Methodological Challenges using Routine Clinical Care Data for Real-World Evidence: a Rapid Review utilizing a systematic literature search and focus group discussion

Michelle Pfaffenlehner[1,2*], Max Behrens[1,2], Daniela Zöller[1,2], Kathrin Ungethüm[3,4], Kai Günther[3,4], Viktoria Rücker[4], Jens-Peter Reese[4], Peter Heuschmann[3,4,5], Miriam Kesselmeier[6], Flavia Remo[6], André Scherag[6], Harald Binder[1,2], Nadine Binder[7,2] – for the EVA4MII project

[1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Centre – University of Freiburg, Freiburg, Germany

[2] Freiburg Centre for Data Analysis and Modelling, University of Freiburg, Freiburg, Germany

[3] Institute for Medical Data Sciences, University Hospital Würzburg, Würzburg, Germany

[4] Institute for Clinical Epidemiology and Biometry, University Würzburg, Würzburg, Germany

[5] Clinical Trial Centre, University Hospital Würzburg, Würzburg, Germany

[6] Institute of Medical Statistics, Computer and Data Sciences, Jena University & Jena University Hospital, Jena, Germany

[7] Institute of General Practice/Family Medicine, Faculty of Medicine and Medical Centre - University of Freiburg, Freiburg, Germany

*corresponding author: michelle.pfaffenlehner@uniklinik-freiburg.de

Abstract

Background

The integration of real-world evidence (RWE) from real-world data (RWD) in clinical research is crucial for bridging the gap between clinical trial results and real-world outcomes. Analyzing routinely collected data to generate clinical evidence faces methodological concerns like confounding and bias, similar to prospectively documented observational studies. This study focuses on additional limitations frequently reported in the literature, providing an overview of the challenges and biases inherent to analyzing routine clinical care data, including health claims data (hereafter: routine data).

Methods

We conducted a literature search on routine data studies in four high-impact journals based on the Journal Citation Reports (JCR) category “Medicine, General & Internal” as of 2022 and three oncology journals, covering articles published from January 2018 to October 2023. Articles were screened and categorized into three scenarios based on their potential to provide meaningful RWE: (1) Burden of Disease, (2) Safety and Risk Group Analysis, and (3) Treatment Comparison. Limitations of this type of data cited in the discussion sections were extracted and classified according to different bias types: main bias categories in non-randomized studies (information bias, reporting bias, selection bias, confounding) and additional routine data-specific challenges (i.e., operationalization, coding, follow-up, missing data, validation, and data quality). These classifications were then ranked by relevance in a focus group meeting of methodological experts. The search was pre-specified and registered in PROSPERO (CRD42023477616).

Results

In October 2023, 227 articles were identified, 69 were assessed for eligibility, and 39 were included in the review: 11 on the burden of disease, 17 on safety and risk group analysis, and 11 on treatment comparison. Besides typical biases in observational studies, we identified additional challenges specific to RWE frequently mentioned in the discussion sections. The

focus group had varied opinions on the limitations of Safety and Risk Group Analysis and Treatment Comparison but agreed on the essential limitations for the Burden of Disease category.

Conclusion

This review provides a comprehensive overview of potential limitations and biases in analyzing routine data reported in recent high-impact journals. We highlighted key challenges that significantly impact analysis results, emphasizing the need for thorough consideration and discussion for meaningful inferences.

Keywords: rapid review, limitation, bias, routine clinical care data, real-world evidence, EHR

Background

Real-world evidence (RWE) derived from real-world data (RWD) becomes increasingly important to support clinical evidence. The growing availability of such data opens up new research opportunities to improve our understanding of clinical practice. A recent definition of real-world data includes data sources such as routine clinical care data, frequently called electronic health records (EHRs), disease-specific registries, administrative data, such as claims data or death registries, and data collected through personal devices [1–4].

This review is limited to routine clinical care data (hereafter: routine data) derived from the health care system, such as EHR and administrative data including insurance and claims data. We focus on methodological challenges and biases that researchers may face when analyzing routine data. Although routine data are not primarily collected for research purposes, their use and longitudinal linkage with data from other sources such as registries or biobanks has the potential to improve health care and regulatory decision making [2,5]. However, it is not only the linkage to other data sources that is important but also the ability to aggregate data from different hospitals or even across different countries and health care systems given a common data model.

Randomized controlled trials (RCTs) are the gold standard for answering questions about treatment efficacy and safety. However, analyzing RWD may be useful to bridge evidence gaps at the interface to clinical practice. Utilizing routine data offers several advantages, including a large number of observations, especially when leveraging and linking multiple clinical data sources. It allows the coverage of different locations, patient populations (e.g., different age distribution or disease severity), and practice patterns in routine health care [1,3,6]. Routine data analysis can also be valuable when RCTs are not feasible due to ethical or practical reasons. Particularly in the context of treatment comparisons, the use of the target trial emulation framework allows to obtain comparable results to those observed in RCTs when carefully and fully emulated [7–9]. An alternative approach, rather than relying solely on routine data or their linkage for treatment comparison, is to use the data as an external control [2,10].

Yet, as routine data are documented for reimbursement or clinical care purposes, the quality of the data from a research perspective is typically lower than that of other prospectively planned studies – including RCTs. Routine data may lack harmonization and interoperability, may often be incomplete and some relevant information may be missing [4]. For instance, body mass index is generally irrelevant for reimbursement purposes but might be a potential risk factor, confounder or effect modifier for several research questions, especially in the field of non-communicable diseases, such as those investigated by Zöller et al. [11] on chronic obstructive pulmonary disease (COPD). The analysis of such data is therefore fraught with methodological challenges, including confounding and several potential biases which are already well-known from clinical epidemiology. Beyond these common concerns, it remains unclear which additional limitations related to routinely collected data, particularly challenges at measurement level – such as how data is collected and translated into variables used for analysis – appear frequently [12].

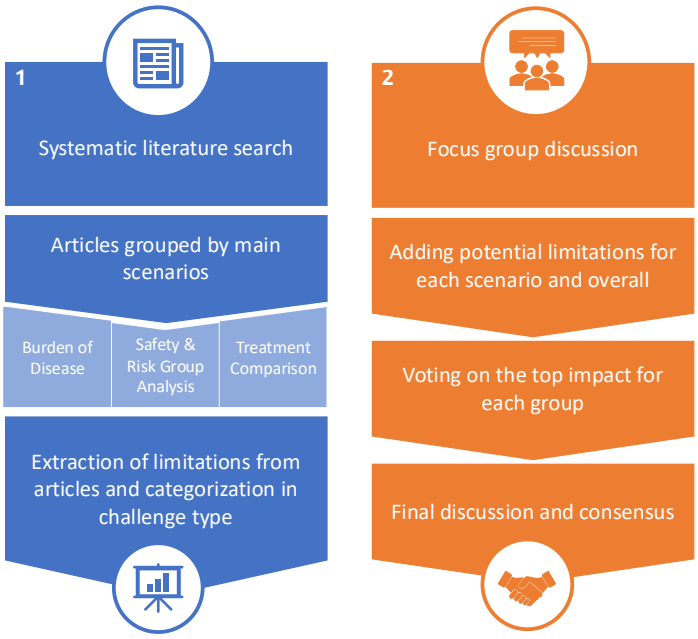
For structuring an investigation on limitations, we suggest three scenarios, derived from the field of clinical epidemiology where routine data hold high potential for generating RWE: (1) *Burden of Disease*, (2) *Safety and Risk Group Analysis*, and (3) *Treatment Comparison*. *Burden of Disease* describes the impact of a disease or health problem on a specified

population, quantified by metrics such as incidence, prevalence, mortality, morbidity, quality of life, and economic impact [13]. As this scenario typically includes the given health care setting in the definition of the study population, we consider it as the most natural application of routine data. *Safety and Risk Group Analysis* covers adverse events of various medical interventions such as treatments, medications, devices and procedures, with a potential focus on identifying and characterizing subgroups with a higher risk profile in the respective population (e.g., due to comorbidities, genetic predispositions, or general demographic characteristics) [14]. Here, routine data offer the possibility of long-term observations, to study rare subgroups and adverse events, as well as to observe patients with different comorbidities, and co-medications. As this scenario is characterized by time-sensitive and potentially unobservable or undocumented information, it is more complex than the *Burden of Disease* scenario. *Treatment Comparison* deals with the causal effects of medical treatments. This area is and will continue to be dominated by evidence from RCTs. However, with the beginning of 2025, a new EU regulation for health technology assessment will enter into force. Although the process is still in development, health technology developers will need to address a high number of PICOs with diverse comparators [15]. Combined with the tight timelines in the joint clinical assessment process, clinical studies will not be available for all PICOs, creating a high need for additional data sources like routine data. In addition, routine data offer the opportunity to study the effects in a routine setting, including effects of noncompliance, rare subgroups or subgroups not typically eligible for clinical studies, confounding by indication as well as potential effects of site-specific impact factors on treatment outcome. Here, similar issues as in *Safety and Risk Group Analysis* complicate the statistical analysis, especially due to the causal interpretation of the primary outcome.

