viswanathan's taxonomy [30]). In this taxonomy, non-RCTs are defined as prospective studies in which individuals or groups are assigned to an intervention or control group using a non-random method (e.g., date of birth, date of inclusion, or investigator judgement). Including non-RCTs extends beyond the Pufulete 2022 approach. However, according to Table 2 of the external report, the listed confounders were only identified from observational studies or randomized trials. As according to the above definition, non-RCTs are not observational studies, considering them in this case does not provide additional information.

When establishing the inclusion criteria and selecting the relevant studies, besides the population and interventions, the external experts also considered the outcomes of interest. Additionally, for the inclusion of the study, 90% of the study participants had to fulfil the inclusion criterion “RRMS”. These inclusion criteria were more detailed than those in Pufulete 2022, which is understandable, particularly given the need to ensure that the study adequately represented the population investigated in the research question.

Extraction and saturation criterion

Following Pufulete's approach, the external experts extracted the characteristics of the studies and patients, the interventions, and all factors used for adjustment in the statistical analyses. For RCTs, the patient characteristics reported in tabular form in the study publications (“Table 1”) were extracted. The external experts also extracted the percentage of women analysed, as well as the mean age at baseline with the standard deviation. For each potential confounder identified, they noted the outcome for which it was used in the analysis. However, in the interest of saving resources, it seems unnecessary to extract additional information on age and sex, especially since the external experts did not draw any conclusions from it.

Pufulete 2022 used a saturation criterion when extracting confounders identified in the literature review. They terminated confounder extraction if no new potential confounders were identified within 10 consecutive studies. However, the external experts only applied a saturation criterion to studies prior to 2020. These studies were randomly selected and extraction stopped using the saturation criterion “no new potential confounders identified

within 15 consecutive studies”. By contrast, all identified confounders were extracted from relevant studies published since 2020 (n=40, out of 351 identified in the search). The external experts justified this division based on the publication dates of the studies, stating that the field of multiple sclerosis develops much faster than the field of cardiovascular diseases (the subject of the 2022 Pufulete study), the disease mechanisms of which have been relatively well understood for decades. In order to adequately consider the recent developments in multiple sclerosis reported in recent literature and new observational data sources, the external experts divided the studies based on their publication date. Their justification for dividing the search according to publication date is understandable. However, the external report does not clarify how many additional potential confounders were identified by not applying a saturation criterion to studies published after 2020, or by applying a saturation criterion of 15 consecutive studies (instead of 10) to studies published before 2020. Therefore, based on the information in the external report, it cannot be determined whether the additional investment of resources was justified by the additional confounders identified (57/125 [46%] of which were identified exclusively in studies published in 2020 or later). While the external experts assume that they identified a more complete list, they also describe their approach as very resource-intensive.

The external experts searched the full texts for keywords (mediator, modifier, causal, mediate, mediating, acyclic, and dag). This keyword search was unsuccessful for the current indication of RRMS (“causal" was mentioned 3 times, but in a different context, and "acyclic" was mentioned only once, see Table 3 in the external report). Nevertheless, this approach seems plausible. To simplify this step, the keyword search could be replaced with simple scrutinization of the publication to identify a directed acyclic graph (DAG). If a DAG is identified, its variables should be extracted as potential confounding variables (see Section 5.3).

Qualitative clinician interviews

Clinician involvement is the second component of systematic confounder identification in Pufulete 2022. This follows the Cochrane recommendations, which have also been adopted by the ROBINS-I tool to assess bias in non-randomized studies [12,13].

The external experts also conducted semi-structured clinician interviews (n=8). They developed a vignette-based topic guide similar to that in Pufulete 2022, with 5 scenarios based on clinical cases (see Appendix B in the external report). The clinicians were selected based on their experience in prescribing the interventions of interest (dimethyl fumarate or glatiramer acetate) to patients with multiple sclerosis.

The interviews identified 35 additional potential confounders (22% of the total of 160) that were not found in the systematic literature review. In particular, these included patient-

centred factors, such as “patient wish” or “need for safety”. These findings are consistent with those of Pufulete 2022.

Pufulete 2022 assumed that saturation could be achieved with 12 experts (6 in cardiology and 6 in cardiac surgery). The external experts had originally planned to recruit 10 participants. However, due to difficulties with recruitment, only 8 clinicians ultimately participated. The external experts state that the 8 contributions were saturated and adequately reflected the factors considered in the treatment decision. Overall, it is considered appropriate to have involved 8 clinicians for the qualitative interviews. It is not possible to assess whether fewer interviews would have also achieved saturation.

The external report only states whether a confounding variable was mentioned during an interview; it does not specify how many times this occurred. This is appropriate, given that the sole purpose of this step is to identify new potential confounders. The relevance of a potential confounder cannot be inferred from the number of clinicians who mention it.

Quantitative clinician survey

The third component involved external experts conducting a quantitative clinician survey, following the approach taken by Pufulete 2022. For this purpose, the confounders identified in the literature review and interviews were compiled. Furthermore, for the survey, the external experts summarized and reduced the identified 160 confounders (13 were covered by similar confounders, 2 could not be assessed without a study context, and 8 had not yet been identified at the start of the survey). Ultimately, 137 confounders were presented to the survey participants for evaluation. Unlike in Pufulete 2022, participants were not asked to name the 5 most relevant confounders. Instead, they were asked to rate each of the 137 potential confounders on a 5-point scale ranging from “1 – no consideration” to “5 – strong consideration”, indicating the extent to which each confounder influenced their decision to prescribe dimethyl fumarate or glatiramer acetate. Initially, the external experts had planned to ask the clinicians to name the 10 most relevant confounders, similar to the approach of Pufulete 2022. However, they ultimately decided against this in order to improve the survey's feasibility and avoid additional hurdles during recruitment. The approach of the external experts also gave the clinicians the opportunity to suggest further potential confounders in a free-text field.

The external experts invited 79 people to participate in the survey. Of these, 21 started the survey and 7 completed it. The low response rate may have been due to the considerable resources required from the clinicians to assess the large number of potential confounders. The external experts also assumed this. It is unclear from the publication what the response rate was in Pufulete 2022. It merely states that 110 clinicians completed the survey.

The planned approach of the external experts, which involves having the clinicians score the individual confounders rather than naming the 10 most relevant confounders, represents an alternative approach to assessing the relevance of confounders. This approach is understandable. Although scoring is more resource-intensive than the other possible assessment method, the assessment of each individual confounder is preferable overall to the approach chosen by Pufulete 2022, which involves naming the 5 most relevant confounders. The scores can be used to rank all potential confounders according to their relevance. This ranking can then be used as a decision-making aid (e.g. in cases where the statistical model does not converge, the scores can be used to justify which potential confounder can be eliminated first when adjusting the model). To reduce the effort involved in this confounder assessment, Section 5.3 suggests a possible approach to summarizing the list of the potential confounders of interest more intensively.

No additional potential confounders were identified by free-text field entries. This option is not included in the approach described by Pufulete 2022, and the external experts did not justify this additional step. As minimal additional information was gained, we consider this step to be unnecessary.

Categorization and classification

In the fourth and final step, the external experts assessed whether the potential confounders influenced both the treatment decision and treatment result. Of the 160 potential confounders, 148 (93%) were categorized as “likely predictors of the treatment decision”. 128 (80%) of the potential confounders were also likely to influence the treatment result. Eleven (7%) potential confounders were categorized as “unclear”, which the external experts interpreted as “unlikely to be a predictor of either the treatment decision or the outcomes”. The external experts did not categorize the potential confounder “type of MS” because only RRMS was investigated. Similarly, in a last step Pufulete 2022 categorized potential confounding variables according to whether they predicted both the treatment and the outcome, only one of them, or neither. They categorized 34 of the 70 identified potential confounders (49%) as predictors of both.

The purpose of such categorization is to identify “true confounders”. However, as establishing the relationship between variables requires strong assumptions and is therefore subject to uncertainty, it is usually impossible to identify confounders in this way. Consequently, confounding variables whose influence on the treatment decision and treatment result is uncertain should not be excluded. Therefore, this step of categorizing each individual variable is unnecessary. Only a literature-based justification can explain the exclusion of a potential confounder (see Section 5.3).

In this final step, external experts summarized the 160 potential confounders into 136 potential confounders and assigned them to different categories such as comorbidities and clinical status. For example, the “lifestyle” cluster comprises the potential confounders “tobacco use”, “type of professional work activity”, and “height and weight”. The latter comprises the 3 potential confounders “body mass index”, “overweight or obesity” and “weight”. This type of summarization was not performed by Pufulete 2022, but it was suggested in the publication's discussion as a useful step because these variables are highly correlated, and summarization could reduce the number of confounders to be considered in the statistical analysis. Section 5.2.1 describes the results of our in-house investigations regarding options for further summarization of the potential confounders identified by the external experts.

5.1.3 Feasibility of confounder identification based on Pufulete 2022

The review of the external report showed that the approach essentially met the specifications outlined in Pufulete 2022. Through a systematic literature review and clinician interviews, the external experts identified a total of 160 potential confounders. Therefore, implementing confounder identification according to the Pufulete 2022 approach is feasible in principle.

The external experts spent 62 person-days (8 hours per day) identifying these 160 confounders (see Table 6 in the external report). The estimated resources did not include project planning, management, supervision or report preparation.

The timeframe for these project components, which form part of the overall identification process, is therefore unclear. In their report, the external experts describe how their team's methodological and clinical expertise could have facilitated several steps, suggesting that the reported timeframe may be an underestimation.

To assess feasibility, it was also to be evaluated whether the timeframe specified in the commission is achievable. As subsequent consented adjustments to the procedure were made without updating the timeframe, it is not appropriate to compare the reported timeframe with that specified in the initial commission. For this reason, the timeframe reported in the external report is referred to primarily below.

The systematic literature review is the most resource-intensive component of confounder identification, accounting for an estimated 33 person-days. It should be noted that RRMS was selected on the basis that a sufficient body of evidence was assumed to be available. After removing duplicates, the search yielded 1202 hits for screening. 80 studies were identified from which relevant information was extracted. This is similar in scope to Pufulete 2022, who screened 2544 publications without duplicates and ultimately extracted information from 47 eligible publications. In indications where the expected body of evidence is smaller, substantially fewer resources will be required for this component, as there will be fewer hits

to screen and less data to extract in the literature review. Additionally, the approach adopted by the external experts not to select a saturation criterion for confounder identification in publications after 2020 is highly sensitive. Therefore, resources could be saved by applying a saturation criterion to all publications in this step.

