Can serious adverse event rates be used as a metric of trial representativeness for pharmacological interventions? Glucagon-like peptide-1 receptor agonist trials as an exemplar

Saleh Ali M Almazam

Bsc. Msc.

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

School of Health and Wellbeing

College of Medical, Veterinary and Life Sciences

University of Glasgow

January 2025

Acknowledgement

First and foremost, all praise is due to God for giving me the strength and determination to complete my thesis. I express my deepest gratitude to my supervisors, Professor David McAllister, Professor James Lewsey, Dr Claire Niedzwiedz and Dr Peter Hanlon, whose expertise, understanding, and patience considerably added to my PhD experience. Your guidance helped me navigate my research challenges, and your encouragement made this achievement possible. I am deeply honoured to be one of your students. Special thanks to the University of Glasgow, Umm Al Qura University, and the Saudi Arabia Cultural Bureau in London for providing the necessary resources and valuable assistance in my journey. I must acknowledge my colleagues for all the stimulating discussions that inspired me; your support and friendship made this journey enjoyable and memorable. Finally, I want to express my genuine gratitude to my family: my mother, my wife, my brothers, and my sisters, who supported and encouraged me with their kind words and prayers. This thesis is dedicated to the memory of my father and his enduring influence on my life.

Author’s Declaration

I declare that I wrote this thesis entirely by myself. This thesis has not been submitted in part or in whole for a degree or any other qualification. This work is all my own unless otherwise acknowledged or referenced in this thesis.

Abstract

Background: Randomised controlled trials (RCTs) are considered the gold standard for assessing the efficacy and safety of interventions because they introduce the concept of randomisation, which effectively deals with unmeasured confounders. Trial representativeness is crucial for determining the generalisability of trial findings to the target population. However, various groups, such as older people and those with multiple long-term conditions, are often under-represented in trials. Moreover, there is no gold standard measure of trial representativeness. Current measures are complex, subjective and time-consuming, suggesting the need for new and better measures of trial representativeness. Serious adverse events (SAEs) are likely to be reported in trials and routine care. Moreover, they are reasonably objective, tangible and predictable trial outcomes. Consequently, the SAE rate has been proposed as a potential metric of trial representativeness. However, little is known about its feasibility and validity in measuring trial representativeness. This thesis investigates whether the SAE rate can be used to measure trial representativeness, using glucagon-like peptide-1 receptor agonists (GLP-1 RA) trials as an exemplar.

Methodology: This thesis builds on an ongoing systematic review of novel antidiabetics. GLP-1 RA trials were used as an exemplar for the analysis. Data sources used in this thesis were ClinicalTrials.gov, clinical study reports (CSRs), study protocols, and journal publications. I conducted several approaches to examine whether the SAE rate can be used as a proxy for trial representativeness. First, I examined SAE reporting in RCTs in the published literature to explore the feasibility of using the SAE rate as a metric of trial representativeness. Second, I explored SAE capturing in GLP-1 RA trials to enable the calculation of SAE rates. Third, I compared SAE rates between intervention and control arms to determine if combining SAE rates of trial arms is feasible to increase statistical precision. Fourth, I assessed GLP-1 RA trials using the PRECIS-2 (PRagmatic Explanatory Continuum Indicator Summary) tool to enable fair comparison with the SAE rate and explore the challenges of using this tool. I protocolised and operationalised the PRECIS-2 tool to score GLP-1 RA trials as objectively as possible. Fifth, I compared the SAE rate with the PRECIS-2 tool based on the differences in their associations with several markers of trial representativeness that serve as fair umpires to examine the validity of the SAE rate as a metric of trial representativeness. Finally, I examined the association between

eligibility criteria and the SAE rate to examine whether eligibility criteria, a possible driver of trial representativeness, is associated with the SAE rate.

Results: Firstly, I found that SAEs were reported for most GLP-1 RA trials. However, SAE timeframes were not explicitly reported for nearly half of the trials. Major adverse cardiovascular event (MACE) trials had inconsistent reporting of MACE counts in SAE total counts. Secondly, I found no difference in SAE rates between intervention and control arms. Furthermore, the retrospective assessment of GLP-1 RA trials using the PRECIS-2 tool was challenging and time-consuming. The missingness of information required to score recruitment and organisation domains was high. I found no correlations between the domains of the PRECIS-2 score, except for modest correlations between eligibility criteria and recruitment domains and between setting and primary outcome domains. Moreover, I found no association between the SAE rate and the PRECIS-2 score. Additionally, I found that all fair umpires did not strongly favour the SAE rate over the PRECIS-2 tool. However, the direction of the difference in associations for half of the umpires favoured the SAE rate. Finally, I found that trials with permissive eligibility criteria were positively associated with higher SAE rates. Trials with increased continuous eligibility criteria scores were also positively associated with increased SAE rates.

Conclusion: Most GLP-1 RA trials reported sufficient SAE data indicating the ability to calculate SAE rates. SAE rates were similar across trial arms, suggesting the feasibility of combining SAE rates from the intervention and control arms. The SAE rate was not associated with the PRECIS-2 score, and none of the fair umpires strongly favoured the SAE rate. However, the directions of half of the fair umpires favoured the SAE rate and trials with permissive eligibility criteria were associated with increased SAE rates. Therefore, the SAE rate may be useful as a quick-to-measure proxy of the restrictiveness of eligibility criteria. However, given that the triangulation of evidence is inconsistent in supporting the use of the SAE rate as a standalone metric of trial representativeness, examination of the SAE rate would need to be considered carefully in combination with other metrics before making overall judgements about trial representativeness.

Table of contents

Acknowledgement ..... **II**

Author's Declaration ..... **III**

Abstract ..... **IV**

Table of contents ..... **VI**

List of tables ..... **XIII**

List of figures ..... **XV**

List of abbreviations ..... **XVI**

Chapter 1 Introduction **1**

1.1 Chapter overview 1

1.2 Randomised controlled trials 1

1.3 Trial representativeness 1

1.4 Current measures of trial representativeness 2

1.5 The rationale for this thesis 4

1.5.1 Using serious adverse events as a metric of trial representativeness 4

1.5.2 Choosing T2DM and GLP-1 RAs as an exemplar in this thesis 6

1.6 Thesis research question and approach to analysis 7

1.7 Aims and objectives of the thesis 8

1.8 Outline of the thesis 8

Chapter 2 The feasibility of using SAE rates as a marker of trial representativeness, a literature review of SAE reporting in RCTs **10**

2.1 Chapter overview 10

2.2 Background 10

2.2.1 SAEs in RCTs 10

2.2.2 SAE reporting in RCTs 11
2.3 Aims and objectives 12
2.4 Methods 12
2.5 Literature review 14
2.5.1 SAE reporting in RCTs 14
2.5.2 Harm reporting practice in journal publications of RCTs 25
2.5.3 Methods of SAE reporting in RCTs 31
2.6 Discussion of literature findings 35
2.7 Conclusion 37
Chapter 3 Capturing SAEs in GLP-1 RA trials 38
3.1 Chapter overview 38
3.2 Background 38
3.2.1 SAE reporting in RCTs 38
3.2.2 Other factors associated with the reported SAE rates 39
3.3 Aim and objectives 40
3.4 Methods 40
3.4.1 Study selection 40
3.4.2 Data sources 42
3.4.3 SAEs data extraction and harmonisation 44
3.4.4 Statistical analysis 45
3.4.5 Data management 47
3.5 Results 47
3.5.1 Summary of trials 47
3.5.2 SAE reporting in GLP-1 RA trials 48
3.5.3 Factors that associate with SAE reporting 50
3.6 Discussion 52
3.6.1 Summary of findings 52
3.6.2 Interpretation 52

3.6.3 Comparison with previous literature 53
3.6.4 Strengths and limitations 54
3.6.5 Implications 55
3.7 Conclusion 55

Chapter 4 Comparison of SAE rates between trial arms: GLP-1 RA trials as an exemplar 56

4.1 Chapter overview 56
4.2 Background 56
4.2.1 The difference in reported SAE rates between trial arms 56
4.3 Aim and objectives 57
4.4 Methods 57
4.4.1 Study selection and data extraction 57
4.4.2 Data sources 57
4.4.3 Statistical analysis 58
4.5 Results 58
4.5.1 Summary of trials 58
4.5.2 SAE reporting in GLP-1 RA trials 59
4.5.3 SAE rates across GLP-1 RA trial arms 59
4.6 Discussion 60
4.6.1 Summary of findings 60
4.6.2 Interpretation 61
4.6.3 Comparison with previous literature 61
4.6.4 Strengths and limitations 62
4.6.5 Implications 63
4.7 Conclusion 63

Chapter 5 The pragmatism GLP-1 RA trials: A retrospective analysis using the PRECIS-2 tool 64

5.1 Chapter overview 64

5.2 Background 64
5.2.1 The PRECIS-2 tool 64
5.2.2 The suitability of the PRECIS-2 tool as a comparator with the SAE rate 65
5.3 Aim and objectives 65
5.4 Methods 66
5.4.1 Study selection 66
5.4.2 Data sources 66
5.4.3 Protocolisation and operationalisation of the PRECIS-2 tool for GLP-1 RA trials 66
5.4.4 PRECIS-2 score data extraction and harmonisation 78
5.4.5 Statistical analysis 78
5.4.6 Missing data 79
5.5 Results 79
5.5.1 Retrospective assessment of GLP-1 RAs using the PRECIS-2 tool 79
5.5.2 Correlation between the domains of the PRECIS-2 tool 83
5.6 Discussion 84
5.6.1 Summary of findings 84
5.6.2 Interpretation 84
5.6.3 Comparison with previous literature 85
5.6.4 Strengths and limitations 86
5.6.5 Implications 87
5.7 Conclusion 87
Chapter 6 Comparison between the PRECIS-2 tool and the SAE rate as measures of trial representativeness: GLP-1 RA trials as an exemplar 88
6.1 Chapter overview 88
6.2 Background 88
6.2.1 The lack of a gold standard measure of trial representativeness 88

6.2.2 The need for fair umpire tests 89
6.2.3 Markers of trial representativeness 91
6.2.4 Prior expectations 91
6.3 Aims and objectives 94
6.4 Methods 94
6.4.1 Study selection and SAE data extraction 94
6.4.2 Data sources 94
6.4.3 Baseline and trial characteristics data extraction and harmonisation 95
6.4.4 Statistical analysis 95
6.4.5 Missing data 97
6.5 Results 97
6.5.1 Descriptive statistics 97
6.5.2 SAE reporting in GLP-1 RA trials 99
6.5.3 The association between the SAE rate and the PRECIS-2 score 99
6.5.4 The association between the fair umpires with the normalised SAE rate and the normalised mean PRECIS-2 score 100
6.6 Discussion 103
6.6.1 Summary of findings 103
6.6.2 Interpretation 104
6.6.3 Comparison with previous literature 108
6.6.4 Strengths and limitations 109
6.6.5 Implications 109
6.7 Conclusion 110
Chapter 7 The association between eligibility criteria and the SAE rate in GLP-1 RA trials 111
7.1 Chapter overview 111
7.2 Background 111
7.2.1 Eligibility criteria, a driver of trial representativeness 111

7.2.2 The association between eligibility criteria and the SAE rate 111
7.3 Aim and Objectives 112
7.4 Methods 113
7.4.1 Study selection and data extraction 113
7.4.2 Data sources 113
7.4.3 Eligibility criteria data extraction and harmonisation 113
7.4.4 Protocolisation and assessment of the restrictiveness of the eligibility criteria 119
7.4.5 Statistical analysis 126
7.5 Results 127
7.5.1 SAE reporting in GLP-1 RA trials 127
7.5.2 Descriptive statistics 127
7.5.3 The association between eligibility criteria and the SAE rate 131
7.6 Discussion 133
7.6.1 Summary of findings 133
7.6.2 Interpretation 134
7.6.3 Comparison with previous literature 134
7.6.4 Strengths and limitations 135
7.6.5 Implications 135
7.7 Conclusion 136
Chapter 8 Discussion 137
8.1 Review of background and main findings 137
8.1.1 Knowledge gap 137
8.1.2 Approach 138
8.1.3 Main findings and interpretation 138
8.1.4 The feasibility of using the SAE rate as a metric of trial representativeness 138

8.1.5 Triangulation of evidence to determine whether the SAE rate can reflect trial representativeness 139
8.2 Contribution to the literature 141
8.3 Implications 142
8.4 Strengths and limitations 143
8.5 Future research 144
8.6 Conclusion 145
References 146
Appendices 169
Appendix A: Details of literature review on AE reporting in RCTs 169
SAE reporting in RCTs 169
Harm reporting in journal publications 175
Methods of SAE reporting in RCTs 180
Appendix B: The PRECIS-2 rationale extraction template 183
Appendix C: Sensitivity analysis for the correlation between the PRECIS-2 domains using MICE imputation 184
Appendix D: Sensitivity analysis for the association between the normalised SAE rate and the PRECIS-2 domains using MICE imputation 185
Appendix E: Complete case analysis of the estimated differences between the SAE rate and the mean PRECIS-2 score 186

List of tables

Table 2.1 Search strategy of the literature review 13

Table 2.2 SAE reporting in RCTs..... 15

Table 2.3 Harm reporting practice in journal publications of RCTs 26

Table 2.4 Methods of SAE reporting in RCTs 32

Table 3.1 Eligibility criteria of the current systematic review 41

Table 3.2 Summary of SAE count reporting for GLP-1 RA trials 49

Table 3.3 Summary of SAE timeframe reporting for GLP-1 RA trials 49

Table 3.4 Reporting of MACE outcomes count in total reported SAE count..... 49

Table 4.1 Summary of the reported SAE data for GLP-1 RA trials 59

Table 5.1 The operationalisation of the PRECIS-2 for GLP-1 RA trials 68

Table 5.2 Summary of mean scores for domains of the PRECIS-2 tool..... 80

Table 6.1 The selection and eligibility of fair umpire tests 90

Table 6.2 The prior expected associations of fair umpire tests with the SAE rate and the mean PRECIS-2 score..... 93

Table 6.3 Summary of continuous markers of trial representativeness 98

Table 6.4 Summary of categorical markers of trial representativeness 98

Table 6.5 The association between the normalised SAE rate and PRECIS-2 domains 99

Table 6.6 The estimated differences between the SAE rate and the mean PRECIS-2 score 101

Table 6.7 The actual associations of fair umpire tests with the SAE rate and the mean PRECIS-2 score..... 106

Table 7.1 The protocol for the extraction of eligibility criteria..... 115

Table 7.2 Binary restrictiveness score based on highly restrictive cutoffs or levels 120
Table 7.3 Scores for each variable based on specific cutoffs 122
Table 7.4 Descriptive statistics of continuous criteria 128
Table 7.5 The distributions of categorical criteria of GLP-1 RA trials 129
Table 7.6 The association between highly restrictive eligibility criteria and the SAE rate 132
Table 7.7 The association between continuous eligibility criteria score and the SAE rate 132
Table 7.8 The associations between specific highly restrictive criteria and the SAE rate 133

List of figures

Figure 2.1 PRISMA flow diagram of the literature selection process..... 13

Figure 3.1 The distribution of SAE rates of GLP-1 RA trials..... 48

Figure 3.2 The association between trial characteristics and the SAE rate in GLP-1 RA trials 51

Figure 4.1 The difference in SAE rates between intervention and control arms.... 60

Figure 5.1 The PRECIS-2 mean score for GLP-1 RA trials..... 81

Figure 5.2 The distribution of scores for each domain of the PRECIS-2 tool 82

Figure 5.3 The PRECIS-2 wheel showing the mean score for each of the nine domains of the PRECIS-2 tool (1 is very explanatory, 5 is very pragmatic) 82

Figure 5.4 The correlation between the domains of the PRECIS-2 tool..... 83

Figure 7.1 Example of scoring eligibility criteria using the binary restrictiveness score 121

Figure 7.2 The distributions of continuous eligibility criteria of GLP-1 RA trials... 130

Figure 7.3 The association between eligibility criteria and the SAE rate..... 132

List of abbreviations

Abbreviation	Definition
AAA	Abdominal Aortic Aneurysm
AACT	Aggregate Analysis of ClinicalTrials.gov
ACTTION	Analgesic, Anaesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks
AERQS	The Adverse Event Reporting Quality Score
AEs	Adverse Events
Anti-CCP	Anti-cyclic Citrullinated Peptide
AUC	Area Under the Curve
BMI	Body Mass Index
BNF	British National Formulary
CDUS	Clinical Data Update System
CHD	Coronary Heart Disease
ChiCTR	Chinese Clinical Trial Register
CKD	Chronic Kidney Disease
CONSORT	Consolidated Standards of Reporting Trials
CrCl	Creatinine Clearance
CRT	cardiac resynchronisation therapy
CSRs	Clinical Study Reports
CV	Cardiovascular
DBP	Diastolic Blood Pressure
DMARD	Disease-modifying Antirheumatic Drug
DPP-4i	Dipeptidyl Peptidase-4 inhibitors
eGFR	estimated Glomerular Filtration Rate
EMA	European Medicines Agency
EU-CTR	EU Clinical Trials Register
EXSCEL	Exenatide Study of Cardiovascular Event Lowering
FBG	Fasting Blood Glucose
FDA	The United States Food and Drug Administration
GIST 2.0	Generalisability Index of Study Traits 2.0
GLP-1	Glucagon-Like Peptide-1

GLP-1 RA	Glucagon-Like Peptide-1 Receptor Agonist
HbA1C	Glycated Haemoglobin
HR	Hazard Ratio
ICMJE	International Committee of Medical Journal Editors
IRB	Institutional Review Board
IRR	Incidence Rate Ratio
ITT	Intent-to-Treat
MACE	Major Adverse Cardiovascular Events
MAR	Missing at Random
MCAR	Missing Completely at Random
MI	Myocardial Infarction
MICE	Multivariate Imputation by Chained Equations
MTC	Medullary Thyroid Cancer
NCI	The National Cancer Institute
NIH	National Institutes of Health
NLM	National Library of Medicine
NSAIDs	Non-steroidal anti-inflammatory drugs
NYHA	New York Heart Association
PP	Per-Protocol
PRECIS	PRagmatic Explanatory Continuum Indicator Summary
RAAS	Renin-Angiotensin-Aldosterone System
RCTs	Randomised Controlled Trials
RF	Rheumatoid Factor
ROC	Receiver Operating Characteristic
SAEs	Serious Adverse Events
SBP	Systolic Blood Pressure
SD	Standard Deviation
SE	Standard Error
SGLT-2i	Sodium-Glucose Co-Transporter-2 inhibitors
SORT	Standards for Reporting in Trials
T2DM	Type 2 Diabetes Mellitus
TMDs	Temporomandibular disorders
UMIN	University Hospital Medical Information Network

Chapter 1 Introduction

1.1 Chapter overview

This chapter provides some background about randomised controlled trials (RCTs), their representativeness and current measures of trial representativeness. Furthermore, it discusses serious adverse events (SAEs) in RCTs, focusing on how they can serve as a key metric for assessing trial representativeness. Additionally, it provides the rationale for choosing type 2 diabetes mellitus (T2DM) and glucagon-like peptide-1 receptor agonist (GLP-1 RA) trials as an exemplar.

1.2 Randomised controlled trials

RCTs have played an essential role in changing the view of evidence-based medicine and healthcare practice. The RCT is a study design that randomly assigns participants to intervention or control groups (Kendall, 2003). RCTs are considered the gold standard for assessing the efficacy and safety of interventions, as they introduce the concept of randomisation, which almost uniquely deals with unmeasured confounders and reduces the chances of biased assessments of interventions (Kabisch et al., 2011; Misra, 2012; Bhide, Shah and Acharya, 2018). Moreover, they incorporate blinding and control groups in their designs, which can isolate the effect of interventions from confounding factors, providing firm conclusions on the efficacy and safety of interventions (Malay and Chung, 2012; Spieth et al., 2016). Additionally, RCTs have clearly defined interventions and specified outcomes to ensure the internal validity and reliability of their results (Sørensen, Lash and Rothman, 2006).

1.3 Trial representativeness

Trial representativeness is crucial for determining the generalisability of trial outcomes (Braslow et al., 2005). Trial representativeness refers to the extent to which the characteristics of the trial participants reflect those of the people in the general population

who may be considered candidates for treatment (target population) (Qi et al., 2021). The greater the discrepancy between the trial participants and the people in the general population, the lower the likelihood that trial outcomes can be generalised (Kennedy-Martin et al., 2015; Usman et al., 2022). Achieving a representative trial population is often challenging due to the stringent eligibility criteria of RCTs, where vulnerable patients such as older people or those with multiple long-term conditions are frequently excluded (Clark et al., 2019). Moreover, trials that claim to be pragmatic and have broad eligibility criteria often still do not enrol older people and those with comorbidities (Sedrak et al., 2021). The exclusion of older adults and people with comorbidities is highly problematic because they represent a significant portion of patients in routine care (Herrera et al., 2010). It may result in an incomplete understanding of intervention effects, leading to unexpected adverse events (AEs) or suboptimal efficacy in routine care (Nijsten et al., 2009; Garcia-Doval et al., 2012). Therefore, recruiting patients typically seen in routine care may enhance the generalisability of trial findings.

1.4 Current measures of trial representativeness

The measurement of trial representativeness is crucial to improve the quality of decision-making in clinical practice (Malmivaara, 2021). However, assessing the representativeness of RCTs poses significant challenges, given the variability in study designs, eligibility criteria and patient characteristics (Rothwell, 2005). These factors, alongside data availability, have constrained the assessment of trial representativeness (Cahan, Cahan and Cimino, 2017). Various approaches have been used to measure trial representativeness, including score-based assessments. For example, the PRECIS (PRagmatic Explanatory Continuum Indicator Summary) tool provides a scoring system and a visual summary to assess the pragmatism of trials (Thorpe et al., 2009). It was developed to aid researchers in matching trial designs with their aims, whether pragmatic or explanatory (Loudon et al., 2013). Moreover, it may indirectly reflect trial representativeness. However, the lack of sufficient information to assess trials and subjective assessment are limitations of the retrospective assessment using the PRECIS tool (Dal-Ré, 2020).

Other approaches include quantifying the effect of eligibility criteria on trial representativeness. These methods aim to identify who would be excluded from the target

population based on the trial eligibility criteria (He et al., 2020). Although so-called pragmatic trials may have permissive eligibility criteria, many eligible subjects may not enrol in these trials (Howard et al., 2009; Oude Rengerink et al., 2017; Clapp et al., 2023). Healthcare practitioners may choose not to refer eligible patients to trials due to safety concerns or treatment preferences (Go et al., 2006). Moreover, patients may refuse to participate in trials because they want to choose their treatment or have concerns about the safety of these trials (Kemeny et al., 2003; Nipp, Hong and Paskett, 2019). Additionally, these methods are either exclusively applied to a single trial or not generalisable to other target populations because trial designs and contexts are unique, and the target populations differ (Stuart, Ackerman and Westreich, 2018; He et al., 2020). Another limitation of these approaches is the lack of high-quality benchmarking real-world data (Dagenais et al., 2022). Therefore, recognising these factors alongside the impact of eligibility criteria is crucial for the validity of these approaches.

Another approach for assessing trial representativeness involves comparing key patient characteristics such as age, sex, comorbidities and disease severity between the trial participants and the general population (Bernhardt et al., 2015; He et al., 2020). Large differences between both populations indicate limited trial representativeness (Stuart, Bradshaw and Leaf, 2015). However, this approach has potential weaknesses. Patient characteristics often do not show significant differences due to other unmeasured characteristics, which may lead to misleading conclusions about trial representativeness (Kurki et al., 2024). Moreover, the lack of robust real-world data that includes various patient characteristics makes comparing the trial and target populations challenging (Tan et al., 2022). Therefore, there is growing interest in exploring new and better methods to assess trial representativeness (Hanlon et al., 2021, 2022; Qi et al., 2021; Sun et al., 2021).

