

# 1 Methodological Challenges using Routine Clinical 2 Care Data for Real-World Evidence: a Rapid 3 Review utilizing a systematic literature search and 4 focus group discussion

5

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

9 <sup>1</sup> Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Centre – University of  
10 Freiburg, Germany

11 <sup>2</sup> Freiburg Centre for Data Analysis and Modelling, University of Freiburg, Freiburg, Germany

12 <sup>3</sup> Institute for Medical Data Sciences, University Hospital Würzburg, Würzburg, Germany

13 <sup>4</sup> Institute for Clinical Epidemiology and Biometry, University Würzburg, Würzburg, Germany

14 <sup>5</sup> Clinical Trial Centre, University Hospital Würzburg, Würzburg, Germany

15 <sup>6</sup> Institute of Medical Statistics, Computer and Data Sciences, Jena University & Jena University  
16 Hospital, Jena, Germany

17 <sup>7</sup> Institute of General Practice/Family Medicine, Faculty of Medicine and Medical Centre - University of  
18 Freiburg, Freiburg, Germany

19 \*corresponding author

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

## 20 Abstract

### 21 Background

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

### 29 Methods

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

### 42 Results

43 In October 2023, 227 articles were identified, 69 were assessed for eligibility, and 39 were  
44 included in the review: 11 on the burden of disease, 17 on safety and risk group analysis, and  
45 11 on treatment comparison. Besides typical biases in observational studies, we identified  
46 additional challenges specific to RWE frequently mentioned in the discussion sections. The

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

## 50 Conclusion

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

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

56

## 57 Background

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

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

73 Randomized controlled trials (RCTs) are the gold standard for answering questions about  
74 treatment efficacy and safety. However, analyzing RWD may be useful to bridge evidence  
75 gaps at the interface to clinical practice. Utilizing routine data offers several advantages,  
76 including a large number of observations, especially when leveraging and linking multiple  
77 clinical data sources. It allows the coverage of different locations, patient populations (e.g.,  
78 different age distribution or disease severity), and practice patterns in routine health care  
79 [1,3,6]. Routine data analysis can also be valuable when RCTs are not feasible due to ethical  
80 or practical reasons. Particularly in the context of treatment comparisons, the use of the target  
81 trial emulation framework allows to obtain comparable results to those observed in RCTs when  
82 carefully and fully emulated [7–9]. An alternative approach, rather than relying solely on routine  
83 data or their linkage for treatment comparison, is to use the data as an external control [2,10].

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

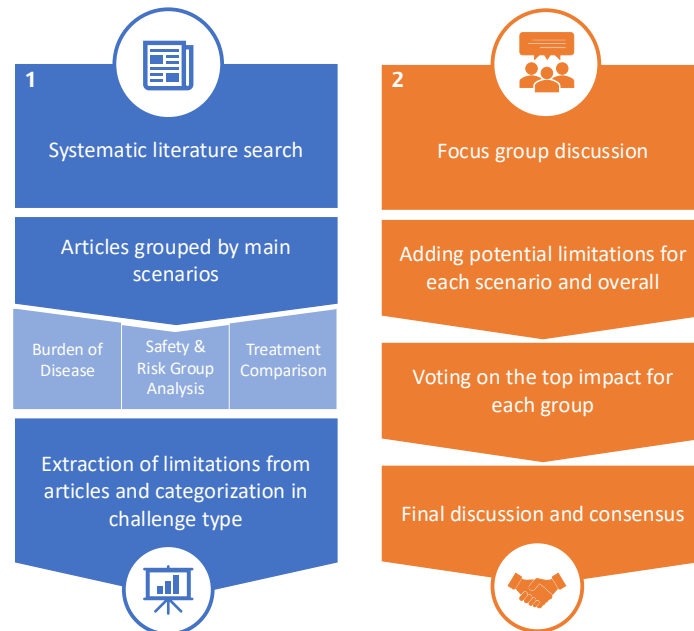
97 For structuring an investigation on limitations, we suggest three scenarios, derived from the  
98 field of clinical epidemiology where routine data hold high potential for generating RWE: (1)  
99 *Burden of Disease*, (2) *Safety and Risk Group Analysis*, and (3) *Treatment Comparison*.  
100 *Burden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

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

124 In this work, we aim to provide an overview of the reported challenges and biases inherent in  
125 routine data analysis with respect to their potential impact in the three main scenarios. This is  
126 accomplished in a step-wise procedure as illustrated in Figure 1. First, a systematic literature  
127 search is conducted to extract the limitations outlined in the respective discussion sections.  
128 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

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

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

167

## 168 [Data extraction and categorization](#)

169 Methodological challenges and limitations mentioned in the discussion sections of the included  
170 publications were extracted as the main outcome. In addition, study characteristics such as  
171 design, the underlying data sources, the country from which data was obtained, the methods  
172 employed, and the purpose of the published studies were retrieved. Publications were  
173 categorized into the three predefined main areas of application (Burden of Disease, Safety and  
174 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 specificities 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