The external experts reported that 16 person-days were required for conducting the component “qualitative interview”. Eight interviews were conducted. It is unclear how long each interview lasted. Pufulete 2022 states that each interview lasted between 26 and 45 minutes. Therefore, developing the interview guideline, as well as transcribing the interviews and analysing the information, would be the most resource-intensive parts of this component.

The quantitative survey and categorization/classification require the least resources, at 8 and 5 person-days respectively. While these 2 components require the least resources from external experts, clinicians need to invest substantial resources. However, these can be reduced markedly by summarizing the confounding variables prior to the quantitative survey (see Section 5.3.2).

5.2 In-house investigations

5.2.1 Potential confounders for the indication of RRMS

The external experts identified 160 potential confounding variables and summarized those that could be measured or assessed in a similar way, reducing them to 136 potential confounders. According to the external experts, this (small) reduction demonstrates that many of these more than 100 factors are clearly different conceptual aspects that must be considered in order to obtain unbiased estimates of the impact on treatment decisions.

Summarizing the confounders identified through the literature review and clinician interviews was not part of the approach in Pufulete 2022. However, the authors discuss this as a logical next step. Currently, however, so far there are no recommendations for this step.

With the aim of replicating a randomized study based on registry data, the complete recording of 160 or 136 confounders does not appear reasonable. On the one hand, this means that all these confounders must be recorded in the data source (e.g., a registry) for all patients included in the study, with high data quality. On the other hand, such a large number of adjustment variables is also likely to lead to problems in the statistical analyses, depending on the number of patients available in the data set to be examined [31,32]. Despite the summarization of confounders that are measured or assessed in a similar way, the list of 136 summarized confounders presented in Table 5 of the external report shows several overlapping concepts in terms of content (e.g., country and region; health-related quality of life, specific to the indication and generic).

For this reason, we examined the extent to which it is possible to summarize the identified potential confounding variables more intensively, without having to accept a relevant loss of information regarding the potential confounders.

During processing, it also became clear that some of the identified confounding variables could be excluded based on the present research question of the replication of the CONFIRM RCT (see section below). The steps taken are described below and presented in Figure 2.

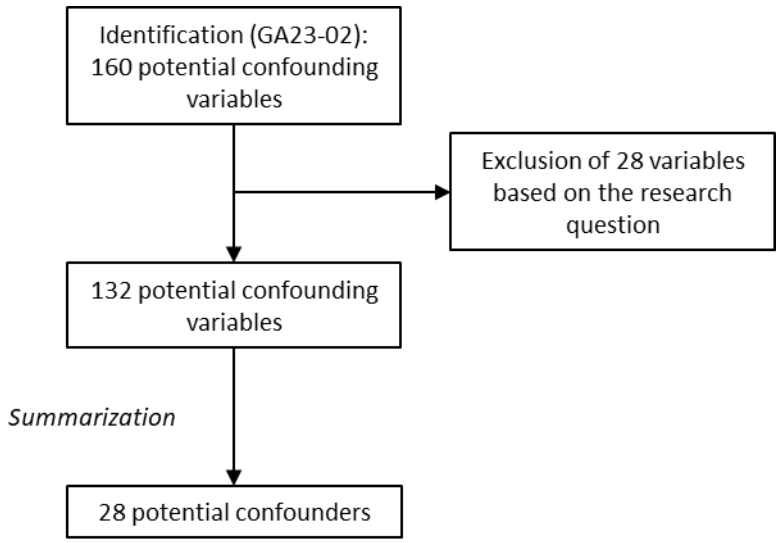

Figure 2: Summarization of potential confounding variables: results

In the following text, the term “confounding variable” refers to the variables extracted in Step 1, to ensure standardized terms. The term “confounder” refers to the results of summarizing these variables.

The results of summarizing these variables in a confounder list, including proposals for operationalization, were presented to the external experts at a workshop for discussion (see the section below).

Exclusion of variables based on the present research question

The systematic confounder identification presented here is intended to be used to conduct a non-randomized comparison of dimethyl fumarate vs. glatiramer acetate in patients aged 18 years and older with RRMS. In the following text, it is assumed that the CONFIRM RCT [24] will be replicated as a model study as an NRSI using target trial emulation. This allowed the specific inclusion and exclusion criteria of the CONFIRM RCT to be used to exclude variables that are not relevant to the present question. In addition, further methodological reasons were identified that are not relevant in terms of compliance with the methodological principles of target trial emulation. A total of 28 variables were excluded. The reasons for exclusion can be summarized as follows:

- Exclusion due to lack of relevance in the German healthcare context (n = 4)
- Exclusion due to the inclusion/exclusion criteria of the CONFIRM RCT study protocol [24] (n = 14)
- Exclusion due to methodological reasons (e.g., characteristics of the course of the study) (n = 4)
- Exclusion because the original publication was beyond the scope of the research question (n = 6)

Examples of excluded variables are provided below to improve comprehensibility. All 28 identified variables that were excluded, along with the reasons for their exclusion, are listed in Appendix B, Table 3.

Variables such as health insurance status and the cost of hospital treatment (e.g., previous hospital care costs) were excluded due to their lack of relevance in the context of the German healthcare system.

A total of 14 variables were not considered further based on their lack of relevance for the replication of the CONFIRM RCT. These are, on the one hand, variables that must have the same manifestation for all patients in accordance with the inclusion criteria. An example of this is the McDonald diagnostic criterion for the classification of multiple sclerosis (McDonald criteria, MS classification), which all patients must meet. On the other hand, these are variables that do not apply to any patient due to the exclusion criteria. For example, the presence of hepatitis is an exclusion criterion in the CONFIRM RCT. However, it should be noted that in the case of specific exclusion criteria or contraindications of the drug, it must be checked whether the confounder includes other manifestations that would contradict exclusion. For example, it is not appropriate to exclude the confounder “absolute lymphocyte count, lymphopenia” (continuous operationalization) based on the contraindication “severe lymphopenia” of dimethyl fumarate. This can only be done after restricting the operationalization of the confounder to severe lymphopenia.

Two of the 4 variables excluded for methodological reasons - time on treatment and time on study - are time-dependent and represent characteristics of the course of the study. However, for the present research question, it is assumed that only baseline values are considered as confounders. Therefore, the 2 variables (time since last visit and EDSS scores/improvement at the end of the study treatment/last check-up) were also excluded.

For a further 6 variables, the original publication considered was outside the scope of the present research question. For example, the following 3 variables were extracted from a publication investigating pregnancy-related relapses: Disease-modifying therapy (DMT) re-initiation postpartum; the number of visits 1 year post-partum; and the number of visits in

pregnancy. An existing pregnancy was an exclusion criterion for the CONFIRM RCT. Accordingly, this specific patient population is not considered further based on the present research question, which is to replicate the CONFIRM RCT.

Summarization

The remaining 132 potential confounding variables were summarized with the aim of reducing duplicates and overlaps. This should result in no more than an acceptable loss of information regarding the identified potential confounders. To this end, confounding variables that were identical or overlapped in terms of content were summarized. Ultimately, the 132 potential confounding variables were reduced to 28 potential confounders. The set of 28 potential confounders, including the proposed operationalizations, is shown in Table 1 below. The choice of operationalizations is discussed in the following section. Appendix B, Table 3 provides a list of the potential confounding variables summarized under each respective potential confounder.

For example, the following potential confounding variables were assigned to the confounder (health-related) “Quality of Life”: EQ-5D, SF-36 scores, the Multiple Sclerosis Impact Scale, the Multiple Sclerosis International Quality of Life Questionnaire, and the Sexual Satisfaction Scale Score. The following potential confounding variables were assigned to the confounder “comorbidities”: Comorbidities (general); diagnosis of depression; Hamilton Anxiety Rating Scale (HAM-A); Hamilton Depression Rating Scale (HAM-D); allergy; cardiovascular comorbidities; Charlson Comorbidity Index (CCI) score; chronic infections; dermatological diseases; diabetes; headache; skin problems; thyroid dysfunction; tuberculosis; bowel symptoms; depression (self-rated); gastrointestinal disorders; and hospitalizations in the previous year.

Some identified potential confounding variables can be assigned to different confounders. For instance, the confounding variable “magnetic resonance imaging” (MRI) status was assigned to the confounder “number of T2 lesions”. Assigning a variable to one potential confounder is sufficient, as this ensures coverage.

Table 1: Identified set of potential confounders, including proposed operationalizations, in the indication of RRMS (multipage table)

No.	potential confounder[a]	Proposed operationalization
1	Gender	Female vs. male vs. diverse
2	Age	Continuous
3	EDSS Score	Continuous
4	EDSS change in previous year	Continuous
5	Multiple Sclerosis Functional Composite (MSFC) Score	Continuous
6	Fatigue	Using a validated instrument: continuous
7	Comorbidities	Number of comorbidities[b]: continuous
8	Absolute lymphocyte count; lymphopenia	Absolute lymphocyte count: continuous
9	Study centre; Centre for Multiple Sclerosis	MS-specialized vs. non-specialized centre[b]
10	BMI	Continuous
11	Tobacco consumption	Yes vs. no or Smoker vs. former smoker vs. never smoker
12	Infratentorial lesions	Yes vs. no
13	Number of contrast agent-enriching lesions	Continuous
14	Volume of T1 hypointense lesions	Continuous
15	Number of T2 lesions	Continuous
16	Volume of T2 lesions	Continuous
17	Tolerance of previous therapy and expectations of treatment[c]	No adequate operationalization possible, see explanations in the following text
18	Previous treatment with another disease-modifying therapy	Yes vs. no
19	Patient wish	No adequate operationalization possible, see explanations in the following text
20	Health-related quality of life	Using a validated instrument: continuous
21	Number of previous relapses	Relapses in the last year: continuous
22	Severity of the relapses	Relapse treated with steroids in the last year: yes vs. no
23	Time since the diagnosis of multiple sclerosis	Time since MS diagnosis: continuous
24	Cerebrospinal fluid (CSF) status	CSF-specific oligoclonal bands: yes vs. no
25	Type of professional activity	Physical labour: yes vs. no
26	Level of education	Education time in years: continuous or No degree or no higher education vs. degree or higher education
27	Ethnicity	White, Asian, black or African-American, others
28	Recent pregnancy (in the last 3 months)	Yes vs. no

Table 1: Identified set of potential confounders, including proposed operationalizations, in the indication of RRMS (multipage table)

No.	potential confounder[a]	Proposed operationalization
a. In the original German working paper, the confounders reduced by summarization were translated from English into German. All identified potential confounding variables, including the allocation to the respective potential confounders, can be found in Appendix B, Table 2 and Table 3. In these tables, the confounding variables were not translated in the original German working paper in order to avoid translation errors. b. This operationalization was discussed in the workshop with the external experts; further detailed information can be found in the following section on the operationalization of potential confounders. c. This variable contains the components: reasons for discontinuation of previous treatment, efficacy (known or expected), tolerance, side effects of other medications BMI: body mass index; EDSS: Expanded Disability Status Scale; MS: multiple sclerosis		

Of the 28 potential confounders, 26 were identified through the literature review and 18 through clinician interviews. The potential confounders "patient wish" and "type of professional activity" were only identified through the interviews, not the literature.