1.5 The rationale for this thesis

1.5.1 Using serious adverse events as a metric of trial representativeness

SAE refers to the occurrence of any unwanted health condition in human subjects while using a trial intervention that leads to hospitalisation, congenital defects, disability, a health condition that requires intervention to prevent permanent damage, or death, irrespective of the cause (FDA., 2023). The recording, documentation and reporting of SAEs are crucial to ensure the safety of participants and the transparency of trial outcomes (James et al., 2020). Beyond their importance in maintaining trial safety, SAEs may be related to trial representativeness. Trials often recruit healthier and younger patients, who are less likely to experience SAEs than those in the target population (Garcia-Doval et al., 2012). Hanlon et al. (2021, 2022) proposed using the SAE rate as a potential metric for assessing trial representativeness. They found that the observed SAE rates in trials were significantly lower than expected SAE rates (based on hospitalisations and deaths) in people who may be candidates for treatment within routine care, indicating issues with the representativeness of the trials (Hanlon et al., 2021). Therefore, SAEs in trials may reflect trial representativeness as 'healthier' trial populations in less representative trials are likely to have lower rates of SAEs.

Examining the use of the SAE rate as a metric of trial representativeness is valuable for the following reasons:

1) Any SAE must be reported, regardless of the cause, including in the placebo arms (FDA., 2023). Consequently, SAEs are likely to be reported, ensuring their reliability for assessing trial representativeness.

2) The SAE rate provides a reasonably objective, quantifiable and tangible measure, allowing researchers to assess trial representativeness more precisely. SAEs are numeric health-related outcomes recorded based on clearly defined criteria, which minimises subjective assessment of trial representativeness.

3) The SAE rate is a time-efficient metric, as the time required to obtain the data and perform the assessment is much lower than other measures.

4) SAEs are predictable trial outcomes, making them valuable in the prior assessment of trial representativeness. The prediction of SAEs can be performed using previous trial data to train machine-learning models to forecast SAEs (Ménard et al., 2019).

5) The SAE rate does not rely solely on real-world data to perform the assessment, especially when real-world data are incomplete or not readily available. The assessment can be performed by comparing SAE rates with rates from landmark pragmatic trials (Harrington et al., 2023). For example, the SAE rate of a new GLP-1 RA trial can be compared to the SAE rate of the EXSCEL (Exenatide Study of Cardiovascular Event Lowering) landmark pragmatic trial.

However, there are challenges to studying the SAE rate as a metric of trial representativeness, including:

1) For some treatments, the SAE rate may be influenced by the study treatment. Therefore, the SAE rate needs to be examined first for safe intervention trials rather than toxic intervention trials to avoid complicating the analysis. For this reason, I will use GLP-1 RA trials as an exemplar.

2) It is unclear whether SAE data are consistently reported across all trials, especially SAE counts and timeframes. Therefore, I will explore SAE reporting in published literature. Moreover, I will explore SAE capturing in GLP-1 RA trials.

3) It is unclear whether SAE rates are similar across trial arms. Therefore, I will compare SAE rates across intervention and control arms. This comparison will be valuable in showing that SAEs are unrelated to the intervention of interest and can be combined from all arms to increase statistical precision.

4) The lack of a universally accepted gold standard measure of trial representativeness makes it challenging to compare the SAE rate with another measure of representativeness. Furthermore, the lack of information on multimorbidity and frailty makes it difficult to examine whether SAE reflects trial representativeness, as this

information is crucial to assess trial participation. Alternatively, a triangulation of evidence will be used by comparing the SAE rate with the PRECIS-2 tool, which may indirectly reflect trial representativeness and by studying the association between SAE rates and eligibility criteria, which may determine trial representativeness.

1.5.2 Choosing T2DM and GLP-1 RAs as an exemplar in this thesis

This thesis builds on an existing systematic review of novel antidiabetics (Butterly et al., 2022). I will use T2DM trials that examined GLP-1 RAs as an exemplar. T2DM accounts for 90% of diabetes cases worldwide (Tripathi and Srivastava, 2006). T2DM is highly prevalent among older adults and those with comorbidities (Pearson-Stuttard et al., 2022). However, they are often under-represented in diabetes trials (O'Shea, Teeling and Bennett, 2013; Miklavcic et al., 2020). For example, an analysis of diabetes trials registered on ClinicalTrials.gov found that 1364 (54.9%) trials excluded patients older than 75 years, while only 15 trials exclusively recruited patients older than 65 years (Lakey et al., 2013). Therefore, it is challenging to treat people with diabetes who are older or have multiple comorbidities. This gap has led clinicians to extrapolate the evidence from trials of younger and healthier participants, raising concerns about the representativeness and the generalisability of T2DM trials (Bethel et al., 2017).

GLP-1 RAs hold particular importance among novel antidiabetics, as they showed promising results not only for the management of T2DM but also for multiple cardiovascular risk factors, including obesity, dyslipidaemia, and arterial hypertension (Del Olmo-Garcia and Merino-Torres, 2018; Michos, Lopez-Jimenez and Gulati, 2023). GLP-1 RAs mimic the effect of glucagon-like peptide-1 (GLP-1), a hormone that stimulates insulin production and inhibits glucagon release (Madsbad, 2016). They are generally safe and well-tolerated interventions compared to toxic interventions such as chemotherapeutic agents, which minimise the occurrence of SAEs related to the intervention and do not complicate this analysis (Harris and McCarty, 2015; Wolff et al., 2022; Wu et al., 2022). Moreover, trials of novel interventions mostly exclude older adults and people with comorbidities (Hurria et al., 2015; Hanlon et al., 2019). Therefore, there are concerns about their representativeness (Rothwell, 2006).

1.6 Thesis research question and approach to analysis

Thesis Research Question: Can the SAE rate be used as a metric of trial representativeness? Glucagon-like peptide-1 receptor agonist trials as an exemplar

To investigate whether the SAE rate can be used to measure trial representativeness, I propose:

1) To explore the feasibility of using the SAE rate as a metric of trial representativeness through:

   - Examining SAE reporting in published literature.

   - Exploring SAE reporting in GLP-1 RA trials.

   - Examining the differences in SAE rates between trial arms.

2) To explore whether SAEs reflect trial representativeness. I will use triangulation of evidence due to the absence of a gold standard measure of trial representativeness through:

   - Studying the association between the SAE rate and the PRECIS-2 score.

   - Examining the difference in the association between potential markers of trial representativeness that serve as fair umpires with the SAE rate and the PRECIS-2 score. A fair umpire is an imperfect measure, yet it can still distinguish between two measures without bias toward either (Glasziou, Irwig and Deeks, 2008).

   - Studying the association between SAE rates and eligibility criteria.

1.7 Aims and objectives of the thesis

This thesis aims to investigate whether the SAE rate can be used to measure trial representativeness, using GLP-1 RA trials as an exemplar.

The objectives of this thesis are to:

1) Study the literature on SAE reporting to examine how frequently SAE data is reported.

2) Explore SAE capturing in GLP-1 RA trials to ensure the reporting of sufficient information for the calculation of SAE rates.

3) Examine the difference in SAE rates between intervention and control arms of GLP-1 RA trials to assess if SAE rates from the intervention and control arms can be combined to estimate SAE rates with greater precision.

4) Assess GLP-1 RA trials using the PRECIS-2 tool as objectively as possible to enable a fair comparison with the reported SAE rate as metrics of trial representativeness.

5) Compare the SAE rate with the PRECIS-2 tool based on the differences in their associations with the fair umpire tests to explore the validity of the SAE rate as a metric of trial representativeness.

6) Examine the association between eligibility criteria and SAE rates to assess if a potential driver of trial representativeness, such as eligibility criteria, is associated with the SAE rate.

1.8 Outline of the thesis

Chapter 1 outlines the importance of RCTs and their representativeness. Moreover, it discusses the current measures of trial representativeness. Additionally, this chapter provides the rationale for choosing the SAE rate as a metric of trial representativeness and for choosing T2DM and GLP-1 RA trials as an exemplar.

Chapter 2 examines SAE reporting in published literature. It explores the reporting of SAE counts and timeframes, the practice of harm reporting in journal publications and the current methods of SAE reporting.

Chapter 3 explores SAE capturing in GLP-1 RA trials to ensure the feasibility of calculating SAE rates. Moreover, it explores the factors that may be associated with the reported SAE rates.

Chapter 4 examines the difference in SAE rates between intervention and control arms in trials of GLP-1 RA to know if SAE rates of intervention and control arms can be combined as this will overcome low numbers and increase the statistical power.

Chapter 5 retrospectively assesses GLP-1 RA trials using the PRECIS-2 tool to enable fair comparison with the SAE rate. Moreover, it explores the challenges of the retrospective assessment using the PRECIS-2 tool.

Chapter 6 examines the association between the SAE rate and the PRECIS-2 score. Moreover, it compares the SAE rate with the PRECIS-2 score based on the differences in their associations with baseline and trial characteristics, which serve as fair umpires, to explore the validity of the SAE rate as a metric of trial representativeness.

Chapter 7 examines the association between highly restrictive eligibility criteria and SAE rates. Additionally, it examines the association between the continuous score of eligibility criteria and SAE rates. The aim is to determine whether eligibility criteria, as a driver of trial representativeness, may be associated with the SAE rate.

Chapter 8 discusses the main findings of this thesis, showing the contribution of this work to the research literature. Furthermore, it indicates the implications of the findings of this thesis to clinical research. Additionally, it discusses the strengths and weaknesses of this work. Finally, it suggests future research based on the findings and research gaps identified through this work.

Chapter 2 The feasibility of using SAE rates as a marker of trial representativeness, a literature review of SAE reporting in RCTs

2.1 Chapter overview

This chapter will explore the feasibility of using the reported SAE rates in RCTs as a metric of trial representativeness. First, this review will provide an overview of SAE reporting in RCTs, focusing on reporting SAE frequencies, SAE timeframes and discrepancies between SAE data sources. Second, it will study the quality of harm reporting practices in journal publications of RCTs, focusing on their adherence to harm reporting recommendations and guidelines. Third, it will present an overview of current SAE reporting methods.

2.2 Background

2.2.1 SAEs in RCTs

The healthcare sector must provide safe and effective therapeutic interventions that ultimately help improve health outcomes (Talbot and Nilsson, 1998; Haleem et al., 2015). Moreover, these interventions must undergo a process of evaluation, ranging from laboratory studies to RCTs, to ensure their safety and efficacy (Kabisch et al., 2011). When RCTs test a new therapy, they aim to detect and collect safety data as part of their practice protocol (Singh and Loke, 2012). Monitoring, detection, and documentation of SAEs are critical parts of safety data in RCTs due to the importance and impact of this data on the outcomes of trials and, in turn, on clinical decisions (Singh and Loke, 2012).

2.2.2 SAE reporting in RCTs

Reporting SAEs to authorities, trial registries, and publications is essential, as underreporting may lead to a biased assessment of the intervention (Schroll, Penninga and Gøtzsche, 2016; Abdel Shaheed et al., 2022). Moreover, sufficient SAE reporting is important for using the SAE rate as a metric for trial representativeness. Specifically, if SAEs are consistently recorded and monitored across RCTs, they can provide insights into how well the trial participants reflect the general population (Hanlon et al., 2022). Moreover, detailed SAE reporting helps identify any patterns or trends in SAEs that might be specific to subgroups. Therefore, robust and sufficient SAE reporting is a prerequisite for using the SAE rate as a metric for trial representativeness.

Several guidelines and recommendations, such as the Consolidated Standards of Reporting Trials (CONSORT), aim to guide investigators to improve the quality of harm reporting in RCT reports (Pitrou et al., 2009). The CONSORT statement, which includes a checklist and a flowchart that guide researchers in comprehensively reporting the design, findings and interpretations of RCTs, issued an extension specifically concerning harm reporting to improve the quality of harm reporting in RCTs (Moher et al., 2010; Hunsinger et al., 2014). This extension includes items on reporting event counts and timeframes (Ioannidis et al., 2004). The CONSORT extension has been widely applied and has helped enhance the quality of SAE reporting in RCTs (Anderson et al., 2024). However, the adherence to this extension has been inadequate and affected by several factors, such as space limits in publications and the slow uptake of the extension by investigators (Dimairo et al., 2018).

2.3 Aims and objectives

This chapter will explore the feasibility of using the reported SAE rate as a proxy for trial representativeness. Specifically, it will examine the literature on SAE reporting in RCTs.

The objectives of this literature review are:

1) To explore the literature on SAE reporting, focusing on the reporting of SAE counts and timeframes, and discrepancies between different SAE data sources to investigate their reliability and highlight areas where data might be incomplete or biased.

2) To examine the literature concerning the quality of harm reporting in journal publications of RCTs, showing the adherence to harm reporting recommendations and guidelines.

3) To explore methods of SAE reporting to underline potential areas of improvement for the accuracy and inclusiveness of SAE data.

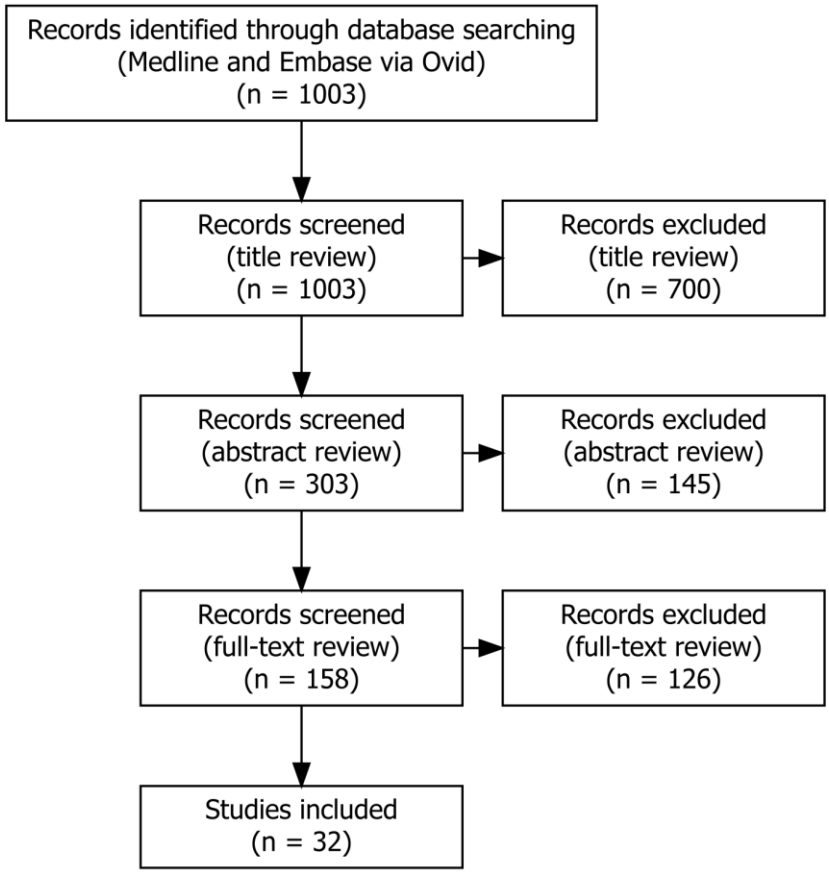
2.4 Methods

Medline and Embase databases were accessed via Ovid for the literature search. As illustrated in Table 2.1, different keywords for RCTs, SAEs and reporting were used. The search started in 2019 and updated in 2024, retrieving 1003 articles. The studies included in this review must have examined SAE reporting in RCTs. Studies that were not relevant to this topic were excluded. In the screening process, 700 studies were excluded after the title review and 145 articles were excluded after reading the abstract. Following a full-text screening, 126 more studies were excluded. Ultimately, this resulted in 32 studies being included in this review (see Figure 2.1).

Table 2.1 Search strategy of the literature review

#	Searches
1	RCT or Randomised controlled trial or clinical study or clinical research or phase III/IV trial or clinical experimentation or clinical testing or randomised controlled trials as a topic/ or randomised controlled trials as a topic/ or clinical trials as a topic/ or phase III as a topic/
2	SAE or serious adverse event or serious side effect or serious reaction or serious harm or fatal adverse event or fatal reaction or death or hospitalisation or adverse event or adverse effect or adverse reaction or harm or side effect
3	Reporting or posting or publishing or underreporting or declaring or investigating or capturing
4	1 and 2 and 3

Figure 2.1 PRISMA flow diagram of the literature selection process



2.5 Literature review

2.5.1 SAE reporting in RCTs

Table 2.2 summarises the findings from 18 studies that examined SAE reporting in RCTs across different medical conditions. These studies focused on reporting SAE counts and timeframes. Moreover, they explored inconsistencies in SAE reporting between journal publications and other data sources, such as trial registries and CSRs. SAE data were reported across a majority of studies (nearly 80 to 100%), enabling the calculation of SAE rates and facilitating comparisons between trials. Moreover, explicit SAE timeframes were not frequently reported in these studies. Furthermore, inconsistency in reporting SAE data was apparent between journal publications and other data sources.

Table 2.2 SAE reporting in RCTs

Study	Medical condition	Type of study	Primary data topics	Search strategy	Main findings
Scharf and Colevas (2006)	Oncology	Comparative analysis	Data from a sponsor's database with relevant published trials data	The National Cancer Institute (NCI) Clinical Data Update System (CDUS) electronic database of RCTs and PubMed	• Low-grade (mild and moderate) AEs were underreported in publications; only (58%) of grade 1 and 2 AEs were reported in articles. • 305 of 423 (72%) of drug-related grade ≥ 3 AEs (severe, medically significant, life-threatening or death) cited on CDUS were reported in publications. • Reported SAEs in databases were higher than in publications; a total of grade ≥ 3 AEs cited on CDUS was 611, while 413 in publications. • This study was limited to phase II trials, which may not reflect the safety reporting in later trial phases (III and IV). • This study focused on trials in the NCI CDUS database, which may limit the generalisability of its results.

Seruga et al. (2011)	Oncology	Systematic review	Updated labels and publications of targeted anticancer RCTs	The US Food and Drug Administration (FDA) website, updated label references and Medline	• Out of 76 serious adverse drug reactions (SADRs) reported on updated drug labels, 30 (39%) of them were not reported on publications of RCTs. • This study focused on assessing the reporting of SADRs, which are rare and less likely to be captured within the trial duration.
Smith et al. (2013)	Pain	Systematic review	Publications of pain treatment journals	Three Journals of pain (European Journal of Pain, Journal of Pain, and PAIN)	• SAEs were not reported for 33 (41.2%) trials. • Moreover, 70 (87.5%) studies did not report AE timeframes. • This study was limited, as publications were the only assessed source of AE data, and they only selected three journals.
Riveros et al., (2013)	Trials listing drugs as an intervention	Systematic review	Trials summaries and journal publications of RCTs of different diseases	Trials with results posted on ClinicalTrials.gov and PubMed for publications	• Trials with results posted on ClinicalTrials.gov had better AE and SAE reporting than those posted on publications (73% versus 45%) and (99% versus 63%), respectively.

					• This study included only the first publication of trial findings, missing essential safety data in the following publications.
Belknap et al. (2013)	Oncology	Systematic review	Institutional Review Board (IRB) SAEs reports and research records, and electronic or paper medical records of six clinical trials involving bevacizumab or oxaliplatin.	IRB SAE reports and research records and electronic or paper medical records of NCI-designated Comprehensive Cancer Centre	• IRB reports were inadequate; 182 (75%) out of 205 AEs were not reported to IRB. • However, 24 (80%) SAEs were reported to IRB. • This analysis was focused on six RCTs conducted in the same setting, limiting the generalisability of the findings. Moreover, two RCTs included in this study were Phase II, which may not effectively capture rare SAEs due to its small sample size.
Maund et al. (2014)	Major depressive disorder	Systematic review	Reports of clinical studies and publications of nine RCTs for duloxetine	PubMed and Cochrane	• Journal publications did not include adequate safety data compared to CSRs. • Deaths and suicides were adequately reported in only 2 out of 9 trials in their relevant publications.

					• The occurrence or non-occurrence of SAEs was not reported in 3 out of 9 trials in their relevant publications. • However, SAEs were reported in 8 of 9 CSRs.
Hughes, Cohen, and Jaggi (2014)	Depression and Psychosis	A cross-sectional study of trial summaries for antidepressant and antipsychotic therapy	Registries of RCTs for antidepressant and antipsychotic therapy and their correspondent publications	clinicalstudyresults.org registry and bibliography of registries	• The frequencies of SAEs were reported in 125 (88%) trial summaries, 85 (59.9%) journal articles and 95 (93.1%) unpublished trial summaries. • Inconsistencies between registries and publications were found in the reported number of SAEs; 694 (43.2%) SAEs on registries of RCTs were not reported in publications of RCTs. • Nearly half of the included trials did not have correspondent publications, which may have impacted the applicability of the outcomes of this study to other drug classes and led to biased findings.

Tang et al. (2015)	Different diseases	Systematic review	All records were exported from ClinicalTrials.gov. RCTs with SAE were published at ClinicalTrilas.gov, and relevant publications were identified.	ClinicalTrials.gov and Medline	• SAEs were more frequently reported at ClinicalTrials.gov than in journal publications in 31 trials. • SAEs were not reported in 26 (13%) publications. • The total number of SAEs per treatment arm was not reported in 33 (16%) publications. • This study only searched a single database for publications, which may have affected the study sample and led to biased findings.
Hodkinson, Gamble, and Smith (2016)	Obesity	Case study	RCT summaries and RCT publications of orlistat	RCTs summaries, Medline and Cochrane	• Inconsistencies in AE reporting were found between RCT summaries and publications; 31 (51%) AEs were not found in publications of RCTs. • Inconsistencies between RCT summaries and publications were also found in the reported SAEs; 311 (95%) SAEs were not found in publications of RCTs.

					• AE timeframes were reported for 4 out of 5 trials. • This study was based on a meta-analysis of only five RCTs, which may limit the generalisability of its findings. Moreover, obtaining CSRs for all trials was not feasible, which may have led to biased results.
Maillet et al. (2016)	Oncology	Systematic review	Publications of RCTs of oncology	Medline	• Frequencies of grade 3/4 AEs (severe, medically significant or life-threatening) were reported in 312 (96%) journal publications. • Frequencies of grade 5 AEs (death) were reported in 237 (73%) journal publications. • This study focused on publications as its source of safety data, which may inadequately reflect the actual practice of RCTs.
De Vries et al. (2016)	Depression	Meta-analysis	Data of review of FDA-registered trials of second-generation	review of FDA and corresponding publications	• Inconsistencies were found in reported SAEs. • The reported number of SAEs was present in 9 (43%) articles compared to FDA reviews.

			antidepressants and relevant publications		• This study was limited, as the FDA database may miss important SAE data, which may have affected the outcomes of this comparison. Moreover, this study was performed on a single drug class, so its results may not apply to other pharmacological interventions.
Tfelt-Hansen, Lindqvist, and Do (2018)	Migraine	Systematic review	Publications of migraine RCTs	Medline and PubMed	• 50 (69%) studies did not report the number of patients with any SAE. • Data used in this study were obtained from journal publications, which may inadequately reflect the actual practice of RCTs.
Phillips et al. (2019)	Different diseases	Systematic review	Publications of Phase III and IV RCTs of pharmacological interventions	The BMJ, the JAMA, the Lancet, and the NEJM	• SAEs were adequately reported; 132 (72%) publications reported the number of SAEs. • AE frequencies were adequately reported; 160 (87%) publications reported the frequencies of AEs.

					• SAE timeframe was not reported in 95 (57.2%) publications. • This study included high-impact journals, which may have led to better outcomes and increased the likelihood of bias. The authors also limited their search period to one year, which may not reflect contemporary practice.
Hodkinson et al. (2021)	Schizophrenia or bipolar disorder	Systematic review	CSRs, Individual participant data (IPD), trials summaries and journal publications	MEDLINE, Central, EMBASE, PsycINFO and ClinicalTrials.gov	• CSRs reported approximately eight times more SAEs than journal publications. • SAE reporting was complete on CSRs 35 (100%). • However, SAE reporting on journal publications and trial registry was 20 (61%) and 17 (49%), respectively. • This study assessed SAE reporting across all registries collectively rather than showing the reporting for each trial registry.

Yao et al. (2021)	Oncology	Systematic review	RCTs publications	PubMed, Embase, Medline, and NEJM	• Only 41 (25.5%) publications reported SAEs. • Publications of industry-sponsored trials reported higher SAEs than trials with other funding sources (57.6% vs 20.7%). • Publications of trials published in high-impact journals reported higher SAEs than those published in other journals (31.9% vs 16.7%). • However, SAE reporting improved overtime. • This study only examined SAE reporting in journal publications, which may inadequately reflect the actual practice of SAE reporting in RCTs.
Paludan-Müller, Créquit and Boutron (2021)	Oncology	Systematic review	CSRs, trial summaries on registries and journal publications	European Medicines Agency (EMA) and trial registries	• CSRs had more complete harm reporting compared to other sources. • SAE reporting was complete on CSRs 36 (100%) and ClinicalTrials.gov 37 (100%). • Moreover, 19 (95%) trials on other trial registries reported SAEs.