In this work, we aim to provide an overview of the reported challenges and biases inherent in routine data analysis with respect to their potential impact in the three main scenarios. This is accomplished in a step-wise procedure as illustrated in Figure 1. First, a systematic literature search is conducted to extract the limitations outlined in the respective discussion sections. Second, a subsequent focus group discussion with experts is held to supplement and to

129 evaluate the identified list of challenges. Finally, we identified challenges that have a
130 comparably high potential to affect the analysis findings and require thorough consideration
131 and discussion in order to draw meaningful conclusions.

132 *Figure 1 Process Flow Diagram*



133

134 Overview of the process used in this work to identify and evaluate challenges and biases in
135 routine care data analysis. This includes (1) a systematic literature search to extract reported limitations
136 followed by (2) expert focus group discussions to supplement and discuss these challenges regarding
137 their impact on real-world evidence.

138 Methods

139 Data sources and search strategies

140 In October 2023, we conducted a systematic search of MEDLINE via PubMed. The search
141 focused on English-language publications in the following top-ranked journals based on the
142 Journal Citation Reports (JCR) category "Medicine, General & Internal" as of 2022 [16]: (i) The
143 Lancet, (ii) New England Journal of Medicine, (iii) Journal of the American Medical Association
144 (JAMA) and (iv) British Medical Journal (BMJ). In addition, we included the following oncology
145 journals, based on our experience, as RWE is particularly prevalent in this therapeutic field
146 [17]: (v) JAMA Oncology, (vi) The Lancet Oncology, and (vii) Journal of Clinical Oncology. In

order to ensure recency and thereby relevance of the articles, the search included all original articles published between January 2018 and October 2023. The following terms were queried in the title or abstract: “real-world evidence”, “RWE”, “Real-world data”, “real-world”, “routine data”, “routine care data”, “Emulation” and “Electronic health data”. A study was defined as being eligible if routine clinical care data collected by the healthcare system, such as longitudinal claims data or EHR, was analyzed. Other types of publications, such as comments, letters, perspectives or reviews were excluded. Studies were also excluded, in which only registry data or manually collected data were analyzed. In addition, we assessed if any aspect from the three pre-defined scenario categories were investigated in the studies: (1) Burden of Disease, (2) Safety and Risk Group Analysis, and (3) Treatment Comparison. If this was not the case, the study was excluded. The detailed search strategy is described in the Supplementary Material (Supplementary S1). All articles were indexed and organized using Zotero.

Three reviewers (MP, KG, KU) independently screened each publication through all stages of the review process. First, titles and abstracts of the search results were screened to ensure relevance and adherence to the inclusion and exclusion criteria. In a second step, the full texts of the included abstracts were assessed to obtain a final decision on inclusion in the review. In case of any discrepancy between two reviewers regarding the eligibility of specific studies, an additional independent reviewer (MB) was consulted to resolve the issue. The review was pre-specified and registered in PROSPERO (CRD42023477616).

Data extraction and categorization

Methodological challenges and limitations mentioned in the discussion sections of the included publications were extracted as the main outcome. In addition, study characteristics such as design, the underlying data sources, the country from which data was obtained, the methods employed, and the purpose of the published studies were retrieved. Publications were categorized into the three predefined main areas of application (Burden of Disease, Safety and Risk Group Analysis, Treatment Comparison), based on their research question. The extracted

175 limitations were subsequently assigned to the main bias categories in non-randomized studies
176 as defined in the Cochrane Handbook: confounding, selection bias, information bias and
177 reporting bias [18]. We identified five specific categories to highlight potential biases that are
178 specific to the use of routine data. First, as clinical routine data is primarily collected for
179 reimbursement purposes based on clinic-specific coding practices, the category *Coding*
180 *Challenges* is introduced. This category includes for example discrepancies in coding practices
181 between clinics or specificalities of the ICD10 (the International Classification of Diseases and
182 Related Health Problems) coding system [19]. This challenge may lead to information bias,
183 specifically misclassification or detection bias, regardless of the already well-known detection
184 bias, which typically result from varying quality in detection methods. Second, as one of the
185 main drivers of documentation quality is again reimbursement as well as patient care rather
186 than research, the category *Operationalization or Availability of Variables* is established to
187 address to which extent studies were restricted by the availability of the data for answering the
188 research question. This type of missing data may lead to unmeasured confounding and
189 potential selection bias. Third, routine data may suffer from other types of missing data, i.e.,
190 missing records in certain variables, as typically known from observational studies, resulting in
191 the category *Missing Data*. Fourth, routine data may also suffer from different lengths of follow-
192 up largely varying between patients, for which we introduce a further category called *Follow-*
193 *up Challenges*. Finally, the category *Validation & Data Quality* is added, which leads to
194 potential bias as large amounts of routine data typically cannot be carefully validated. This also
195 includes the variability of data quality over time arising e.g., due to changes in coding systems,
196 such as transitions between versions of the ICD coding standard, or the ongoing digitalization
197 of hospitals.

198 We did not assess the quality or risk of bias of individual studies since we were not extracting
199 data or outcomes of the studies for subsequent analysis and we were only interested in their
200 stated limitations. Still, we expected that there were challenges not reported in the study
201 publications, e.g., because of limited relevance or lack of awareness. Therefore, a subsequent

focus group meeting was initiated for complementing and ranking the identified list of challenges.

Focus Group

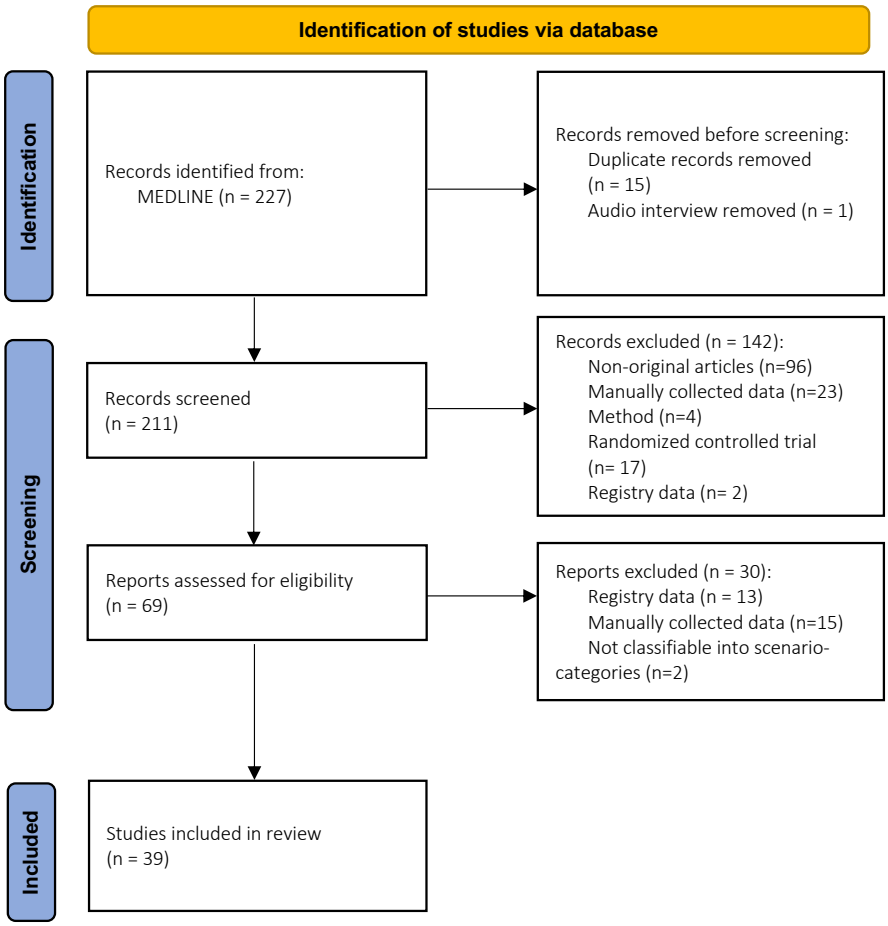
At a four-hour workshop, the results of the systematic literature search were presented in the form of a slide presentation for each scenario to a selected group of co-authors of this manuscript plus one independent expert from the area of health technology assessments. After each scenario was presented, focus group members were asked to add potential additional challenges in a moderated joint discussion. In addition, through a *General* category, we left room for limitations and challenges that can be found in RWD studies not specific to any of the three scenarios. All identified challenges were listed for a subsequent voting on their potential impact within each scenario. Every focus group member had two votes for each of the three main scenarios, plus one additional vote for the *General* category. All challenges with at least one vote were summarized and considered as key challenges deemed to have a high potential to influence the results requiring thorough consideration in analyses.