Operationalization of potential confounders

Finally, suitable operationalizations of the potential confounders were proposed and are presented in Table 1. Where possible, a continuous operationalization was favoured as this provides the greatest amount of information. For example, a continuous operationalization of the EDSS score was favoured over a dichotomous operationalization based on a cut-off value of 3.5 (≤3.5 vs. >3.5). Additionally, it can be evaluated whether ordinal-scaled confounders can be classified as continuous variables.

If continuous operationalizations are not possible, a categorical operationalization is selected instead. Due to the higher information content, manifestations should not be summarized if possible. Summarizations of manifestations (e.g., dichotomizations) can only be considered if statistical models do not converge. These summarizations must be pre-specified. For example, if there are too few patients in a category (smoker vs. former smoker vs. never smoker) and the statistical model does not converge, a dichotomization of the confounder “tobacco use ” (yes vs. no) could be pre-specified.

It was not possible to adequately operationalize the 2 confounders “patient wish” and “tolerance of previous therapy and expectations of the treatment”. See the following section for further information.

Results of the workshop with the external experts

The list of confounders and the selected operationalizations were discussed at a workshop with external experts. The points discussed at the workshop, along with the conclusions drawn regarding the confounders shown in Table 1, are presented below.

Unmeasurable confounders

One of the key topics discussed at the workshop was how to operationalize the confounder "patient wish". This confounder is complex and takes into account personal expectations and previous individual experiences of patients. According to the external experts, interactions between the various parameters that determine this confounder are to be expected, and "patient wish" should be operationalized as granularly as possible. It does not seem possible in this form to adequately reflect this confounder in the sense of shared decision making, with all the different individual points of view. This also applies in part to the factors that influence the physician's treatment decision. For example, the physician's decision may be influenced by reasons for previous discontinuation of therapy, suspected or known drug effects, drug tolerance or side effects of other medications. The potential confounder "tolerance of previous therapy and expectations of treatment" represents such a factor. Similar to the patient wish, adequate operationalization is not possible here either due to the complexity and multifaceted nature of the issue. The suggestion by the external experts to potentially break down the reasons for the next treatment cannot be taken into account for reasons of feasibility due to the large number of additional different variables involved.

Since it is not possible to operationalize and thus record the 2 identified potential confounders ("patient wish", "tolerance of previous therapy and expectations of treatment"), these 2 relevant potential confounders cannot be included in the investigation (unmeasurable confounders, see **Glossary**). These potential confounders cannot be adjusted for in the comparative NRSI. This must then be considered when interpreting the results [10].

Any potential confounders that cannot be measured or operationalized must always be named, and which variables this refers to and the reasons why operationalization is not possible must be transparently and comprehensively explained (see Section 5.3 for recommendations). The implications of failing to consider these confounders should also be discussed when interpreting the results.

Comorbidities

The workshop with external experts also discussed the confounder "comorbidities". Instead of a single summarized confounder for comorbidities, the external experts recommended including important comorbidities (such as psoriatic arthritis or gastrointestinal diseases) as separate confounders. The collective category would result in a loss of information. For example, dimethyl fumarate is prescribed more often in cases of psoriatic arthritis, and glatiramer acetate more often in cases of gastrointestinal complaints or lymphopenia. This recommendation by the external experts is in principle understandable and should be implemented where possible in order to lose as little information as possible. This is because, in principle, a summarization assumes that all comorbidities considered are equally relevant in terms of the extent and direction of bias if they are not taken into account. It makes sense to consider comorbidities individually, especially if only a few comorbidities are relevant to a particular question. Nevertheless, the recommendation should be discussed in light of the number (n = 18) of comorbidities identified in the present case. Depending on the number of patients available for statistical analysis, it must be weighed up whether the loss of information resulting from summarization is acceptable. In the present case, summarizing the various comorbidities into a confounder "number of comorbidities" is considered acceptable overall.

Study centre

In addition, with regard to the chosen operationalization of the potential confounder "study centre" (specialized MS centres vs. non-specialized MS centres), the external experts commented that the study centres should be considered as a stratification characteristic. They considered each individual study centre to be a relevant confounder, as the capacities available (e.g., for infusions or MRI) vary greatly depending on the centre and, in some cases, different treatment philosophies are applied. This could have a potential influence on the treatment result. This recommendation does not appear to be feasible in view of the number of German centres treating patients with multiple sclerosis and the resulting sharp increase in the number of manifestations to be taken into account in the statistical model. Depending on the number of patients available, it should be discussed whether there are other suitable operationalizations in addition to those proposed in Table 1.

Summary

A total of 160 potential confounding variables identified by external experts were summarized into 28 potential confounders. Particular emphasis was placed on the methodological approach and the necessary transparent presentation of this approach. The potential confounding variables were summarized with the aim of reducing duplicates and overlaps. This should result in no more than an acceptable loss of information regarding the identified potential confounders. Based on the experience gained, overarching recommendations are derived and described in Section 5.3.

Similar to Pufulete 2022, selecting confounders based on a causal understanding of the relationships between the various variables was not the aim of the confounder identification and summarization described here. Confounding variables with overlapping content that are likely to be highly correlated with each other should be summarized. Whether the identified potential confounders are possibly mediators or colliders (see **Glossary**) was not checked. The exclusion of such variables should be justified based on the literature, and the interrelations between the potential confounders should be visualized using a DAG [1,2].

5.2.2 Further suggestions for systematic confounder identification in the literature

Figure 3 shows the results of the focused information retrieval for relevant publications on systematic confounder identification according to the study inclusion criteria (see Section 4.2.2). The MEDLINE search strategy can be found in Section A.2. The search was conducted on 13 December 2024.

The citations of the hits that were checked in full text, but excluded, can be found in Section A.1.

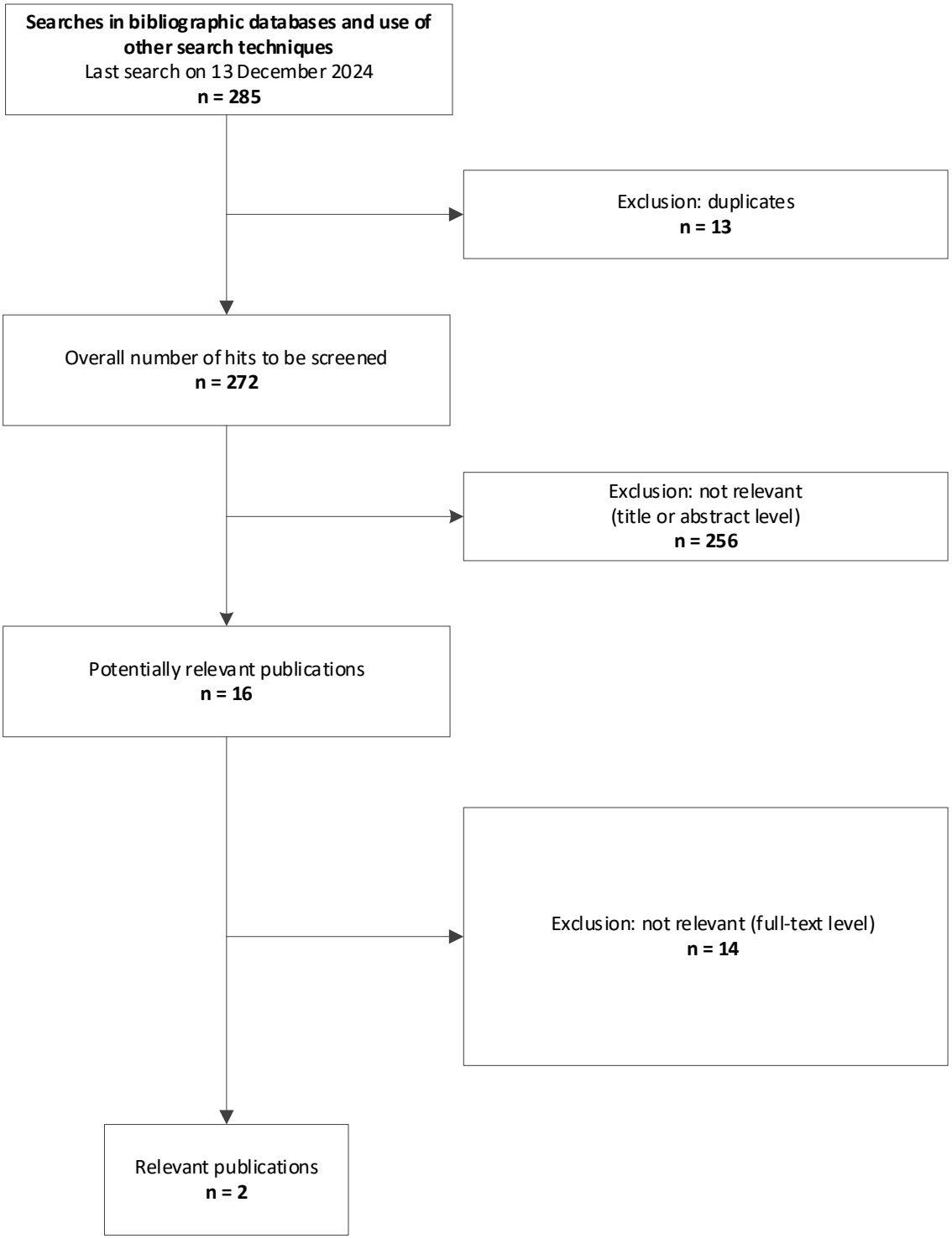

Figure 3: Results of focused information retrieval and study selection for publications on systematic approaches to confounder identification

The 2 publications rated as relevant are the journal publication by Pufulete 2022 [11] and the 2023 HTA report by Harris and Pufulete [27], which describes the methods for confounder

identification of Pufulete 2022. No additional publications on systematic approaches to confounder identification were identified.

5.3 Proposal for requirements for systematic confounder identification

In order to evaluate intervention effects based on a non-randomized comparison, a set of potential confounders should be identified beforehand as part of the study planning process. To this end, concrete recommendations for systematic confounder identification were derived based on the results of a confounder identification by external experts and on further investigations conducted in-house. These represent a proposal for the specific requirements for systematic confounder identification.