					• However, 16 (50%) journal publications did not report SAEs.
Taillefer de Laportalière et al. (2023)	Depression	Systematic review	Trials summaries and journal publications	Medline and ClinicalTrials.gov	• SAEs were reported in 9 (90%) trials. • Journal publications reported that only 94 (41.5%) of SAEs were reported on ClinicalTrials.gov. • This study focused on RCTs of a single drug from a single trial registry, which may limit the generalisability of the findings.
Madi et al. (2023)	Covid-19	Systematic review	Trials summaries and journal publications	PubMed and ClinicalTrials.gov	• Journal publications reported 364 (51%) of SAEs compared to trial summaries posted on ClinicalTrials.gov. • SAEs were reported in 18 (90%) trial summaries.

2.5.2 Harm reporting practice in journal publications of RCTs

Table 2.3 summarises the findings from various studies that examined the quality of harm reporting practices in journal publications of RCTs regarding adherence to reporting guidelines and recommendations. Studies indicated that harm reporting was often inadequate, with many publications failing to adhere to most CONSORT harm checklist items. However, items concerning the reporting of event counts were sufficiently fulfilled. These studies only focused on assessing harm reporting in journal publications, which may not fully capture the actual reporting practice of RCTs.

Table 2.3 Harm reporting practice in journal publications of RCTs

Study	Medical condition	Type of study	Primary data topics	Search strategy	Main findings
Smith et al. (2012)	Pain	Systematic review	Publications of pain treatment journals	Three Journals of pain (European Journal of Pain, Journal of Pain, and PAIN)	● The mean CONSORT harms score improved over time from 5.4 (2.5) in the first epoch to 6.5 (2.7) in the second epoch. ● Industry-funded trials showed a higher CONSORT harms total score than other fund resources, 0.22 95% CI (0.09-0.36). ● 88% of trials fulfilled items concerned with reporting SAE counts and timeframes. Additionally, the fulfilment of this item improved over time from 74% to 97%. ● This study was focused on RCTs published in three pain journals, which may limit its generalisability.
Péron et al. (2013)	Oncology	Systematic review	Systemic solid tumours therapy	Medline	● Based on the AE reporting quality score (AERQS), better quality of AE reporting was observed in trials that received industrial funding; they scored 10.68 points on a 16-point scale.

					• 296 (91%) trials fulfilled the reporting of the item in AERQS that concerns SAE reporting. • 171 (53%) trials reported AE timeframes.
Hodkinson et al. (2013)	Hypertension, urology, epilepsy, complementary medicine	Systematic review	Seven studies that assessed the quality of harm reporting in 800 RCTs	Reviews of published and unpublished research that assessed the quality of harm reporting in RCTs	• Harm reporting was inadequate; (50%) of CONSORT harm checklist items were not adhered to in 6 of 7 studies. • This study did not include studies of other harm reporting guidelines. Moreover, the number of included studies was limited to only seven studies. • Some of the included studies were published before the issue of CONSORT harm extension, which may lead to biased findings.
Sivendran et al. (2014)	Oncology	Systematic review	Publications of oncology trials	PubMed, Embase, and Medline	• AE reporting was selective and heterogeneous; the median AE reporting score derived from CONSORT harm extension was 8 out of 14, and the range was 3 to 12. • However, 135 (77%) articles reported AE counts.

					• Furthermore, deaths were reported in 132 (75%) articles. • 109 (62%) articles reported AE timeframes. • The search time interval of three years and the probability of poor adherence in the past may have limited the analysis of these results.
Mahinbakht, Lavasani, and Guirguis (2014)	Oncology	Systematic review	Publications of early-phase breast cancer using adjuvant trastuzumab trials	Medline and Cochrane	• AE reporting was inadequate; 4 out of 5 studies showed a total adherence to CONSORT harms checklist items of less than (50%). • However, 80% of studies reported the frequency of deaths due to cardiac events. • This study only focused on the early phases of breast cancer. Thus, its results may not apply to all RCT reporting practices. Moreover, this study only evaluated five RCTs, limiting its generalisability.
Chen et al. (2015)	Oncology	Systematic review	Publications trials of immune checkpoint inhibitors	Medline, Embase, and Cochrane	• AE reporting was inadequate. The mean quality score adopted from CONSORT harm extension was 11.21 out of 21, with a range of 3.5 to 17.5.

					• AE reporting improved over time; the mean score increased from 9.09 to 11.81 points. • Grade 3/4 AEs were reported for 96% of studies • AE timeframes were reported for 30% of studies. • This study scoring tool was adopted from CONSORT harm recommendations instead of using the same items, which may have impacted the quality of their assessment.
Gewandter et al. (2015)	Temporomandibular disorders (TMDs)	Systematic review	Publications of RCTs of TMDs therapy	PubMed	• The fulfilment of most CONSORT harm extension items was between (10%) to (23%), and only one item exceeded (36%). • AE reporting improved over time. • The authors used CONSORT harm items to assess the quality of reporting, which was published in 2004, while they included studies between 1969 and 2013, which may have led to biased outcomes.

Williams et al. (2016)	Pain	Systematic review	Publication in journals of anaesthesiology and journals of intravenous and invasive pain management	Major anaesthesiology and pain journals	• CONSORT harm recommendation items 4, 7, and 8 were fulfilled by more than 75% of RCTs, whereas 50 to 70% met CONSORT harm recommendation items 3, 6, and 10. • Items 1, 2, and 5 were fulfilled by less than 50% of the included trials, and only less than 2% of RCTs met CONSORT harm recommendation item 9. • 140 (85%) fulfilled the reporting of the item in the CONSORT harm extension that concerns SAE reporting. • 95 (72%) of studies reported AE timeframes. • This study included only six journals, which may limit its generalisability.
Westergren, Narum, and Klemp (2018)	Gastrointestinal bleeding or perforation	Systematic review	Publications of trials comparing corticosteroid to placebo	PubMed, Embase and Cochrane	• Harm reporting was inadequate; the mean CONSORT score was 5.25 out of 10. • 130 (81.8%) fulfilled the reporting of the item in the CONSORT harm extension that concerns SAE reporting.

2.5.3 Methods of SAE reporting in RCTs

Table 2.4 reviews methods of SAE reporting in RCTs. These studies showed limitations of current methods, which may lead to underreporting, such as selective reporting and incomplete reporting forms. However, new methods demonstrated promising results in terms of accelerating the reporting process and predicting the underreporting of SAEs.

Table 2.4 Methods of SAE reporting in RCTs

Study	Medical condition	Type of study	Primary data topics	Search strategy	Main findings
(London et al., 2009)	Oncology	case analysis	an electronic system to report SAEs called eSAEy	Thomas Jefferson University	• eSAEy reduced the reporting period and enhanced the reporting precision. • 588 SAEs were reported using eSAEy, and the median time of the whole reporting process was <2 days (mean of 7±0.2 days), while it was 24 days (mean of 45 ± 5.7 days) for the paper-based system. • This system was designed for RCTs based at Thomas Jefferson University and their affiliation members' institutions. As such, the applicability of this system requires more validation to be implemented in other RCTs.
Bolland et al. (2013)	Osteoporosis in postmenopausal women	Secondary analysis of RCT data	Patient-reported and investigator-verified SAEs in a RCT of a 1g calcium supplement	participant's medical records, hospital records or death certificates	• Discrepancies between patient-reported and investigator-verified SAEs; 25 of the 58 verified MIs and 13 of the 63 verified strokes were underreported. • 50% of MIs and 42% of strokes were not reported to investigators.

					• The participants in this RCT were older (mean age 74 years) and had comorbidities and cognitive impairment, which may have contributed to underreporting and biased outcomes.
Crépin, Villeneuve, and Merle (2016)	Different diseases (trials of investigational drugs or medical strategies)	A cross-sectional study of all SAE case report forms was reported in 2012 to the sponsor (Limoges University Hospital) from all clinical trials.	A standardised data quality evaluation form	All SAE case report forms reported to sponsor from all clinical trials	• The quality of SAE sponsor reports was low; most forms were not filled. • (5.7%) of reports did not report the date of onset of the SAE, while assessments of causality were not reported in (9.3%) of reports. • 36% of the included RCTs were not pharmacological, which may have affected the outcomes of this study. • This study included trials from a single sponsor, which may limit its generalisability.
Ménard et al. (2019)	Different diseases	Proof of concept	A data set of RCTs of different diseases	Roche/Genentech sponsored RCTs	• The prediction tool is an efficient method of detecting underreporting.

					• The model scored 0.67 in the area under the curve (AUC) of receiver operating characteristic (ROC) for the statistical scenario and 0.97 for the zero scenario. • For 25%, 50%, 67%, and 75% scenarios, the AUC was 0.62, 0.79, 0.89, and 0.92, respectively. • The model may require further validation because it was only trained on data from a single sponsor (Roche/Genentech).
Mayo-Wilson et al. (2019)	Neuropathic pain and bipolar depression	Systematic review	CSRs and corresponding publications of RCTs of gabapentin for neuropathic pain and quetiapine for bipolar depression	Cochrane, PubMed, Embase, LILACS, CINAHL, International Clinical Trials Registry Platform Search Portal and ClinicalTrials.gov	• CSRs did not apply selective reporting criteria for reporting all AEs, while publications of RCTs applied selection criteria that missed extensive safety data.

2.6 Discussion of literature findings

SAEs were sufficiently reported, especially for trials registered on ClinicalTrials.gov and CSRs (Scharf and Colevas, 2006; Belknap et al., 2013; Riveros et al., 2013; Hughes, Cohen and Jaggi, 2014; Maund et al., 2014; Tang et al., 2015; Hodkinson, Gamble and Smith, 2016; Maillet et al., 2016; Phillips et al., 2019; Hodkinson et al., 2021; Paludan-Müller, Créquit and Boutron, 2021; Taillefer de Laportalière et al., 2023; Madi et al., 2023). The sufficient SAE reporting on ClinicalTrials.gov and CSRs ensures the reliability of data used to measure trial representativeness. Furthermore, SAE reporting is notably better than reporting eligibility criteria and baseline patient characteristics. The reporting of eligibility criteria and baseline characteristics of RCTs in journal publications is generally inadequate, inconsistent, and lacks clear justification (Van Spall et al., 2007; Wertli et al., 2013). Moreover, the details of reported eligibility criteria on trial registries such as ClinicalTrials.gov often vary between trials (Ross et al., 2010). Consequently, assessing trial representativeness based on reported eligibility criteria or baseline characteristics is more challenging than the reported SAEs. However, in journal publications, SAE reporting was insufficient (Smith et al., 2013; de Vries et al., 2016; Tfelt-Hansen, Lindqvist and Do, 2018; Yao et al., 2021). This insufficient SAE reporting may be attributed to the space limitations in journal publications, which may lead to selective reporting (Tang et al., 2015). Therefore, SAE data should be mainly extracted from trial registries such as ClinicalTrials.gov or relevant CSRs.

Few studies examined SAE timeframe reporting and found it is frequently underreported (Péron et al., 2013; Smith et al., 2013; Sivendran et al., 2014; Chen et al., 2015; Phillips et al., 2019). However, two studies found that SAE timeframes were sufficiently reported (Hodkinson, Gamble and Smith, 2016; Williams et al., 2016). These inconsistent findings could be attributed to differences in methodology and sample sizes. Furthermore, inconsistencies in the reported number of SAEs were found between RCTs summaries, CSRs and publications (Scharf and Colevas, 2006; Seruga et al., 2011; Riveros et al., 2013; Hughes, Cohen and Jaggi, 2014; Maund et al., 2014; de Vries et al., 2016; Hodkinson, Gamble and Smith, 2016; Hodkinson et al., 2021; Paludan-Müller, Créquit and Boutron, 2021; Taillefer de Laportalière et al., 2023; Madi et al., 2023). These inconsistencies can be attributed to differences in reporting standards, selective reporting, or inadequate methodological consistency. For example, the elimination of events from publications that

were not assumed to be linked to the intervention or that were infrequent and related to chance (Scharf and Colevas, 2006; Hughes, Cohen and Jaggi, 2014; Maund et al., 2014; Tang et al., 2015; de Vries et al., 2016; Hodkinson, Gamble and Smith, 2016; Phillips et al., 2019). Therefore, journal publications may have limitations when used as a source of safety data.

Studies that have examined the adherence of RCT publications to guidelines and recommendations of harm reporting found that this was poor, inadequate, and heterogeneous (Smith et al., 2012; Hodkinson et al., 2013; Péron et al., 2013; Mahinbakht, Lavasani and Guirguis, 2014; Chen et al., 2015; Gewandter et al., 2015; Williams et al., 2016; Westergren, Narum and Klemp, 2018). However, items concerning the reporting of event counts were sufficiently fulfilled (Smith et al., 2012; Péron et al., 2013; Chen et al., 2015; Gewandter et al., 2015; Williams et al., 2016; Westergren, Narum and Klemp, 2018). Furthermore, the quality of harm reporting in RCTs has improved over time (Smith et al., 2012; Chen et al., 2015; Gewandter et al., 2015), which may be related to the adoption of journals to recommendations and guidelines, leading authors to adhere more too. Moreover, publications of industry-funded RCTs demonstrated better reporting quality than other funding resources (Péron et al., 2013; Williams et al., 2016). The impact of funding resources on the quality of harm reporting is apparent and worthy of consideration in evaluating RCT outcomes.

Additionally, few studies have examined different SAE reporting methods used in the current practice of RCTs. The quality of SAE reporting forms was poor; most forms were incomplete, and the fulfilment of items was also low (Crépin, Villeneuve and Merle, 2016). CSR did not apply selective reporting criteria, reporting all events, while publications of RCTs applied selection criteria missing extensive safety data (Mayo-Wilson et al., 2019). eSAEy reduced the reporting period and enhanced precision (London et al., 2009). The prediction tool is an efficient means of detecting underreporting (Ménard et al., 2019). Furthermore, it could be useful for assessing trial representativeness as it may give a prior comparison of expected SAE rates in the trial participants with SAE rates in the target population.

2.7 Conclusion

Studies demonstrate that SAE reporting was generally sufficient, especially on CSRs and trial registries. However, SAE reporting in publications of RCTs was insufficient and inconsistent with databases and other data resources. Given this variability, it is crucial to recognise that the reliability of SAE data can differ depending on the source. Therefore, the SAEs reported on trial registries and CSRs are most reliable and can be used in the measurement of trial representativeness. However, exploring SAE reporting in an exemplar, namely GLP-1 RA trials, is essential to ensure that SAE reporting is sufficient to calculate SAE rates for trials. Moreover, the differences in SAE reporting between data sources will be considered in this thesis.

Chapter 3 Capturing SAEs in GLP-1 RA trials

3.1 Chapter overview

This chapter will explore SAE capturing in GLP-1 RA trials to enable the calculation of SAE rates for measuring trial representativeness. Moreover, it will examine the factors that may be associated with the reported SAE rates.

3.2 Background

3.2.1 SAE reporting in RCTs

SAE reporting in RCTs is essential for assessing the safety of new drugs, medical devices, or other interventions (Ioannidis and Lau, 2001). Furthermore, SAEs may be related to trial representativeness (Hanlon et al., 2022). SAEs are monitored and recorded in a specific timeframe that typically starts from the randomisation of participants and continues to the end of the follow-up period, although this duration may differ depending on the study design or regulatory requirements (Unkel et al., 2019; James et al., 2020). Although the reporting of SAEs and their timeframes has been investigated across different diseases and therapeutic areas, it remains unclear whether GLP-1 RA trials report sufficient information to calculate SAE rates.

Trial registries such as ClinicalTrials.gov represent the primary sources of reported SAEs; they often include detailed descriptions of SAE counts and timeframes (Aslam et al., 2013; Hartung et al., 2014; Adam et al., 2018). However, some registered trials do not post their results in registries (Jones et al., 2021; Negoro et al., 2023). Journal publications are another source of SAE data, but they often lack adequate details and are prone to biased SAE reporting. For example, 110 trials registered on ClinicalTrials.gov were examined for publication bias and found that 38 trials had inconsistent SAE reporting; 87% of these trials reported more SAEs on ClinicalTrials.gov than publications (Hartung et al., 2014). Consequently, some trials are expected to underreport critical trial outcomes such as SAEs

(Tang et al., 2015; Rosati et al., 2016). This issue is particularly evident in unregistered trials or those without results on registries. Furthermore, the guidelines for reporting on registries don't specify criteria for including outcomes that are SAEs in the total reported SAEs. As a result, it is unclear whether trials with outcomes that are SAEs, such as major adverse cardiovascular events (MACE), would report outcome counts in total SAEs. This lack of clarity could result in inconsistent SAE reporting.

3.2.2 Other factors associated with the reported SAE rates

The reported SAE rates in RCTs may be affected by the type of outcome measured (Hindiyeh et al., 2022). The outcomes measured in RCTs can be classified as soft or hard outcomes. Soft outcomes are usually subjective, such as patient-reported symptoms, or surrogate measures, such as laboratory measurements and radiographic findings, whereas hard outcomes are reasonably objective and have a tangible impact on patient health, such as hospitalisation and death (Asmar and Hosseini, 2009; Medeiros, 2017; Van Lieshout and Wijffels, 2020). Furthermore, RCTs with hard outcomes are more likely to have larger sample sizes and longer durations, and they are conducted in later-phase RCTs, while RCTs with soft outcomes are usually conducted in early-phase RCTs (Ciani et al., 2022; Verghis et al., 2023). Therefore, the discrepancies between the types of trial outcomes may lead to differences in reported SAE rates between RCTs.

Additionally, the reported SAE rates in RCTs may be associated with the type of population analysis (Phillips et al., 2019). Intention-to-treat (ITT) analysis includes all randomised participants, whereas per-protocol (PP) only includes those who completed the trial (Gupta, 2011). Older adults and people with comorbidities are more likely to withdraw from trials (Herrera et al., 2010; Pitkala and Strandberg, 2022). Moreover, trials with PP analysis often conduct run-in periods to examine participants' adherence and exclude those who may drop out or violate their protocols (Pablos-Méndez, Barr and Shea, 1998; Grayek et al., 2022). Therefore, the ITT analysis trials are expected to report higher SAE rates than trials that only conduct PP analysis (Hernán and Robins, 2017).

Furthermore, trial setting (single-centre or multicentre) and trial country-level (national or multinational) may affect the reported SAE rates. Trials conducted in multicentre and

multi-national settings are expected to be larger, have broader eligibility criteria, and recruit older adults and people with comorbidities (Girbes and de Grooth, 2020; Sedrak et al., 2021; Gumber et al., 2024). Consequently, they may report higher SAE rates than trials conducted only in a single centre or single country.

3.3 Aim and objectives

This chapter will examine the feasibility of using the reported SAE rates to measure trial representativeness. Specifically, it will determine whether GLP-1 RA trials report sufficient SAE information to calculate SAE rates.

The objectives of this chapter are:

1) To examine SAE reporting across all GLP-1 RA trials to ensure the capture and availability of sufficient SAE data to calculate SAE rates.

2) To compare SAE reporting between trial registries and trial publications to investigate their reliability as sources of SAE data and highlight areas where data might be insufficient.

3) To study the difference in reported SAE rates between GLP-1 RA trials based on their database registration, result reporting, primary outcome, population analysis, trial setting and trial country-level, to highlight issues related to transparency and consistency in reporting and understand how various trial characteristics may influence SAE rates.

3.4 Methods

3.4.1 Study selection

This thesis builds on an existing systematic review of novel antidiabetics (Butterly et al., 2022). The search and the selection of studies were completed. I subsequently applied an

additional selection of studies based on the drug class of GLP-1 RAs. 196 GLP-1 RA trials were identified for this thesis. 172 trials were registered on ClinicalTrials.gov, 12 were registered on the University Hospital Medical Information Network (UMIN) Clinical Trials Registry, EU Clinical Trials Register (EU-CTR) and Chinese Clinical Trial Register (ChiCTR), and 12 were unregistered. The eligibility criteria of the systematic review are described in Table 3.1.

Table 3.1 Eligibility criteria of the current systematic review

Types of participants	Inclusion criteria	• Aged 18 years or older. • Recently diagnosed with T2DM. • All countries. • All subgroups, such as older people or those with long-term conditions.
	Exclusion criteria	• Diagnosed with other types of diabetes other than T2DM. • Diagnosed with pre-diabetes. • At risk of T2DM but not currently diagnosed.
Types of interventions	Inclusion criteria	• Any GLP-1 RA, Sodium-Glucose Co-Transporter-2 inhibitor (SGLT-2i) or Dipeptidyl Peptidase-4 inhibitor (DPP-4i) as the trial intervention. • Short-acting or modified drug preparations can be included as monotherapy, dual therapy or triple therapy with other antidiabetic medications.
	Exclusion criteria	• Trials that deliver the trial intervention as a single dose only, for example, perioperative GLP-1 RA trials. • Trials were performed under fasting conditions.

Types of comparators	Inclusion criteria	• Comparators were either a placebo or one or more other antidiabetic medications. • Trials with the same drug class and the same drug comparisons.
	Exclusion criteria	• Trials with any non-pharmacological comparators.
Outcomes of interest	• Trials that used glycated haemoglobin (HbA1c) in either % or mmol/mol to measure glycaemic control. • Trials that used weight in kilograms or change in body mass index (BMI) to measure weight change. • Trials that used composite outcomes, such as MACE or singular outcomes, such as cardiovascular death, non-fatal myocardial infarction (MI), and hospitalisation for heart failure, to measure the cardiovascular outcome. • Trials that measure at least one of the above outcomes. • These outcomes were not required to be the primary outcome of the trial. • Non-inferiority trials.	

3.4.2 Data sources

This section describes different data sources used to extract data of GLP-1 RA trials. Sources include trial registries, CSRs, trial protocols and journal publications. The use of different data sources may provide more complete datasets as each source may contain unique information.

3.4.2.1 ClinicalTrials.gov

ClinicalTrials.gov is a United States public database for clinical trials and observational studies (Hartung et al., 2014). It is administered by the National Library of Medicine (NLM) at the National Institutes of Health (NIH) in collaboration with the FDA (Gresham

et al., 2022). ClinicalTrials.gov aims to improve the transparency of clinical studies by providing a central registry of their information to researchers, medical practitioners, and patients (Adam et al., 2018). It contains information about designs, conduct and results of clinical studies (Zarin et al., 2011). All trials involving pharmacological, device and biological interventions must register and submit their results on ClinicalTrials.gov according to a US federal mandate (DeVito, Bacon and Goldacre, 2020). It involves recording details of trials before they start, such as their objectives, designs, and outcome measures (Bhaskar, 2018). However, not all trials report their results on registries (McCord et al., 2022).

Information about all registered studies on ClinicalTrials.gov is downloaded and loaded on the Aggregate Analysis of ClinicalTrials.gov (AACT), a relational database managed by Duke University (Tasneem et al., 2012). Moreover, data from all trials registered on ClinicalTrials.gov are aggregated and structured in the AACT in an accessible format, enhancing their utility and public availability (Hirsch et al., 2013). However, some elements are not standardised. For example, eligibility criteria are structured in a free-text format, limiting their usefulness for secondary analysis (Kavalci and Hartshorn, 2023). Data used in this thesis were pulled from the AACT database.

3.4.2.2 Journal publications

Journals of medical research usually require researchers to register and post their results on trial registries in alignment with the recommendations of the International Committee of Medical Journal Editors (ICMJE) (De Angelis et al., 2004). RCTs publish their results to ensure the transparency of their findings (Bhide, Shah and Acharya, 2018). Moreover, published results of RCTs allow researchers to conduct further research and systematically meta-analyse their findings (Moher and Olkin, 1995). Data used in this thesis were obtained from journal publications of RCTs that did not post their results or were unregistered on ClinicalTrials.gov.