Results

Study selection and characteristics

In total, 227 records were identified from MEDLINE of which 15 duplicate publications and 1 audio interview were removed, leading to 211 records eligible for title and abstract screening. From this first screening, we excluded 142 publications because they were not original research articles, were only methods articles, or did not use routinely collected data from clinical care. In a second step, 69 records were evaluated in a full-text screening for verifying the data source as coming from electronic health records or administrative data, and to be categorizable into the three main scenarios. Finally, 39 studies were included in the review. An overview of the review process and study selection is depicted in Figure 2.

228 *Figure 2 PRISMA flow diagram [20]*



229

230 A brief summary of the study characteristics is shown in Table 1. Most of the studies were
231 published in BMJ (36%), with 55% being published in the past two years (2022 and 2023). This
232 is supporting our decision to limit the search to the past five years. The majority of extracted
233 publications (46%) was based on data from the US. While some publications used only EHR
234 data (59%), other studies linked this information to additional sources, such as biobanks or
235 registries (28%). The use of data sources varied between studies, from single-country to
236 multinational studies. The majority of the publications used nation-wide (46%) or multi-center
237 data (38%), while the remaining studies used multi-national (10%), single-center (3%) and
238 territory-wide (3%) data.

239 *Table 1 Summary of study characteristics*

Study Characteristics	n (%)
Journals	
BMJ	14 (35.9)
JAMA	9 (23.1)
New England Journal of Medicine	4 (10.2)
The Lancet	5 (12.8)
JAMA Oncology	2 (5.1)
Journal of Clinical Oncology	5 (12.8)
Year	
2018	1 (2.6)
2019	5 (12.8)
2020	4 (10.3)
2021	7 (18.0)
2022	10 (25.6)
2023	12 (30.8)
Country	
Canada	1 (2.6)
France	1 (2.6)
Hong Kong	1 (2.6)
Israel	3 (7.7)
South Korea	1 (2.6)
UK	5 (12.8)
UK + Canada	1 (2.6)
UK + USA	2 (5.1)
USA	18 (46.2)
USA + South Korea	1 (2.6)
Qatar	1 (2.6)
>2 countries	4 (10.2)
Data Source	
EHR	23 (59.0)
EHR + additional linked data*	11 (28.2)
Administrative data**	5 (12.8)
Scope of Data	
single-center	1 (2.6)
multi-center	15 (38.5)
territory-wide	1 (2.6)
nation-wide	18 (46.2)
multi-national	4 (10.3)

In this table, the number of studies *n*, along with its percentage in brackets *(%)*, is presented for each category. EHR= Electronic health records

* additional linked data include exome sequencing data, hospital admission data, mortality data, data from biobanks, claims data, questionnaire data, and data from registries

** administrative data include insurance information and claims data

Figure 3 provides a summary of the limitations mentioned in the discussion section of the extracted studies stratified into the data source types and according to their assigned scenario category. In total, there are 11 publications in the category *Burden of Disease*, 17 publications in *Safety and Risk Group Analysis* and 11 publications in *Treatment Comparison*. As expected, the main bias types are typically mentioned in the publications. However, it is important to recognize the additional biases that frequently appear. A detailed overview for each publication is provided in the Supplementary Material S2 – Table S3, including a broad overview of the study design including the main outcomes and their respective methods, objectives, and the scope of the analysis in terms of data use.

Figure 3 Summary of biases extracted from the studies

		Main Bias Categories				Additional Bias Categories				
		Confounding	Selection Bias	Information Bias	Reporting Bias	Coding Challenges	Operationalization or Availability of Variables	Missing Data	Follow-up Challenges	Validation & Data Quality
Burden of Disease	EHR	3	4	5	2	2	2	3		
	Linked EHR	2	3	2	1			1	2	2
	Admin. Data		1							
Safety & Risk Group	EHR	7	4	5	2		3	2	3	2
	Linked EHR	4	4	6	4	2	1	2		2
	Admin. Data	2	2	2	1		2	1	2	
Treatment Comparison	EHR	9	9	7		1	3	2		1
	Linked EHR									
	Admin. Data	2	2	1					1	

Number of studies facing the main biases and additional identified bias categories grouped by data source types and the main scenarios: *Burden of Disease, Safety and Risk Group Analysis, and Treatment Comparison*. Abbreviations: EHR…Electronic Health Records, Admin. Data…Administrative Data

In the following, we summarize the findings on methodological challenges and limitations from both the review and the focus group workshop for each main scenario, separately. In Figure 4, all challenges that were added and ranked according to their relevance in the focus group discussion are summarized.

265 *Figure 4 Overview of Workshop Results*

Scenario 1: Burden of Disease

- Selection bias (+++): Transferability of study results to larger populations
- Missing data (+++)
- Misclassification bias (+++): Challenges in tracking and disease classification
- Data quality (+): Change in data quality over time

- Missing information due to death
- Heterogeneity and specialization of hospitals
- Calibration of algorithms: Adjusting to dynamic data changes.

Scenario 2: Safety and Risk Group

- Case vs. patient (+++)
- Methodological challenges for safety endpoint (++)
- Missing follow-up (++)
- Reverse p-hacking (+)
- Information bias (+): observer bias, measurement bias, detection bias

- Heterogeneity of patient's treatment pathways
- Off-label medication use
- Missing information regarding the reason for therapy switch

Scenario 3: Treatment Comparison

- Residual Confounding (++)
- Immortal time bias (++)
- Adherence bias (+): Treatment switcher
- Sector boundaries and data availability (+): challenges in data transferability between different data sectors
- Confounding by indication (+)
- Availability of variables (+)
- Selection bias (+): Transferability of study results to larger population

- Uncertainty about timing of intervention
- Control group definition
- Verifiability of inclusion and exclusion criteria
- Follow-up duration (some drugs have a long lag in efficacy)

General Limitations

- Record Linkage (+++)
- Validation of data (++)
- Selection of measurements:(+)

- Data quality and hospital documentation improvement
- Case overlap and data overlap when using different database
- Estimator vs. Estimand: Differentiation and meaning of these terms

266

Findings of the focus group discussion divided into the three main scenarios: *Burden of Disease, Safety and Risk Group Analysis*, and *Treatment Comparison*, as well as the additional category *General Limitations*. The content stated within the blue colored areas are the voting result, with the plus signs (+) indicating the importance weighting. The blue font highlights challenges and limitations already occurring in the systematic search. The black font and particularly the content outside the blue area are limitations and challenges added by the focus group.