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

## 204 Focus Group

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

## 216 Results

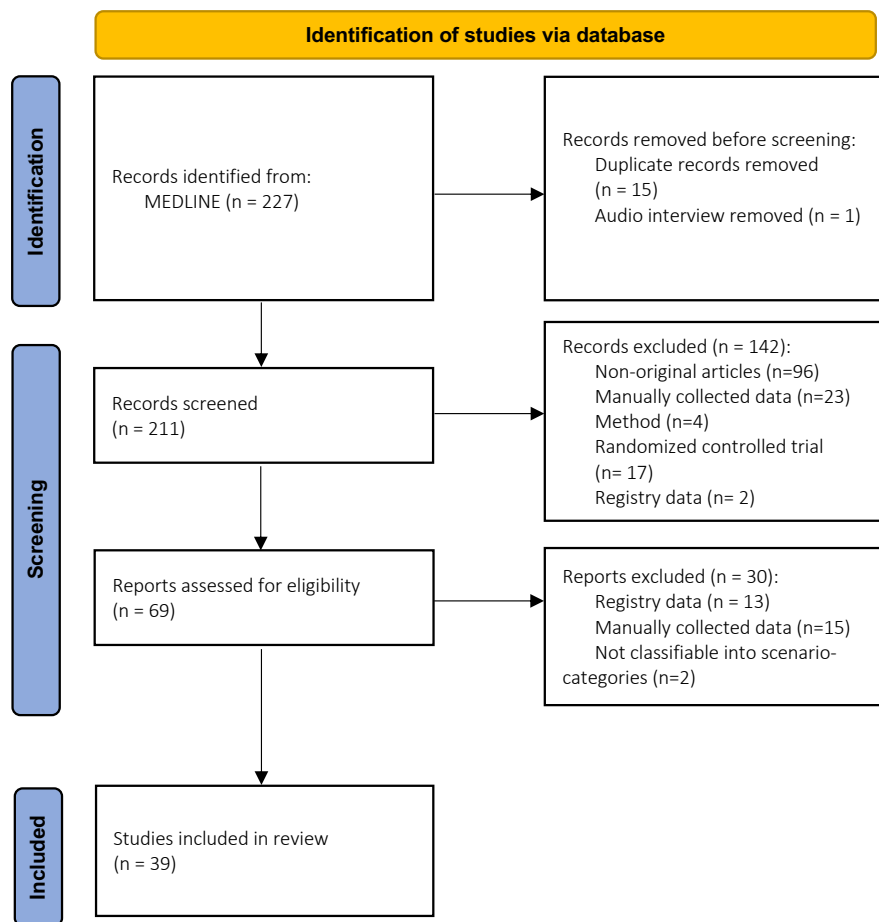
### 217 Study selection and characteristics

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

226

227

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 (%)
<b>Journals</b>	
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)
<b>Year</b>	
2018	1 (2.6)
2019	5 (12.8)
2020	4 (10.3)
2021	7 (18.0)
2022	10 (25.6)
2023	12 (30.8)
<b>Country</b>	
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)
<b>Data Source</b>	
EHR	23 (59.0)
EHR + additional linked data*	11 (28.2)
Administrative data**	5 (12.8)
<b>Scope of Data</b>	
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)

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

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

244 \*\* administrative data include insurance information and claims data

245

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

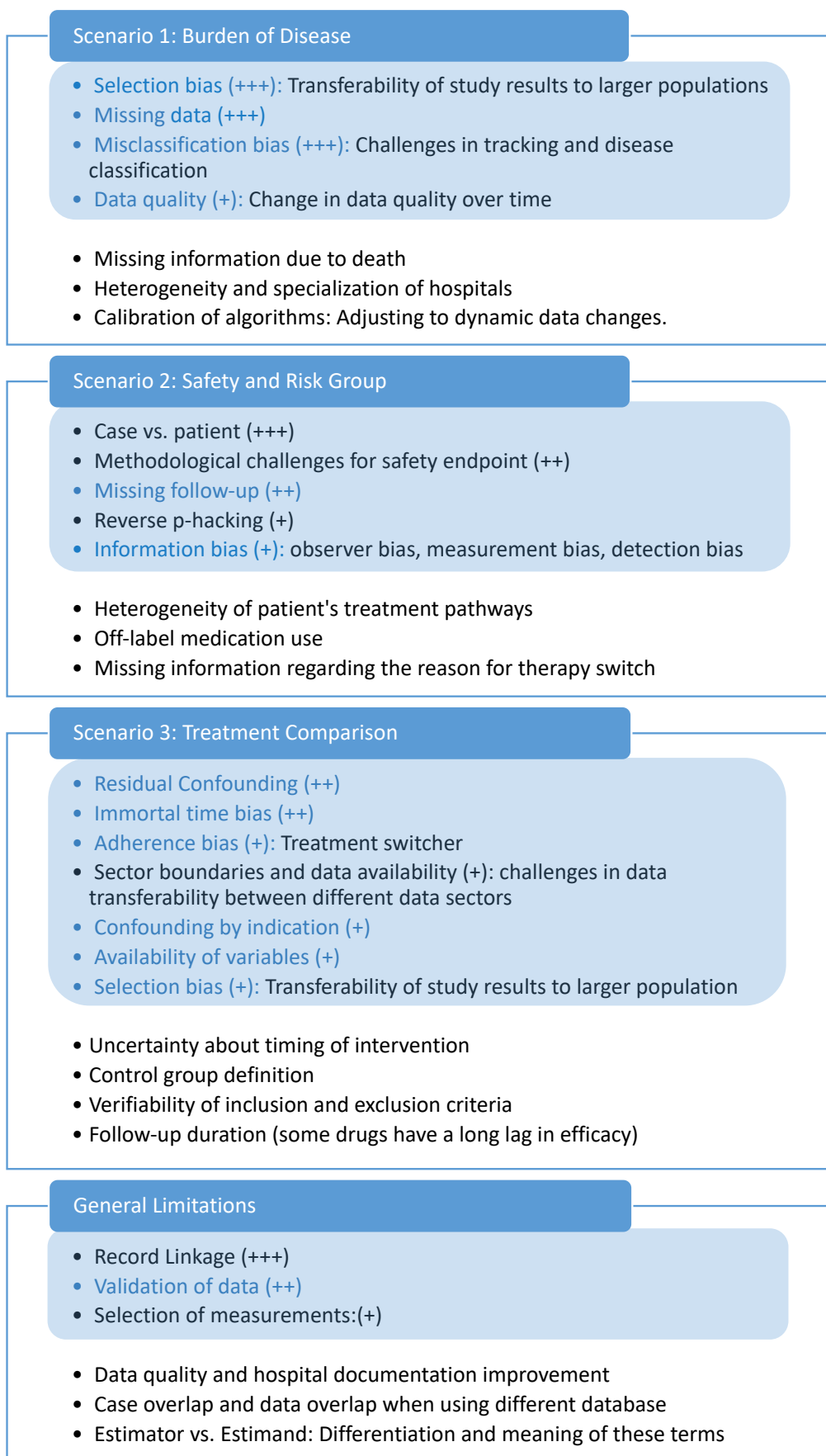
255 *Figure 3 Summary of biases extracted from the studies*