Although both the external report and the publication by Pufulete 2022 refer to an indication with an extensive body of evidence, the situation of rare(r) diseases was also considered when deriving the recommendations, as this represents the usual situation for areas such as routine practice data collection.

The developed proposal for systematic confounder identification comprises 3 main steps, as shown in Figure 4.

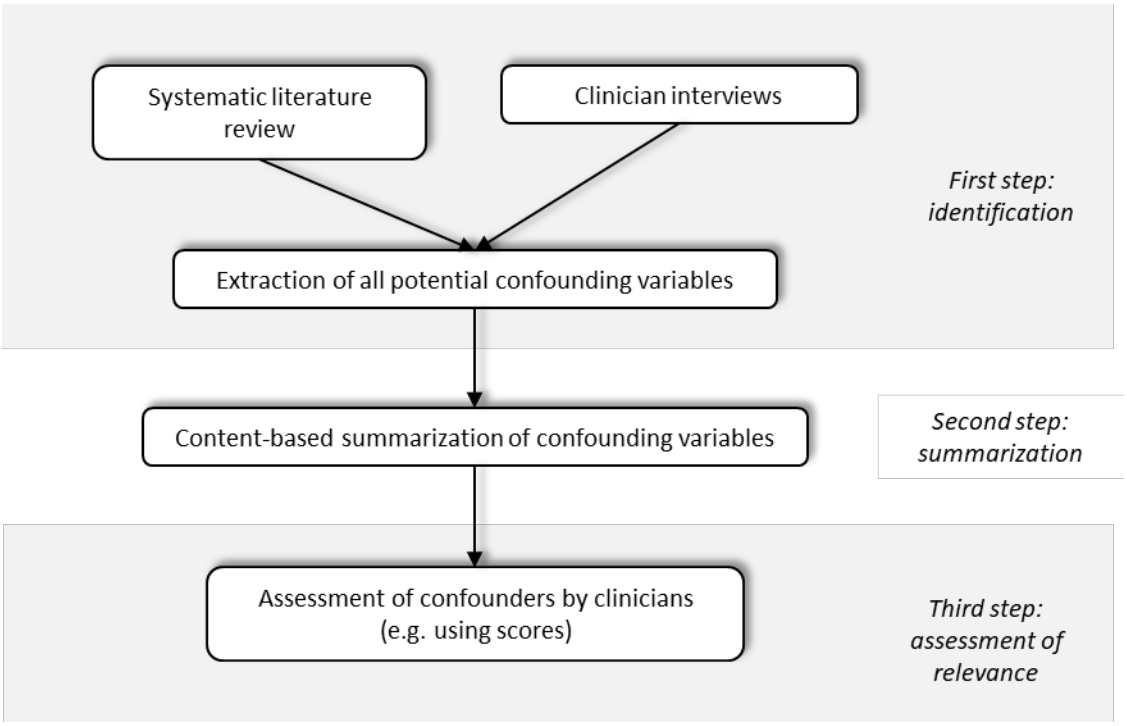

Figure 4: Proposed 3-step systematic approach to confounder identification

Following Pufulete 2022, the first step (identification) comprises 2 components: a systematic literature review and clinician interviews. Once all potential confounding variables have been extracted, the second step involves summarizing any confounding variables that overlap in

content in order to save resources. This step builds upon the methods described in Pufulete 2022. In the third step, clinicians assess the relevance of the summarized potential confounders.

Detailed and transparent documentation of the approach and results is required to enable the confounder identification process to be understood. This documentation should be published as part of the study protocol of the planned comparative NRSI.

The recommendations formulated for each step of the process are explained below. These recommendations can also be found as a checklist in Appendix C.

5.3.1 Step 1: Identification

Primary publications on RCTs and cohort studies should be considered for systematic information retrieval. Guidelines or single-arm studies may subsequently be considered

Depending on the quality and quantity of the available literature on the therapeutic indication of interest, a decision should be made regarding which study type to include. Initially, suitable RCTs and cohort studies should be included, as recommended by Pufulete 2022 and the external experts. Particularly for questions relating to indications with an expected smaller body of evidence (e.g., rare diseases), if suitable RCTs and cohort studies are unavailable, single-arm studies (including before-and-after comparisons) and the patient characteristics listed in tabular form in the study publications may be considered. In these cases, guidelines can also be considered in the search, and screened for prognosis or risk factors and factors that determine treatment decisions.

However, it is not advisable to rely solely on secondary literature. This would only be possible if the secondary literature is of a high quality and up to date, and if it covers the entire research question and includes a sufficient number of suitable primary studies. It must also provide enough information to assess confounders (e.g., adjustment variables). Previous experience of reviewing study documents in the context of routine practice data collection shows that this is generally not the case [21,22].

The inclusion and exclusion criteria for the study pool used for confounder extraction should closely align with the research question

The inclusion and exclusion criteria for the literature review should be tailored to the specific research question. These criteria should be defined based on the population, intervention, comparison and outcome of the research question (PICO). In particular, the definition of the population (P) should closely align with the research question. For instance, in the present investigation, publications investigating relapses during pregnancy are outside the scope of the research question of the RCT to be replicated (CONFIRM) and therefore do not need to be considered for confounder extraction.

With regard to the intervention and the comparator, only one of the two components usually needs to be fulfilled for inclusion in the study. This means that studies analysing only the comparator ("C" of PICO) and not the intervention ("I" of PICO) can also be included.

Including a specific outcome (O) in the inclusion criteria is particularly recommended if the research question explicitly focuses on this outcome. However, such a restriction is generally unnecessary for questions relating to routine practice data collection or the benefit assessment of drugs, since many patient-relevant outcomes are typically examined.

For comparative observational studies, all adjustment variables should be extracted, regardless of statistical significance. For randomized or non-comparative studies, all patient characteristics should be extracted.

Following the approach of Pufulete 2022 and the external experts, all factors used for adjustment in statistical analyses should be extracted as potential confounders. For RCTs - or, depending on the available evidence, for single-arm studies - the patient characteristics reported in tabular form ("Table 1") in the study publication should be extracted. It is particularly important here to clearly indicate which potential confounding variable was extracted from which source (literature reference, ideally including the specific citation). This information can be included in an appendix to the main document on confounder identification.

If a DAG is identified in a publication, the variables included in the DAG should be extracted. This is because a DAG is expected to contain more information. For example, a DAG may contain variables that were not included in the statistical model for adjustment because they were not recorded in the dataset.

According to the classic definition of a confounder, it influences both the treatment and the outcome of interest [21,22]. Based on this definition, a separate set of potentially relevant confounders should be identified for each outcome of interest. In the external report, these were the annual relapse rate, confirmed disability progression using EDSS, health-related quality of life and AEs. Especially for analyses in which confounder adjustment using propensity score methods (such as matching or weighting) is planned, it is sufficient for pragmatic reasons to define a superordinate set of all potentially relevant confounders across all outcomes, which can be used for the analysis of all outcomes.

When extracting information from primary publications, it is advisable that a suitable saturation criterion is applied.

In line with the approach adopted by Pufulete 2022 and external experts, it makes sense to define an objective and plausible saturation criterion in advance for data extraction in primary publications. In Pufulete 2022, extraction ceased if no new potential confounder was identified within 10 consecutive studies. However, the external experts only applied a

saturation criterion to studies prior to 2020, stopping extraction from the randomly selected studies if no new confounders were identified within 15 consecutive studies. It is particularly useful to apply a saturation criterion when the number of identified publications is high, as this saves resources. Since other sources are also used to identify potential confounders, such as qualitative clinician interviews, it is reasonable to assume that using 10 consecutive studies as a saturation criterion is appropriate.

Clinician interviews should be conducted.

The identification of potential confounders by clinicians is an important component of the proposed 3-step process.

The interview process should be transparently documented. For example, Pufulete 2022 and the external experts developed vignette-based topic guides that describe various clinical case scenarios (see the supplement to Pufulete [11] or Appendix B of the external report). It should be emphasized that this step should not only involve validating the potential confounders already identified in the literature review, but also identifying further potential confounders.

It is not possible to make a clear recommendation regarding the number of clinicians to be involved. Pufulete 2022 interviewed 6 clinicians per indication (i.e. 12 in total) and assumed that this would be sufficient to achieve saturation. The external experts involved 8 clinicians. The number of interviews actually required to achieve (theoretical) saturation should be estimated depending on the indication. As a guide, 9 to 17 interviews have been found to achieve saturation for various types of research questions in qualitative research [33]. This aligns with the approach of Pufulete 2022 and the external experts. Decision criteria for the number of interviews to be conducted include the indication, the expected body of evidence and the variability in the course of the disease.

In the current indication of RRMS, 35 out of 160 potential confounding variables (22%) were identified through clinician involvement and were not found in the literature. Of the summarized list of potential confounders, 2 (7%) remained that were named exclusively in the interviews. In Pufulete 2022, 11/70 (16%) of potential confounding variables were not identified through a literature review. Of these, 7 (10%) were identified exclusively through qualitative interviews.

Even though the proportion of confounders identified through interviews appears to be rather small in the examples provided, it should not be concluded that this step is unnecessary. Firstly, it is foreseeable that interviews will be more important in indications where fewer publications are available. Secondly, both Pufulete 2022 and the external experts identified patient-centred factors in the interviews that have not yet been taken into account in the literature. Even if these are variables that are difficult or impossible to operationalize,

knowledge of the relevance of these factors contributes to a more differentiated description of the black box of "unmeasured confounding" (see **Glossary**). This should be considered when interpreting the results of the adjusted non-randomized comparison.

The names of the clinicians involved, along with any potential conflicts of interest, should be disclosed.

The number of clinicians involved and their speciality must be clearly documented. For documentation purposes, the names of the clinicians and details of any potential conflicts of interest must be provided (see Appendix C [Recommendations checklist]). This information can be included in an appendix to the main document on confounder identification. Knowledge of the study protocol of the planned study for which the confounder identification is being performed, or of the dataset of the data source used, can influence the clinicians' assessment. Ideally, in the case of routine practice data collection, the clinicians (or at least some of them) should therefore have no knowledge of the dataset of the data source used (e.g., the disease registry).

5.3.2 Step 2: Summarization

Any confounding variables with overlapping content should be summarized. This should be done after extraction. Transparent and comprehensible documentation must be provided.

Once all identified potential confounding variables have been extracted from 1) the systematic literature review and 2) qualitative clinician interviews, it is advisable that duplicated or strongly overlapping variables are summarized where appropriate. In this case, potential confounding variables that measure the same construct or whose content overlaps should be summarized. Appropriate expertise in the therapeutic indication should be taken into account. This approach means that potential confounding variables that are strongly correlated with each other are considered together.

During the summarization process, identified potential confounding variables may fall under different potential confounders. Assigning them to one of the potential confounders is sufficient, as this ensures coverage.