Burden of disease

Five publications reported limitations potentially leading to confounding. Of these, two studies had limited availability of variables included in the analysis resulting in the possibility of unmeasured confounding [21,22]. However, Kamran et al. [22] argued that with the use of a limited number of variables in the model a reliable identification and validation of all variables even across different institutions was feasible. Limitations related to the risk of selection bias were present in eight studies. On the one hand, in Canavan et al. [21] a volunteer bias could not be ruled out because the data used for the analyses included individuals who opted to be included in the study database. On the other hand, in Witberg et al. [23] the absence of simultaneously enrolled comparison groups was mentioned which could lead to a potential selection bias. Moreover, some studies highlighted the fact that the results might not be transferable to other clinical practices or populations [21,24–26]. The primary type of information bias (n=7) observed was misclassification bias, affecting either the exposure or outcome [21,24,27,28]. Beck et al. [24] highlighted that clinical diagnoses were probably not fully captured due to challenges in diagnostic and physician coding. On the other hand, Manz et al. [28] noted a limitation indicating that they were limited to use only patients with a coded classifier variable for comparing the machine learning model with the commonly used prognostic reference. This limitation resulted from differences in the characteristics of the patients with and without the coded variable. Kamran et al. [22] additionally emphasized that even though a common EHR provider across facilities was implemented, it remains crucial to have in-depth knowledge of the local deployment. Two studies reported limited follow-up time [26,29]. However, Biccler et al. [29] noted that despite the longer follow-up period compared

295 to clinical trials, it was still insufficient to evaluate very long-term effects. The lack of health
296 record validation was reported in two studies [29,30]. While all studies acknowledged
297 limitations in their discussions, only three studies [22,25,27] took an additional step to assess
298 these limitations in the context of their specific study.

299 During the focus group discussion, further limitations were added to those identified in the
300 literature review and are presented in the first part of Figure 4. Some relevant and reasonable
301 limitations added by the experts were the inconsistency of the data quality over time. This
302 challenge is interconnected to the adjustment or calibration of algorithms to dynamic data
303 changes, especially when using a model that continuously uses recent data from routine
304 clinical care. Additionally, the heterogeneity and specialization of hospitals substantially impact
305 the burden of disease. Including only hospitals with specializations in a certain disease
306 treatment would overestimate the burden. Another added limitation was *missing information*
307 *due to death* in situations with death as competing risk, where it would be impossible to
308 diagnose the disease even though the patient may have had it [31].

309 The rating of the limitations of *Burden of Disease* is indicated by the blue area in Figure 4. The
310 transferability of study results to a larger population, which goes hand in hand with the
311 possibilities of selection bias, missing data and misclassification of data, was weighted by the
312 experts to be most important. The importance of data quality in EHR data was also recognized,
313 specifically the variability of data quality over time due to e.g., changes in coding systems,
314 such as transitions between versions of the ICD coding standard, or individual coding behavior.
315 This could potentially introduce bias to the results concerning the burden of a disease,
316 particularly when comparing diagnoses from the present to the past or the burden over an
317 extended period of time.

318 Safety and Risk Group Analysis

319 In this category, the information bias was predominantly manifested as the risk for
320 misclassification bias [32–40] with only one study addressing immortal time bias [38].
321 Specifically, certain limitations were connected with coding challenges, since the classification

322 was based on diagnosis codes from the EHR systems [33,35]. In addition, some form of
323 potential information bias occurred because the drug use in an EHR system had been defined
324 as the prescription of the medication and not as the actual drug intake [36,41]. Aspects like the
325 quantification and analysis of information concerning the physician's decision-making process
326 for treatment selection, treatment switching or the effect of pre-treatment were often missing
327 or inadequately explored [38,40,42] which could lead to confounded results. Details regarding
328 adherence or dosage, as well as unstructured free-text information, were not accessible
329 [40,42,43]. Additionally, significant risk factors, such as smoking, were not available, creating
330 a potentially high risk of confounding in the analysis. Selection bias limitations were evident in
331 ten studies including volunteer bias due to the involvement of individuals who volunteered to
332 join the database used for analysis [35,44]. Moreover, several studies highlighted site
333 heterogeneity, noting that certain hospitals with robust programs may have had a higher
334 burden of disease [33–35]. The limitation of missing data was mentioned in five studies
335 [33,36,38,39,45], for instance, in Li et al. [36] who stated that EHR data sources lacked
336 comprehensive coverage of medical events recorded in other healthcare facilities. While only
337 three studies [38,43,44] explicitly stated their limitations without providing further elaboration,
338 the remaining studies either addressed how they mitigated certain limitations through selected
339 methodologies or sensitivity analyses [34,36,37,39–42,45,46]. Alternatively, they presented
340 arguments explaining why certain limitations were unlikely or invalid in the context of their
341 specific study [32,33,35,39,45,47].

342 A few specific challenges were added by the experts in the focus group discussion as displayed
343 in the second section of Figure 4. First, it is considered important to correctly distinguish
344 between the case and patient definition. Given that patients may have multiple cases
345 associated with them in routine clinical data, it is essential for the research question to specify
346 whether the data refers to the patient level or the individual case level. The lack of clarity could
347 potentially violate the i.i.d-assumptions (independent and identically distributed assumption),
348 which is the basis for many commonly used statistical methods.

Second, another challenge mentioned is the problem of reverse p-hacking for safety endpoints, where the analysis is manipulated to intentionally favor non-significant results [48]. Third, the focus group emphasized the importance of the limitation related to the lack of follow-up data, especially in health care systems where no linkage with routine outcome information is available. A longer follow-up period is crucial for conducting thorough safety analyses to observe adverse events. This aspect is closely linked to the issue of missing information due to death, as death prevents the observation of subsequent adverse events.

Moreover, the experts added several limitations related to medication use. These included insufficient information on the reasons for change in medication and its effect, as well as information on dosage. Additionally, concerns were raised about the off-label use of medication. Unlike in controlled clinical trials, the treatment paths of patients are quite heterogeneous due to the decisions of the clinician or the different specializations of the hospitals.

Treatment Comparison

Confounding remained a particular area of concern, specifically in achieving balance between the groups. Kim et al. [49] reported persistent imbalance in variables even after applying matching techniques. However, they addressed this issue by adjusting for these variables in further analyses. Likewise, Xie et al. [50] stated that individuals treated with the study medication had a higher baseline health burden, potentially leading to underestimated findings due to residual confounding. Only two studies explicitly report missing risk factors that could subsequently not be used for balancing [51,52]. Wang et al. [8] highlighted that claims data had not recorded medication use in hospital, therefore, they needed to use alternative index date and follow-up definition. Similar to the *Safety and Risk Group Analysis*, most potential information biases were a type of misclassification [50,53–56]. Rentsch et al. [52] reported that the identification of outcome events was not provided by a validated algorithm, which could also potentially lead to information bias. Same holds true for Wong et al. [57] as they did not distinguish the reason for deaths. Kim et al. [49] emphasized the general issues of immortal

time and time lag biases in the discussion section, but elaborated on the methods used to address and avoid these biases. One study noted the concern of potential upcoding, i.e. the intentional use of ICD-10 and OPS (Operation and Procedure Classification System) codes with the greatest reimbursement rather than those with the greatest clinical relevance, resulting in overstating the patient's severity of disease [58]. Xie et al. [55] highlighted that the effectiveness of the investigated drug could be overestimated in individuals with poor health characteristics who opt not to receive treatment, while those with better health characteristics who decide not to be treated could lead to underestimation. The studies predominantly stated their open limitations including those that persisted despite implementing causal inference methods or conducting sensitivity analyses aimed at addressing them [50,52,53,55]. Beyond that, Kim et al. [49] detailed how they mitigated their limitations.

The focus group discussion resulted in the following additions. First, data availability with regard to the transferability within different health care sectors, e.g., ambulatory and stationary sector, was added as challenge, especially for Germany [59]. Second, the timing of an intervention is also critical to assess its effectiveness accurately, understand its impact on patient outcomes, and make informed clinical decisions. In EHR or administrative databases, the timing may not be explicitly recorded or may be subject to errors or omissions. Third, the difficulty in defining control groups for comparison was highlighted as well as the verifiability of the inclusion or exclusion criteria [9,60]. Fourth, similar to the *Safety and Risk Group Analysis* scenario, follow-up duration is critical. However, in the context of treatment comparison, the criticality arises also from the possibility that drugs can have a longer delay in efficacy.

Compared to the previous scenarios the ratings of the *Treatment Comparison* category were more widely distributed, with votes spread across numerous limitations. This distribution highlighted the wide range of challenges to consider. Residual confounding and immortal time bias received the majority of votes. Not only the data availability of different data sectors, but also the availability of variables was considered important which is connected to residual confounding, e.g. absence of known confounders.

General limitations and challenges

The focus group viewed the linkage of records, the validation of data and the selection of measurements as critical to conduct routine care data analysis. If there are multiple measurements, it can be challenging to specify which measurement is considered as the relevant one. In connection with record-linkage, both with other hospitals and other data sources, the cases and patient data may overlap between hospitals and other data sources. In addition, the challenges with the definition of estimands and estimators was highlighted [14,61,62].