		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
<i>Burden of Disease</i>	EHR	3	4	5	2	2	2	3		
	Linked EHR	2	3	2	1			1	2	2
	Admin. Data		1							
<i>Safety &amp; Risk Group</i>	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	
<i>Treatment Comparison</i>	EHR	9	9	7		1	3	2		1
	Linked EHR									
	Admin. Data	2	2	1					1	

256  
 257 Number of studies facing the main biases and additional identified bias categories grouped  
 258 by data source types and the main scenarios: *Burden of Disease*, *Safety and Risk Group Analysis*,  
 259 *and Treatment Comparison*. Abbreviations: EHR...Electronic Health Records, Admin. Data...  
 260 Administrative Data

261 In the following, we summarize the findings on methodological challenges and limitations from  
 262 both the review and the focus group workshop for each main scenario, separately. In Figure  
 263 4, all challenges that were added and ranked according to their relevance in the focus group  
 264 discussion are summarized.

265 *Figure 4 Overview of Workshop Results*



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

### 273 **Burden of disease**

274 Five publications reported limitations potentially leading to confounding. Of these, two studies  
275 had limited availability of variables included in the analysis resulting in the possibility of  
276 unmeasured confounding [21,22]. However, Kamran et al. [22] argued that with the use of a  
277 limited number of variables in the model a reliable identification and validation of all variables  
278 even across different institutions was feasible. Limitations related to the risk of selection bias  
279 were present in eight studies. On the one hand, in Canavan et al. [21] a volunteer bias could  
280 not be ruled out because the data used for the analyses included individuals who opted to be  
281 included in the study database. On the other hand, in Witberg et al. [23] the absence of  
282 simultaneously enrolled comparison groups was mentioned which could lead to a potential  
283 selection bias. Moreover, some studies highlighted the fact that the results might not be  
284 transferable to other clinical practices or populations [21,24–26]. The primary type of  
285 information bias (n=7) observed was misclassification bias, affecting either the exposure or  
286 outcome [21,24,27,28]. Beck et al. [24] highlighted that clinical diagnoses were probably not  
287 fully captured due to challenges in diagnostic and physician coding. On the other hand, Manz  
288 et al. [28] noted a limitation indicating that they were limited to use only patients with a coded  
289 classifier variable for comparing the machine learning model with the commonly used  
290 prognostic reference. This limitation resulted from differences in the characteristics of the  
291 patients with and without the coded variable. Kamran et al. [22] additionally emphasized that  
292 even though a common EHR provider across facilities was implemented, it remains crucial to  
293 have in-depth knowledge of the local deployment. Two studies reported limited follow-up time  
294 [26,29]. However, Bicler 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.



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

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

## 362 [Treatment Comparison](#)

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

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

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

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

## 403 General limitations and challenges

404 The focus group viewed the linkage of records, the validation of data and the selection of  
405 measurements as critical to conduct routine care data analysis. If there are multiple  
406 measurements, it can be challenging to specify which measurement is considered as the  
407 relevant one. In connection with record-linkage, both with other hospitals and other data  
408 sources, the cases and patient data may overlap between hospitals and other data sources.  
409 In addition, the challenges with the definition of estimands and estimators was highlighted  
410 [14,61,62].

## 411 Discussion

### 412 *Major findings of this work briefly summarized*

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

### 419 *Strength and limitations of this work*

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

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

436 Our study assessed various limitations faced across different study scenarios of interest,  
437 resulting in a ranked list of potentially relevant challenges. While our focus was specifically on  
438 limitations stated in the discussion section of publications, we did not provide explicit methods  
439 to address these limitations. However, this has already been partially discussed in other works  
440 [10,63]. The choice of the applied methods of causal inferences, such as propensity score  
441 techniques or the method of target trial emulation for established treatments, serves to reduce  
442 potential biases and preventing fundamental errors in the study design. In the next step, we  
443 aim to explore the requirements and concerns of various stakeholders regarding the use of  
444 routine clinical data to achieve different objectives, particularly in therapeutic evaluations. This  
445 will involve understanding the diverse expectations, needs, and potential challenges faced by  
446 stakeholders in leveraging routine clinical data for meaningful and accurate therapeutic  
447 assessments.