The summarization process should take place after all identified potential confounding variables have been extracted in order to ensure transparent and comprehensible documentation. It must be possible to understand which potential confounding variables have been summarized under which potential confounders.

It makes sense to conduct the "summarization of overlapping potential confounding variables" step before the clinicians' assessment of the relevance of potential confounders.

This minimizes the resources required to assess each confounder individually, without having to accept a relevant loss of information.

For each identified confounding variable, it should be checked whether it falls within the scope of the research question.

Due to the intensive examination of the content of each identified potential confounding variable, it makes sense to check for each variable during the summarization step whether it falls within the scope of the research question and whether it meets the methodological and content-related criteria of a confounder for the research question:

- For each confounder identification that is carried out, for example, as part of routine practice data collection, there is a previously defined research question for which inclusion and exclusion criteria are defined in the same way as for an RCT. If a variable is an exclusion criterion in the context of the present research question, it does not need to be considered further.
- In order to replicate a randomized study, only variables that are already known at the start of the study (baseline) may be considered as confounders. Any variables that represent characteristics of the course of the study must be excluded.
- Variables that can only take on one manifestation in the research question (e.g., all patients must have been identified on the basis of a specific diagnostic criterion) do not constitute confounders for this question. To identify these variables, it is advisable to compare them with the specific inclusion and exclusion criteria of the planned study.
- If the research question is to be investigated using routine practice data from German study centres, variables such as previous treatment costs can be excluded based on their lack of relevance in the German healthcare context.
- If prospective data collection is planned, all variables that are only relevant for retrospective data collection (e.g., time since the last visit) can be excluded.

Any potential confounder excluded due to content-related considerations must be justified based on the literature.

The assumptions made to identify a variable as a potential confounder are usually subject to uncertainty. Similarly, knowledge of the relationships between the various identified variables will never be complete. It is therefore necessary to provide a literature-based justification for excluding a variable from the list of identified potential confounding variables. It is not sufficient for a variable to be classified as irrelevant by the clinicians involved. Similarly, the lack of a suitable instrument for recording a potential confounder is not a reason for excluding a potential confounder (see also the comments below on operationalization).

As a rule, adjustments for potential mediators or colliders should be avoided in the statistical model (see, for example, Williams 2018 [2]). Variables should generally only be excluded based on literature. If there is merely a suspicion that a variable is a collider, for example, it can be excluded from the model by means of a sensitivity analysis.

Confounders whose relevance is unclear should be recorded and considered in the statistical analysis.

An operationalization must be specified for all identified potential confounders.

The study protocol or statistical analysis plan must specify a suitable operationalization for all identified potential confounders a priori.

Based on the information content, continuous operationalizations are to be preferred. If continuous operationalizations are not possible, a categorical operationalization is chosen. Due to the higher information content, it is preferable not to summarize the manifestations. Only if the statistical model does not converge can summarizations of manifestations (e.g., dichotomizations) be considered. These must be specified in advance.

If variables have been identified as potential confounders that are measured using an instrument for patient-reported outcomes, such as health-related quality of life or fatigue, the operationalization should ensure that these are assessed using a validated instrument.

In principle, the appropriate operationalization should also be chosen in terms of feasibility. For example, it is understandable if the operationalization of a confounder only covers a limited period of time (e.g., number of relapses in the previous year instead of number of relapses in the previous 3 years). Another example is the operationalization of comorbidities. In the present example of RRMS, a whole range of comorbidities were identified as potential confounders. If these are summarized as one superordinate confounder, “comorbidities”, this results in the greatest loss of information. An alternative would be to select the most important comorbidities or to consider each individual comorbidity separately as a confounder. Depending on the number of patients available, it must be weighed up whether the loss of information resulting from summarization is acceptable.

Although potential confounders that cannot be measured or operationalized cannot be included in the analysis, they should be named and considered when interpreting the results.

There are potential confounders, such as patient wish or physician’s decision, which cannot be adequately recorded or operationalized due to their complexity. These potential confounders should nevertheless be clearly named and reasons given as to why operationalization is not possible. This also applies to potential confounders based on complex

examination methods that are not routinely performed in everyday care and are therefore not recorded (e.g., regular examination of cerebrospinal fluid outside the MS diagnosis).

Although these potential confounders cannot be included in the analysis, they must be considered when interpreting the results [10].

5.3.3 Step 3: Assessment of relevance

As the third step, clinicians should assess the relevance of the potential confounders. It is advisable that they use a scoring system.

There are various approaches to assessing the relevance of potential confounders, such as ranking the 5 most relevant confounders or having clinicians assign a score to each identified confounding variable based on how strongly it influences the treatment decision (e.g., on a 5-point scale from 1 [no consideration] to 5 [strong consideration]). The advantage of a scoring system is that each individual potential confounder receives a score, providing substantially more information than simply naming the 5 most relevant confounders. As this type of assessment is more resource-intensive, it is advisable to summarize the identified confounders in advance wherever possible to save resources.

The scores assigned should not be used to exclude potential confounders. All identified potential confounders should be recorded. However, the result of the assessment step can be used in the data analysis for content-based variable selection in cases where the statistical model does not converge. Based on the score, a decision can then be made as to which confounder is not included in the model. For example, the potential confounder with the lowest score is the first to be excluded from the model.

Ideally, the group of clinicians involved in this step should not consist exclusively of the clinicians involved in step 1 (“identification”). Here, too, it is necessary to name the clinicians and disclose any potential conflicts of interest.

6 Discussion

In comparative NRSI, systematic identification of potential confounders is a prerequisite for adequate confounder control. The aim of confounder control is to obtain an estimate of the causal effect of interest that is as unbiased as possible, despite the lack of randomization. Such non-randomized study designs are playing an increasingly important role in benefit assessments, for example in the context of routine practice data collection pursuant to §35a (3b) of the German Social Code Book V (SGB V), which, according to legal requirements, must be carried out without randomization. In addition, it is crucial to adequately identify and adjust for potential confounders when only single-arm studies are available for the intervention of interest and a comparison of the single arms of different studies (or with an external control) is planned.

Despite the number of comparative NRSI, Pufulete 2022 [11] is the only published proposal for systematic confounder identification to date. Furthermore, focused information retrieval did not identify any further publications on systematic approaches.

Pufulete 2022 propose 3 methodological components for systematic confounder identification: a systematic literature review, clinician interviews, and a clinician survey. The present studies in the indication of RRMS confirmed the importance of the 3 components. A purely literature-based approach would not have identified all important known confounders (such as patient-centred factors). The sometimes inadequate reporting of potential confounders that are identified and considered in the statistical model is a problem [34]. It can be assumed that clinician interviews are even more important when there are only a small number of studies in the available literature that can be used to identify possible confounders. This is to be expected for a number of indications for which routine practice data collection is conducted, largely involving rare diseases.

The Pufulete 2022 approach, which incorporates the involvement of clinicians as well as a systematic literature review, is considered to be very resource-intensive by the authors themselves. This was also confirmed by external experts with regard to confounder identification in the indication of RRMS. Identifying the 160 potential confounding variables took 62 person-days. Nevertheless, the approach is feasible in principle. The systematic literature review accounted for the largest proportion of resources used, but also for the greatest number of confounders identified. It is unclear to what extent this level of resource investment can be transferred to other indications. Both Pufulete 2022 and the present investigation focused on indications with an extensive body of evidence.

The 160 variables identified by external experts included several that were similar or overlapping in content. In the present investigation, these variables were summarized to

create 28 potential confounders. Summarizing these potential confounding variables is a central step in terms of feasibility, not only for data collection but also for statistical analyses.

Even if summarization was successful, it cannot generally be assumed that an existing data source (e.g., a registry) already contains a comprehensive recording of all potential confounders. In particular in cases with a comparison of single arms of different studies or with an external control, retrospective data that have already been collected are used. These data sources often do not contain information on the identified potential confounders, or if information is available, it is often incomplete [22,35-37]. Retrospective data collection is therefore often not an option. This is in line with previous experiences from routine practice data collection, where prospective data collection has been necessary in all cases to ensure the required recording of confounders.

The set of confounders identified in the external report and in our in-house investigations in this working paper is to be used to replicate a specific RCT. Such a replication of an RCT can subsequently be used, for example, to answer transferability issues (e.g., whether the effects observed in an RCT in younger adults [18 to 55 years] are transferable to older adults [> 55 years]). If, for example, the replication of the results of an RCT in younger adults using registry data was successful, these registry data can then be used to investigate the research question in older adults. If such a question is the reason for confounder identification, this should be taken into account when defining the inclusion criteria for the literature review. For instance, the inclusion criteria should reflect both populations (in the example mentioned, younger and older adults). If some of the identified potential confounders are only relevant for 1 of the 2 questions, they can be excluded for sub-questions or included for only one of the research questions, provided this is justified.

Whether a particular variable constitutes a confounder is a question of causality and cannot be statistically tested in the data analysed [38]. In certain cases, purely prognostic factors can also become confounders due to the protocol [39]. Similarly, there is usually insufficient information available to reliably specify the causal relationships between the various variables that determine the treatment decision and the treatment result [38]. Similarly, Pufulete 2022 describe that there is no certainty that the variables they selected are sufficient to control for bias caused by confounding. Accordingly, the decisions made in this report, particularly with regard to the performed summarization or the proposed operationalizations, are merely assumptions that cannot be verified. Whether the identified potential confounders are possibly mediators or colliders was not verified. The exclusion of such variables should be justified based on the literature, and the interrelations between potential confounders should be visualized using a DAG [1,2].

Even when confounders are carefully and systematically identified and recorded in high quality and fully considered in the analysis, residual or unmeasured confounding remains a

risk due to unknown or further excluded or unmeasurable confounders. These uncertainties must be considered when interpreting the results. One option is to apply a shifted null hypothesis, as suggested for routine practice data collection [10]. An effect (positive or negative) is only derived if the confidence interval for the observed effect exceeds a defined threshold.

In summary, the systematic approach to confounder identification proposed in this working paper is considered feasible. However, the present investigations also highlight the resource-intensive nature of confounder identification in comparative NRSI. The next step is to examine other possible options for saving resources. Firstly, this refers to ways of reducing the resources required for confounder identification (e.g., by restricting the literature search). Secondly, the number of potential confounders can be reduced based on their relevance (e.g., by means of their potential influence on effect estimates), thereby reducing the resources required to record them. Any options for saving resources must be weighed against the associated loss of information and potential loss of certainty of results.