3.4.2.3 Clinical study reports

Data reported in journal publications may be incomplete, leading to biased assessments of the safety and efficacy of interventions (Li et al., 2018). Therefore, regulatory bodies require companies to submit the CSRs of their studies (Davis and Miller, 2017). CSRs

contain detailed information about the designs, conduct and analysis of RCTs (Mayo-Wilson et al., 2017). Although CSRs provide extensive information that surpasses other types of publications, they may be disadvantaged by their length, complexity and accessibility (Doshi and Jefferson, 2013). Data used in this thesis were obtained from CSRs of RCTs that did not post their results or were unregistered on ClinicalTrials.gov, and the required information was unavailable in journal publications.

3.4.2.4 Study protocols

Researchers produce protocols to describe the design, conduct, organisation, and analysis of their trials (Chan and Hróbjartsson, 2018). Study protocols must include ethical considerations to ensure the safety of their participants (Nardini, 2014). Moreover, they include standardisation of procedures to minimise the variability of collected data (Evans, 2010). Data used in this thesis were obtained from protocols of trials that did not post their results, were unregistered on ClinicalTrials.gov, and the required information was unavailable in publications or CSRs.

3.4.3 SAEs data extraction and harmonisation

The SAE data were extracted for all GLP-1 RA trials and used to calculate the SAE rate for each trial to compare them, accounting for SAE timeframe and sample size variations. The extracted SAE data included SAE counts (the number of subjects who experienced SAEs and subjects at risk) and SAE timeframes. SAE counts were extracted for all trial arms in two steps. First, I extracted the reported SAE counts for trials with results published on ClinicalTrials.gov from the reported event totals AACT table. Second, I extracted the reported SAE counts from journal publications or CSRs into a standardised CSV file for the trials without results posted on ClinicalTrials.gov, trials registered on other registries and unregistered trials. Then, I reviewed the reported SAEs for trials with MACE outcomes for inconsistent reporting of MACE counts in the total SAE counts and updated them accordingly.

SAE timeframes were extracted for all trials following a multi-step approach. First, I extracted SAE timeframes from the reported event totals AACT table. For example, trial NCT02288273 reported the timeframe of SAEs in the reported event totals AACT table.

Second, I obtained SAE timeframes from the result groups AACT table (reported event type) if there was no information in the reported event totals table. For instance, the timeframe of the trial NCT00082381 was unavailable in the reported event totals AACT table; therefore, I obtained the timeframe from the result groups AACT table. Third, I obtained timeframe data from the primary outcome AACT table if there was no information in the result groups AACT table. For example, I obtained the timeframe from the primary outcome table for trial NCT01029886 because the timeframe of SAE data was not in the reported event totals AACT table or the result groups AACT table. Finally, I manually extracted timeframe data for the primary outcome from journal publications into a standardised CSV file for trials without results posted on ClinicalTrials.gov, trials registered on other registries and unregistered trials. SAE timeframes from the AACT tables were extracted as text. I harmonised all extracted SAE timeframes to numeric values and transformed them from days, weeks and months to years.

The primary outcome, population analysis, trial setting, and trial country-level data were extracted as follows. I extracted the primary outcome and the population analysis data as text from the primary outcome AACT table or relevant publications. These data were then harmonised into binary variables called “primary outcome” (MACE or none) and “population analysis” (ITT or PP). Moreover, I manually extracted the trial setting and country-level for all trials from ClinicalTrials.gov and journal publications as binary variables called “trial setting” (single-centre or multicentre) and “country-level” (national or multinational).

3.4.4 Statistical analysis

3.4.4.1 Descriptive statistics

Descriptive statistics were used to summarise data and results. Frequencies and percentages of RCTs were used to describe SAE reporting across registries and journal publications. Moreover, they were used to describe sources of SAE timeframe reporting. SAE rates were calculated for all included trials to compare SAEs between different studies. Initially, the person-time was calculated for all trials using the following formula:

person-year = (years * subjects not experiencing event) + ((years * 0.5) * subjects experiencing event)

The first part of the formula accounts for the total potential exposure time for all subjects over the study duration. The second part of the formula adjusts for those subjects who are no longer at risk once they have had an initial incident event. The factor of 0.5 is an assumption that subjects experiencing SAEs contribute half of a year of exposure on average, suggesting that the SAE rate is constant throughout the trial period. This adjustment is necessary because I am calculating person-years for incident SAEs, and simply multiplying the number of subjects at risk by follow-up time would overestimate the time at risk. The second part of the formula ensures that the time at risk is reduced for subjects who experience an initial incident SAE, as they are no longer considered at risk for subsequent incident SAEs. However, these assumptions may not account for the subjects who may continue participation after experiencing an SAE. Moreover, it may not capture the variability in the timing, severity, and effect of the SAEs.

Then, SAE rates were calculated by dividing the subjects affected by the calculated person-year using the following formula:

SAE rate = subjects affected / person-year

Additionally, a density plot was used to visualise the distribution of SAE rates.

3.4.4.2 Regression analysis

The absence of overdispersion was initially checked by comparing the variance to the mean, where overdispersion occurs if the variance is greater than the mean. Overdispersion was also assessed by calculating the ratio of the residual deviance to the residual degrees of freedom, which indicates overdispersion if the ratio is significantly greater than 1. Overdispersion was found with generalised linear regression models using the Poisson distribution.

Therefore, negative binomial models were fitted to analyse the difference in reported SAE rates between trial registries, trials with results on ClinicalTrials.gov or in publications, trials with MACE outcomes or none, types of population analysis, trial setting and trial

country-level. The outcome variable in these models was the SAE count. An offset was included to account for the variation in person-time.

3.4.4.3 Analysis software

R software version 4.2.2 was used for all analyses of the thesis. R software was used to create tables, plots, and summary statistics. All correlations and regression analyses in the thesis were performed using R software.

3.4.5 Data management

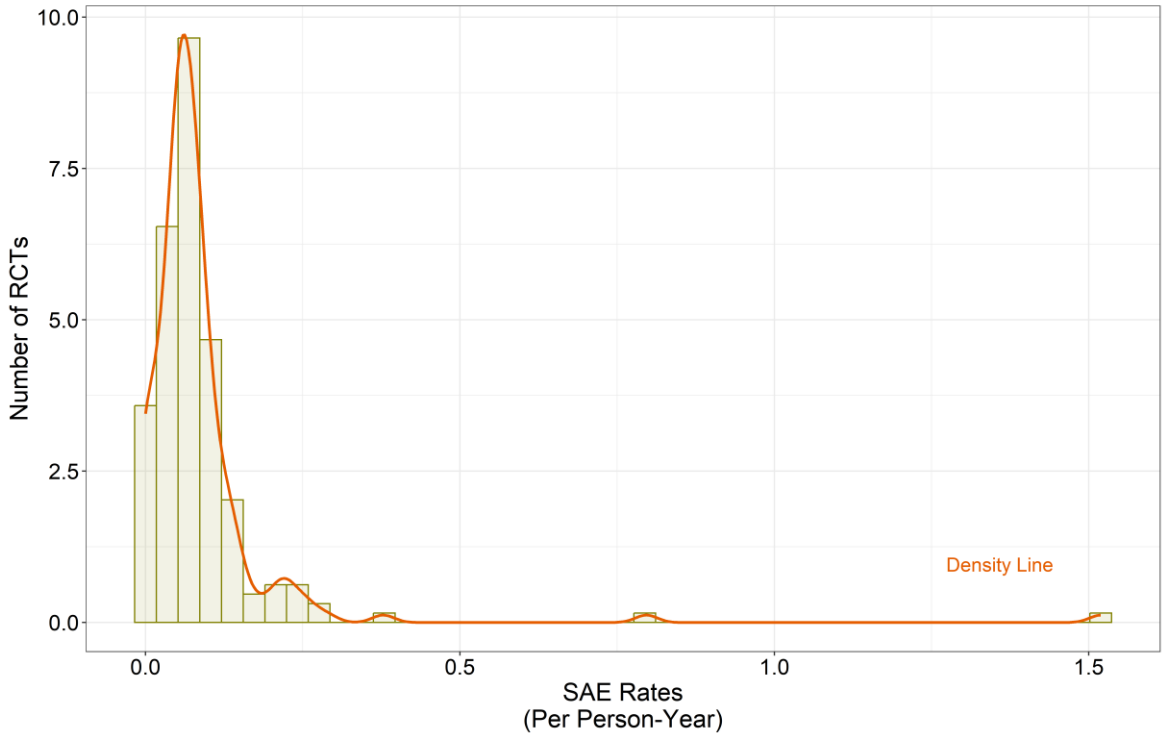
Three data sets were used in this thesis. The first data set was published RCT data, including SAEs, baseline, and trial characteristics data. The second data set was synthesised to score GLP-1 RA trials using the PRECIS-2 tool. The third data set was created to score the eligibility criteria. Excel and CSV files were used to store the data sets. Text files were used to store analysis results. These data sets are maintained by myself and the supervisors of this project. All data sets will be handled following the rules of the data documentation. The data sets are stored in the University of Glasgow cloud storage, local drive storage and the GitHub repository. They are backed up by IT backup and the university cloud storage.

3.5 Results

3.5.1 Summary of trials

In this analysis of GLP-1 RA trials, a total of 196 were assessed for SAE reporting. Figure 3.1 shows that SAE rates varied widely across trials, ranging from zero to 1.5 events per person-year; notably, 21 (10.7%) trials reported zero SAE. Furthermore, the distribution of SAE rates followed a negative binomial distribution. Additionally, 9 (4.6%) trials had MACE as the primary outcome, while the remaining trials had surrogate outcomes, including HbA1C and body weight.

Figure 3.1 The distribution of SAE rates of GLP-1 RA trials



3.5.2 SAE reporting in GLP-1 RA trials

As illustrated in Table 3.2, SAE counts were reported in 186 (95%) of the included trials. Furthermore, they were reported for all GLP-1 RA trials registered on ClinicalTrials.gov, including trials that did not post their results in the registry. Additionally, of the 12 trials registered on other registries, only 3 (25%) trials did not report SAE counts, while only 5 (41.7%) out of 12 unregistered trials reported SAE counts in their publications. As demonstrated in Table 3.3, all SAE timeframes for trials registered on ClinicalTrials.gov were obtained from AACT tables except for trials that did not post results in the registry. However, only 94 (48%) trials reported the timeframes in the reported event totals AACT table. As described in Table 3.4, trials had inconsistent reporting of MACE counts in the total SAE counts. 7 MACE trials (77.8%) reported MACE count in SAE total, whereas two trials (22.2%) did not. Of these two trials, one was registered on ClinicalTrials.gov, and the other was unregistered.

Table 3.2 Summary of SAE count reporting for GLP-1 RA trials

Registration type	Trials	Reported in registry (%)	Only reported in publication (%)	Not reported (%)
ClinicalTrials.gov	172	133 (77.3%)	39 (22.7%)	0
Other registries	12	0	9 (75 %)	3 (25%)
Unregistered	12	0	5 (41.7 %)	7 (58.3%)
Total	196	133 (67.9 %)	53 (27%)	10 (5.1%)

Table 3.3 Summary of SAE timeframe reporting for GLP-1 RA trials

SAE timeframe source	Trials (%)
AACT reported event totals table	94 (48%)
AACT results group (reported event type) table	29 (14.8%)
AACT primary outcome table	10 (5.1%)
Journal publication (primary outcome)	63 (32.1%)
Total	196

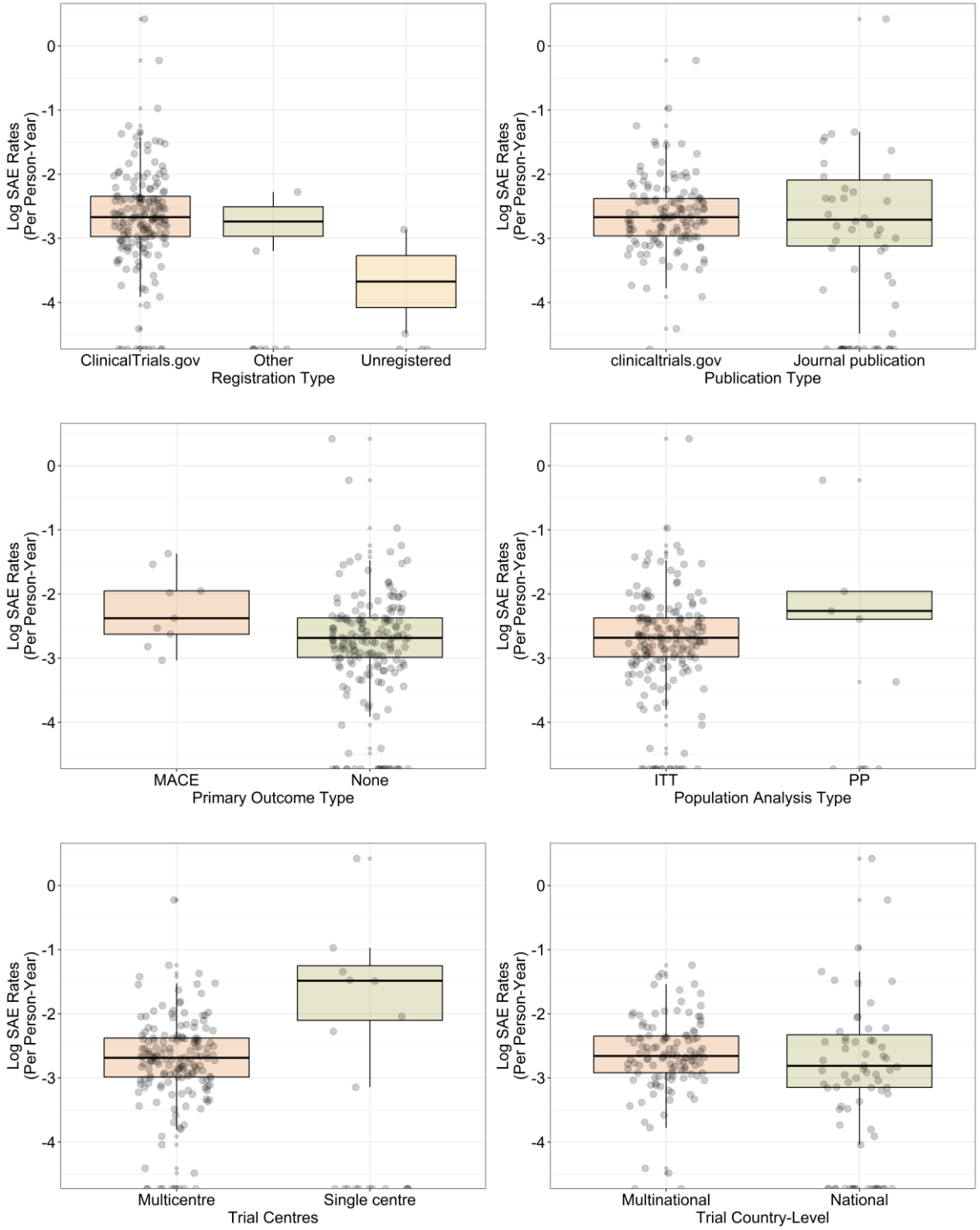
Table 3.4 Reporting of MACE outcomes count in total reported SAE count

Trial ID	MACE count	MACE at risk	SAE count	SAE at risk	MACE reported in total SAEs	Results posted on ctgov
NCT01179048	1303	9340	4674	9340	Yes	Yes
NCT02692716	137	3183	659	3182	Yes	Yes
NCT01394952	1257	9901	4035	9892	Yes	Yes
NCT02465515	493	9463	1954	9432	Yes	Yes
NCT01147250	805	6068	1294	6063	Yes	Yes
NCT01720446	256	3297	1192	3297	Yes	Yes
NCT01144338	1744	14752	2456	14716	No	Yes
NCT01455896	832	4156	633	4144	No	No
NCT01018173	22	2110	103	2110	Yes	No

3.5.3 Factors that associate with SAE reporting

As shown in Figure 3.2, the reported SAE rates for trials registered on other registries and unregistered trials were lower than ClinicalTrials.gov by 69% (IRR=0.31, 95% CI 0.12 – 0.73) and 70% (IRR=0.30, 95% CI 0.12 – 0.76), respectively. However, no difference in SAE rates was found between trials that posted results on ClinicalTrials.gov and those only published in journals (IRR=1.11, 95% CI 0.78 – 1.40). Trials with MACE outcomes reported SAE rates 49% higher than trials with soft outcomes (IRR=1.49, 95% CI 1.45 – 1.54). Surprisingly, Trials with ITT analysis reported SAE rates 47% less than trials with PP analysis (IRR=0.53, 95% CI 0.33 – 0.84). Moreover, multicentre trials reported SAE rates 44% less than single-centre trials (IRR=0.56, 95% CI 0.37 – 0.85). However, no difference in SAE rates was found between multinational and national trials (IRR=1.03, 95% CI 0.82 – 1.29).

Figure 3.2 The association between trial characteristics and the SAE rate in GLP-1 RA trials



3.6 Discussion

This chapter analysed SAE reporting in 196 GLP-1 RA trials. SAE reporting was explored for the included GLP-1 RA trials and compared between registries. Furthermore, the factors associated with the reported SAE rate were examined.

3.6.1 Summary of findings

SAE counts were reported for most trials; however, nearly half of the included trials did not report SAE timeframes within the registry data. Instead, these timeframes had to be estimated based on outcome results. Trials registered on ClinicalTrials.gov had more complete SAE reporting and higher SAE rates than other registries and unregistered trials. Additionally, trials with hard outcomes (i.e. MACE) had higher SAE rates than those with soft outcomes (i.e. HbA1C and body weight). Surprisingly, trials that conducted ITT analysis reported lower SAE rates than trials with PP analysis, and multicentre trials had lower SAE rates than single-centre trials. No difference in SAE rates was found between trials with results on ClinicalTrials.gov and those only published in journals. Moreover, no difference in SAE rates was found between multinational and national trials.

3.6.2 Interpretation

Most trials reported SAEs, indicating sufficient reporting of SAE counts. However, nearly half of the included trials did not explicitly report SAE timeframes. Therefore, timeframes of result groups (reported event type) and primary outcomes were used as a proxy of SAE timeframes. Although this is not ideal, as some trials may have longer SAE timeframes than for result groups and primary outcomes, it provides a practical approach that enables SAE rate calculation. Sufficient SAE reporting indicates that SAE rates can be calculated for most trials. However, two trials did not include MACE count in the total SAEs count, indicating inconsistent reporting across trials with hard serious outcomes. This discrepancy could be due to the lack of clear guidance on registries and journals reporting hard serious outcomes. To address this discrepancy in reporting, I included the MACE count in the total SAEs count for these two trials. While checking the inconsistency will avoid bias for this

thesis, it presents a challenge for the broader use of the SAE rate as a metric; all trials with MACE outcomes should be checked for inconsistencies in reporting total MACE counts in the total counts.

The higher SAE rates of trials registered on ClinicalTrials.gov could be attributed to better SAE reporting practices in these trials. Furthermore, trials registered on ClinicalTrials.gov could be larger and recruit a more diverse population than unregistered trials. Trials with MACE outcomes reported higher SAE rates, likely due to the inclusion of patients with cardiovascular comorbidities (Ferreira-González et al., 2007; Vestergaard Kvist et al., 2021). MACE trials include such patients partly to improve power to detect the effect on MACE (i.e. they have higher SAE rates by design) (Wise et al., 2019; Wang et al., 2022).

Additionally, this study found that multicentre trials reported lower SAE rates than single-centre trials. Moreover, trials with PP analysis reported higher SAE rates than those with ITT analysis. These findings could be attributed to that some single-centre trials and trials with PP analysis may focus on a specific high-risk population, such as patients with chronic kidney disease (CKD). Furthermore, the sample sizes were disproportionate, with 15 (7.7%) for PP trials and 26 (13.3%) for single-centre trials, which may have led to biased results due to high variance in the PP and single-centre trials. Therefore, larger sample sizes would produce different findings.

3.6.3 Comparison with previous literature

The finding that SAEs were reported for most trials was consistent with other studies that examined SAE reporting in clinical trials of various diseases and interventions, such as cancer drugs, COVID-19 vaccines and anaesthesia (Riveros et al., 2013; Paludan-Müller, Créquit and Boutron, 2021; Yuniar et al., 2022; Taillefer de Laportalière et al., 2023). A previous study assessed SAE reporting of trials registered on ClinicalTrials.gov and found that 199 (99%) of 202 trials reported SAEs, whether on ClinicalTrials.gov or journal publications (Riveros et al., 2013). Similarly, a study of harm reporting in cancer trials found that 37 (100%) studies registered on ClinicalTrials.gov rereported SAEs (Paludan-Müller, Créquit and Boutron, 2021). Moreover, another study reported that SAEs in COVID-19 vaccine clinical trials were available in 100% of 61 trials (Yuniar et al., 2022).

Furthermore, a systematic review of harm reporting in esketamine trials found that 9 (90%) trials reported SAEs (Taillefer de Laportalière et al., 2023). However, a study found that SAEs were available in 378 (48.9%) of 773 exercise therapy trials, indicating a low SAE reporting rate (Niemeijer et al., 2020). This may be due to a lack of standardisation of SAEs reporting in exercise therapy trials or because exercise therapy trials may not be viewed as high-risk interventions.

The finding that SAE timeframes were only reported for 94 (48%) of included studies was consistent with the previous literature. A study of AE reporting in solid tumour trials found that 109 (63%) trials reported AE timeframes (Sivendran et al., 2014). Similarly, a systematic review of AE reporting in cancer therapy trials found that 171 (53%) studies reported AE timeframes (Péron et al., 2013). Moreover, a study of AE reporting in pharmacological intervention trials found that 95 (57.2%) trials did not report AE timeframes (Phillips et al., 2019). However, a systematic review of AE reporting in invasive pain treatment found that AE timeframes were reported for 95 (72%) studies (Williams et al., 2016). Underreporting explicit AE timeframes may be more widespread across different therapeutic areas, suggesting the need for consistent AE timeframe reporting.

3.6.4 Strengths and limitations

This analysis was based on trials from different registries, which provided a variety of RCTs. This is the first study to examine SAE reporting in GLP-1 RA trials. One limitation is that this study focuses on GLP-1 RAs, which may limit the generalisability of the findings to trials of other drug classes or diseases, especially trials of toxic interventions. Another potential limitation is that I only looked at the journal publication when I could not find the SAE data on ClinicalTrials.gov; I did not assess for any inconsistency in reporting between the different data sources.

3.6.5 Implications

Most trials reported sufficient SAE data, indicating the feasibility of calculating SAE rates. The outcomes of trials should be considered when analysing SAEs because types of outcomes could be associated with different levels of SAE rates. The findings of this chapter have implications for the analysis of the following chapters. SAE counts and relevant timeframes will be calculated and used in subsequent analysis. The calculated SAE rates will be compared across trial arms as well as with the PRECIS-2 tool and eligibility criteria. Future studies may explore the reporting of specific SAEs, such as cardiovascular events.

3.7 Conclusion

SAE reporting across all GLP-1 RA trials was reasonably sufficient, indicating that SAE rates could be calculated and compared across trials. Moreover, it is reliable regarding data availability, especially if the trials were registered on ClinicalTrials.gov.

Chapter 4 Comparison of SAE rates between trial arms: GLP-1 RA trials as an exemplar

4.1 Chapter overview

This chapter will compare SAE rates between intervention and control arms. This is important as, if using SAE as a marker of representativeness, a decision has to be made around whether to analyse all participants or only those from a single arm. The former gives greater statistical power but could lead to bias if the rates of SAE differ between arms. Therefore, it is important to assess if combining SAE rates of intervention and control arms is a justifiable choice or if this could lead to bias. This chapter assesses this issue.

4.2 Background

4.2.1 The difference in reported SAE rates between trial arms

The documentation, recording and reporting of SAE data in RCTs and real-world records indicate its potential as a measure of trial representativeness. However, whether intervention and control arms would report similar SAE rates or may differ is unclear as they are likely to depend on the nature of the intervention and control arms. Control arms in RCTs, including placebo and active comparator, provide the baseline for comparison with intervention groups and ensure that the results are accurate and unbiased (Gupta and Verma, 2013; Sil et al., 2019). For some interventions, where the intervention itself rarely directly causes SAEs, the intervention and control arms are likely to report similar SAE rates. However, trials of toxic interventions may report more SAEs in intervention arms compared to placebo (Wolff et al., 2022). Therefore, the choice between the arm level or total SAEs to measure trial representativeness depends on the magnitude of the difference between the arms. If the difference between arms is fairly narrow, total SAEs could be used in the measurement, increasing the statistical power.