## 448 Conclusion

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

453

## 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.

## 470 Authors' contributions

471 MP, DZ, HB and NB developed the study conception and design. MP, MB, KU and KG were  
472 involved in the systematic search process. MP extracted the data from the publications. KU  
473 and KG critically reviewed the extraction. MP wrote the first draft of the manuscript, NB  
474 supervised this process. All authors critically reviewed and approved the final version.

## 475 Acknowledgments

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

477

## 478 References

- 479 [1] Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-  
480 World Evidence — What Is It and What Can It Tell Us? *N Engl J Med* 2016;375:2293–7.  
481 <https://doi.org/10.1056/NEJMsb1609216>.
- 482 [2] Rahman R, Vantz S, McDunn J, Louv B, Reyes-Rivera I, Polley M-YC, et al.  
483 Leveraging external data in the design and analysis of clinical trials in neuro-oncology.  
484 *Lancet Oncol* 2021;22:e456–65. [https://doi.org/10.1016/S1470-2045\(21\)00488-5](https://doi.org/10.1016/S1470-2045(21)00488-5).
- 485 [3] Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses  
486 Substitute for Randomized Controlled Trials? *Clin Pharmacol Ther* 2017;102:924–33.  
487 <https://doi.org/10.1002/cpt.857>.
- 488 [4] Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications,  
489 challenges and opportunities. *BMC Med Res Methodol* 2022;22:287.  
490 <https://doi.org/10.1186/s12874-022-01768-6>.
- 491 [5] Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health  
492 records to facilitate clinical research. *Clin Res Cardiol* 2017;106:1–9.  
493 <https://doi.org/10.1007/s00392-016-1025-6>.
- 494 [6] Sheldrick RC. Randomized Trials vs Real-world Evidence: How Can Both Inform  
495 Decision-making? *JAMA* 2023;329:1352–3. <https://doi.org/10.1001/jama.2023.4855>.
- 496 [7] Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal  
497 Inference From Observational Data. *JAMA* 2022;328:2446.  
498 <https://doi.org/10.1001/jama.2022.21383>.
- 499 [8] Wang SV, Schneeweiss S, Franklin JM, Desai RJ, Feldman W, Garry EM, et al.  
500 Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of  
501 32 Clinical Trials. *JAMA* 2023;329:1376–85. <https://doi.org/10.1001/jama.2023.4221>.
- 502 [9] Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a  
503 Randomized Trial Is Not Available: Table 1. *Am J Epidemiol* 2016;183:758–64.  
504 <https://doi.org/10.1093/aje/kwv254>.
- 505 [10] Burger HU, Gerlinger C, Harbron C, Koch A, Posch M, Rochon J, et al. The use of

- 506 external controls: To what extent can it currently be recommended? *Pharm Stat*  
507 2021;20:1002–16. <https://doi.org/10.1002/pst.2120>.
- 508 [11] Zöller D, Haverkamp C, Makoudjou A, Sofack G, Kiefer S, Gebele D, et al. Alpha-1-  
509 antitrypsin-deficiency is associated with lower cardiovascular risk: an approach based on  
510 federated learning. *Respir Res* 2024;25:38. <https://doi.org/10.1186/s12931-023-02607-y>.
- 511 [12] Schneeweiss S. Von Real-World-Daten zur Real-World-Evidenz: eine praktische  
512 Anleitung. *Prävent Gesundheitsförderung* 2023. <https://doi.org/10.1007/s11553-023-01026-7>.
- 513 [13] Hessel F. Burden of Disease Burden of disease(s). In: Kirch W, editor. *Encycl. Public*  
514 *Health*, Dordrecht: Springer Netherlands; 2008, p. 94–6. [https://doi.org/10.1007/978-1-4020-](https://doi.org/10.1007/978-1-4020-5614-7_297)  
515 [5614-7\\_297](https://doi.org/10.1007/978-1-4020-5614-7_297).
- 516 [14] Unkel S, Amiri M, Benda N, Beyersmann J, Knoerzer D, Kupas K, et al. On  
517 estimands and the analysis of adverse events in the presence of varying follow-up times  
518 within the benefit assessment of therapies. *Pharm Stat* 2019;18:166–83.  
519 <https://doi.org/10.1002/pst.1915>.
- 520 [15] Buchholz I. PT28 Lost in PICO? a Simulation of the EU HTA Scoping Process n.d.
- 521 [16] 2022 Journal Impact Factor, Journal Citation Reports (Clarivate, 2023) n.d.
- 522 [17] Purpura CA, Garry EM, Honig N, Case A, Rassen JA. The Role of Real-World  
523 Evidence in FDA-Approved New Drug and Biologics License Applications. *Clin Pharmacol*  
524 *Ther* 2022;111:135–44. <https://doi.org/10.1002/cpt.2474>.
- 525 [18] Sterne JAC, Hernán MA, McAleenan A, Reeves BC, Higgins JPT. Chapter 25:  
526 Assessing risk of bias in a non-randomized study. In: Higgins JPT, Thomas J, Chandler J,  
527 Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic*  
528 *Reviews of Interventions version 6.4 (updated August 2023)*. Cochrane, 2023. Available from  
529 : [www.training.cochrane.org/handbook](http://www.training.cochrane.org/handbook).
- 530 [19] World Health Organization. ICD-10 : international statistical classification of diseases  
531 and related health problems : tenth revision, 2nd ed. World Health Organization. 2004.  
532 <https://iris.who.int/handle/10665/42980>.
- 533 [20] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The  
534 PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst Rev*  
535 2021;10:89. <https://doi.org/10.1186/s13643-021-01626-4>.
- 536 [21] Canavan M, Wang X, Ascha M, Miksad R, Showalter TN, Calip G, et al. End-of-Life  
537 Systemic Oncologic Treatment in the Immunotherapy Era: The Role of Race, Insurance, and  
538 Practice Setting. *J Clin Oncol Off J Am Soc Clin Oncol* 2023;41:4729–38.  
539 <https://doi.org/10.1200/JCO.22.02180>.
- 540 [22] Kamran F, Tang S, Otles E, McEvoy DS, Saleh SN, Gong J, et al. Early identification  
541 of patients admitted to hospital for covid-19 at risk of clinical deterioration: model  
542 development and multisite external validation study. *BMJ* 2022;376:e068576.  
543 <https://doi.org/10.1136/bmj-2021-068576>.
- 544 [23] Witberg G, Barda N, Hoss S, Richter I, Wiessman M, Aviv Y, et al. Myocarditis after  
545 Covid-19 Vaccination in a Large Health Care Organization. *N Engl J Med* 2021;385:2132–9.  
546 <https://doi.org/10.1056/NEJMoa2110737>.
- 547 [24] Beck DB, Bodian DL, Shah V, Mirshahi UL, Kim J, Ding Y, et al. Estimated  
548 Prevalence and Clinical Manifestations of UBA1 Variants Associated With VEXAS Syndrome  
549 in a Clinical Population. *JAMA* 2023;329:318–24. <https://doi.org/10.1001/jama.2022.24836>.