Discussion

Major findings of this work briefly summarized

This review aimed to present a comprehensive overview of the limitations and biases inherent in the analysis of routine clinical care data. Additionally, we pinpointed the challenges with the greatest potential to influence the results of the analysis emphasizing thorough consideration and discussion for the derivation of meaningful inference. We have intentionally provided a comprehensive overview of the study's features to serve as a guide, categorically organized into different scenarios.

Strength and limitations of this work

Particular strengths of this work are, first, the inclusion of high impact journals in the review to focus on the databases with highest potential for routine data analyses and use of a broad search strategy to cover the work analyzing data routinely collected within the healthcare system. Second, we extracted and categorized the limitations and potential biases identified in the publications. In addition, our two-step procedure, which included a subsequent focus group evaluation, resulted in a more comprehensive overview of potential challenges. Finally, this review only included studies that were successfully published. Therefore, we could not identify challenges that prevented successful data analyses and publication.

Naturally, this study is also subject to further limitations. Firstly, by restricting our search strategy to high-impact publications for the sake of recency and relevance, there is a possibility that we overlooked certain publications and their associated limitations. However, the incorporation of expert experiences and knowledge helped to cover any missed challenge. Second, given that our panel consisted solely of German experts in the field of statistics and clinical epidemiology, the added limitations might be influenced by local challenges specific to this context and might miss the important aspects such as the international or the clinical perspective.

Our study assessed various limitations faced across different study scenarios of interest, resulting in a ranked list of potentially relevant challenges. While our focus was specifically on limitations stated in the discussion section of publications, we did not provide explicit methods to address these limitations. However, this has already been partially discussed in other works [10,63]. The choice of the applied methods of causal inferences, such as propensity score techniques or the method of target trial emulation for established treatments, serves to reduce potential biases and preventing fundamental errors in the study design. In the next step, we aim to explore the requirements and concerns of various stakeholders regarding the use of routine clinical data to achieve different objectives, particularly in therapeutic evaluations. This will involve understanding the diverse expectations, needs, and potential challenges faced by stakeholders in leveraging routine clinical data for meaningful and accurate therapeutic assessments.

Conclusion

This review provides a comprehensive examination of potential limitations and biases in analyses of routine clinical care data reported in recent high-impact journals. We highlighted challenges that could significantly influence analysis results, stressing the necessity for thorough consideration and discussion to derive meaningful conclusions.

454 ## Abbreviations

COPD	Chronic Obstructive Pulmonary Disease
EHR	Electronic Health Records
ICD	International Classification of Diseases and Related Health Problems
JCR	Journal Citation Report
OPS	Operation and Procedure Classification System
RCT	Randomized Controlled Trial
RWD	Real-World Data
RWE	Real-World Evidence

455 ## Ethics approval and consent to participate

456 Not applicable

457 ## Consent for publication

458 Not applicable

459 ## Availability of data and materials

460 Not applicable

461 ## Competing interests

462 The authors declare that they have no competing interests.

463 ## Funding

464 This work was funded by the Federal Ministry of Education and Research (BMBF) in Germany
465 in the framework of the EVA4MII project (FKZ 01ZZ2308A, 01ZZ2308B, 01ZZ2308C). The
466 funding agency had no role in the design, data collection, analyses, interpretation, and
467 reporting of the study. The work of HB, MB, and NB has also been funded by the Deutsche
468 Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 499552394 –
469 SFB 1597.

Authors' contributions

MP, DZ, HB and NB developed the study conception and design. MP, MB, KU and KG were involved in the systematic search process. MP extracted the data from the publications. KU and KG critically reviewed the extraction. MP wrote the first draft of the manuscript; NB supervised this process. MB, DZ, HB, VR, JRP, PH, MK, FR and AS provided additional intellectual content to the manuscript. All authors critically reviewed and approved the final version.

Acknowledgments

We would like to thank the participants of the focus group workshop for the joint discussion.

References

[1] Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence — What Is It and What Can It Tell Us? N Engl J Med 2016;375:2293–7. https://doi.org/10.1056/NEJMsb1609216.

[2] Rahman R, Ventz S, McDunn J, Louv B, Reyes-Rivera I, Polley M-YC, et al. Leveraging external data in the design and analysis of clinical trials in neuro-oncology. Lancet Oncol 2021;22:e456–65. https://doi.org/10.1016/S1470-2045(21)00488-5.

[3] Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther 2017;102:924–33. https://doi.org/10.1002/cpt.857.

[4] Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022;22:287. https://doi.org/10.1186/s12874-022-01768-6.

[5] Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017;106:1–9. https://doi.org/10.1007/s00392-016-1025-6.

[6] Sheldrick RC. Randomized Trials vs Real-world Evidence: How Can Both Inform Decision-making? JAMA 2023;329:1352–3. https://doi.org/10.1001/jama.2023.4855.

[7] Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. JAMA 2022;328:2446. https://doi.org/10.1001/jama.2022.21383.

[8] Wang SV, Schneeweiss S, Franklin JM, Desai RJ, Feldman W, Garry EM, et al. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. JAMA 2023;329:1376–85. https://doi.org/10.1001/jama.2023.4221.

[9] Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a

Randomized Trial Is Not Available: Table 1. Am J Epidemiol 2016;183:758–64. https://doi.org/10.1093/aje/kwv254.

[10] Burger HU, Gerlinger C, Harbron C, Koch A, Posch M, Rochon J, et al. The use of external controls: To what extent can it currently be recommended? Pharm Stat 2021;20:1002–16. https://doi.org/10.1002/pst.2120.

[11] Zöller D, Haverkamp C, Makoudjou A, Sofack G, Kiefer S, Gebele D, et al. Alpha-1-antitrypsin-deficiency is associated with lower cardiovascular risk: an approach based on federated learning. Respir Res 2024;25:38. https://doi.org/10.1186/s12931-023-02607-y.

[12] Schneeweiss S. Von Real-World-Daten zur Real-World-Evidenz: eine praktische Anleitung. Prävent Gesundheitsförderung 2023. https://doi.org/10.1007/s11553-023-01026-7.

[13] Hessel F. Burden of DiseaseBurdenof disease(s). In: Kirch W, editor. Encycl. Public Health, Dordrecht: Springer Netherlands; 2008, p. 94–6. https://doi.org/10.1007/978-1-4020-5614-7_297.

[14] Unkel S, Amiri M, Benda N, Beyersmann J, Knoerzer D, Kupas K, et al. On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. Pharm Stat 2019;18:166–83. https://doi.org/10.1002/pst.1915.

[15] Buchholz I. PT28 Lost in PICO? a Simulation of the EU HTA Scoping Process n.d.

[16] 2022 Journal Impact Factor, Journal Citation Reports (Clarivate, 2023) n.d.

[17] Purpura CA, Garry EM, Honig N, Case A, Rassen JA. The Role of Real-World Evidence in FDA-Approved New Drug and Biologics License Applications. Clin Pharmacol Ther 2022;111:135–44. https://doi.org/10.1002/cpt.2474.

[18] Sterne JAC, Hernán MA, McAleenan A, Reeves BC, Higgins JPT. Chapter 25: Assessing risk of bias in a non-randomized study. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023). Cochrane, 2023. Available from : www.training.cochrane.org/handbook.

[19] World Health Organization. ICD-10 : international statistical classification of diseases and related health problems : tenth revision, 2nd ed. World Health Organization. 2004. https://iris.who.int/handle/10665/42980.

[20] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10:89. https://doi.org/10.1186/s13643-021-01626-4.

[21] Canavan M, Wang X, Ascha M, Miksad R, Showalter TN, Calip G, et al. End-of-Life Systemic Oncologic Treatment in the Immunotherapy Era: The Role of Race, Insurance, and Practice Setting. J Clin Oncol Off J Am Soc Clin Oncol 2023;41:4729–38. https://doi.org/10.1200/JCO.22.02180.

[22] Kamran F, Tang S, Otles E, McEvoy DS, Saleh SN, Gong J, et al. Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study. BMJ 2022;376:e068576. https://doi.org/10.1136/bmj-2021-068576.

[23] Witberg G, Barda N, Hoss S, Richter I, Wiessman M, Aviv Y, et al. Myocarditis after Covid-19 Vaccination in a Large Health Care Organization. N Engl J Med 2021;385:2132–9. https://doi.org/10.1056/NEJMoa2110737.