4.3 Aim and objectives

This chapter will examine the feasibility of using SAE rates as a measure of trial representativeness. Specifically, it will determine whether SAE rates of intervention and control arms are similar in trials of GLP-1 RA and, therefore, whether, for the purposes of analysing SAE rates, they can be combined. This approach will overcome low event numbers and increase the statistical power, but it would only be valid if SAE rates in intervention and control arms were similar.

The objectives of this chapter are:

1) To examine the difference in reported SAE rates between placebo and intervention arms across all GLP-1 RA trials.

2) To examine the difference in reported SAE rates between active comparator and intervention arms across all GLP-1 RA trials.

4.4 Methods

4.4.1 Study selection and data extraction

196 GLP-1 RA studies were identified for this thesis. The study selection and SAE data extraction are described in detail in 3.4.1 and 3.4.3.

4.4.2 Data sources

Data sources used for this analysis are described in 3.4.2.

4.4.3 Statistical analysis

4.4.3.1 Descriptive statistics

Descriptive statistics were used to summarise data and results. SAE data were summarised for all arms, including subjects at risk, subjects affected, and person-years.

4.4.3.2 Regression analysis

As previously explained in 3.4.4.2, overdispersion was examined by comparing the variance to the mean and calculating the ratio of the residual deviance to the residual degrees of freedom. Overdispersion was found with generalised linear regression models using the Poisson distribution.

Therefore, separate negative binomial models were fitted to analyse the difference in reported SAE rates between placebo and, intervention arms and active comparator and intervention arms. At the beginning of this analysis, trials were grouped according to their control arms as trials with placebo arms and trials with active comparator arms. The outcome variable in these models was the SAE count. A categorical variable indicated the arm group, either placebo or intervention for the trials with placebo arms, and active comparator or intervention for the trials with active comparator arms. An offset was included to account for the variation in person-time.

4.5 Results

4.5.1 Summary of trials

In this analysis of GLP-1 RA trials, 10 (5.1%) trials were excluded from this analysis due to missing SAE data. Regarding treatment arms of the 186 trials included for arms analysis, 90 trials had placebo arms, while 96 trials had only active comparator arms (see Table 4.1 for SAE data of each arm).

Table 4.1 Summary of the reported SAE data for GLP-1 RA trials

Arms	Subjects affected	Subjects at risk	Person years
Intervention	11160	53853	140293.5
Placebo	10576	41661	125116.5
Intervention	1618	23596	27563.04
Active comparator	2523	36160	49608.04
Total	25877	155270	342581.1

4.5.2 SAE reporting in GLP-1 RA trials

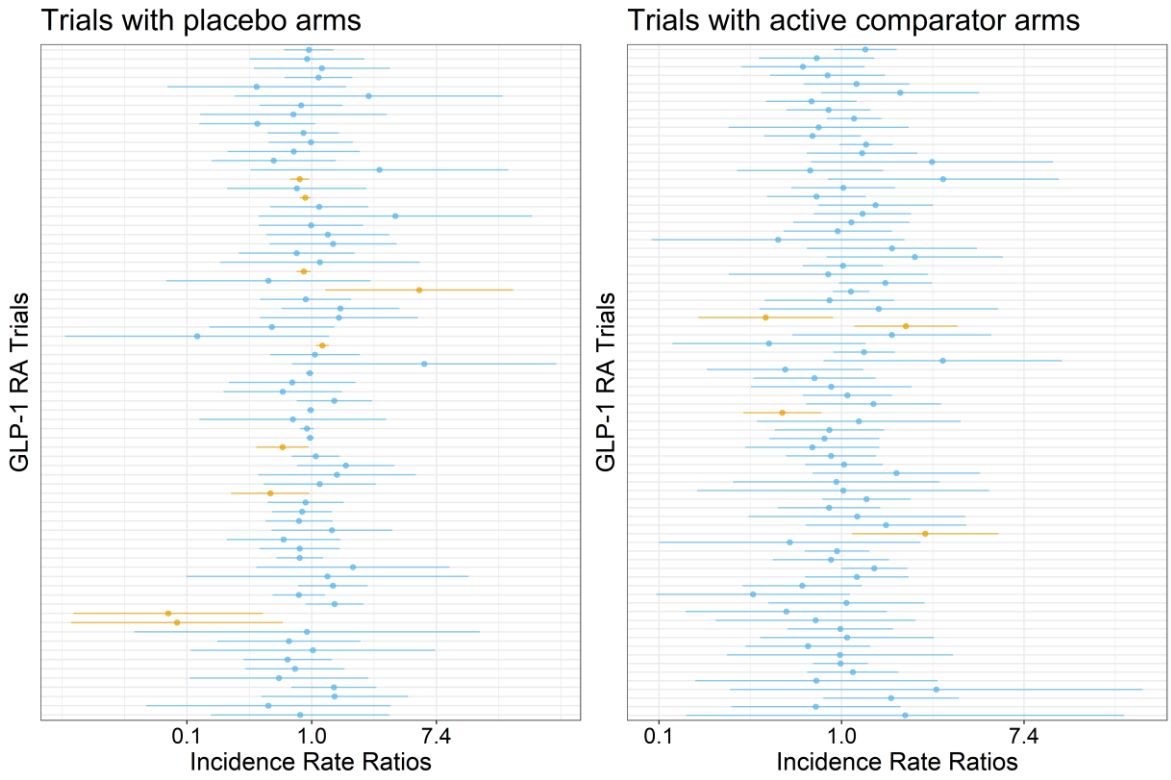
SAE reporting in GLP-1 RA trials is described in 3.5.2.

4.5.3 SAE rates across GLP-1 RA trial arms

The difference in SAE rates between placebo and intervention arms across all GLP-1 RA trials with placebo arms was analysed. SAE rates were 4% lower in the intervention arms than in the placebo arms. However, the 95% confidence interval included the null (IRR=0.96, 95% CI 0.78 – 1.18). Moreover, there was generally little difference between trials (see Figure 4.1).

The difference in SAE rates between active comparator and intervention arms across all GLP-1 RA trials with active comparator arms was estimated. SAE rates were 6% lower in the active comparator arms than in the intervention arms. However, the 95% confidence interval included the null (IRR=0.94, 95% CI 0.78 – 1.15). In addition, there was generally little difference between active comparator and intervention arms across trials (see Figure 4.1).

Figure 4.1 The difference in SAE rates between intervention and control arms



4.6 Discussion

This chapter analysed the difference in reported SAE rates between placebo and intervention arms and active comparator and intervention arms across all GLP-1 RA trials to know if SAE rates of intervention and control arms can be combined.

4.6.1 Summary of findings

SAE rates were compared between control and intervention arms across all GLP-1 RA trials. No difference in SAE rates was found between placebo and intervention arms or between active comparator and intervention arms.

4.6.2 Interpretation

The finding that SAE rates did not differ between trial arms indicated that most SAEs were not directly attributed to trial interventions. They could be mainly attributed to other factors, such as older age and multiple long-term illnesses. Therefore, SAE rates of intervention arms, control arms, or combined can be used to measure trial representativeness. However, using the total SAE rate may be more appropriate because it overcomes low event numbers and increases the statistical power. Moreover, not all trials use placebo or active comparator arms in their design. However, using SAE rates from the intervention arms in trials of potentially toxic drugs may not be a reliable approach because SAEs are usually higher in intervention arms than placebo arms. SAE rates from the control arms for trials of toxic interventions may provide a better comparator to SAE rates in routine care because they are usually active treatments generally used in routine care (Hanlon et al., 2022). Although combining SAE rates from intervention and control arms may increase the statistical precision, careful implication is required, especially when there is evidence that the intervention is likely to cause SAEs.

4.6.3 Comparison with previous literature

Several studies examined the differences between placebo and intervention arms and found no difference in SAE rates, similar to the findings of this study. A previous study compared SAE rates between placebo and intervention arms across naltrexone trials and found no difference (IRR=0.84, 95% CI 0.66 – 1.06) (Bolton et al., 2019). Furthermore, a study estimated the difference in SAE rates between placebo and intervention arms across different vaccines trials and found no difference between arms across all types, inactivated vaccine (IRR=0.84, 95% CI 0.68 – 1.06), mRNA vaccines (IRR=1.10, 95% CI 0.91 – 1.33), Protein-subunit vaccines (IRR=1.01, 95% CI 0.66 – 1.55), Viral vector vaccines (IRR=0.82, 95% CI 0.64 – 1.05) (Kwasi et al., 2022). Another study estimated the difference in SAE rates between control and intervention arms in cannabinoids trials and found no significant difference between arms (IRR=1.04, 95% CI 0.78 – 1.39) (Tongtong Wang, Jean-Paul Collet, Stan Shapiro, 2008).

Additionally, the finding that SAE rates did not differ between intervention and active comparator arms was consistent with other studies. Another study compared SAE rates between trial arms (intervention vs placebo and intervention vs active comparator) across 21 index conditions and found no difference in SAE rates (IRR men 0.91; 95% CI 0.81 – 1.02, IRR women 0.99; 95% CI 0.87 – 1.10) (Hanlon et al., 2022). Moreover, a study assessed the difference in SAE rates between trial arms (active comparator vs intervention and placebo vs intervention) in celecoxib in osteoarthritis trials and found no difference in SAE rates between active comparator and intervention arms (IRR=1.02, 95% CI 0.91 – 1.15). However, intervention arms showed lower SAE rates than placebo (IRR=0.67, 95% CI 0.46 – 0.98) (Moore et al., 2005). This difference in SAE rates could be attributed to the use of non-selective nonsteroidal anti-inflammatory (NSAID) such as diclofenac in the placebo arm due to insufficient symptom control (Varga, Sabzwari and Vargova, 2017). Another study found that SAEs were 1.3 times greater in cancer patients who received cancer drugs than those who received a placebo (Wolff et al., 2022). This notable large difference could be associated with drug-related effects because cancer drugs are relatively toxic and could lead to more SAEs compared to other interventions. Moreover, it could be related to the disease complexity and patients' vulnerability, where the participants are likely to have more SAEs compared to healthy participants, especially during the administration of cancer drugs.

4.6.4 Strengths and limitations

This analysis was based on trials from different registries, which provided a variety of RCTs. Moreover, this study examined the difference in SAE rates between intervention and control, and between intervention and active comparator, ensuring that the SAE rates are not primarily attributed to the intervention and that SAE rates can be combined to increase the statistical precision. One limitation is that this study focuses on GLP-1 RAs, which may limit the generalisability of the findings to trials of other drug classes or diseases, particularly trials of toxic interventions.

4.6.5 Implications

The comparison of SAE rates between trial arms provided valuable information on the accuracy of the measurement, particularly regarding the use of intervention or control arms or total SAE rates. It suggests that, for the purposes of this specific analysis, it may be reasonable to combine SAE rates from intervention and control arms of trials. These findings have implications for the analysis of the following chapters, which will compare the SAE rate with the PRECIS-2 tool as measures of trial representativeness and study the association between SAE rates and eligibility criteria.

4.7 Conclusion

No differences in SAE rates were found between intervention and placebo arms or between intervention and active comparator arms. These comparisons concluded that SAEs are not primarily attributed to the trial interventions, and combining SAE rates of intervention and control arms is feasible for the purposes of using it as a measure of trial representativeness.

Chapter 5 The pragmatism GLP-1 RA trials: A retrospective analysis using the PRECIS-2 tool

5.1 Chapter overview

This chapter will retrospectively assess the pragmatism of GLP-1 RA trials using the PRECIS-2 tool. It will also explore the challenges of using this tool, which is important to understand if it is to be used as a proxy for representativeness and compared to other measures, such as the SAE rate, later in this thesis.

5.2 Background

5.2.1 The PRECIS-2 tool

There has been an increased recognition that more attention needs to be paid to differentiate between RCTs based on their aims and applicability (Akobeng, 2005; Gartlehner et al., 2006). The pragmatic approach of RCTs aims to assess the effectiveness of interventions under usual care conditions, whereas the explanatory design of RCTs aims to investigate interventions under ideal conditions to examine their efficacy (Zwarenstein et al., 2008). The original PRECIS tool and the current PRECIS-2 tool were developed to help trials match their designs with their aims, explanatory or pragmatic (Loudon et al., 2017). Although the original PRECIS tool was broadly used to assess the pragmatism of RCTs, it was criticised for the lack of interrater reliability and the need for a numeric scale (Koppenaal et al., 2011; Glasgow et al., 2012). Moreover, users of PRECIS requested an additional explanation of the domains. Therefore, in 2015, the PRECIS-2 tool was introduced to fill the gaps and add the required adjustments (Loudon et al., 2015).

The PRECIS-2 tool consists of 9 domains: eligibility criteria, recruitment, setting, organisation, flexibility (delivery), flexibility (adherence), follow-up, primary outcome, and primary analysis. Each domain gets a score based on a rating scale of 5; 1 refers to a very explanatory design, and 5 indicates a very pragmatic design of the domain (Loudon et

al., 2015). The PRECIS-2 tool is intended to assess the design of trials prospectively. Still, investigators used the tool for retrospective analysis of trial pragmatism (Loudon et al., 2017).

5.2.2 The suitability of the PRECIS-2 tool as a comparator with the SAE rate

The absence of a universally accepted gold standard measure for trial representativeness is a fundamental challenge, making evaluating and comparing new measures inherently difficult (Sun et al., 2021). Current measures of trial representativeness are either difficult to apply or lead to inconclusive findings, which can limit their regular use to assess trial representativeness (Dekkers et al., 2010). The PRECIS-2 tool offers a more straightforward approach by evaluating nine dimensions that reflect the pragmatism of the trial (Norton et al., 2021). Although the PRECIS-2 tool was not originally designed as a representativeness tool, it can indirectly contribute to understanding trial representativeness by focusing on the pragmatism of the trial. For example, a trial that is considered pragmatic according to the PRECIS-2 score is likely to be representative due to its broad eligibility criteria for participation and settings that reflect usual care (Lipman et al., 2017; Usman et al., 2022).

Therefore, The PRECIS-2 tool could provide the best possible comparison with the SAE rate as measures of trial representativeness. However, using the PRECIS-2 tool for retrospective assessment requires adaptation and protocolisation of its domains to minimise subjectivity in the assessment and enable a fair comparison with the SAE rate. In addition, the PRECIS-2 tool could be limited by the availability of the rationale for scoring domains, which may affect the accuracy of the measurement. Therefore, information required to score the PRECIS-2 domains will be obtained from different sources to mitigate this limitation.

5.3 Aim and objectives

In the attempt to explore whether the SAE rate can reflect trial representativeness, the SAE rate will be compared to an existing proxy measure related to trial representativeness, the

PRECIS-2 tool. This chapter aims to assess the pragmatism of GLP-RA trials using the PRECIS-2 as objectively as possible to ensure a fair comparison with the SAE rate.

The objectives of this chapter are:

1) To retrospectively assess the pragmatism of GLP-1 RA trials using the PRECIS-2 tool.

2) To identify the challenges associated with using the PRECIS-2 tool for retrospective assessment.

5.4 Methods

5.4.1 Study selection

196 GLP-1 RA studies were identified for this thesis. The study selection is described in detail in 3.4.1.

5.4.2 Data sources

Data sources used for this analysis are described in 3.4.2.

5.4.3 Protocolisation and operationalisation of the PRECIS-2 tool for GLP-1 RA trials

Loudon et al., 2015 and the PRECIS-2 website described the PRECIS-2 tool, its domains and assessment guidance. The nine components of the tool are eligibility criteria, recruitment, setting, organisation, flexibility (delivery), flexibility (adherence), follow-up, primary outcome, and primary analysis. Each domain is rated on a scale from 1 to 5: 1-very explanatory, 2-rather explanatory, 3-equally explanatory/pragmatic, 4-rather pragmatic and 5-very pragmatic. The "eligibility criteria" domain assesses how restrictive the eligibility criteria are for those who would be candidates for the intervention in routine

care. The "recruitment" domain assesses how and where trial participants are recruited. The "Setting" domain examines where the trial is conducted. The "organisation" domain assesses how the intervention is organised and implemented in the trial. The "flexibility (delivery)" domain assesses the flexibility in how the intervention is delivered in the trial. The "flexibility (adherence)" domain examines the flexibility of the participant's adherence to the intervention. The "follow-up" domain assesses the nature and extent of follow-up procedures in the trial. The "primary outcome" domain considers the type of the primary outcome and its relevance to the patient and the provider. The "primary analysis" domain assesses the trial based on the type of population analysis used in the trial.

The PRECIS-2 tool must be applied as objectively as possible to ensure a fair comparison with the SAE rate. Therefore, the PRECIS-2 website and recommendations for assessment were reviewed, and an assessment protocol was drafted based on the recommendations of Loudon et al., 2015. This protocol was developed and refined based on weekly meetings with my supervisors and the research team. The protocolisation and operationalisation of the PRECIS-2 tool were challenging because the tool was developed and designed to assess trials prospectively. As described in Table 5.1, the protocol for assessing trials using the PRECIS-2 tool included clarification and interpretation for each of the nine domains in the context of T2DM to ensure standardised and objective scoring across all trials. Moreover, this protocol provides detailed guidance on what constitutes a pragmatic or explanatory approach for each domain. Table 5.1 includes examples for each domain to show how different trial designs may lead to different PRECIS-2 scores.

Following the protocolisation and operationalisation of the PRECIS-2 domains, a pilot assessment with a small set of trials was conducted to investigate the feasibility of using the tool to retrospectively assess the pragmatism of GLP-1 RA trials. The pilot assessment indicated issues with the availability of information required for the assessment and protocol interpretation. Moreover, the pilot assessment suggested that a diabetologist's insight is needed to review and update the protocol for assessing clinical aspects of some domains, particularly eligibility criteria, flexibility (delivery) and follow-up. Dr Elaine Butterly, a diabetologist in NHS Greater Glasgow and Clyde, was approached and agreed to contribute her clinical knowledge to the operationalisation of the clinical aspects of some domains of the PRECIS-2 tool.

The involvement of the diabetologist was crucial to ensure the standardisation of scoring clinical aspects across all trials and the alignment of assessment with the usual care of T2DM. Moreover, the diabetologist reviewed the scoring of clinical aspects of some domains for all the included studies to ensure that the assessment of these domains aligns with the usual care. Furthermore, a sample of 10 trials was drawn and scored by another PhD student, Khalid Alsallumi, to ensure that the scoring was consistent across trials.

Table 5.1 The operationalisation of the PRECIS-2 for GLP-1 RA trials

Domain	Criteria of assessment
Eligibility criteria	• If no information is available about the eligibility criteria, the score will be "missing". • Score 1 for trials with highly restrictive criteria such as specific age groups or specific comorbidity (e.g., including diabetic patients with only CKD). • Score 1 for trials that apply four or more of the below criteria that reduce the score by 1 point. • Score 2 for trials that apply three of the below criteria that reduce the score by 1 point. • Score 3 for trials that apply two of the below criteria that reduce the score by 1 point. • Score 4 for trials that include anyone with the condition of interest but only exclude subjects likely to have poor adherence to the intervention (e.g. mental disability). • Score 4 for trials that include anyone with the condition of interest but only exclude subjects with comorbid conditions (e.g. cardiovascular disease). • Score 4 for trials that include anyone with the condition of interest but only exclude subjects using interventions other than the product of the trial (e.g. antihypertensive drugs). • Score 4 for trials that include anyone with the condition of interest but only exclude subjects eligible for the trial intervention in usual care due

	to challenges irrelevant to the delivery of the intervention (e.g. inaccessible trial centre). ● Score 4 for trials that include anyone with the condition of interest but only exclude subjects eligible for the intervention of the trial in usual care and likely to not respond to the intervention (e.g. those with sever disease conditions). ● Score 4 for trials that include anyone with the condition of interest but only use measures or tests to exclude subjects (e.g. exclude based on C-peptide or MRI). ● Score 4 for trials that include anyone with the condition of interest but only exclude subjects dependent on others' help. ● Score 5 for trials that include anyone with the condition of interest but only exclude severely comorbid subjects who are not likely to receive the intervention in routine care (e.g. renal failure). ● Score 5 for trials that include anyone with the condition of interest and likely to receive the intervention in routine care.
Example of pragmatic eligibility criteria	● NCT02229396 "Inclusion criteria: Has a diagnosis of T2DM. Has HbA1c of 8.0% to 12.0%, inclusive, at Visit 1 and Visit 2. Treated with a stable dose of metformin ≥1500 mg/day for at least 2 months prior to Screening. Exclusion criteria FPG ≥280 mg/dL (15.6 mmol/L). Serum calcitonin concentration ≥40 pg./mL (≥40 ng/L) at Visit 1 (Screening) Clinically significant abnormal free T4 values or patients needing initiation or adjustment of thyroid treatment according to the investigator. Abnormal thyroid stimulating hormone (TSH) value at Screening will be further evaluated by free T4. Patients with clinically significant abnormal free T4 values will be excluded. Known active proliferative retinopathy. History of, or currently have, acute or chronic pancreatitis, or have triglyceride concentrations ≥500 mg/dL (≥5.65 mmol/L) at Visit 1 History or presence of inflammatory bowel disease or other severe GI diseases, particularly those which may impact gastric emptying, such as gastroparesis or pyloric stenosis. History of gastric bypass surgery or gastric banding surgery, or either procedure is planned during the time period of the study. Current use of gastric balloons is also excluded."

Example of explanatory eligibility criteria	• NCT01617434 “Inclusion Criteria: Diagnosed with type 2 diabetes for at least 180 days prior to screening and treated with stable basal insulin analogue dose of minimum 20 U/day with or without stable metformin equal to or above 1500 mg/day for at least 8 weeks prior to screening (defined as insulin adjustments less than 10% during the past 8 weeks as assessed by the investigator). HbA1c (glycosylated haemoglobin A1c) 7.0-10.0% (both inclusive). Body mass index (BMI) 20-45 kg/m^2 (both inclusive). Exclusion Criteria: Female of child-bearing potential who is pregnant, breast-feeding or intending to become pregnant. Recurrent severe hypoglycaemic episodes or hypoglycaemic unawareness. Treatment with glucose-lowering agent(s) other than stated in the inclusion criteria in a period of 12 weeks prior to screening. Impaired liver or renal function. Uncontrolled treated or untreated hypertension (systolic blood pressure (SBP) equal to or above 180 mmHg and/or diastolic blood pressure (DBP) equal to or above 100 mmHg). Any clinically significant disorder, except for conditions associated with type 2 diabetes history which in the investigator's opinion could interfere with results of the trial. Known or suspected abuse of alcohol or narcotics.”
Recruitment	• If no information is available about the method or the nature of the recruitment site (usual care or other than usual care), the score will be "missing". • Score 1 for trials that are single centres and use other clinics (for example, hospitals, speciality clinics and research centres) to recruit participants for the intervention that will be mainly used in primary clinics. • Score 1 for trials that use advertisements or telephone calls to recruit subjects. • Score 1 for trials that use incentives such as money to recruit participants. • Score 2 for trials that are national, multicentre, and use other clinics (for example, hospitals, speciality clinics and research centres) to recruit their participants for the intervention that will be mainly used in primary care clinics. • Score 2 for trials that search medical records and then mail eligible participants.