- 550 [25] Forrest IS, Petrazzini BO, Duffy Á, Park JK, Marquez-Luna C, Jordan DM, et al.  
551 Machine learning-based marker for coronary artery disease: derivation and validation in two  
552 longitudinal cohorts. *Lancet Lond Engl* 2023;401:215–25. [https://doi.org/10.1016/S0140-](https://doi.org/10.1016/S0140-6736(22)02079-7)  
553 [6736\(22\)02079-7](https://doi.org/10.1016/S0140-6736(22)02079-7).
- 554 [26] Vasileiou E, Simpson CR, Shi T, Kerr S, Agrawal U, Akbari A, et al. Interim findings  
555 from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in  
556 Scotland: a national prospective cohort study. *Lancet Lond Engl* 2021;397:1646–57.  
557 [https://doi.org/10.1016/S0140-6736\(21\)00677-2](https://doi.org/10.1016/S0140-6736(21)00677-2).
- 558 [27] Haas EJ, Angulo FJ, McLaughlin JM, Anis E, Singer SR, Khan F, et al. Impact and  
559 effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19  
560 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an  
561 observational study using national surveillance data. *Lancet Lond Engl* 2021;397:1819–29.  
562 [https://doi.org/10.1016/S0140-6736\(21\)00947-8](https://doi.org/10.1016/S0140-6736(21)00947-8).
- 563 [28] Manz CR, Chen J, Liu M, Chivers C, Regli SH, Braun J, et al. Validation of a Machine  
564 Learning Algorithm to Predict 180-Day Mortality for Outpatients With Cancer. *JAMA Oncol*  
565 2020;6:1723–30. <https://doi.org/10.1001/jamaoncol.2020.4331>.
- 566 [29] Biccler JL, Glimelius I, Eloranta S, Smeland KB, Brown P de N, Jakobsen LH, et al.  
567 Relapse Risk and Loss of Lifetime After Modern Combined Modality Treatment of Young  
568 Patients With Hodgkin Lymphoma: A Nordic Lymphoma Epidemiology Group Study. *J Clin*  
569 *Oncol Off J Am Soc Clin Oncol* 2019;37:703–13. <https://doi.org/10.1200/JCO.18.01652>.
- 570 [30] Chang AR, Moore BS, Luo JZ, Sartori G, Fang B, Jacobs S, et al. Exome Sequencing  
571 of a Clinical Population for Autosomal Dominant Polycystic Kidney Disease. *JAMA*  
572 2022;328:2412–21. <https://doi.org/10.1001/jama.2022.22847>.
- 573 [31] Binder N, Blümle A, Balmford J, Motschall E, Oeller P, Schumacher M. Cohort  
574 studies were found to be frequently biased by missing disease information due to death. *J*  
575 *Clin Epidemiol* 2019;105:68–79. <https://doi.org/10.1016/j.jclinepi.2018.09.010>.
- 576 [32] Cohen-Stavi CJ, Magen O, Barda N, Yaron S, Peretz A, Netzer D, et al. BNT162b2  
577 Vaccine Effectiveness against Omicron in Children 5 to 11 Years of Age. *N Engl J Med*  
578 2022;387:227–36. <https://doi.org/10.1056/NEJMoa2205011>.
- 579 [33] Damrauer SM, Chaudhary K, Cho JH, Liang LW, Argulian E, Chan L, et al.  
580 Association of the V122I Hereditary Transthyretin Amyloidosis Genetic Variant With Heart  
581 Failure Among Individuals of African or Hispanic/Latino Ancestry. *JAMA* 2019;322:2191–202.  
582 <https://doi.org/10.1001/jama.2019.17935>.
- 583 [34] Filion KB, Lix LM, Yu OH, Dell’Aniello S, Douros A, Shah BR, et al. Sodium glucose  
584 cotransporter 2 inhibitors and risk of major adverse cardiovascular events: multi-database  
585 retrospective cohort study. *BMJ* 2020;370:m3342. <https://doi.org/10.1136/bmj.m3342>.
- 586 [35] Forrest IS, Chaudhary K, Vy HMT, Petrazzini BO, Bafna S, Jordan DM, et al.  
587 Population-Based Penetrance of Deleterious Clinical Variants. *JAMA* 2022;327:350–9.  
588 <https://doi.org/10.1001/jama.2021.23686>.
- 589 [36] Li X, Ostropolets A, Makadia R, Shoaibi A, Rao G, Sena AG, et al. Characterising the  
590 background incidence rates of adverse events of special interest for covid-19 vaccines in  
591 eight countries: multinational network cohort study. *BMJ* 2021;373:n1435.  
592 <https://doi.org/10.1136/bmj.n1435>.
- 593 [37] Lyu H, Zhao SS, Zhang L, Wei J, Li X, Li H, et al. Denosumab and incidence of type 2  
594 diabetes among adults with osteoporosis: population based cohort study. *BMJ*