The only way to reduce the increased requirements for the extent of data collection caused by the necessary identification and recording of potential confounders and to avoid the existing residual uncertainty (e.g., regarding unmeasurable confounders and their lack of consideration in the analysis) is randomization. Conducting an RCT eliminates the need to identify and record potential confounders, markedly reduces the required number of patients and ultimately increases feasibility. At the same time, it substantially increases the certainty of the results. RCTs are also in principle feasible for rare diseases [40]. The frequent criticism that RCTs do not have good external validity could be addressed by broad inclusion and exclusion criteria in the sense of a pragmatic RCT [41-43]. Within RCTs, registry-based RCTs represent a meaningful option because the use of an existing data structure results in further efficiency gains [44].

7 Conclusion

The external experts applied the approach suggested by Pufulete 2022 to identify a total of 160 potential confounding variables. These were summarized into 28 potential confounders in further in-house investigations. This demonstrated that confounder identification based on Pufulete 2022 was feasible in principle, for example in the context of routine data collection.

The following recommendations for systematic confounder identification in 3 steps were derived from the experiences gained during the investigations:

1) Confounder identification

Potential confounding variables should be identified through a systematic literature review and clinician interviews.

2) Confounder summarization

Confounders that measure the same construct or overlap in terms of content should be summarized. A clear operationalization must be defined for each potential confounder.

3) Confounder assessment

The relevance of each potential confounder should be assessed in a clinician survey. The result can inform a content-based selection of variables.

The second step in particular builds on the methods proposed by Pufulete 2022 and aims to save resources without losing relevant information.

The systematic approach to confounder identification is resource-intensive, despite the described resource-saving options. The next step is to examine other possible ways to save resources in both the process of confounder identification and the associated recording of potential confounders. Any resource-saving options must be weighed against the associated loss of information and potential reduction in the certainty of results.

In order to minimize confounding bias in a comparative NRSI, systematic confounder identification is required. However, there will always be uncertainty as to whether all relevant confounders have actually been identified and can be considered in the analysis. Both unknown and unmeasurable confounders remain an unresolved problem, meaning that in comparative NRSI, the added benefit of a treatment can only be inferred if the effects are sufficiently large. The only sensible alternative is to conduct an RCT, such as a registry-based RCT, which does not require confounder identification and control.

References

1. Feeney T, Pires Hartwig F, Davies NM. How to use directed acyclic graphs: guide for clinical researchers. BMJ 2025; 388: e078226. https://doi.org/10.1136/bmj-2023-078226.

2. Williams TC, Bach CC, Matthiesen NB et al. Directed acyclic graphs: a tool for causal studies in paediatrics. Pediatr Res 2018; 84(4): 487-493. https://doi.org/10.1038/s41390-018-0071-3.

3. Catalogue of bias collaboration, Lee H, Aronson JK et al. Catalogue Of Bias; Collider bias [online]. 2019 [Accessed: 25.02.2025]. URL: https://catalogofbias.org/biases/collider-bias/.

4. Sendor R, Sturmer T. Core concepts in pharmacoepidemiology: Confounding by indication and the role of active comparators. Pharmacoepidemiol Drug Saf 2022; 31(3): 261-269. https://doi.org/10.1002/pds.5407.

5. 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.

6. Desai RJ, Wang SV, Sreedhara SK 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 2024; 384: e076460. https://doi.org/10.1136/bmj-2023-076460.

7. Hernan 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.

8. Wang SV, Schneeweiss S, RCT-DUPLICATE Initiative. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. JAMA 2023; 329(16): 1376-1385. https://doi.org/10.1001/jama.2023.4221.

9. Mathes T. Anforderung 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.

10. 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: 02.06.2020]. URL: https://www.iqwig.de/download/A19-43_Versorgungsnahe-Daten-zum-Zwecke-der-Nutzenbewertung_Rapid-Report_V1-1.pdf.

11. 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.

12. Reeves BC, Deeks JJ, Higgins JPT et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.5; Chapter 24: Including non-randomized studies on intervention effects [last updated October 2019] [online]. 2024 [Accessed: 12.03.2025]. URL: https://training.cochrane.org/handbook/current/chapter-24.

13. 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.

14. Vanier A, Fernandez J, Kelley S et al. Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health. BMJ Evid Based Med 2024; 29(1): 1-5. https://doi.org/10.1136/bmjebm-2022-112091.

15. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden; Version 7.0 [online]. 2023 [Accessed: 02.09.2024]. URL: https://www.iqwig.de/methoden/allgemeine-methoden_version-7-0.pdf.

16. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Exagamglogen Autotemcel (Sichelzellkrankheit); Bewertung gemäß § 35a SGB V; AbD-Konzept [online]. 2023 [Accessed: 23.01.2025]. URL: https://www.iqwig.de/download/a23-49_exagamglogen-autotemcel_abd-konzept_v1-0.pdf.

17. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Etranacogen Dezaparvovec (Hämophilie B) – Bewertung gemäß § 35a SGB V; AbD-Konzept [online]. 2023 [Accessed: 11.07.2023]. URL: https://www.iqwig.de/download/a22-83_etranacogen-dezaparvovec-haemophilie-b_abd-konzept_v1-1.pdf.

18. CSL Behring. Study Protocol; 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 G-BA [online]. 2024 [Accessed: 24.03.2025]. URL: https://www.g-ba.de/downloads/40-268-10667/2024-07-18_AM-RL-XII_Etranacogen-Dezaparvovec_2022-AbD-005_Studienunterlagen.pdf.

19. Novartis Gene Therapies. Routine data collection and evaluations of onasemnogene abeparvovec in Germany, Study Protocol [online]. 2024 [Accessed: 24.03.2025]. 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.

20. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anwendungsbegleitende Datenerhebung zu Onasemnogen-Abeparvovec: Prüfung des Studienprotokolls und des statistischen Analyseplans; Addendum zum Auftrag A20-61 [online]. 2021 [Accessed: 17.03.2025]. URL: https://www.iqwig.de/download/a21-107_anwendungsbegleitende-datenerhebung-onasemnogen-abeparvovec-pruefung-protokoll-und-sap_addendum-zum-auftrag-a20-61_v1-0.pdf.

21. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anwendungsbegleitende Datenerhebung zu Brexucabtagen Autoleucel: Prüfung des Studienprotokolls (Version 1.0) und des statistischen Analyseplans (Version 1.0); 2. Addendum zum Projekt A21-130 [online]. 2023 [Accessed: 17.03.2025]. URL: https://www.iqwig.de/download/a22-136_brexucabtagen-autoleucel_addendum-zum-projekt-a21-130_v1-0.pdf.

22. 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: 17.03.2025]. URL: https://doi.org/10.60584/A23-99.

23. 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: 17.03.2025]. URL: https://doi.org/10.60584/A24-57.

24. Fox RJ, Miller DH, Phillips JT et al. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. N Engl J Med 2012; 367(12): 1087-1097. https://doi.org/10.1056/NEJMoa1206328.

25. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clin Epidemiol 2018; 10: 771-788. https://doi.org/10.2147/CLEP.S166545.

26. Zeng J, Gensheimer MF, Rubin DL et al. Uncovering interpretable potential confounders in electronic medical records. Nat Commun 2022; 13(1): 1014. https://doi.org/10.1038/s41467-022-28546-8.

27. Harris J, Pouwels KB, Johnson T et al. Bleeding risk in patients prescribed dual antiplatelet therapy and triple therapy after coronary interventions: the ADAPTT retrospective population-based cohort studies. Health Technol Assess 2023; 27(8): 1-257. https://doi.org/10.3310/MNJY9014.

28. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol 2019; 34(3): 211-219. https://doi.org/10.1007/s10654-019-00494-6.

29. Hirt J, Hemkens LG. Systematische Confounderidentifikation in der Indikation schubförmig remittierende multiple Sklerose (RRMS); Supplement (Anhang F) zur externen Expertise zum IQWiG-Arbeitspapier GA23-02. [Demnächst verfügbar unter: https://www.iqwig.de/download/ga23-02_supplement-zur-externen-expertise_anhang-f.xlsx].

30. Viswanathan M, Berkman ND, Dryden DM et al. Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank [online]. 2013 [Accessed: 13.03.2025]. URL: https://www.ncbi.nlm.nih.gov/books/n/methresbias/pdf/.

31. Harrell FE Jr. Regression Modeling Strategies; with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. New York: Springer; 2001.

32. Peduzzi P, Concato J, Kemper E et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49(12): 1373-1379. https://doi.org/10.1016/s0895-4356(96)00236-3.

33. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. Soc Sci Med 2022; 292: 114523. https://doi.org/10.1016/j.socscimed.2021.114523.

34. Petersen JM, Barrett M, Ahrens KA et al. The confounder matrix: A tool to assess confounding bias in systematic reviews of observational studies of etiology. Res Synth Methods 2022; 13(2): 242-254. https://doi.org/10.1002/jrsm.1544.

35. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Amivantamab (NSCLC) – Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung [online]. 2022 [Accessed: 17.03.2025]. URL: https://www.iqwig.de/download/a22-05_amivantamab_nutzenbewertung-35a-sgb-v_v1-0.pdf.

36. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Ravulizumab (Neuromyelitis-optica-Spektrum-Erkrankungen); Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung [online]. 2023 [Accessed: 17.03.2025]. URL: https://www.iqwig.de/download/a23-50_ravulizumab_nutzenbewertung-35a-sgb-v_v1-0.pdf.

37. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Lanadelumab (hereditäres Angioödem) – Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung [online]. 2021 [Accessed: 17.03.2025]. URL: https://www.iqwig.de/download/a21-63_lanadelumab_nutzenbewertung-35a-sgb-v_v1-0.pdf.

38. Brookhart MA, Sturmer T, Glynn RJ et al. Confounding control in healthcare database research: challenges and potential approaches. Med Care 2010; 48(6 Suppl): S114-120. https://doi.org/10.1097/MLR.0b013e3181dbebe3.

39. Brookhart MA, Schneeweiss S, Rothman KJ et al. Variable selection for propensity score models. Am J Epidemiol 2006; 163(12): 1149-1156. https://doi.org/10.1093/aje/kwj149.

40. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Evidenz zu Orphan Drugs; Arbeitspapier [online]. 2021 [Accessed: 11.07.2023]. URL: https://www.iqwig.de/download/ga21-01_evidenz-zu-orphan-drugs_arbeitspapier_v1-0.pdf.

41. Mathes T, Buehn S, Prengel P et al. Registry-based randomized controlled trials merged the strength of randomized controlled trails and observational studies and give rise to more pragmatic trials. J Clin Epidemiol 2018; 93: 120-127. https://doi.org/10.1016/j.jclinepi.2017.09.017.

42. Lange S, Lauterberg J. Pragmatischere randomisierte Studien mit Fokus auf Registerbasierung. Präv Gesundheitsf 2022. https://doi.org/10.1007/s11553-022-00974-w.