	• Score 3 for trials that are multinational, multicentre and use other clinics (for example, hospitals, speciality clinics and research centres) to recruit their participants for the intervention that will be mainly used in primary care clinics. • Score 3 for trials that are single centres and use usual care clinics to recruit participants. • Score 4 for trials that are national, multicentre, and use usual care clinics to recruit their participants. • Score 5 for trials that are multinational, multicentre, and use usual care clinics to recruit their participants without any additional effort.
Example of pragmatic recruitment	• NCT00641056 "Patients were identified, under direction from the site principal investigators, from patient populations at all trial sites. Potential participants were subsequently recruited according to standard local practices."
Example of explanatory recruitment	• NCT01744236 "Volunteers with type 2 diabetes will be recruited using established recruitment methods: (1) participants in the previous studies of the VU University Diabetes Centre will be contacted (if informed consent was obtained); (2) advertisements in local newspapers, folders and posters; (3) affiliated healthcare workers (internal medicine, general practitioners) will inform patients of the existence of this study; and (4) websites."
Setting	• If no information is available about the nature of the setting (usual care or other than usual care), the score will be "missing". • Score 1 for trials conducted in a single centre and site other than the usual care clinic (for example, hospitals, speciality clinics and research centres) for the intervention that will be mainly used in usual care clinics. • Score 2 for trials conducted in national, multicentre, and sites other than usual care clinics (for example, hospitals, speciality clinics and research centres) for the intervention that will be mainly used in usual care clinics. • Score 3 for trials conducted in a single centre and usual care clinics. • Score 3 for trials conducted in multinational, multicentre, and sites other than usual care clinics (for example, hospitals, speciality clinics and

	research centres) for the intervention that will be mainly used in usual care clinics. ● Score 4 for trials conducted in national, multicentre, and usual care clinics. ● Score 5 for trials conducted in settings that mimic the usual care. ● Score 5 for trials conducted in multinational, multicentre, and usual care clinics.
Example of pragmatic setting	● NCT02730377 "The study was conducted in conducted in the primary care setting at 232 multinational sites."
Example of explanatory setting	● NCT01744236 "The study was performed at the VU University Medical Center (VUmc), the Netherlands between July 2013 and August 2015. All examinations will be performed at the Clinical Research Unit (CRU) of the Department of Internal Medicine/Diabetes Centre of the VU University Medical Center in Amsterdam, the Netherlands."
Organisation	● If no information is available about the organisation, the score will be "missing". ● Score 1 for trials that apply two of the below criteria that reduce the score by 2 points. ● Score 3 for trials that provide training or education that is not required in usual care. ● Score 3 for trials that require certification or experience that is not required in usual care (e.g. specialised nurses). ● Score 3 for trials that use additional resources or tests that are not used in usual care (e.g. MRI). ● Score 3 for trials that use an additional staff member not part of the usual care (e.g., radiologists). ● Score 4 for trials that increase their resources to have more frequent assessments or procedures than usual care (e.g. additional assessment sites).

	• Score 5 for trials conducted without using additional resources, staff or require certification that is not required in usual care.
Example of pragmatic organisation	• NCT01394952 “Investigators were advised to promote a healthy lifestyle and to manage glucose concentrations according to local guidelines and were free to add any glucose-lowering drug apart from another GLP-1 receptor agonist or pramlintide. Management of blood pressure, lipids, other cardiovascular risk factors, and medical conditions was at the discretion of either the study investigator or the patient’s usual physician(s) as informed by current country guidelines.”
Example of explanatory organisation	• NCT01744236 “Primary outcome measure is resting heart rate variability assessed with a beat-to-beat heart rate monitor and spectral analyses software. For the acute intervention study, this is measured by exocrine secreted volume assessed by secretin-enhanced MR cholangiopancreatography. For the 12-week study, this is measured by faecal elastase-1 levels. Postvoiding bladder residue will be assessed using ultrasonic bladder scan.”
Flexibility (delivery)	• If no information is available about the delivery of the intervention, the score will be "missing". • Score 1 for trials that apply a highly defined protocol for the delivery of the intervention. • Score 1 for trials that apply four of the below criteria that reduce the score by 1 point. • Score 2 for trials that apply three of the below criteria that reduce the score by 1 point. • Score 3 for trials that apply two of the below criteria that reduce the score by 1 point. • Score 4 for trials that apply restrictions on the number or the type of co-interventions. • Score 4 for trials that provide specific direction on managing potential events related to the intervention. • Score 4 for trials that provide specific direction to enhance the delivery of the intervention.

	• Score 4 for trials that strictly define the timing of the intervention. • Score 4 for trials that use additional interventions unavailable in usual care. • Score 5 for trials that do not specify permitted co-interventions. • Score 5 for trials that leave the details of the delivery procedure to the health care provider. • Score 5 for trials that only apply a specified protocol of dose escalation to avoid or reduce side effects (e.g. dose escalation for GLP-1 RAs to minimise gastrointestinal side effects).
Example of pragmatic flexibility (delivery)	• NCT02305381 "Patients received semaglutide (0.5 or 1.0 mg subcutaneously) or volume-matched placebo once weekly for 30 weeks followed by a 5-week follow-up period. Study medication was administered following a fixed dose-escalation regimen. For 0.5 mg, the maintenance dose was reached after 4 weeks of 0.25 mg semaglutide or matching placebo once weekly. For 1.0 mg, the maintenance dose was reached after 4 weeks of 0.25 mg, followed by 4 weeks of 0.5 mg semaglutide or matching placebo once weekly."
Example of explanatory flexibility (delivery)	• UMIN000005331 "Patients were randomly assigned to receive once daily liraglutide (0.3 mg, 0.6 mg for 7 days followed by 0.9 mg/day) or glargine according to their age and basal FMD using computer software. All patients were encouraged to continue diet and exercise therapy during the study. Treatment was performed at each respective medical care centre for 14 weeks, then endothelial function and serum biomarkers were measured again at the end of the study at Hokkaido University Hospital using the same parameters as at study entry."
Flexibility (adherence)	• If no information is available about the adherence to the intervention, the score will be "missing". • Score 1 for trials that apply a pre-screening (run-in or lead-in period) to assess adherence and exclude non-adherent patients. • Score 1 for trials that exclude non-adherent participants. • Score 2 for trials that measure and monitor the adherence to the intervention.

	• Score 5 for trials that do not apply measures or enforce the adherence of subjects to the intervention. • Score 5 for trials that permit full flexibility on how and when subjects take the intervention.
Example of pragmatic flexibility (adherence)	• NCT02863419 “Subjects should stay in the trial irrespective of lack of adherence to randomised treatment, lack of adherence to visit schedule, missing assessments or trial product discontinuation for any reason.”
Example of explanatory flexibility (adherence)	• NCT02597049 “Compliance with study drug (dulaglutide or placebo) was determined by review of the patient diary and return of unused study drug at every visit. Poorly compliant patients received additional training and instructions as required. 1 subject was excluded for non-compliant with Study Visits.”
Follow-up	• If no information is available about the follow-up, the score will be "missing". • Score 1 for trials that apply two of the below criteria that reduce the score by 3 points. • Score 2 for trials with longer visits or more extensive data collection than the usual care. • Score 2 for trials with follow-up visits scheduled based on the occurrence of an event that may result in the primary end point. • Score 2 for trials that contact patients if they do not attend their scheduled visits. • Score 5 for trials with follow-up visits and intervals of no more than usual care. • Score 5 for trials with no follow-up contact with participants to obtain data. • Trials that do not report the follow-up frequencies or intervals will be assessed based on the frequency and the intervals of the primary outcome measurement. • Trials with follow-up visits and intervals more than usual care will be assessed based on the intervals and the frequency of the visits:

	• Score 1 for visits every 2 weeks or less. • Score 2 for visits every 3 weeks or less. • Score 3 for visits every 4-6 weeks. • Score 4 for visits every 7-12 weeks. • Score 5 for visits every 8 weeks or more.
Example of pragmatic follow-up	• NCT01394952 “Participants were seen at 2 weeks, 3 months, and 6 months and then every 6 months for detailed assessments.”
Example of explanatory follow-up	• NCT01755572 “Subjects entered the baseline test period, which consisted of a 24-hour ambulatory blood pressure monitoring (ABPM) and a 24-hour urine collection (1–7 days prior to day 1). Subjects attended four clinic visits (days 1, 21, 42, and 63) at the start and end of each treatment period”
Primary outcome	• If no information is available about the primary outcome, the score will be "missing". • Score 1 for trials that apply two of the below criteria that reduce the score by 3 points. • Score 2 for trials that have an outcome that requires central adjudication or special training to measure it (e.g. retinopathy). • Score 2 for trials that have an outcome that is not of importance to the health care provider and the patients (e.g. anti-drug antibodies). • Score 3 for trials that have a surrogate outcome of importance only to the health care provider (e.g. HbA1C and BMI). • Score 4 for trials that measure the outcome at a time earlier than the usual care (less than 4 weeks). • Score 5 for trials that have hard serious outcomes (composite outcome), such as MACE. • Score 5 for trials that have outcomes that are important to the health care provider and patients (e.g. AEs).

Example of pragmatic primary outcome	● NCT02692716 “The primary endpoint is time from randomization to first occurrence of a major adverse CV event (MACE) composite endpoint consisting of CV death, non-fatal MI or non-fatal stroke (the classic three-point MACE).
Example of explanatory primary outcome	● NCT01755572 “The primary outcome was the within-subject hourly and baseline change in plasma ANP between liraglutide or placebo assessed after the first injection or after 3 weeks of daily administration.”
Primary analysis	● If no information is available about the primary analysis, the score will be "missing". ● Score 1 for trials that use per-protocol or as-treated analysis. ● Score 2 for trials that include data from all patients who were randomised and received a trial product with efficacy data. ● Score 3 for trials that report the use of modified intention-to-treat analysis without details of the modification. ● Score 4 for trials that analyse their primary outcome based on an intention-to-treat analysis using all available data for subjects who received at least one dose of the study drug. ● Score 5 for trials that analyse their primary outcome based on an intention-to-treat analysis using all available data for all randomised subjects.
Example of pragmatic primary analysis	● NCT01394952 “All efficacy and safety analyses will be conducted using an intention-to-treat approach that includes all randomized participants regardless of adherence.”
Example of explanatory primary analysis	● NCT01755572 “Eighteen participants were included in the analysis (2 subjects were excluded due to significant protocol deviations).”

5.4.4 PRECIS-2 score data extraction and harmonisation

The PRECIS-2 website was contacted to check if the online tool can be used for data extraction and retrospective assessment of GLP-1 RA trials. Although the website granted access to the online tool, it was unsuitable for this kind of retrospective assessment. Therefore, a standardised template adapted from the PRECIS-2 online tool was used to extract rationale and score all trials (see Appendix B: The PRECIS-2 rationale extraction template). The template included trial id, study title, author, rationale for the assessment, text score (Very Explanatory to Very Pragmatic) and numeric score (1-5).

Information needed to assess trials using the PRECIS-2 tool was manually extracted from ClinicalTrials.gov, journal publications, CSRs and/or study protocols of included studies. First, it was collected from ClinicalTrials.gov. Second, information was extracted from relevant publications, CSRs, or publicly available protocols if insufficient information was available on ClinicalTrials.gov or the trial was unregistered on ClinicalTrials.gov. The information required for the assessment was extracted in the standardised template. The domains were defined as Missing if the information required to assess the domain was unavailable in any data sources.

5.4.5 Statistical analysis

Descriptive statistics were used to summarise data and results. Frequencies and percentages were used to describe the missingness of information required to score each domain of the PRECIS-2 tool. The mean score of all domains of the PRECIS-2 tool were calculated for each trial. I used the mean score across all domains for which data were available, assuming that all domains contribute equally to the overall score. Although this approach aligns with some previous studies that used the PRECIS-2 score for retrospective assessment, it assumes that each domain is equally important. If this is not the case, it may bias the overall score. A histogram was used to demonstrate the mean PRECIS-2 scores, while bar plots were used to show the frequencies of the domains. A radial plot (the PRECIS-2 wheel) was used to show the overall pragmatism of included GLP-1 RA trials; the mean score of each domain of the PRECIS-2 tool was calculated and used for this plot.

In addition, a correlation plot and correlation coefficients were used to illustrate associations between domains.

5.4.6 Missing data

Information needed to score recruitment and organisation domains was frequently not reported. For the main analysis in this thesis, the mean PRECIS-2 score for each trial was calculated without accounting for the missing domain to ensure the analysis reflects the actual data available for each trial. However, the complete case analysis was not feasible for the analysis of correlation between domains due to the high missingness. Therefore, the missing domains were scored with the median score for each trial because it deals with outliers and maintains the central tendency, ensuring that the imputed scores are reasonable and do not skew the results.

5.5 Results

5.5.1 Retrospective assessment of GLP-1 RAs using the PRECIS-2 tool

GLP-1 RA trials were retrospectively scored using the PRECIS-2 tool. Considerable efforts were dedicated to extracting the information needed to assess the PRECIS-2 domains because they were not consistently reported across trials. Different data sources were used for each trial to extract the required information for the assessment, resulting in extended durations for the extraction process. The extraction and the scoring process took considerable time due to the depth and precision required to extract the information and apply the assessment criteria. The extraction of information for each trial required 30 to 60 minutes and an additional 30 minutes to score all domains, resulting in a cumulative time of approximately 300 hours.

Table 5.2 demonstrates the missing information required to assess the domains of the PRECIS-2 tool. Missing information was high for recruitment, organisation, and flexibility

(adherence) domains, 149 (76%), 85 (43.4%), and 43 (21.9%), respectively. Information for assessing eligibility criteria and primary outcomes domains were reported for all trials.

As illustrated in Figure 5.1, the mean PRECIS-2 scores for GLP-1 RA trials indicated a medium level of pragmatism for most trials. The minimum mean score of all included RCTs was 1.3 (1) SD, and the maximum was 4.8 (0.6) SD. As shown in Figure 5.2, most domains showed a medium level of pragmatism. Figure 5.3 shows the PRECIS-2 wheel that describes the pragmatism level of GLP-1 RA trials using the mean score of each domain of the PRECIS-2 tool. GLP-1 RA trials were generally equally pragmatic/explanatory. The mean scores for all domains are summarised in Table 5.2.

Table 5.2 Summary of mean scores for domains of the PRECIS-2 tool

PRECIS-2 tool domain	Mean score	Level of Pragmatism	Missing values (%)
Eligibility criteria	3.1 (0.9) SD	Equally pragmatic/explanatory	0
Recruitment	3.1 (1.5) SD	Equally pragmatic/explanatory	149 (76%)
Setting	3.5 (1) SD	Equally pragmatic/explanatory	8 (4.1%)
Organisation	3.5 (1.3) SD	Equally pragmatic/explanatory	85 (43.4%)
Flexibility (delivery)	4.2 (0.7) SD	Pragmatic	13 (6.6%)
Flexibility (adherence)	1.4 (0.5) SD	Explanatory	43 (21.9%)
Follow-up	2.4 (0.9) SD	Explanatory	5 (2.6%)
Primary outcome	2.8 (1) SD	Equally pragmatic/explanatory	0

Primary analysis	4.1 (1) SD	Pragmatic	1 (0.5%)

Figure 5.1 The PRECIS-2 mean score for GLP-1 RA trials

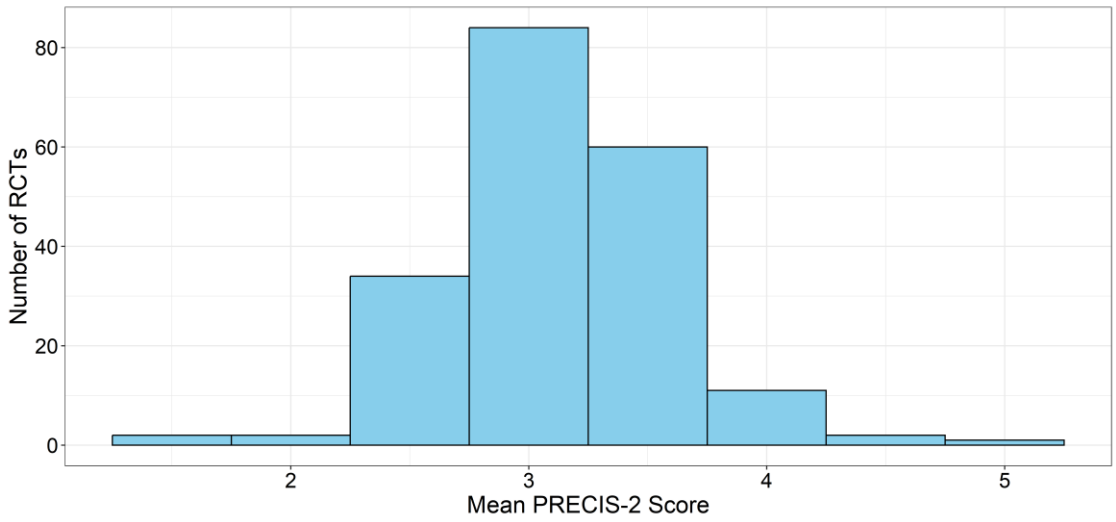


Figure 5.2 The distribution of scores for each domain of the PRECIS-2 tool

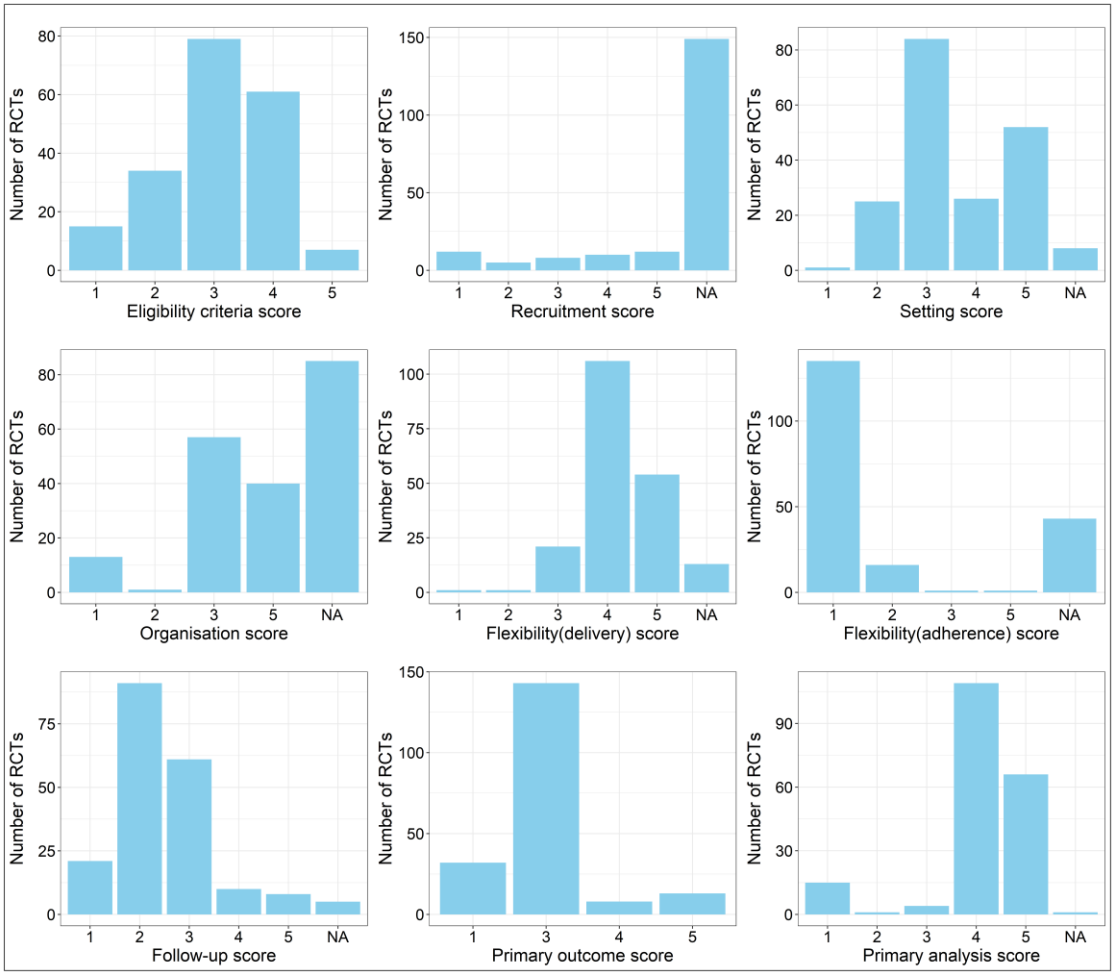
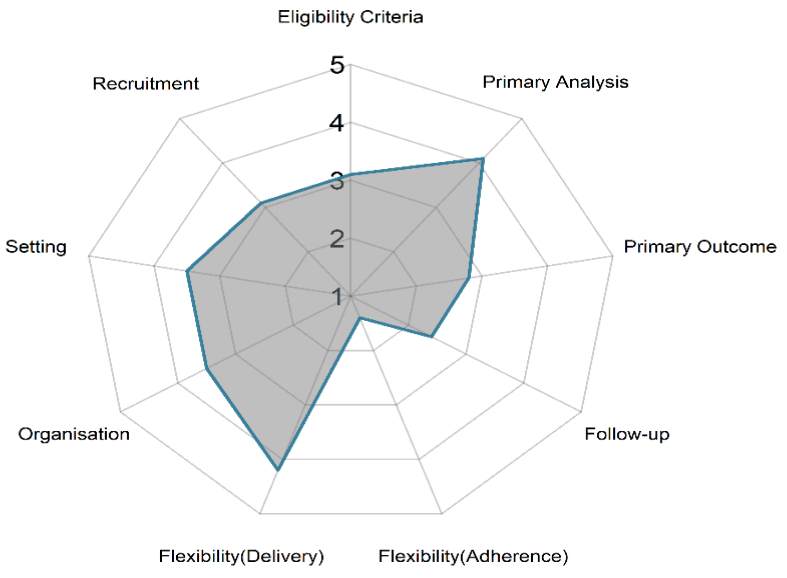


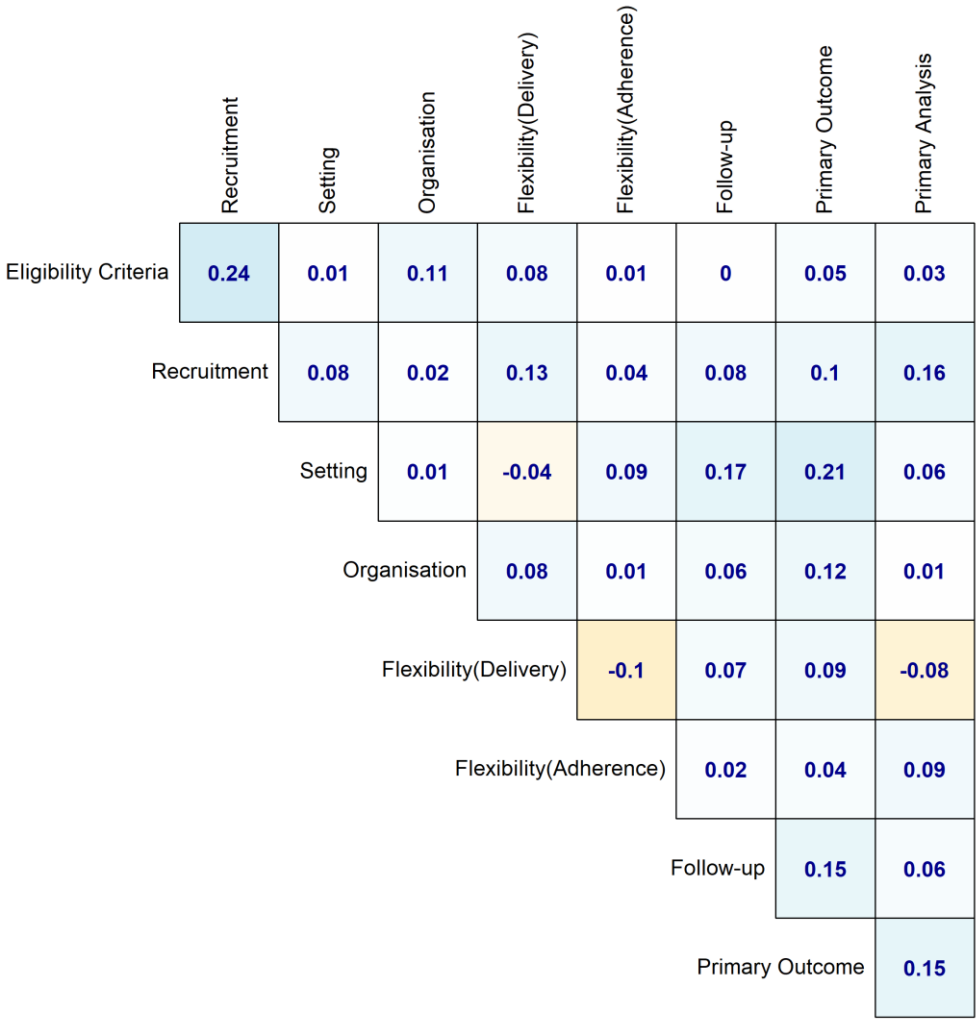
Figure 5.3 The PRECIS-2 wheel showing the mean score for each of the nine domains of the PRECIS-2 tool (1 is very explanatory, 5 is very pragmatic)



5.5.2 Correlation between the domains of the PRECIS-2 tool

Figure 5.4 illustrates the correlations between the PRECIS-2 domains using median imputation. There were modest correlations between eligibility criteria and recruitment domains, r=0.24 and between setting and primary outcome domains, r=0.21. However, correlations between other domains were weak or negligible. Moreover, almost all correlations were positive except for flexibility (delivery) with setting, flexibility (adherence) and primary analysis. As expected for a well-formed score, the domains were only weakly correlated. However, it is expected to see associations between those weakly correlated domains.