- 595 2023;381:e073435. <https://doi.org/10.1136/bmj-2022-073435>.
- 596 [38] Martin P, Cohen JB, Wang M, Kumar A, Hill B, Villa D, et al. Treatment Outcomes  
597 and Roles of Transplantation and Maintenance Rituximab in Patients With Previously  
598 Untreated Mantle Cell Lymphoma: Results From Large Real-World Cohorts. *J Clin Oncol Off*  
599 *J Am Soc Clin Oncol* 2023;41:541–54. <https://doi.org/10.1200/JCO.21.02698>.
- 600 [39] Seymour CW, Kennedy JN, Wang S, Chang C-CH, Elliott CF, Xu Z, et al. Derivation,  
601 Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis.  
602 *JAMA* 2019;321:2003–17. <https://doi.org/10.1001/jama.2019.5791>.
- 603 [40] You SC, Rho Y, Bikdeli B, Kim J, Siapos A, Weaver J, et al. Association of Ticagrelor  
604 vs Clopidogrel With Net Adverse Clinical Events in Patients With Acute Coronary Syndrome  
605 Undergoing Percutaneous Coronary Intervention. *JAMA* 2020;324:1640–50.  
606 <https://doi.org/10.1001/jama.2020.16167>.
- 607 [41] Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al.  
608 Comprehensive comparative effectiveness and safety of first-line antihypertensive drug  
609 classes: a systematic, multinational, large-scale analysis. *Lancet Lond Engl* 2019;394:1816–  
610 26. [https://doi.org/10.1016/S0140-6736\(19\)32317-7](https://doi.org/10.1016/S0140-6736(19)32317-7).
- 611 [42] Vinogradova Y, Coupland C, Hill T, Hippisley-Cox J. Risks and benefits of direct oral  
612 anticoagulants versus warfarin in a real world setting: cohort study in primary care. *BMJ*  
613 2018;362:k2505. <https://doi.org/10.1136/bmj.k2505>.
- 614 [43] Chavez-MacGregor M, Lei X, Zhao H, Scheet P, Giordano SH. Evaluation of COVID-  
615 19 Mortality and Adverse Outcomes in US Patients With or Without Cancer. *JAMA Oncol*  
616 2022;8:69–78. <https://doi.org/10.1001/jamaoncol.2021.5148>.
- 617 [44] Berry ASF, Finucane BM, Myers SM, Abril A, Kirchner HL, Ledbetter DH, et al.  
618 Association of Supernumerary Sex Chromosome Aneuploidies With Venous  
619 Thromboembolism. *JAMA* 2023;329:235–43. <https://doi.org/10.1001/jama.2022.23897>.
- 620 [45] Deng Y, Polley EC, Wallach JD, Dhruva SS, Herrin J, Quinto K, et al. Emulating the  
621 GRADE trial using real world data: retrospective comparative effectiveness study. *BMJ*  
622 2022;379:e070717. <https://doi.org/10.1136/bmj-2022-070717>.
- 623 [46] Chemaitelly H, AlMukdad S, Ayoub HH, Altarawneh HN, Coyle P, Tang P, et al.  
624 Covid-19 Vaccine Protection among Children and Adolescents in Qatar. *N Engl J Med*  
625 2022;387:1865–76. <https://doi.org/10.1056/NEJMoa2210058>.
- 626 [47] Harstad E, Shults J, Barbaresi W, Bax A, Cacia J, Deavenport-Saman A, et al.  $\alpha$ 2-  
627 Adrenergic Agonists or Stimulants for Preschool-Age Children With Attention-  
628 Deficit/Hyperactivity Disorder. *JAMA* 2021;325:2067–75.  
629 <https://doi.org/10.1001/jama.2021.6118>.
- 630 [48] Chuard PJC, Vrtílek M, Head ML, Jennions MD. Evidence that nonsignificant results  
631 are sometimes preferred: Reverse P-hacking or selective reporting? *PLOS Biol*  
632 2019;17:e3000127. <https://doi.org/10.1371/journal.pbio.3000127>.
- 633 [49] Kim NH, Han KH, Choi J, Lee J, Kim SG. Use of fenofibrate on cardiovascular  
634 outcomes in statin users with metabolic syndrome: propensity matched cohort study. *BMJ*  
635 2019;366:l5125. <https://doi.org/10.1136/bmj.l5125>.
- 636 [50] Xie Y, Bowe B, Al-Aly Z. Molnupiravir and risk of hospital admission or death in adults  
637 with covid-19: emulation of a randomized target trial using electronic health records. *BMJ*  
638 2023;380:e072705. <https://doi.org/10.1136/bmj-2022-072705>.