43. Ford I, Norrie J. Pragmatic Trials. N Engl J Med 2016; 375(5): 454-463. https://doi.org/10.1056/NEJMra1510059.

44. Lauer MS, D'Agostino RB Sr. The randomized registry trial -the next disruptive technology in clinical research? N Engl J Med 2013; 369(17): 1579-1581. https://doi.org/10.1056/NEJMp1310102.

The full report (German version) is published under
https://www.iqwig.de/projekte/ga23-02.html

Appendix A Information retrieval

A.1 List of publications excluded from the focused information retrieval for the systematic approach to confounder identification

1 Cobzaru R, Jiang S, Ng K et al. State of the Art Causal Inference in the Presence of Extraneous Covariates: A Simulation Study. AMIA Annual Symposium Proceedings/AMIA Symposium 2021; 2021: 334-342.

2 Huang B, Szczesniak R. Can't see the wood for the trees: confounders, colliders and causal inference - a statistician's approach. Thorax 2019; 74(4): 323-325. https://doi.org/10.1136/thoraxjnl-2018-212780.

3 Inoue K, Sakamaki K, Komukai S et al. Confounder Selection and Sensitivity Analyses to Unmeasured Confounding from Epidemiological and Statistical Perspectives. J Epidemiol 2024. https://doi.org/10.2188/jea.JE20240082.

4 Keen R, Tiemeier H. Covariate Selection from Data Collection Onwards: A Methodology for Neurosurgeons. World Neurosurg 2022; 161: 245-250. https://doi.org/10.1016/j.wneu.2021.11.057.

5 Luijken K, Groenwold RHH, van Smeden M et al. A comparison of full model specification and backward elimination of potential confounders when estimating marginal and conditional causal effects on binary outcomes from observational data. Biom J 2024; 66(1): e2100237. https://doi.org/10.1002/bimj.202100237.

6 Malec SA, Wei P, Bernstam EV et al. Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance. J Biomed Inform 2021; 117: 103719. https://doi.org/10.1016/j.jbi.2021.103719.

7 Petersen JM, Barrett M, Ahrens KA et al. The confounder matrix: A tool to assess confounding bias in systematic reviews of observational studies of etiology. Res Syn Meth 2022; 13(2): 242-254. https://doi.org/10.1002/jrsm.1544.

8 Quint JK, Minelli C. Can't see the wood for the trees: confounders, colliders and causal inference - a clinician's approach. Thorax 2019; 74(4): 321-322. https://doi.org/10.1136/thoraxjnl-2018-212488.

9 Steiner PM, Cook TD, Shadish WR et al. The importance of covariate selection in controlling for selection bias in observational studies. Psychol Methods 2010; 15(3): 250-267. https://doi.org/10.1037/a0018719.

10 Tang W, Spiegelman D, Liao X et al. Causal Selection of Covariates in Regression Calibration for Mismeasured Continuous Exposure. Epidemiology 2024; 35(3): 320-328. https://doi.org/10.1097/ede.0000000000001706.

11 Tanner-Smith EE, Lipsey MW. Identifying Baseline Covariates for Use in Propensity Scores: A Novel Approach Illustrated for a Non-randomised Study of Recovery High Schools. Peabody Journal of Education 2014; 89(2): 183-196. https://doi.org/10.1080/0161956x.2014.895647.

12. van Zwie10 A, 10nant PWG, Kelly-Irving M et al. Avoiding overadjustment bias in social epidemiology through appropriate covariate selection: a primer. J Clin Epidemiol 2022; 149: 127-136. https://doi.org/10.1016/j.jclinepi.2022.05.021.

13 Zawadzki RS, Grill JD, Gillen DL. Frameworks for estimating causal effects in observational settings: comparing confounder adjustment and instrumental variables. BMC Med Res Methodol 2023; 23(1): 122. https://doi.org/10.1186/s12874-023-01936-2.

14 Zeng J, Gensheimer MF, Rubin DL et al. Uncovering interpretable potential confounders in electronic medical records. Nature communications 2022; 13(1): 1014. https://doi.org/10.1038/s41467-022-28546-8.

A.2 Search strategies

A.2.1 Bibliographic databases

1 MEDLINE

Search interface: Ovid

Ovid MEDLINE(R) ALL 1946 to December 12, 2024

#	Searches
1	(confounder* or covariate*).ti.
2	((covariate$1 or confounder$1) adj3 (identif* or select* or bias or control)).ti,ab.
3	(causal adj (inference* or effect*)).ti,ab.
4	(target* adj3 trial* adj3 emulat*).ti,ab.
5	3 or 4
6	((observational* or cohort or non-randomized) adj3 (stud* or data)).ti,ab.
7	and/1-2.6
8	and/1.5
9	or/7-8

Appendix B Summarization of the identified potential confounders

Table 2: Assignment of the identified potential confounding variables to the respective potential confounders (multipage table)

No.[a]	potential confounders[b]	No.[c]	Identified potential confounding variables[b]	n (Lit[d])	Int[e]	Score[f]
1	Gender/sex	1	Gender/sex	61	Yes	2.43
2	Age	2	Age	70	Yes	3.86
3	EDSS score	3	EDSS score (mean or median)	61	No	2.88
		4	Disability status	1	Yes	3.25
		5	Neurological symptoms	0	Yes	3.5
		6	Cognition (self-rated)	1	No	3.43
		7	Functional System (FS) score	1	No	2.29
		8	Numbness	1	No	2.57
		9	Patient Determined Disease Steps (PDDS) score	2	No	2.29
4	EDSS change preceding year	10	EDSS change preceding year	1	No	3.88
		11	Deterioration index (defined as the EDSS score divided by the disease duration)	2	No	4
		12	EDSS in the past 3 months	1	No	3
		13	MS Severity Score (MSSS)	1	No	1.5
		14	Disease activity	2	No	4.5
		15	Disease course	1	No	na
		16	Remit of relapse	0	Yes	4.43
		17	Time since last relapse	5	No	3.71
5	MS Functional Composite (MSFC) score	18	MS Functional Composite (MSFC) score	3	No	1.75
		19	Ambulation index	3	No	2.88
		20	Clinical condition	0	Yes	4.25
6	Fatigue	21	Fatigue	1	No	2.86
		22	Modified Fatigue Impact Scale (MFIS) score	1	No	2.14

Table 2: Assignment of the identified potential confounding variables to the respective potential confounders (multipage table)

No.[a]	potential confounders[b]	No.[c]	Identified potential confounding variables[b]	n (Lit[d])	Int[e]	Score[f]
7	Comorbidities	23	Comorbidities (general)	2	Yes	4
		24	Depression diagnosis	2	No	2
		25	Hamilton Anxiety Rating Scale (HAM-A)	1	No	na
		26	Hamilton Depression Rating Scale (HAM-D)	1	No	na
		27	Allergy	1	No	na
		28	Cardiovascular comorbidities	1	Yes	2.5
		29	Charlson Comorbidity Index (CCI) score	3	No	1.86
		30	Chronic infections	0	Yes	4.13
		31	Dermatological diseases	0	Yes	2.88
		32	Diabetes	1	No	1.75
		33	Headache	1	No	na
		34	Skin problems	0	Yes	2.25
		35	Thyroid dysfunction	1	No	na
		36	Tuberculosis	0	Yes	2.38
		37	Bowel symptoms	1	No	3.86
		38	Depression (self-rated)	1	No	3
		39	Gastrointestinal disorders	0	Yes	4
		40	Hospitalizations in previous year	1	No	2
8	Absolute lymphocyte count; Lymphopenia	41	Absolute lymphocyte count; Lymphopenia	1	Yes	4.57
9	Study centre; MS centre	42	Study centre; MS centre	3	No	na
		43	Country	6	No	1.14
		44	Region	5	No	1.14
		45	Within-country region	2	No	1.14
		46	Type and speciality of prescriber	2	No	2.71
10	BMI	47	BMI; or height and weight	6	No	2.29
		48	Overweight or obesity	1	No	2.29
		49	Weight	2	No	2
11	Tobacco use	50	Tobacco use	1	No	2.14
12	Infratentorial lesion	51	Localization of lesions	0	Yes	3.57
		52	Pyramidal impairment	1	Yes	3.38
		53	Spinal lesions	0	Yes	4.14

Table 2: Assignment of the identified potential confounding variables to the respective potential confounders (multipage table)

No.[a]	potential confounders[b]	No.[c]	Identified potential confounding variables[b]	n (Lit[d])	Int[e]	Score[f]
13	Number of contrast-enhancing lesions	54	Number of gadolinium enhancing lesions; Number of contrast-enhancing lesions; Number of patients with gadolinium enhancing lesions; Patients with enhancement	21	Yes	4.57
		55	Status of gadolinium enhancing T1 activity	1	No	4.29
14	T1 hypointense lesion volume	56	T1 hypointense lesion volume	7	No	3
15	Number of T2 lesions	57	Number of T2 lesions	12	No	4.14
		58	MRI status	1	Yes	3
		59	MRI brain lesions; number of lesions	2	Yes	3.86
		60	New lesions	0	Yes	4.29
		61	Presence of at least MRI T2 lesions and at least one contrast-enhancing MRI lesion	2	No	3.43
16	Volume of T2 lesions	62	Volume of T2 lesions	7	No	3.14
		63	Enhancing lesion volume	3	No	3
		64	Lesion load	0	Yes	3.71
		65	Size of lesions	0	Yes	3.14
		66	Brain volume; normalized brain volume	7	Yes	2.17
		67	Normalized grey matter volume	1	No	2.29
		68	Normalized white matter volume	1	No	2.43
17	Tolerance of previous therapy and expectations of treatment	69	Reason for discontinuing prior treatment	2	No	4.67
		70	Drug efficacy (known or suspected)	0	Yes	4.57
		71	Drug tolerance	0	Yes	4.57
		72	Side effects of other medications (MS and other indications)	0	Yes	4.57

Table 2: Assignment of the identified potential confounding variables to the respective potential confounders (multipage table)