549 [24] Beck DB, Bodian DL, Shah V, Mirshahi UL, Kim J, Ding Y, et al. Estimated
550 Prevalence and Clinical Manifestations of UBA1 Variants Associated With VEXAS Syndrome
551 in a Clinical Population. JAMA 2023;329:318–24. https://doi.org/10.1001/jama.2022.24836.

552 [25] Forrest IS, Petrazzini BO, Duffy Á, Park JK, Marquez-Luna C, Jordan DM, et al.
553 Machine learning-based marker for coronary artery disease: derivation and validation in two
554 longitudinal cohorts. Lancet Lond Engl 2023;401:215–25. https://doi.org/10.1016/S0140-
555 6736(22)02079-7.

556 [26] Vasileiou E, Simpson CR, Shi T, Kerr S, Agrawal U, Akbari A, et al. Interim findings
557 from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in
558 Scotland: a national prospective cohort study. Lancet Lond Engl 2021;397:1646–57.
559 https://doi.org/10.1016/S0140-6736(21)00677-2.

560 [27] Haas EJ, Angulo FJ, McLaughlin JM, Anis E, Singer SR, Khan F, et al. Impact and
561 effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19
562 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an
563 observational study using national surveillance data. Lancet Lond Engl 2021;397:1819–29.
564 https://doi.org/10.1016/S0140-6736(21)00947-8.

565 [28] Manz CR, Chen J, Liu M, Chivers C, Regli SH, Braun J, et al. Validation of a Machine
566 Learning Algorithm to Predict 180-Day Mortality for Outpatients With Cancer. JAMA Oncol
567 2020;6:1723–30. https://doi.org/10.1001/jamaoncol.2020.4331.

568 [29] Biccler JL, Glimelius I, Eloranta S, Smeland KB, Brown P de N, Jakobsen LH, et al.
569 Relapse Risk and Loss of Lifetime After Modern Combined Modality Treatment of Young
570 Patients With Hodgkin Lymphoma: A Nordic Lymphoma Epidemiology Group Study. J Clin
571 Oncol Off J Am Soc Clin Oncol 2019;37:703–13. https://doi.org/10.1200/JCO.18.01652.

572 [30] Chang AR, Moore BS, Luo JZ, Sartori G, Fang B, Jacobs S, et al. Exome Sequencing
573 of a Clinical Population for Autosomal Dominant Polycystic Kidney Disease. JAMA
574 2022;328:2412–21. https://doi.org/10.1001/jama.2022.22847.

575 [31] Binder N, Blümle A, Balmford J, Motschall E, Oeller P, Schumacher M. Cohort
576 studies were found to be frequently biased by missing disease information due to death. J
577 Clin Epidemiol 2019;105:68–79. https://doi.org/10.1016/j.jclinepi.2018.09.010.

578 [32] Cohen-Stavi CJ, Magen O, Barda N, Yaron S, Peretz A, Netzer D, et al. BNT162b2
579 Vaccine Effectiveness against Omicron in Children 5 to 11 Years of Age. N Engl J Med
580 2022;387:227–36. https://doi.org/10.1056/NEJMoa2205011.

581 [33] Damrauer SM, Chaudhary K, Cho JH, Liang LW, Argulian E, Chan L, et al.
582 Association of the V122I Hereditary Transthyretin Amyloidosis Genetic Variant With Heart
583 Failure Among Individuals of African or Hispanic/Latino Ancestry. JAMA 2019;322:2191–202.
584 https://doi.org/10.1001/jama.2019.17935.

585 [34] Filion KB, Lix LM, Yu OH, Dell'Aniello S, Douros A, Shah BR, et al. Sodium glucose
586 cotransporter 2 inhibitors and risk of major adverse cardiovascular events: multi-database
587 retrospective cohort study. BMJ 2020;370:m3342. https://doi.org/10.1136/bmj.m3342.

588 [35] Forrest IS, Chaudhary K, Vy HMT, Petrazzini BO, Bafna S, Jordan DM, et al.
589 Population-Based Penetrance of Deleterious Clinical Variants. JAMA 2022;327:350–9.
590 https://doi.org/10.1001/jama.2021.23686.

591 [36] Li X, Ostropolets A, Makadia R, Shoaibi A, Rao G, Sena AG, et al. Characterising the
592 background incidence rates of adverse events of special interest for covid-19 vaccines in
593 eight countries: multinational network cohort study. BMJ 2021;373:n1435.

https://doi.org/10.1136/bmj.n1435.

[37] Lyu H, Zhao SS, Zhang L, Wei J, Li X, Li H, et al. Denosumab and incidence of type 2 diabetes among adults with osteoporosis: population based cohort study. BMJ 2023;381:e073435. https://doi.org/10.1136/bmj-2022-073435.

[38] Martin P, Cohen JB, Wang M, Kumar A, Hill B, Villa D, et al. Treatment Outcomes and Roles of Transplantation and Maintenance Rituximab in Patients With Previously Untreated Mantle Cell Lymphoma: Results From Large Real-World Cohorts. J Clin Oncol Off J Am Soc Clin Oncol 2023;41:541–54. https://doi.org/10.1200/JCO.21.02698.

[39] Seymour CW, Kennedy JN, Wang S, Chang C-CH, Elliott CF, Xu Z, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. JAMA 2019;321:2003–17. https://doi.org/10.1001/jama.2019.5791.

[40] You SC, Rho Y, Bikdeli B, Kim J, Siapos A, Weaver J, et al. Association of Ticagrelor vs Clopidogrel With Net Adverse Clinical Events in Patients With Acute Coronary Syndrome Undergoing Percutaneous Coronary Intervention. JAMA 2020;324:1640–50. https://doi.org/10.1001/jama.2020.16167.

[41] Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet Lond Engl 2019;394:1816–26. https://doi.org/10.1016/S0140-6736(19)32317-7.

[42] Vinogradova Y, Coupland C, Hill T, Hippisley-Cox J. Risks and benefits of direct oral anticoagulants versus warfarin in a real world setting: cohort study in primary care. BMJ 2018;362:k2505. https://doi.org/10.1136/bmj.k2505.

[43] Chavez-MacGregor M, Lei X, Zhao H, Scheet P, Giordano SH. Evaluation of COVID-19 Mortality and Adverse Outcomes in US Patients With or Without Cancer. JAMA Oncol 2022;8:69–78. https://doi.org/10.1001/jamaoncol.2021.5148.

[44] Berry ASF, Finucane BM, Myers SM, Abril A, Kirchner HL, Ledbetter DH, et al. Association of Supernumerary Sex Chromosome Aneuploidies With Venous Thromboembolism. JAMA 2023;329:235–43. https://doi.org/10.1001/jama.2022.23897.

[45] Deng Y, Polley EC, Wallach JD, Dhruva SS, Herrin J, Quinto K, et al. Emulating the GRADE trial using real world data: retrospective comparative effectiveness study. BMJ 2022;379:e070717. https://doi.org/10.1136/bmj-2022-070717.

[46] Chemaitelly H, AlMukdad S, Ayoub HH, Altarawneh HN, Coyle P, Tang P, et al. Covid-19 Vaccine Protection among Children and Adolescents in Qatar. N Engl J Med 2022;387:1865–76. https://doi.org/10.1056/NEJMoa2210058.

[47] Harstad E, Shults J, Barbaresi W, Bax A, Cacia J, Deavenport-Saman A, et al. α2-Adrenergic Agonists or Stimulants for Preschool-Age Children With Attention-Deficit/Hyperactivity Disorder. JAMA 2021;325:2067–75. https://doi.org/10.1001/jama.2021.6118.

[48] Chuard PJC, Vrtílek M, Head ML, Jennions MD. Evidence that nonsignificant results are sometimes preferred: Reverse P-hacking or selective reporting? PLOS Biol 2019;17:e3000127. https://doi.org/10.1371/journal.pbio.3000127.

[49] Kim NH, Han KH, Choi J, Lee J, Kim SG. Use of fenofibrate on cardiovascular outcomes in statin users with metabolic syndrome: propensity matched cohort study. BMJ 2019;366:l5125. https://doi.org/10.1136/bmj.l5125.