- 639 [51] Mahévas M, Tran V-T, Roumier M, Chabrol A, Paule R, Guillaud C, et al. Clinical  
640 efficacy of hydroxychloroquine in patients with covid-19 pneumonia who require oxygen:  
641 observational comparative study using routine care data. *BMJ* 2020;369:m1844.  
642 <https://doi.org/10.1136/bmj.m1844>.
- 643 [52] Rentsch CT, Beckman JA, Tomlinson L, Gellad WF, Alcorn C, Kidwai-Khan F, et al.  
644 Early initiation of prophylactic anticoagulation for prevention of coronavirus disease 2019  
645 mortality in patients admitted to hospital in the United States: cohort study. *BMJ*  
646 2021;372:n311. <https://doi.org/10.1136/bmj.n311>.
- 647 [53] Andersson NW, Thiesson EM, Baum U, Pihlström N, Starrfelt J, Faksová K, et al.  
648 Comparative effectiveness of bivalent BA.4-5 and BA.1 mRNA booster vaccines among  
649 adults aged ≥50 years in Nordic countries: nationwide cohort study. *BMJ* 2023;382:e075286.  
650 <https://doi.org/10.1136/bmj-2022-075286>.
- 651 [54] Deputy NP, Deckert J, Chard AN, Sandberg N, Moulia DL, Barkley E, et al. Vaccine  
652 Effectiveness of JYNNEOS against Mpox Disease in the United States. *N Engl J Med*  
653 2023;388:2434–43. <https://doi.org/10.1056/NEJMoa2215201>.
- 654 [55] Xie Y, Bowe B, Al-Aly Z. Nirmatrelvir and risk of hospital admission or death in adults  
655 with covid-19: emulation of a randomized target trial using electronic health records. *BMJ*  
656 2023;381:e073312. <https://doi.org/10.1136/bmj-2022-073312>.
- 657 [56] Zheng B, Green ACA, Tazare J, Curtis HJ, Fisher L, Nab L, et al. Comparative  
658 effectiveness of sotrovimab and molnupiravir for prevention of severe covid-19 outcomes in  
659 patients in the community: observational cohort study with the OpenSAFELY platform. *BMJ*  
660 2022;379:e071932. <https://doi.org/10.1136/bmj-2022-071932>.
- 661 [57] Wong CKH, Au ICH, Lau KTK, Lau EHY, Cowling BJ, Leung GM. Real-world  
662 effectiveness of molnupiravir and nirmatrelvir plus ritonavir against mortality, hospitalisation,  
663 and in-hospital outcomes among community-dwelling, ambulatory patients with confirmed  
664 SARS-CoV-2 infection during the omicron wave in Hong Kong: an observational study.  
665 *Lancet Lond Engl* 2022;400:1213–22. [https://doi.org/10.1016/S0140-6736\(22\)01586-0](https://doi.org/10.1016/S0140-6736(22)01586-0).
- 666 [58] Marafino BJ, Escobar GJ, Baiocchi MT, Liu VX, Plimier CC, Schuler A. Evaluation of  
667 an intervention targeted with predictive analytics to prevent readmissions in an integrated  
668 health system: observational study. *BMJ* 2021;374:n1747. <https://doi.org/10.1136/bmj.n1747>.
- 669 [59] Lang C, Gottschall M, Sauer M, Köberlein-Neu J, Bergmann A, Voigt K. „Da kann  
670 man sich ja totklingeln, geht ja keiner ran“ – Schnittstellenprobleme zwischen stationärer,  
671 hausärztlicher und ambulant-fachspezialisierter Patientenversorgung aus Sicht Dresdner  
672 Hausärzte. *Gesundheitswesen* 2019;81:822–30. <https://doi.org/10.1055/a-0664-0470>.
- 673 [60] D’Arcy M, Stürmer T, Lund JL. The importance and implications of comparator  
674 selection in pharmacoepidemiologic research. *Curr Epidemiol Rep* 2018;5:272–83.  
675 <https://doi.org/10.1007/s40471-018-0155-y>.
- 676 [61] Luijken K, van Eekelen R, Gardarsdottir H, Groenwold RHH, van Geloven N. Tell me  
677 what you want, what you really really want: Estimands in observational  
678 pharmacoepidemiologic comparative effectiveness and safety studies. *Pharmacoepidemiol*  
679 *Drug Saf* 2023;32:863–72. <https://doi.org/10.1002/pds.5620>.
- 680 [62] European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity  
681 analysis in clinical trials to the guideline on statistical principles for clinical trials 2020.
- 682 [63] Nguyen VT, Engleton M, Davison M, Ravaud P, Porcher R, Boutron I. Risk of bias in  
683 observational studies using routinely collected data of comparative effectiveness research: a

684 meta-research study. BMC Med 2021;19:279. <https://doi.org/10.1186/s12916-021-02151-w>.

685 [64] Coltin H, Pequeno P, Liu N, Tsang DS, Gupta S, Taylor MD, et al. The Burden of  
686 Surviving Childhood Medulloblastoma: A Population-Based, Matched Cohort Study in  
687 Ontario, Canada. J Clin Oncol Off J Am Soc Clin Oncol 2023;41:2372–81.  
688 <https://doi.org/10.1200/JCO.22.02466>.