Figure 5.4 The correlation between the domains of the PRECIS-2 tool



5.6 Discussion

This chapter assessed trials using the PRECIS-2 tool as objectively as possible to ensure a fair comparison with the SAE rate. Moreover, it explored the challenges of using this tool for retrospective assessment.

5.6.1 Summary of findings

This is the first study to retrospectively examine the pragmatism of GLP-1 RA trials using the PRECIS-2 tool. The extraction and the assessment of all trials required a significant amount of time. The missing information needed to score recruitment and organisation domains was high. The analysis of the mean PRECIS-2 scores of 196 GLP-1 RA trials showed a medium level of pragmatism for most trials. The mean score of each domain showed that the most explanatory domains were flexibility (adherence) and follow-up, whereas the most pragmatic domains were flexibility (delivery) and primary analysis. Eligibility criteria, recruitment, setting, organisation, and primary outcome domains showed a medium level of pragmatism. Modest correlations were found only between eligibility criteria and recruitment and between setting and primary outcome.

5.6.2 Interpretation

The medium level of pragmatism for most trials could be attributed to the mixed aims of measuring the efficacy and effectiveness of the intervention in GLP-1 RA trials. Moreover, heterogeneity between scores of the PRECIS-2 domains for each trial may contribute to the mean score that indicates equally pragmatic/explanatory designs of GLP-1 RA trials. The modest correlation between eligibility criteria and recruitment domains may imply the role of the design of eligibility criteria in determining the recruitment process. Additionally, the modest correlation between setting and primary outcome domains could be attributed to that trial settings may be determined based on the clinical requirements for measuring the primary outcome. However, these correlations may be affected by the missing data.

The use of the PRECIS-2 tool for retrospective assessment was challenging. The missingness of information required to score recruitment and organisation domains was high. Moreover, information on the flexibility (adherence) domain was unavailable on ClinicalTrials.gov, relevant publications or protocols if not implemented or measured. Contacting investigators of the included trials to obtain the missing information was not viable. Despite efforts to operationalise and protocolise the PRECIS-2 tool for GLP-1 RA trials, the assessment was affected by the inherent subjectivity of the tool. Moreover, the retrospective assessment was time-consuming due to the search for information from different sources necessary for assessment. Additionally, most trials reported the use of detailed dose escalation protocols to deliver GLP-1 RAs, which may deduct the score of the flexibility (delivery) domain based on the guidance of the PRECIS-2 tool. However, these trials were scored as very pragmatic because dose escalation is a part of routine care to avoid the gastrointestinal side effects of GLP-1 RAs (Romera et al., 2019).

5.6.3 Comparison with previous literature

Several studies used the PRECIS-2 for retrospective assessment of pragmatism in diabetes, cardiovascular, oncology, rheumatoid arthritis, nursing care, Chinese herbal medicine and behavioural interventions RCTs (Johnson et al., 2016; Aves et al., 2017; Forbes et al., 2017; Lu et al., 2017; Luoma et al., 2017; Choi et al., 2019; Sepehrvand et al., 2019; Devos et al., 2019; Stewart et al., 2020; Ettori-Ajasse et al., 2020; Fitzpatrick et al., 2020; Saesen et al., 2023). Moreover, several studies used the original PRECIS tool for retrospective analysis of the pragmatism in schizophrenia, obesity, nursing, acupuncture and lifestyle interventions RCTs (Koppenaal et al., 2011; Tosh, 2011; Glasgow et al., 2012; Witt et al., 2012; Palese, Bevilacqua and Dante, 2014; Rosas et al., 2015). Three PRECIS-2 studies indicated a high level of pragmatism in their included trials (Johnson et al., 2016; Forbes et al., 2017; Stewart et al., 2020). Five PRECIS-2 studies showed a medium level of pragmatism (Luoma et al., 2017; Choi et al., 2019; Devos et al., 2019; Sepehrvand et al., 2019; Fitzpatrick et al., 2020). One PRECIS-2 study reported a lower level of pragmatism (Ettori-Ajasse et al., 2020). These studies showed variability in the reported level of pragmatism, indicating differences in the pragmatism of RCTs across several therapeutic areas.

Similarly to this study, previous studies reported that information required to score some domains was often missing (Lu et al., 2017; Choi et al., 2019; Ettori-Ajasse et al., 2020; Fitzpatrick et al., 2020; Saesen et al., 2023). Moreover, they found that the assessment relies on subjective interpretations (Johnson et al., 2016; Lu et al., 2017; Choi et al., 2019; Ettori-Ajasse et al., 2020).

Additionally, previous studies that used the PRECIS tool for retrospective assessment used the mean, median and total score to describe the pragmatism of their studies. Additionally, they used the PRECIS wheel to illustrate the overall pragmatism. Similarly to this study, most studies used the mean score of all domains for each trial to describe the level of pragmatism (Koppenaal et al., 2011; Glasgow et al., 2012; Johnson et al., 2016; Aves et al., 2017; Luoma et al., 2017; Choi et al., 2019; Sepehrvand et al., 2019; Fitzpatrick et al., 2020; Stewart et al., 2020). However, few studies used the PRECIS wheel to describe the pragmatism of their studies (Glasgow et al., 2012; Rosas et al., 2015; Forbes et al., 2017; Stewart et al., 2020). Moreover, few studies used the median or total score (Tosh, 2011; Glasgow et al., 2012; Palese, Bevilacqua and Dante, 2014; Lu et al., 2017; Devos et al., 2019; Ettori-Ajasse et al., 2020; Fitzpatrick et al., 2020). Therefore, this study used the mean score of each trial to describe their pragmatism. This method avoids introducing bias from imputed values and maintains the reliability of the actual collected data. Moreover, this study used the PRECIS wheel to describe the overall pragmatism of GLP-1 RA trials.

Furthermore, previous studies scored missing information required for assessment as 3 (equally pragmatic/explanatory), blank (excluded from the mean, median and total calculation), inferred from the trial description or zero. Missing values were mainly scored as 3 or blank (Koppenaal et al., 2011; Aves et al., 2017; Devos et al., 2019; Sepehrvand et al., 2019; Fitzpatrick et al., 2020). However, a score of 3 may skew the results towards the medium level. Therefore, dealing with missing values as blank is the most appropriate approach to handling missing data (Zwarenstein et al., 2020).

5.6.4 Strengths and limitations

A strength of this analysis is that the PRECIS-2 tool was operationalised and protocolised to score GLP-1 RA trials as well as possible, which minimised the subjectivity of the

assessment across all studies and ensured fair comparison with the SAE rate. The use of the PRECIS-2 tool for retrospective assessment may be considered a limitation of this analysis. However, it was used retrospectively in previous studies and was relatively reliable. Another limitation was that the information needed to assess some domains was missing, especially for recruitment and organisation domains. However, missing domains were excluded from the calculation of the mean scores.

5.6.5 Implications

Despite the limitations, the PRECIS-2 tool remains widely used to retrospectively assess the pragmatism of RCTs (Zwarenstein et al., 2020). Given the lack of gold standard measures of trial representativeness, the PRECIS-2 represent an appropriate and accessible comparator with the SAE rate as measures of trial representativeness. It can be applied to a wide range of trials, allowing for meaningful comparisons of representativeness between trials. However, the comparison between the SAE rate and the PRECIS-2 tool may require the use of fair umpires to compare between these metrics in a balanced way. A fair umpire is an imperfect measure, yet it can still distinguish between two measures without bias toward either (Glasziou, Irwig and Deeks, 2008).

5.7 Conclusion

This chapter illustrated the use of the PRECIS-2 tool for retrospective assessment. Moreover, it showed its suitability as a comparator with the SAE rate. It also indicated that most GLP-1 RA trials showed a medium level of pragmatism. Challenges of using the PRECIS-2 tool included inherent subjectivity, time-consuming nature, and lack of information required for assessment. The next chapter will compare the SAE rate and the PRECIS-2 score to determine whether the SAE rate can reflect trial representativeness.

Chapter 6 Comparison between the PRECIS-2 tool and the SAE rate as measures of trial representativeness: GLP-1 RA trials as an exemplar

6.1 Chapter overview

This chapter will compare between the SAE rate and the PRECIS-2 tool as metrics of trial representativeness. First, it will examine the association between the SAE rate and the mean PRECIS-2 score. Second, it will study the association between the SAE rate and the 9 domains of the PRECIS-2 score. Finally, it will compare between the SAE rate and the mean PRECIS-2 score based on the differences in their associations with several markers of trial representativeness.

6.2 Background

6.2.1 The lack of a gold standard measure of trial representativeness

There has been no gold standard measure to examine the representativeness of RCTs due to the complexity and subjectivity of current metrics. A gold standard test is a well-established and widely acknowledged method with high accuracy that is a benchmark for evaluating new tests (Cardoso et al., 2014). In the absence of a gold standard test, several approaches have been used to assess the validity of new tests. One approach is the comparison with existing tests that measure the same underlying construct. Another method uses several imperfect tests that act as a gold standard test (Rutjes et al., 2007). However, these methods may lead to biased and inaccurate estimation of the new test (Whiting et al., 2013). Therefore, using the existing measures of representativeness alone to determine the validity of the SAE rate as a measure of trial representativeness is inadequate and may lead to biased results.

6.2.2 The need for fair umpire tests

The use of a fair umpire test may facilitate a robust and reliable comparison between a new measure and an imperfect existing measure in a balanced way (Ray et al., 2010). A fair umpire is an imperfect measure, yet it can effectively differentiate between two measures without bias towards either. It is required that the umpire test not to be associated with either test except via the true unknown measure (Glasziou, Irwig and Deeks, 2008). However, trial representativeness is a complex concept, and it is difficult to identify a single variable that meets this condition.

As described in Table 6.1, the selection of fair umpires involved assessing whether the potential marker of trial representativeness is neutral and not biased towards the SAE rate or the mean PRECIS-2 score measures. For example, if the marker is part of the assessment of the PRECIS-2 domains, it may introduce bias to the comparison. Therefore, all selected fair umpires have to meet the condition of not being inherently biased toward either metric. Consequently, I chose fair umpires based on their potentiality as markers of trial representativeness and not being biased toward the SAE rate or the mean PRECIS-2 score. Additionally, the process of selecting the umpires involved evaluating whether there is evidence relating them to trial representativeness and ensuring their data is available for analysis. The included markers are baseline sample size, baseline mean age, baseline male percentage, baseline mean HbA1C, baseline mean T2DM duration, trial year difference (the difference between the year of first register and the year since the intervention was first trialled), trial duration, trial phase, trial blinding and trial sponsor. Although the selected fair umpires have met the conditions of a fair umpire test, they are still not ideal as they are not direct measures of trial representativeness.

Table 6.1 The selection and eligibility of fair umpire tests

Markers of trial representativeness	Biased towards PRECIS-2 Biased towards SAE rate Neutral	Eligibility for fair umpire test
Baseline sample size	Neutral	Eligible
Baseline mean age	Neutral	Eligible
Baseline male percentage	Neutral	Eligible
Baseline mean HbA1C	Neutral	Eligible
Baseline mean T2DM duration	Neutral	Eligible
Trial blinding (none/double or more)	Neutral	Eligible
Trial phase (III/IV)	Neutral	Eligible
Trial duration	Neutral	Eligible
Trial sponsorship (industry/other)	Neutral	Eligible
Trial year difference (the difference between the year of first register and the year since the intervention was first trialled)	Neutral	Eligible
Trial sites (single-centre/multicentre)	Biased towards PRECIS-2 (Part of the PRECIS-2 assessment)	Excluded
Trial sites (national/multinational)	Biased towards PRECIS-2 (Part of the PRECIS-2 assessment)	Excluded
Population analysis (ITT/PP)	Biased towards PRECIS-2 (Part of the PRECIS-2 assessment)	Excluded
Primary outcome (MACE/none)	Biased towards PRECIS-2 (Part of the PRECIS-2 assessment)	Excluded

6.2.3 Markers of trial representativeness

The selected fair umpire tests can be used to differentiate between the SAE rate and PRECIS-2 score because they are likely to reflect trial representativeness. For example, baseline mean age, baseline male percentage, baseline mean HbA1C and baseline mean T2DM duration may reflect trial representativeness as they are likely to be affected by how broad or restricted the eligibility criteria are (Averitt et al., 2020; Li et al., 2020). Moreover, trials with baseline characteristics that closely resemble the target population are likely to be more representative (Kennedy-Martin et al., 2015; Averitt et al., 2020).

Additionally, a small baseline sample size is likely to be a marker of a small, earlier or more explanatory trial (Wisniewski et al., 2009; Ramirez et al., 2012; Ciolino, Kaizer and Bonner, 2023). Trials with longer durations are likely to be more representative than trials with shorter durations because they are usually larger and have broader eligibility criteria (Martin et al., 2013; Monti et al., 2018; Sen et al., 2018). Phase IV trials and trials with industry sponsorship could be more inclusive, have larger sample sizes and reflect the real-world population (Emdin et al., 2015; Varma et al., 2021). Earlier trials of a specific intervention often have small sample sizes and restrictive eligibility criteria, while more recent trials tend to be pragmatic and more representative of the real-world population (Treweek and Zwarenstein, 2009; Ciolino, Kaizer and Bonner, 2023). Furthermore, pragmatic trials generally employ less stringent blinding than explanatory trials (Ware and Hamel, 2011; Moustgaard et al., 2020).

6.2.4 Prior expectations

The prior expected associations of fair umpires with each measure are described in Table 6.2. If the fair umpires consistently strongly favoured SAE rates, this suggests that SAE rates may reflect trial representativeness. Conversely, if they favoured the PRECIS-2 tool or neither measure, it indicates that SAE rates may not reflect trial representativeness. However, the umpires that did not favour either metric may indicate that they may not accurately reflect trial representativeness as expected, especially if they show no favouring while not having strong associations with either metric.

It is expected for baseline mean age, baseline male percentage, baseline mean T2DM duration, baseline mean HbA1C, baseline sample size, trial duration, trial year difference and trial industry sponsorship to favour SAE rates as they may reflect trial representativeness in terms of the patient population. For example, trials with high baseline mean age are expected to have higher SAE rates than standard trials, as older participants are generally weaker and more likely to have multiple comorbidities (Hanlon et al., 2021). Moreover, women are more likely to experience SAEs than men because women with T2DM are at increased risk of cardiovascular events (Clemens et al., 2020). Furthermore, increased T2DM duration and elevated HbA1C levels are expected to be associated with increased SAEs because long-term illness and poor glycaemic control may lead to the presence of comorbidities, which increase the risk of SAEs (Fox et al., 2004; Kishimoto et al., 2014; Navarro-pérez et al., 2018; Ghouse et al., 2020).

Additionally, large sample size trials are expected to detect more SAEs than small sample size trials (Yao et al., 2021). Furthermore, trials with longer durations are expected to have higher SAEs than trials with shorter durations, likely because longer trials can capture late-occurring SAEs (Ford and Norrie, 2016; Sen et al., 2017, 2018). Trials with industry sponsorship and trials conducted later after the intervention is first tested are expected to be larger and more inclusive, likely recruiting elderly and patients with comorbidities, who are at increased risk of SAEs (Treweek and Zwarenstein, 2009; Herrera et al., 2010). However, open-label and phase IV trials are expected to favour the mean PRECIS-2 score as they may reflect trial representativeness in terms of design and setting rather than patient population (Dal-Ré, Janiaud and Ioannidis, 2018; Sepehrvand et al., 2019).

Table 6.2 The prior expected associations of fair umpire tests with the SAE rate and the mean PRECIS-2 score

Fair umpires	Expected direction of association		Expected favouring		
	SAE rate	Mean PRECIS-2 score	SAE rate	Mean PRECIS-2 score	Neither
Baseline sample size	Positive association	Positive association	SAE rate		
Baseline mean age	Positive association	Slight positive association	SAE rate		
Baseline male percentage	Negative association	Slight negative association	SAE rate		
Baseline mean T2DM duration	Positive association	Slight positive association	SAE rate		
Baseline mean HbA1C	Positive association	Slight positive association	SAE rate		
Trial duration	Positive association	Slight positive association	SAE rate		
Trial year difference (the difference between the year of first register and the year since the intervention was first trialled)	Positive association	Slight positive association	SAE rate		
Trial sponsorship: Industry	Positive association	Slight positive association	SAE rate		
Trial blinding: Double or more	Slight negative association	Negative association	Mean PRECIS-2 score		
Trial phase: Phase IV	Slight positive association	Positive association	Mean PRECIS-2 score		

6.3 Aims and objectives

This chapter will examine whether SAE rates can reflect trial representativeness. Specifically, it will compare between the SAE rate and the PRECIS-2 score.

The objectives of this chapter are:

1) To examine the association between the SAE rate and the mean PRECIS-2 score.

2) To study the association between the SAE rate and each of the 9 domains of the PRECIS-2 score.

3) To compare the SAE rate with the PRECIS-2 score as measures of trial representativeness based on the differences in associations with the fair umpire tests.

6.4 Methods

6.4.1 Study selection and SAE data extraction

196 GLP-1 RA studies were identified for this project. The study selection and SAE data extraction are described in detail in 3.4.1 and 3.4.3.

6.4.2 Data sources

Data sources used for this analysis are described in 3.4.2.

6.4.3 Baseline and trial characteristics data extraction and harmonisation

The baseline and trial characteristics of the included trials were extracted from ClinicalTrials.gov, journal publications, and/or CSRs. Baseline characteristics included baseline sample size, baseline mean age, baseline gender ratio, baseline mean T2DM duration and baseline mean HbA1C. TableTidier software was used to extract baseline data from journal publications into standard formats, where tables with different structures and terminologies were transformed into a standardised structure. The purpose of this software is to simplify the extraction and analysis of data published in journal articles. Trial characteristics included trial phase, trial blinding, trial sponsor, and trial year difference (the difference between the year of first registration and the year since the intervention was first trialled). Trial characteristics were extracted from ClinicalTrials.gov, CSRs, or journal publications and saved to a standardised CSV file.

Extracted baseline and trial characteristics were harmonised, coded, and prepared for analysis. The baseline sample size was aggregated for all trial arms and then rescaled to the log of base 2 to handle data variability and enhance the interpretability of the results. The weighted mean was calculated for baseline mean age, baseline mean T2DM duration, and baseline mean HbA1C for all trial arms. Baseline mean age, baseline mean T2DM duration and trial year difference (the difference between the year of first register and the year since the intervention was first trialled) were rescaled by dividing them by 10. The baseline male percentage was calculated from the baseline gender ratio. The trial phase was categorised as phase III and IV. Trial blinding was categorised as none and double or more. The trial sponsor type was classified as industry and other.

6.4.4 Statistical analysis

6.4.4.1 Descriptive statistics

Descriptive statistics were used to describe the baseline and trial characteristics of all GLP-1 RA trials. Frequencies and percentages were used to describe categorical characteristics.

Minimum and maximum values were reported for all numeric characteristics. Frequencies and percentages were used to describe missing data for all baseline and trial characteristics.

6.4.4.2 Regression analysis

The SAE rate and the mean PRECIS-2 score originally differed in distribution and scale. To allow the comparison between the mean PRECIS-2 score and the SAE rate, both metrics needed to be normally distributed and on a comparable scale. Therefore, the "orderNorm" function from the R package "bestNormalize" transformed both metrics to be normally distributed by identifying and applying the best-fitting transformation to achieve normality. A linear regression model was fitted to estimate the association between the normalised SAE rate and the normalised mean PRECIS-2 score. Additionally, a linear regression model was fitted to estimate the association between the normalised SAE rate and each of the 9 domains of the mean PRECIS-2 score. Another unadjusted linear model was fitted to estimate the association between the eligibility criteria domain and the normalised SAE rate.

Using bivariate Bayesian regression was the most feasible method to estimate the differences in the association between each umpire with the normalised SAE rate and the normalised mean PRECIS-2 score. It can more effectively capture complex relationships and dependencies than running separate regression models for each metric. Therefore, a bivariate Bayesian regression model was fitted to estimate the association between each umpire with the normalised SAE rate and the normalised mean PRECIS-2 score. The model priors were set to the default priors in the brms package; flat prior for the covariates and Student's t-distribution prior for the metrics to avoid any strong prior assumptions, allowing the data to play a more dominant role in determining the posterior distribution for the covariates and the metrics. Continuous covariates in this model were baseline sample size, baseline mean age, baseline male percentage, baseline mean HbA1C, baseline mean T2DM duration, trial year difference (the difference between the year of first register and the year since the intervention was first trialled) and trial duration. Binary covariates were trial phase, trial blinding and trial sponsor. The differences between the normalised mean PRECIS-2 score and the normalised SAE rate were calculated based on estimates from each draw. Then, the differences were combined to calculate the mean of each difference in the association. Finally, metrics were compared based on the magnitude of these differences.

6.4.5 Missing data

Missing data were imputed using the Multivariate Imputation by Chained Equations (MICE) package in R software because it is expected for missing data not to be missing completely at random (MCAR) but missing at random (MAR). Multiple imputations is a statistical technique used to handle missing data by creating several sets of imputations, accounting for the inherent uncertainty of the imputation process (Jakobsen et al., 2017). SAEs, trial phase, baseline mean age, baseline mean T2DM duration, baseline mean HbA1C and baseline male percentage were imputed, generating five sets of imputations. Then, the brm_multiple() function from the brms package was used to run the bivariate Bayesian regression model separately for each imputed dataset, considering prior and posterior distributions of the covariates. After analysis, the results from these separate models were combined, typically by averaging the coefficients, to create a coherent statistical interpretation of the findings.

6.5 Results

6.5.1 Descriptive statistics

As illustrated in Table 6.3 and Table 6.4, baseline mean age data was unavailable for 9 (4.6%) trials, baseline mean T2DM duration data were missing for 18 (9.9%) trials, and baseline mean HbA1C values were not reported for 13 (6.6%) trials. Baseline male percentages were unavailable for 29 (14.4%) trials, and the trial phase was missing for 17 (8.7%) trials.

Table 6.3 Summary of continuous markers of trial representativeness

Variable	Trials	Missing (%)	Min	Max
Baseline sample size	196	0	18	14752
Baseline mean age (years)	187	9 (4.6%)	33.97	74.20
Baseline T2DM duration (years)	178	18 (9.9%)	0.42	18.18
Baseline mean HbA1C (%)	183	13 (6.6%)	6.30	12.04
Baseline male percentage	167	29 (14.4%)	0.29	0.81
Trial duration (years)	196	0	0.06	6.92
Year difference (difference between the year of first register and the year since the intervention was first trialled) (years)	196	0	0	14.62

Table 6.4 Summary of categorical markers of trial representativeness

Variable	Trials (%)
Trial phase	196
Phase III	150 (76.5%)
Phase IV	29 (14.8%)
Missing	17 (8.7%)
Trial sponsorship	
Industry	157 (80.1%)
Other	39 (19.9%)
Missing	0
Trial blinding	
None	89 (45.4%)
Double or more	107 (54.6%)
Missing	0

6.5.2 SAE reporting in GLP-1 RA trials

SAE reporting in GLP-1 RA trials is described in 3.5.2.

6.5.3 The association between the SAE rate and the PRECIS-2 score

A linear regression analysis was conducted to estimate the association between the normalised SAE rate and the normalised mean PRECIS-2 score. No evidence was found on the association between metrics; the estimated coefficient was of small magnitude, and the CI included the null (β= 0.002, 95% CI -0.14 – 0.14). Furthermore, another linear regression analysis was conducted to estimate the association between the normalised SAE rate and each of the 9 domains of the PRECIS-2 score. As illustrated in Table 6.5, no evidence was found on the association between the 9 domains and the normalised SAE rate; all estimated coefficients were of small magnitudes, and CIs included the null. Additionally, there is no evidence of the association between the eligibility criteria domain and the normalised SAE rates; CI included the null.