[50] Xie Y, Bowe B, Al-Aly Z. Molnupiravir and risk of hospital admission or death in adults with covid-19: emulation of a randomized target trial using electronic health records. BMJ 2023;380:e072705. https://doi.org/10.1136/bmj-2022-072705.

[51] Mahévas M, Tran V-T, Roumier M, Chabrol A, Paule R, Guillaud C, et al. Clinical efficacy of hydroxychloroquine in patients with covid-19 pneumonia who require oxygen: observational comparative study using routine care data. BMJ 2020;369:m1844. https://doi.org/10.1136/bmj.m1844.

[52] Rentsch CT, Beckman JA, Tomlinson L, Gellad WF, Alcorn C, Kidwai-Khan F, et al. Early initiation of prophylactic anticoagulation for prevention of coronavirus disease 2019 mortality in patients admitted to hospital in the United States: cohort study. BMJ 2021;372:n311. https://doi.org/10.1136/bmj.n311.

[53] Andersson NW, Thiesson EM, Baum U, Pihlström N, Starrfelt J, Faksová K, et al. Comparative effectiveness of bivalent BA.4-5 and BA.1 mRNA booster vaccines among adults aged ≥50 years in Nordic countries: nationwide cohort study. BMJ 2023;382:e075286. https://doi.org/10.1136/bmj-2022-075286.

[54] Deputy NP, Deckert J, Chard AN, Sandberg N, Moulia DL, Barkley E, et al. Vaccine Effectiveness of JYNNEOS against Mpox Disease in the United States. N Engl J Med 2023;388:2434–43. https://doi.org/10.1056/NEJMoa2215201.

[55] Xie Y, Bowe B, Al-Aly Z. Nirmatrelvir and risk of hospital admission or death in adults with covid-19: emulation of a randomized target trial using electronic health records. BMJ 2023;381:e073312. https://doi.org/10.1136/bmj-2022-073312.

[56] Zheng B, Green ACA, Tazare J, Curtis HJ, Fisher L, Nab L, et al. Comparative effectiveness of sotrovimab and molnupiravir for prevention of severe covid-19 outcomes in patients in the community: observational cohort study with the OpenSAFELY platform. BMJ 2022;379:e071932. https://doi.org/10.1136/bmj-2022-071932.

[57] Wong CKH, Au ICH, Lau KTK, Lau EHY, Cowling BJ, Leung GM. Real-world effectiveness of molnupiravir and nirmatrelvir plus ritonavir against mortality, hospitalisation, and in-hospital outcomes among community-dwelling, ambulatory patients with confirmed SARS-CoV-2 infection during the omicron wave in Hong Kong: an observational study. Lancet Lond Engl 2022;400:1213–22. https://doi.org/10.1016/S0140-6736(22)01586-0.

[58] Marafino BJ, Escobar GJ, Baiocchi MT, Liu VX, Plimier CC, Schuler A. Evaluation of an intervention targeted with predictive analytics to prevent readmissions in an integrated health system: observational study. BMJ 2021;374:n1747. https://doi.org/10.1136/bmj.n1747.

[59] Lang C, Gottschall M, Sauer M, Köberlein-Neu J, Bergmann A, Voigt K. „Da kann man sich ja totklingeln, geht ja keiner ran“ – Schnittstellenprobleme zwischen stationärer, hausärztlicher und ambulant-fachspezialisierter Patientenversorgung aus Sicht Dresdner Hausärzte. Gesundheitswesen 2019;81:822–30. https://doi.org/10.1055/a-0664-0470.

[60] D’Arcy M, Stürmer T, Lund JL. The importance and implications of comparator selection in pharmacoepidemiologic research. Curr Epidemiol Rep 2018;5:272–83. https://doi.org/10.1007/s40471-018-0155-y.

[61] Luijken K, van Eekelen R, Gardarsdottir H, Groenwold RHH, van Geloven N. Tell me what you want, what you really really want: Estimands in observational pharmacoepidemiologic comparative effectiveness and safety studies. Pharmacoepidemiol Drug Saf 2023;32:863–72. https://doi.org/10.1002/pds.5620.

[62] European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity

683 analysis in clinical trials to the guideline on statistical principles for clinical trials 2020.

684 [63] Nguyen VT, Engleton M, Davison M, Ravaud P, Porcher R, Boutron I. Risk of bias in
685 observational studies using routinely collected data of comparative effectiveness research: a
686 meta-research study. BMC Med 2021;19:279. https://doi.org/10.1186/s12916-021-02151-w.

687 [64] Coltin H, Pequeno P, Liu N, Tsang DS, Gupta S, Taylor MD, et al. The Burden of
688 Surviving Childhood Medulloblastoma: A Population-Based, Matched Cohort Study in
689 Ontario, Canada. J Clin Oncol Off J Am Soc Clin Oncol 2023;41:2372–81.
690 https://doi.org/10.1200/JCO.22.02466.

691 [65] Song Q, Bates B, Shao YR, Hsu F-C, Liu F, Madhira V, et al. Risk and Outcome of
692 Breakthrough COVID-19 Infections in Vaccinated Patients With Cancer: Real-World
693 Evidence From the National COVID Cohort Collaborative. J Clin Oncol Off J Am Soc Clin
694 Oncol 2022;40:1414–27. https://doi.org/10.1200/JCO.21.02419.

695

696 Supplementary Material

697 Supplementary S1

698 *Table S1 Search strategy*

(1) In Title and Abstract		(2) Journals	(3) Time
real-world evidence **OR** Real-world data **OR** real-world **OR** RWE **OR** routine data **OR** routine care data **OR** Emulation **OR** Electronic health record	**AND**	The New England journal of medicine **OR** The Lancet. Oncology **OR** BMJ (Clinical research ed.) **OR** JAMA **OR** Journal of clinical oncology: official journal of the American Society of Clinical Oncology **OR** JAMA oncology **OR** Lancet (London, England)	Filter: 2018 – 2023

699

700 *Table S2 Detailed Search History PubMed on October 31st 2023*

Search	Query	Results
#1	**Search:** ("The New England journal of medicine"[Journal]) OR ("BMJ (Clinical research ed.)" [Journal]) OR ("The Lancet. Oncology" [Journal]) OR ("JAMA" [Journal]) OR ("Journal of clinical oncology : official journal of the American Society of Clinical Oncology" [Journal]) OR ("JAMA oncology"(Journal]) OR ("Lancet (London, England)"[Journal])	435,802
#2	**Search:** "real-world evidence" [tiab] OR "Real-world data" [tiab] OR real-world[tiab] OR RWE[tiab] OR "routine data" [tiab] or "routine care data " [tiab] OR Emulation [tiab] OR "Electronic health record" [tiab]	103,854

#3	**Search:** #1 AND #2	387
#4	**Search:** #1 AND #2 **Filters:** from 2018-2023	227

Supplementary S2

Table S3 provides a comprehensive overview of the extracted studies, including their characteristics and limitations described in their respective discussion section. Note, that for some publications (indicated with an asterisk) the category could not be clearly assigned, especially in the case of *Treatment Comparison* and *Safety and Risk Group Analysis*. These publications did not only compare the effectiveness of treatments but specifically compared the risk of adverse events between two or more treatments [34,37,40–42,45,47].

All studies categorized into the Burden of Disease scenario were cohort studies except for three studies [22,25,28] that were developing or validating a machine learning model to predict certain endpoints using EHR data. Three studies [26,29,64] examined time-to-event outcomes using Cox regression as well as Kaplan-Meier estimator partly with inverse propensity weights. Prevalence and incidences were analyzed in four studies [23,24,27,30]. The studies that were assigned to the category Safety and Risk Group Analysis were mostly cohort studies, in addition to two case-control studies and one cross-sectional cohort study [32,33]. Similar to the category Burden of Disease, predominantly time-to-event outcomes were evaluated using Kaplan-Meier estimation or Cox (proportional hazards) regression models [34,37,38,40–42,44–46]. In contrast, many studies have used propensity score methods such as matching and weighting [32,34,37,40–42,45,46] or applied the concept of target trial emulation [32,37,45]. Similar to the category Safety and Risk Group Analysis, most studies in the Treatment Comparison scenario examined time-to-event outcomes. They employed propensity score methods such as matching and weighting or applied the concept of target trial emulation. In addition to that, two studies used the clone method, a special form of emulating a target trial where each individual is cloned into both treatment groups, censored patients according to their designated treatment strategy and weighted the uncensored to avoid selection bias [9,50,55].