No.[a]	potential confounders[b]	No.[c]	Identified potential confounding variables[b]	n (Lit[d])	Int[e]	Score[f]
18	Previous treatment with other disease modifying therapy (DMT)	73	Previous treatment with other disease modifying therapy (DMT)	4	No	3.86
		74	Type of previous therapy	8	Yes	4.29
		75	Treatment naive; treatment history	10	Yes	4.14
		76	Duration of previous treatment	1	No	3.86
		77	Time on first line treatment	1	No	3.86
		78	Duration of previous natalizumab use	2	No	4
		79	Number of previous therapies	6	No	4.29
		80	Past dalfampridine use	1	No	1.57
		81	Previous treatment with immunosuppressive therapy	1	No	3.86
		82	Previous treatment with interferon	3	No	2.71
		83	Time after natalizumab discontinuation	2	No	4.67
19	Patient wish	84	Patient wish	0	Yes	4.71
		85	Drug route of administration	0	Yes	3.86
		86	Willingness to inject a medication permanently	0	Yes	4.71
		87	Willingness to take pills permanently	0	Yes	4.71
		88	Tolerance of subcutaneous injections	0	Yes	4.71
		89	Patient concerns about drug side effects	0	Yes	4.29
		90	Patient need for safety	0	Yes	4.29

Table 2: Assignment of the identified potential confounding variables to the respective potential confounders (multipage table)

No.[a]	potential confounders[b]	No.[c]	Identified potential confounding variables[b]	n (Lit[d])	Int[e]	Score[f]
20	Quality of Life	91	EuroQol-5D (EQ-5D)	1	No	na
		92	MS Impact Scale (MSIS) physical subscale score	1	No	na
		93	MS Impact Scale (MSIS) psychological subscale score	1	No	2.14
		94	MS International Questionnaire of Quality of Life (MusiQol)	1	No	na
		95	Sexual Satisfaction Scale (SSS) score	1	No	2.43
		96	SF-36 bodily pain subscale score	1	No	3.57
		97	SF-36 general health subscale score	1	No	na
		98	SF-36 mental health subscale score	1	No	na
		99	SF-36 physical functioning subscale score	1	No	na
		100	SF-36 role emotional subscale score	1	No	na
		101	SF-36 role physical subscale score	1	No	na
		102	SF-36 social functioning subscale score	1	No	na
		103	SF-36 vitality subscale score	1	No	na
21	Number of previous relapses	104	Number of previous relapses (no defined index time)	6	Yes	4.67
		105	Disease severity	1	No	na
		106	Disease severity at diagnosis	0	Yes	4.38
		107	Prognosis	0	Yes	4.57
		108	Relapse in prior year	44	No	4.57
		109	Relapses in prior 2 years	23	No	4
		110	Relapses in prior 3 years	2	No	4.14
		111	Relapses in the past 3 months	2	No	4.43
		112	Relapses in the past 6 months	1	No	4.43
22	Severity of relapses	113	Severity of relapses	0	Yes	4.71
		114	Past steroid use	1	No	1.57
		115	Number of treated relapses	1	Yes	3.29
		116	Steroid-treated relapses in prior 2 years	1	No	3.43
		117	Steroid-treated relapses in prior year	1	No	3.83

Table 2: Assignment of the identified potential confounding variables to the respective potential confounders (multipage table)

No.[a]	potential confounders[b]	No.[c]	Identified potential confounding variables[b]	n (Lit[d])	Int[e]	Score[f]
23	Time since diagnosis of MS	118	Time since diagnosis of MS; disease duration	53	No	3.43
		119	Age at onset of first MS symptoms	1	No	3.71
		120	Age at start of disease modifying therapy (DMT); Age at start of treatment	2	No	na
		121	Childhood MS diagnosis	1	No	3.83
		122	Time between symptom onset and disease modifying therapy (DMT) start	2	No	3.29
		123	Time since first MS symptoms	11	No	3.43
		124	Year of treatment initiation	2	No	3
		125	Age at disease onset	5	No	4.14
24	Cerebrospinal fluid status	126	Cerebrospinal fluid status	1	Yes	2.57
		127	Immunoglobulin (IgM) in cerebrospinal fluid	0	Yes	2.43
		128	Neurofilament light chain (NFL)	0	Yes	2.86
25	Type of professional work activity	129	Type of professional work activity	0	Yes	3.57
26	Educational level	130	Educational level	2	No	1.57
27	Ethnicity or race	131	Ethnicity or race	13	No	1.86
28	Pregnancy in the past 3 months	132	Pregnancy in the past 3 months	1	No	na

a. Numbering based on the summarized overlapping confounders.
b. Column not translated in the original German report to avoid translation errors.
c. Continuous numbering of all identified potential confounding variables from 0 to 160, for the excluded variables from 133 to 160 see Table 3.
d. Number of publications in which the respective potential confounding variable was identified.
e. Mention by the clinicians in the qualitative interview (yes or no). No quantitative data were recorded in the external report.
f. Mean value of the survey score, assessment of the influence of each identified potential confounding variable on the treatment decision using a 5-point scale from 1 (no consideration) to 5 (very strong consideration).

BMI: Body Mass Index; EDSS: Expanded Disability Status Scale; Int: (clinician) interview; Lit: literature review MRI: magnetic resonance imaging; MS: multiple sclerosis; n: number; na: not applicable, No: number; SF-36: Short Form-36 Health Survey

Table 3: Identified variables sorted by the respective reason for exclusion (multipage table)

No.[a]	Identified variables[b]	Reason for exclusion	n (Lit[c])	Int[d]	Score[e]
Exclusion based on lack of relevance in the German healthcare context					
133	Health insurance status	No relevance in the German healthcare context	1	No	1.14
134	Health plan type	No relevance in the German healthcare context	2	No	1.14
135	Previous hospital care costs	No relevance in the German healthcare context	1	No	1.5
136	Previous outpatient care costs	No relevance in the German healthcare context	2	No	1.75
Exclusion based on the CONFIRM RCT (based on the inclusion/exclusion criteria in the study protocol)					
137	McDonald criteria; MS classification	Inclusion criterion CONFIRM	5	No	4.57
138	Hepatitis	Exclusion criterion CONFIRM	0	Yes	2.63
139	Human immunodeficiency virus (HIV)	Exclusion criterion CONFIRM	0	Yes	3.25
140	Breastfeeding	Exclusion criterion CONFIRM	1	Yes	3.57
141	Family planning	Exclusion criterion CONFIRM	0	Yes	4.43
142	Pregnancy status	Exclusion criterion CONFIRM	2	Yes	4.86
143	Pregnancy wish	Exclusion criterion CONFIRM	0	Yes	4.43
144	Timeliness of desire to have children	Exclusion criterion CONFIRM	0	Yes	4.43
145	Interferon beta treatment within past 3 months	Exclusion criterion CONFIRM	1	No	3
146	Previous treatment with glatiramer acetate	Exclusion criterion CONFIRM	3	No	3.71
147	Reason for dimethyl fumarate treatment	Exclusion criterion CONFIRM	1	No	4.43
148	Years of glatiramer acetate treatment before study	Exclusion criterion CONFIRM	1	No	3.29
149	Medical uncertainty with MS diagnosis	Exclusion criterion CONFIRM	0	Yes	4
150	Type of MS onset (monofocal or multifocal)	Exclusion criterion CONFIRM	1	No	na
Exclusion based on methodological reasons					
151	EDSS scores/improvement at the end of the study treatment/last check-up	No baseline criterion	1	No	na
152	Time since last visit	No baseline criterion	1	No	2.57
153	Time on study	Characteristics of the study course	5	No	na
154	Time on treatment	Characteristics of the study course	2	No	3.86

Table 3: Identified variables sorted by the respective reason for exclusion (multipage table)

No.[a]	Identified variables[b]	Reason for exclusion	n (Lit[c])	Int[d]	Score[e]
Exclusion because original publication outside the research question					
155	Disease modifying therapy (DMT) re-initiation postpartum	Original publication outside the research question	1	No	4.14
156	Number of visits 1 year postpartum	Original publication outside the research question	1	No	2.71
157	Number of visits in pregnancy	Original publication outside the research question	1	No	2.57
158	Age at time of conception	Original publication outside the research question	1	No	3.43
159	Age of offspring at last follow-up	Original publication outside the research question	1	No	3.14
160	Duration of therapy	Original publication outside the research question	2	No	3.86

a. Continuous numbering of all identified potential confounding variables from 0 to 160, for the potential confounding variables from 0 to 132 see Table 2
b. As in the external report, the potential confounders were not translated in the original German report to avoid translation errors.
c. Number of publications in which the respective variable was identified.
d. Mention by the clinicians in the qualitative interview (yes or no). No quantitative data were recorded in the external report.
e. Mean value of the survey score, evaluation of the influence of each individual variable on the treatment decision using a 5-point scale from 1 (no consideration) to 5 (very strong consideration).

EDSS: Expanded Disability Status Scale; Int: (clinician) interview; Lit.: literature review; MS: multiple sclerosis; n: number; na: not applicable; No.: number

Appendix C Recommendations – checklist

Step 1: Identification

☐ When systematically retrieving information, primary publications on RCTs and cohort studies should be considered first. Guidelines or single-arm studies can then be considered.

☐ The inclusion and exclusion criteria (PICO) for the study pool for confounder extraction should closely align with the research question.

☐ In comparative observational studies, all adjustment variables should be extracted, regardless of statistical significance. In randomized or non-comparative studies, all patient characteristics should be extracted.

☐ When extracting data from primary publications, it is advisable that a suitable saturation criterion is applied.

☐ Clinician interviews should be conducted.

☐ The clinicians should be named, and any potential conflicts of interest disclosed.

Step 2: Summarization

☐ Confounding variables with overlapping content should be summarized. This should be done after extraction. Transparent and comprehensible documentation must be ensured.

☐ For each identified confounding variable, it should be checked whether it lies within the scope of the research question to be analysed**.**

☐ The exclusion of a potential confounder due to content-related considerations must be justified based on the literature.

☐ An operationalization must be defined for all identified potential confounders.

☐ Although potential confounders that cannot be measured or operationalized cannot be included in the analysis, they should be named and considered when interpreting the results.

Step 3: Assessment of relevance

☐ The assessment of potential confounders by clinicians should be carried out in the third step. It is advisable that a scoring system is used.

Required documentation

Literature review

☐ Specification of the search strategy and the sources searched.

☐ Clear designation of the inclusion and exclusion criteria.

☐ Creation of a flow chart for the selection process.

☐ Documentation of the references of the hits that were checked in full, but excluded.

☐ Documentation of the source of each extracted confounding variable (ideally including a specific citation).

Clinician interviews and evaluation using a survey

☐ Specification of names, specialties and institutions of the experts involved, along with disclosure of potential conflicts of interest.

☐ Specification of the interview guideline (or similar) and the documents for assessing the confounders for clinicians.

Summarization

☐ List of potential confounding variables that have been summarized into confounders.

☐ In case of exclusion of identified confounding variables: provision of a literature-based justification, indicating the reference.

Publication of the report

☐ The documentation should be published as part of the study protocol for the planned comparative NRSI.

For Appendix D (External expertise on the IQWiG working paper) and Appendix E (Disclosure of relationships of external experts), please see the full working paper.