689 [65] Song Q, Bates B, Shao YR, Hsu F-C, Liu F, Madhira V, et al. Risk and Outcome of  
690 Breakthrough COVID-19 Infections in Vaccinated Patients With Cancer: Real-World  
691 Evidence From the National COVID Cohort Collaborative. J Clin Oncol Off J Am Soc Clin  
692 Oncol 2022;40:1414–27. <https://doi.org/10.1200/JCO.21.02419>.

693

## 694 Supplementary Material

### 695 Supplementary S1

696 *Table 1 Search strategy*

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

697

698 *Table 2 Detailed Search History PubMed on October 31st 2023*

Search	Query	Results
#1	<b>Search:</b> ("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	<b>Search:</b> "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	<b>Search:</b> #1 AND #2	387
#4	<b>Search:</b> #1 AND #2 <b>Filters:</b> from 2018-2023	227

699

## 700 [Supplementary S2](#)

701 Table 3 provides a comprehensive overview of the extracted studies, including their  
702 characteristics and limitations described in their respective discussion section. Note, that for  
703 some publications (indicated with an asterisk) the category could not be clearly assigned,  
704 especially in the case of *Treatment Comparison* and *Safety and Risk Group Analysis*. These  
705 publications did not only compare the effectiveness of treatments but specifically compared  
706 the risk of adverse events between two or more treatments [34,37,40–42,45,47].

707 All studies categorized into the Burden of Disease scenario were cohort studies except for  
708 three studies [22,25,28] that were developing or validating a machine learning model to predict  
709 certain endpoints using EHR data. Three studies [26,29,64] examined time-to-event outcomes  
710 using Cox regression as well as Kaplan-Meier estimator partly with inverse propensity weights.  
711 Prevalence and incidences were analyzed in four studies [23,24,27,30]. The studies that were  
712 assigned to the category Safety and Risk Group Analysis were mostly cohort studies, in  
713 addition to two case-control studies and one cross-sectional cohort study [32,33]. Similar to  
714 the category Burden of Disease, predominantly time-to-event outcomes were evaluated using  
715 Kaplan-Meier estimation or Cox (proportional hazards) regression models [34,37,38,40–  
716 42,44–46]. In contrast, many studies have used propensity score methods such as matching  
717 and weighting [32,34,37,40–42,45,46] or applied the concept of target trial emulation  
718 [32,37,45]. Similar to the category Safety and Risk Group Analysis, most studies in the  
719 Treatment Comparison scenario examined time-to-event outcomes. They employed  
720 propensity score methods such as matching and weighting or applied the concept of target  
721 trial emulation. In addition to that, two studies used the clone method, a special form of  
722 emulating a target trial where each individual is cloned into both treatment groups, censored  
723 patients according to their designated treatment strategy and weighted the uncensored to  
724 avoid selection bias [9,50,55].

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 trial 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 trial 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 Netherlands			x	x		x			x	



						Ireland Spain UK US									
Lyu et al. [37]*	cohort study	comparative effectiveness	incidence	target trial emulation, PS-matching, Cox regression	nation-wide	UK	x		x		x				
Martin et al. [38]	cohort study	therapy evaluation	time-to-next-treatment, overall survival	KM, Cox regression	nation-wide	USA	x		x			x			x
Seymour et al. [39]	cohort study	phenotype derivation	phenotype frequency, mortality	Multiple imputation, k-means clustering	single-center	USA		x	x			x			
Song et al. [65]	cohort study	comparative effectiveness	disease risk	adj. logistic regression	nation-wide	USA	x			x					
Suchard et al. [41]*	cohort study	comparative effectiveness	relative risk	PS model, matching, Cox regression, meta analysis	multi-national	USA South Korea Japan Germany	x	x	x	x					
Vinogradova et al. [42]*	cohort study	treatment association	hospital admission or death	PS, Cox regression	nation-wide	UK	x		x	x					x
You et al. [40]*	cohort study	treatment association	net adverse clinical events	PS-matching, Cox regression	multi-center	South Korea USA	x		x	x				x	x

Treatment Comparison							Confoundin g	Selec tion bias	Inform ation bias	Reporti ng bias	Follow- up Challen ges	Missing Data	Coding Challen ges	Validati on & Data quality	Operatio nization or availabilit y of variables
Andersson et al. [53]	cohort study	comparative effectiveness	hospital admission , death	target trail emulation, PS- matching, inverse probability of treatment weights, KM	multi- national	Nordi c count ries	x	x	x						
Deputy et al. [54]	case- control study	vaccine effectiveness	associatio n between vaccinatio n status and case patient or control patient status	matching, conditional logistic regression	natio n- wide	USA	x	x	x						
Kim et al. [49]	cohort study	treatment effectiveness	time to cardiovas cular events	PS-matching, stratified Cox regression	natio n- 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- cente r	Franc e	x	x							x
Marafino et al. [58]	cohort study	intervention association	mortality, readmissi on	target-trail emulation, generalized linear mixed effects	multi- cente r	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	natio n-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	natio n-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	territo ry-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	natio n-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	natio n-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

726

727 Abbreviation: KM... Kaplan-Meier, PS... Propensity Score, EOL... End-of-Life, ICU... Intensive Care Unit, ML... Machine Learning, AUROC...

728 Area under the receiver operating curve, AUPRC... Area under the precision-recall curve

729 \* This indicates publications that were not uniquely categorizable into the main scenarios. Specifically, they did not only compare the effectiveness

730 of treatments but compared the risk of adverse events between two or more treatments