727 *Table S3 Overview of study characteristics and limitations*

Author	Study Design	Aim of Study	Outcome	Methods	Scope	Countries	Limitations								
							Main Bias Categories				Additional Bias Category				
Burden of disease							Confounding	Selection bias	Information bias	Reporting bias	Follow-up Challenges	Missing Data	Coding Challenges	Validation & Data quality	Operationalization or availability of variables
Beck et al. [24]	cohort study	genetic association	prevalence	Poisson test	multi-center	USA		x	x				x		
Biccler et al. [29]	cohort study	relapse risk and loss of lifetime after treatment	relapse risks, survival	logistic regression, KM estimation, Aalen-Johansen estimation	multi-national	Nordic countries	x				x			x	
Canavan et al. [21]	cohort study	association of EOL treatment and practice-level factors	use of EOL therapy	adjusted logistic regression with random intercept	nation-wide	USA	x	x	x	x		x			x
Chang et al. [30]	cohort study	genetic association	prevalence	adjusted Firth logistic regression	multi-center	USA		x	x	x				x	
Coltin et al. [64]	cohort study	burden of surviving disease	cumulative incidence of mortality	time-to-event analyses, Cox model	multi-center	Canada		x			x				

Forrest et al. [25]	cohort study	disease-predictive ML	disease prediction	random forest model, model evaluation: AUROC, Sensitivity, Specificity, PPV, NPV, Brier score, adj. linear -, logistic -, and Cox regression	nation-wide	UK		x	x			x			
Haas et al. [27]	cohort study	impact and effectiveness of treatment	prevalence, incidence rate ratio	negative binomial regression	nation-wide	Israel	x	x	x	x					
Kamran et al. [22]	cohort study	ML model development	mortality, treatment events	regularized logistic regression, AUROC, model calibration	multi-center	USA	x								x
Manz et al. [28]	cohort study	Validation of ML Algorithm	mortality	AUROC, AUPRC, scaled Brier score, logistic regression	multi-center	USA			x			x	x		
Vasileiou et al. [26]	cohort study	hospital admission after vaccination	hospital admission	time-dependent Cox model & Poisson regression with inverse propensity weights	nation-wide	UK	x	x			x				
Witberg et al. [23]	cohort study	frequency and severity of disease	(cumulative) incidence	KM estimation	nation-wide	Israel		x	x			x			

		after treatment													
Safety and Risk Group							Confounding	Selection bias	Information bias	Reporting bias	Follow-up Challenges	Missing Data	Coding Challenges	Validation & Data quality	Operationalization or availability of variables
Berry et al. [44]	cohort study	genetic association	incidence, prevalence	Cox regression, adj. logistic regression	multi-center	USA UK	x	x	x					x	
Chavez-MacGregor et al. [43]	cohort study	outcome comparison of risk groups	morality, ventilation, ICU stay, hospitalization	adj. logistic regression	multi-center	USA	x	x							x
Chemaitelly et al. [46]	cohort studies	treatment effectiveness among children and adolescents	incidence rate, cumulative incidence	1:1 matching, KM estimation, Cox regression	nation-wide	Qatar	x	x		x					
Cohen-Stavi et al. [32]	cohort study, case-control study	treatment effectiveness among children	cumulative incidence	target trail emulation, matching, KM estimation	multi-center	Israel	x		x		x				
Damrauer et al. [33]	case-control study, cross sectional study	genetic association	prevalence	adj. logistic regressions	multi-center	USA		x	x			x	x		
Deng et al. [45]*	cohort study	comparative effectiveness	time to event	target trail emulation, inverse	nation-wide	USA	x	x	x		x	x			x

				probability of treatment weighting, generalized boosted models for PS, inverse probability of treatment weighted Kaplan-Meier method, PS-weighted Cox regression											
Filion et al. [34]*	cohort study	adverse event association	major adverse events	time conditional PS using conditional logistic regression, 1:1 matching, Cox regression	multi-center	Canada UK	x	x	x	x	x				x
Forrest et al. [35]	cohort study	disease risk associated with gene variants	risk difference between prevalence	2-sided Fisher's exact tests	multi-center	USA UK		x	x				x		
Harstad et al. [47]*	observational study	treatment effectiveness	frequency, relative risk of Adverse Effects	logistic and Cox regression	multi-center	USA	x	x			x			x	
Li et al. [36]	cohort study	adverse event association	incidence rate	descriptive analyses	multi-national	Australia France Germany Japan Nethe			x	x		x		x	

						rlands Spain UK US									
Lyu et al. [37]*	cohort study	comparative effectiveness	incidence	target trail emulation, PS-matching, Cox regression	natio n-wide	UK	x		x		x				
Martin et al. [38]	cohort study	therapy evaluation	time-to-next-treatment, overall survival	KM, Cox regression	natio n-wide	USA	x		x			x			x
Seymour et al. [39]	cohort study	phenotype derivation	phenotyp e frequency , mortality	Multiple imputation, k-means clustering	single - cente r	USA		x	x			x			
Song et al. [65]	cohort study	comparative effectiveness	disease risk	adj. logistic regression	natio n-wide	USA	x			x					
Suchard et al. [41]*	cohort study	comparative effectiveness	relative risk	PS model, matching, Cox regression, meta analysis	multi-natio nal	USA South Korea Japan Germ any	x	x	x	x					
Vinogrado va et al.[42]*	cohort study	treatment association	hospital admission or death	PS, Cox regression	natio n-wide	UK	x		x	x					x
You et al. [40]*	cohort study	treatment association	net adverse clinical events	PS-matching, Cox regression	multi-cente r	South Korea USA	x		x	x				x	x

Treatment Comparison							Confounding	Selection bias	Information bias	Reporting bias	Follow-up Challenges	Missing Data	Coding Challenges	Validation & Data quality	Operationalization or availability of variables
Andersson et al. [53]	cohort study	comparative effectiveness	hospital admission, death	target trail emulation, PS-matching, İnverse probability of treatment weights, KM	multi-national	Nordic countries	x	x	x						
Deputy et al. [54]	case-control study	vaccine effectiveness	association between vaccination status and case patient or control patient status	matching, conditional logistic regression	nation-wide	USA	x	x	x						
Kim et al. [49]	cohort study	treatment effectiveness	time to cardiovascular events	PS-matching, stratified Cox regression	nation-wide	South Korea	x	x	x	x					
Mahévas et al. [51]	case-control study	treatment effectiveness	survival	inverse probability of treatment weighting, Cox regression, KM, multiple imputation	multi-center	France	x	x							x
Marafino et al. [58]	cohort study	intervention association	mortality, readmission	target-trail emulation, generalized linear mixed effects	multi-center	USA	x	x					x		

				models, difference-in-difference analysis											
Rentsch et al. [52]	cohort study	early treatment association	mortality	inverse probability of treatment weighted KM, Cox regression	nation-wide	USA	x	x	x			x		x	x
Wang et al. [8]	cohort study	RCT comparison	time to event	target trail emulation, PS matching, Cox regression	nation-wide	USA	x	x		x	x				
Wong et al. [57]	cohort study, case-control study	comparative effectiveness	mortality, hospital admission, progression	propensity score matching, Cox regression, conditional logistic regression	territory-wide	Hong Kong	x	x	x			x			
Xie et al. [50]	cohort study	treatment association	relative risk, event rate, absolute risk reduction: hospital admission, death	target trail emulation, clone method, inverse probability of censoring weight	nation-wide	USA	x		x						
Xie et al. [55]	cohort study	treatment effectiveness	relative risk, event rate, absolute risk reduction:	target trail emulation, inverse probability weighting, clone method, weighted KM	nation-wide	USA	x	x	x						

			hospital admission, death												
Zheng et al. [56]	cohort study	comparative effectiveness	hospital admission, death	Cox regression, PS- weighting	nation-wide	UK	x	x	x						x

Abbreviation: KM… Kaplan-Meier, PS... Propensity Score, EOL… End-of-Life, ICU… Intensive Care Unit, ML… Machine Learning, AUROC… Area under the receiver operating curve, AUPRC… Area under the precision-recall curve

* This indicates publications that were not uniquely categorizable into the main scenarios. Specifically, they did not only compare the effectiveness of treatments but compared the risk of adverse events between two or more treatments