Table 6.5 The association between the normalised SAE rate and PRECIS-2 domains

PRECIS-2 domains	β	95% CI	p-value
(Intercept)	0.16	-1.16 – 1.48	0.810
eligibility criteria (unadjusted)	-0.09	-0.24 – 0.05	0.205
eligibility criteria	-0.11	-0.26 – 0.04	0.152
recruitment	0.08	-0.13 – 0.28	0.457
setting	0.02	-0.12 – 0.17	0.745
organisation	0.01	-0.14 – 0.15	0.911
flexibility delivery	-0.001	-0.22 – 0.22	0.993
flexibility adherence	0.02	-0.15 – 0.20	0.802
follow up	-0.003	-0.17 – 0.17	0.974
primary outcome	0.09	-0.08 – 0.25	0.300
primary analysis	-0.11	-0.26 – 0.05	0.172

6.5.4 The association between the fair umpires with the normalised SAE rate and the normalised mean PRECIS-2 score

As illustrated in Table 6.6, fair umpires had differing associations with both metrics. The baseline sample size had positive associations with the normalised mean PRECIS-2 score and the normalised SAE rate; CIs did not include the null (β= 0.2, 95% CI 0.08 – 0.33, β= 0.19, 95% CI 0.08 – 0.31), respectively. Moreover, the baseline mean age and baseline mean T2DM duration showed positive associations with normalised SAE rate; CIs did not include the null (β= 0.46, 95% CI 0.10 – 0.84, β= 0.52, 95% CI 0.03 – 0.98), respectively. However, longer trial durations were negatively associated with the normalised SAE rate; CI did not include the null (β= -0.29, 95% CI -0.44 – -0.13). Other umpires were not associated with either metric.

As shown in Table 6.6, all fair umpires did not strongly favour the normalised SAE rate; all differences crossed zero. However, the direction of the difference for half of the umpires favoured the normalised SAE rate, including baseline mean age, baseline male percentage, baseline mean T2DM duration, phase IV trials, and industry-sponsored trials. Moreover, the directions of these differences were expected, except for phase IV trials. Only trial duration strongly favoured the normalised mean PRECIS-2 score. However, the direction of this umpire was unexpected. Other umpires showed no favouring of either metric.

Table 6.6 The estimated differences between the SAE rate and the mean PRECIS-2 score

Fair umpires	SAE Rate		Mean PRECIS-2		Difference			SE	2.5%	97.5%	Favouring		
	β	CI (95%)	β	CI (95%)	Null	Unexpected	Expected				Neither	PRECIS	SAE
Intercept	-4.8	-7.96 – -1.58	-2.25	-5.74 – 1.17	-2.54			2.44	-7.38	2.20			
Baseline sample size	0.19	0.08 – 0.31	0.20	0.08 – 0.33	-0.01			0.09	-0.19	0.16	Neither		
Baseline mean age	0.46	0.10 – 0.84	-0.09	-0.50 – 0.34	0.55			0.29	-0.01	1.12	Favoured SAE		
Baseline male percentage	-1.15	-2.67 – 0.33	0.61	-0.98 – 2.17	-1.75			1.12	-3.96	0.46	Favoured SAE		
Baseline mean T2DM duration	0.52	0.03 – 0.98	-0.01	-0.54 – 0.54	0.52			0.41	-0.32	1.30	Favoured SAE		
Baseline mean HbA1C	0.08	-0.17 – 0.34	0.09	-0.19 – 0.36	-0.01			0.20	-0.38	0.40	Neither		
Blinding: Double or more	0.15	-0.10 – 0.40	0.15	-0.13 – 0.42	0.00			0.20	-0.39	0.39	Neither		
Trial Phase: Phase IV	0.3	-0.13 – 0.76	-0.19	-0.66 – 0.28	0.49			0.34	-0.18	1.15	Favoured SAE		
Trial duration	-0.29	-0.44 – -0.13	0.06	-0.11 – 0.23	-0.34			0.12	-0.59	-0.10	Favoured PRECIS		

Year difference (difference between the year of first register and the year since the intervention was first trialled)	0.01	-0.41 – 0.44	0.03	-0.42 – 0.47	-0.02	0.33	-0.66	0.64	Neither
Trial Sponsor: Industry	0.2	-0.27 – 0.68	-0.18	-0.71 – 0.34	0.39	0.38	-0.34	1.13	Favoured SAE

6.6 Discussion

This chapter explored whether the SAE rate can reflect trial representativeness. It examined the association between the normalised SAE rate and the normalised mean PRECIS-2 score and with each of the 9 domains of the PRECIS-2 score. Furthermore, it compared the normalised SAE rate with the normalised mean PRECIS-2 score as measures of trial representativeness based on the differences in their associations with several potential markers of trial representativeness.

6.6.1 Summary of findings

No association was found between the normalised SAE rate and the normalised mean PRECIS-2 score. Moreover, none of the nine domains of the PRECIS-2 score was associated with the normalised SAE rate.

Regarding associations between the fair umpires and metrics of trial representativeness, the umpires had varying associations with the normalised SAE rates and the normalised mean PRECIS-2 score. The baseline sample size was positively associated with both measures. The baseline mean age and baseline mean T2DM duration were positively associated with the normalised SAE rate. However, the trial duration was negatively associated with the normalised SAE rate. There was insufficient evidence on the associations of other umpires with both metrics.

Regarding the analysis of the difference between the normalised SAE rate and the normalised mean PRECIS-2 score as metrics of trial representativeness, none of the fair umpires strongly favoured the normalised SAE rate. However, the directions of baseline mean age, baseline mean T2DM duration, baseline male percentage, phase 4 trials and trials with industry-sponsorship favoured the normalised SAE rate. On the other hand, only trial duration strongly favoured the normalised mean PRECIS-2 score. Other fair umpires showed null differences in their associations with the normalised SAE rate and the normalised mean PRECIS-2 score.

6.6.2 Interpretation

Surprisingly, no association was found between the normalised SAE rate and the normalised mean PRECIS-2 score and no associations were found between the normalised SAE rate and PRECIS-2 domains, suggesting that these metrics are independent, as the PRECIS-2 tool may be more focused on the trial design, whereas SAEs mainly reflect baseline health conditions.

As described in Table 6.7, only three fair umpires demonstrated positive associations with the normalised SAE rates, while one umpire showed a negative association. The positive associations between baseline sample size, baseline T2DM duration and baseline mean age with the normalised SAE rate were expected. Trials with large sample sizes may include older adults and comorbid patients, which may explain the association with SAE rates. Furthermore, the larger trials included in this analysis were MACE outcome trials, which have higher SAE rates by design. Moreover, trials with longer mean T2DM durations may recruit subjects with more chronic and advanced stages of the disease. Therefore, including patients with longer disease duration is associated with higher SAE rates. Additionally, the positive association between baseline mean age and the normalised SAE rate indicate that a trial that recruits the elderly, who are likely to be comorbid and weaker, is associated with increased SAE rates. However, the negative association between longer trial duration and the normalised SAE rate was unexpected. This association may be attributed to the tendency of older adults and comorbid patients who are at higher risk of SAEs to decline participation in longer-duration trials. These studies often require long-term commitments and follow-ups, which can be challenging for such individuals. Consequently, longer trials may inadvertently recruit a healthier participant pool (Kaushal et al., 2022; Pitkala and Strandberg, 2022). On the other hand, only one umpire showed a positive association with the normalised mean PRECIS-2 score. These associations indicate that fair umpires did not consistently strongly associate with either metric.

Regarding the differences in associations of fair umpires with the normalised SAE rate and the normalised mean PRECIS-2 score, all fair umpires did not strongly favour the normalised SAE rate. However, baseline mean age, baseline mean T2DM duration, baseline male percentage, phase IV trials and trials with industry-sponsorship favoured the normalised SAE rate. These differences were in the expected direction and favoured as

anticipated, except for phase IV trials (Table 6.7). On the other hand, trial duration strongly favoured the normalised mean PRECIS-2 score. However, this difference was neither in the expected direction nor in the expected favouring (Table 6.7). The remaining differences were null; other umpires did not favour either metric. These findings indicate that the markers of trial representativeness used in this analysis may not be sensitive enough to detect subtle differences between the normalised SAE rate and the normalised mean PRECIS-2 score as they did not consistently strongly favour either metric. Furthermore, it suggests that the SAE rate did not have a clear advantage over the PRECIS-2 score. Therefore, the use of the SAE rate as a metric of trial representativeness is not supported by the findings of this chapter.

Table 6.7 The actual associations of fair umpire tests with the SAE rate and the mean PRECIS-2 score

Fair umpires	Direction of association		Umpire favour			Difference		
	SAE rate	Mean PRECIS-2 score	Neither	PRECIS-2	SAE rate	Null	Unexpected	Expected
Baseline sample size	Positive association	Positive association	Neither			Null		
Baseline mean age	Positive association	Slight negative association	Favoured SAE rate			Expected		
Baseline mean T2DM duration	Positive association	No association	Favoured SAE rate			Expected		
Baseline mean HbA1C	No association	No association	Neither			Null		
Baseline male percentage	Slight negative association	Slight positive association	Favoured SAE rate			Expected		

Trial blinding: Double or more	Slight positive association	Slight positive association	Neither	Null
Trial phase: Phase IV	Slight positive association	Slight negative association	Favoured SAE rate	Unexpected
Trial duration	Negative association	No association	Favoured PRECIS-2	Unexpected
Year difference (difference between the year of first register and the year since the intervention was first trialled)	No association	No association	Neither	Null
Trial sponsorship: Industry	Slight positive association	Slight negative association	Favoured SAE rate	Expected

6.6.3 Comparison with previous literature

Most of the literature on measuring trial representativeness is focused on quantifying the effect of eligibility criteria on trial representativeness and comparing the differences in patient characteristics between the trial population and the target population (He et al., 2020). Furthermore, the feasibility of the SAE rate as a metric of trial representativeness has not been widely studied. This study examined SAE rates using a quite different approach than previous studies by comparing the SAE rate with the PRECIS-2 score based on the differences in their associations with several fair umpire tests. This differs from previous studies, which used expected SAE rates from routine care as a benchmark to compare it with SAE rates observed in trials. For example, a study assessed trial representativeness in RCTs of renin-angiotensin-aldosterone system (RAAS) drugs for hypertension by comparing SAE rates between trials and the community and found that observed SAE rates were lower than expected SAE rates in the community. However, they found that older people trials had higher SAE rates than standard trials, consistent with the finding that increased mean age was associated with higher SAE rates (Hanlon et al., 2021). Another study assessed trial representativeness in RCTs of 21 disease conditions by comparing observed SAE rates in RCTs with expected SAE rates in routine care and found that observed SAE rates were lower than expected SAE rates (Hanlon et al., 2022).

Additionally, several studies used the fair umpire method in the absence of gold standard measures to compare between new and old tests in a balanced way, showing the ability of the fair umpire method to resolve the disagreement between measures. For example, a study examined the validity of new classification criteria for early diagnosis of rheumatoid arthritis compared to existing classification criteria. They compared between the two classification criteria based on symptom duration, rheumatoid factor (RF), anti-cyclic citrullinated peptide (anti-CCP) antibody and the requirement of disease-modifying antirheumatic drug (DMARD) therapy. They found that the new criteria resulted in overdiagnosis of early cases (Cader et al., 2011). Another study examined the validity of the abdominal aortic aneurysm (AAA) screening method compared to non-screening in reducing disease-specific mortality. They compared them based on the AAA incidence, elective surgery, rupture and overtreatment and found that the AAA screening did not significantly reduce disease-specific mortality (Johansson et al., 2018). Although these studies suggested the reliability of the fair umpire method in resolving the disagreement

between new and existing measures, the assessment of trial representativeness is more complex and may be affected by several factors. The selected umpires in this study may not accurately resolve the disagreement between the normalised SAE rates and the normalised mean PRECIS-2 score because they are not direct measures of trial representativeness.

6.6.4 Strengths and limitations

This is the first study to explore the validity of the SAE rate as a metric of trial representativeness by comparing it with an existing measure in the absence of a gold standard measure. Moreover, this is the first study that used fair umpire tests to compare between the SAE rate and the PRECIS-2 score in a balanced way. Additionally, this study included trials from different registries, which provided a variety of RCTs. This will enhance the generalisability of the study findings and reduce selection bias. On the other hand, this study was conducted on GLP-1 RA trials, which may limit the generalisability of its findings to different diseases and drug classes, especially trials of toxic interventions. Furthermore, this study was limited by the lack of some baseline and trial characteristics. However, missing data was imputed to reduce the impact of missingness on this analysis. The imputation process helps maintain statistical power and prevent bias that can occur due to the absence of data. Additionally, the selected fair umpires were not ideal as they do not directly assess trial representativeness, which may result in biased findings.

6.6.5 Implications

Higher SAE rates were observed with larger baseline sample sizes, increased baseline mean age, and longer baseline mean T2DM durations. Furthermore, half of the umpires, including baseline mean age, baseline male percentage, baseline mean T2DM duration, phase IV trials and industry-sponsored trials, slightly favoured SAE rates, indicating that SAE rates may, at least to some extent, reflect trial representativeness. However, the use of the SAE rate as a metric of trial representativeness is not supported by the findings of this chapter. Further research is required to examine whether the SAE rate can reflect trial representativeness because the fair umpires approach did not resolve the disagreements

between the SAE rate and the PRECIS-2 score. Therefore, the next chapter will study the association between the SAE rate and eligibility criteria.

6.7 Conclusion

The SAE rate demonstrated expected associations with certain fair umpires. Moreover, half of the umpires slightly favoured the SAE rates over the mean PRECIS-2 score. However, fair umpires were neither consistently strongly associated with nor favoured the SAE rate. Therefore, the validity of the SAE rate as a metric of trial representativeness is not supported by the findings of this chapter. The next chapter will examine the association between the SAE rate and eligibility criteria to know whether the SAE rate can reflect trial representativeness.

Chapter 7 The association between eligibility criteria and the SAE rate in GLP-1 RA trials

7.1 Chapter overview

This chapter will study the association between the SAE rate and eligibility criteria to examine whether the SAE rate can reflect trial representativeness.

7.2 Background

7.2.1 Eligibility criteria, a driver of trial representativeness

Eligibility criteria are crucial in determining the trial population and can potentially driver trial representativeness (Averitt et al., 2020). Trials with overly restrictive eligibility criteria may exclude subjects who would benefit from the intervention in the real-world, thereby limiting their representativeness (Rothwell, 2005; Tan et al., 2022). However, the role of eligibility criteria in determining trial participation may be moderated by two main factors. First, eligibility criteria are usually replicated from previous studies and ambiguously defined, leading to subjective interpretation and application of criteria (Van Spall et al., 2007; Hao et al., 2014). Second, healthcare practitioners may exclude eligible patients during the informal referral to trials in usual care, weakening the role of eligibility criteria in driving trial participation (Howard et al., 2009; Oude Rengerink et al., 2017).

7.2.2 The association between eligibility criteria and the SAE rate

The eligibility criteria for T2DM trials contain elements that possibly shape the trial population, thereby influencing their representativeness (Sen et al., 2018). Highly restrictive age criteria may result in excluding older people who often have a higher prevalence of T2DM and higher SAEs (Kirkman et al., 2012). Moreover, trials with restrictive BMI ranges, HbA1C levels, renal measurements, disease durations, heart failure

classes and history of comorbidities may skew the trial population toward specific body composition or disease severity that may not reflect the target population (Dennis, 2020; Tan et al., 2022; Shields et al., 2023). Trials with these restrictions may be associated with lower SAE rates compared to permissive trials. Furthermore, the exclusion of patients based on the use of concomitant medications or a history of alcohol may be associated with low SAE rates due to the selection of patients with fewer health complications (Knox et al., 2019) (Moberg and Humphreys, 2017; Knox et al., 2019; Kim et al., 2021). Additionally, trials that require treatment stability or restrict the allowed number of medications used in conjunction with the trial intervention are expected to experience lower SAE rates as they may inadvertently exclude older and sicker patients (Hamaker, Stauder and van Munster, 2014; Lichtman et al., 2017). These criteria, collectively or individually, may drive trial representativeness as they would not drive ineligibility for treatment in routine care. Therefore, the association between eligibility criteria and the SAE rate would indicate that the SAE rate may reflect trial representativeness.

7.3 Aim and Objectives

This chapter will examine whether SAE rates can reflect trial representativeness. Specifically, it will study the association between the SAE rate and eligibility criteria.

The objectives of this chapter are:

1) To study the association between highly restrictive eligibility criteria and the SAE rate.

2) To examine the association between the continuous score of eligibility criteria and the SAE rate.

7.4 Methods

7.4.1 Study selection and data extraction

196 GLP-1 RA studies were identified for this project. The study selection and SAE data extraction are described in detail in 3.4.1 and 3.4.3.

7.4.2 Data sources

Data sources used for this analysis are described in 3.4.2.

7.4.3 Eligibility criteria data extraction and harmonisation

The Identification of key eligibility criteria for GLP-1 RA trials

The eligibility criteria of GLP-1 RA trials were reviewed to identify key criteria. Common criteria for GLP-1 RA trials were identified for extraction following a pilot screening of a random sample of 10 trials. The identified criteria were reviewed by a diabetologist, Dr Elaine Butterly, to ensure that these criteria would not drive ineligibility for treatment in routine care and not part of safety assessment for prescribing GLP-1 RAs in routine care such as calcitonin levels, a hormone that acts as a marker of medullary thyroid cancer (MTC).

The extraction eligibility criteria of GLP-1 RA trials

The extraction of the eligibility criteria was protocolised to ensure standardised and consistent extraction. As described in Table 7.1, age, BMI, fasting blood glucose (FBG), glycated haemoglobin (HbA1C), T2DM duration, systolic blood pressure (SBP), diastolic blood pressure (DBP) and renal function range criteria (estimated glomerular filtration rate (eGFR) or creatinine clearance (CrCl) whichever is available) were extracted as continuous

minimum and maximum values. The requirement of treatment stability was extracted as stable and unspecified. The allowed ancillary regimen was extracted as mono, dual and triple+. The heart failure (HF) criteria were extracted as the New York Heart Association (NYHA) HF classes (Classes I-IV). The criteria of the exclusion based on the use of other medications were extracted as steroids, other, other with steroids, or none. The non-cardiovascular comorbidities criteria were extracted as comprehensive, any other, caution and contraindication, or none. The investigation criteria were extracted as the investigation name or none. Alcohol and drugs, carers, mental disability, physical disability and investigator discretion criteria were extracted as yes or no. Eligibility criteria were mainly extracted from ClinicalTrials.gov. However, they were obtained from journal publications, CSRs, or protocols for unregistered trials or where sparse information was available on ClinicalTrials.gov (i.e. such as not reporting key eligibility details for GLP-1 RA trials including HbA1C range, BMI range, renal measurement range and T2DM duration). For more details on the extraction of eligibility criteria, see Table 7.1.

Table 7.1 The protocol for the extraction of eligibility criteria

Eligibility criteria	Value to be extracted	Definition
Minimum BMI	Continuous value(s) for cutoffs	Less than BMI values were excluded
Maximum BMI	Continuous value(s) for cutoffs	More than BMI values were excluded
Minimum age	Continuous value(s) for cutoffs	Less than age values were excluded
Maximum age	Continuous value(s) for cutoffs	More than age values were excluded
Minimum HbA1C	Continuous value(s) for cutoffs	Less than HbA1C values were excluded
Maximum HbA1C	Continuous value(s) for cutoffs	More than HbA1C values were excluded
Minimum FBG	Continuous value(s) for cutoffs	Less than FBG values were excluded
Maximum FBG	Continuous value(s) for cutoffs	More than FBG values were excluded
Minimum T2DM duration	Continuous value(s) for cutoffs	Less than T2DM durations were excluded
Maximum T2DM duration	Continuous value(s) for cutoffs	More than T2DM durations were excluded
Minimum renal function eGFR	Continuous value(s) for cutoffs	Less than eGFR values were excluded
Maximum renal function eGFR	Continuous value(s) for cutoffs	More than eGFR values were excluded
Minimum renal function CrCl	Continuous value(s) for cutoffs	Less than CrCl values were excluded

Maximum renal function CrCl	Continuous value(s) for cutoffs	More than CrCl values were excluded
Minimum SBP	Continuous value(s) for cutoffs	Less than SBP values were excluded
Maximum SBP	Continuous value(s) for cutoffs	More than SBP values were excluded
Minimum DBP	Continuous value(s) for cutoffs	Less than DBP values were excluded
Maximum DBP	Continuous value(s) for cutoffs	More than DBP values were excluded
Heart failure	Excluded NYHA HF Class, e.g. Class II, Class I-III	Excludes based on the NYHA HF Class
Heart failure	Does not exclude based on NYHA HF Class	Does not exclude based on NYHA HF Class
Ancillary regimen category	mono	Allow the use of single ancillary medications
Ancillary regimen category	dual	Allow the use of two ancillary medications
Ancillary regimen category	triple+	Allow the use of three or more ancillary medications
Treatment stability	stable	Require treatment stability
Treatment stability	unspecified	Do not specify the requirement for treatment stability
Alcohol and drugs	yes	Excludes based on the history of drinking alcohol or using drugs
Alcohol and drugs	no	Does not exclude based on the history of drinking alcohol or using drugs

Other medicines	other	Excludes based on other non-antidiabetic medications, e.g. beta-blockers or ACEIs
Other medicines	steroids	Excludes based on the use of steroids
Other medicines	other and steroids	Excludes based on other non-antidiabetic medications and steroids
Other medicines	none	Does not exclude based on other non-antidiabetic medications or steroids
Non-cardiovascular comorbidities	comprehensive	Excludes based on the presence of all/comprehensive comorbidities
Non-cardiovascular comorbidities	any other	Excludes based on the presence of other comorbidities
Non-cardiovascular comorbidities	cautions and contraindications	Excludes only comorbidities listed in the BNF as cautions or contraindications
Non-cardiovascular comorbidities	none	Does not exclude based on the presence of comorbidities
Carers	yes	Excludes if the participant needs a carer
Carers	no	Does not exclude if the participant needs a carer
Mental disability	yes	Excludes based on the presence of any mental disability

Mental disability	no	Does not exclude based on the presence of any mental disability
Physical disability	yes	Excludes based on the presence of any physical disability
Physical disability	no	Does not exclude based on the presence of any physical disability
Investigator discretion	yes	Excludes based on unspecified investigator discretion
Investigator discretion	no	Does not exclude based on unspecified investigator discretion
Investigations	Text of investigation name, e.g., MRI	Excludes based on tests or investigations that are not part of the routine care, e.g., MRI, glycaemic clamps.
Investigations	none	Does not exclude based on tests or investigations that are not part of the routine care

Abbreviations: ACEIs, Angiotensin-converting-enzyme inhibitors. BMI, Body Mass Index, HbA1C, Glycated Haemoglobin, FBG, Fasting Blood Glucose, T2DM, Type 2 Diabetes Miletus, CV, Cardiovascular, eGFR, estimated Glomerular Filtration Rate, CrCl, Creatinine Clearance, SBP, systolic blood pressure, DBP, diastolic blood pressure, HF, Heart Failure, BNF, British National Formulary, MRI, Magnetic Resonance Imaging, NYHA, New York Heart Association.

7.4.4 Protocolisation and assessment of the restrictiveness of the eligibility criteria

High dimensionality of data often leads to redundancy among variables, where multiple variables convey similar information. Eliminating redundant variables is crucial for the statistical analysis and the interpretability of the findings. Therefore, addressing the high dimensionality of eligibility criteria was essential due to the large number of extracted variables and the volume of the data. The extracted eligibility criteria were checked using tables and visualisations to examine if they make sense singly or in combination. Eligibility criteria elements were dropped if they were completely uniform or perfectly correlated with another element. For example, the DBP criteria was dropped because it was perfectly correlated with the SBP criteria. Moreover, two approaches were used to consolidate these criteria into a single variable and reduce dimensionality. As described in Table 7.2, the first approach was to assign highly restrictive cutoffs and levels for each criteria according to a predefined protocol. For example, if a trial excluded patients with age >60 years, the trial was defined as highly restrictive; if not, proceed to BMI. If the trial did not exclude patients with BMI <45, then proceed to other highly restrictive cutoffs or levels. A trial was defined as permissive if it did not have any of the highly restrictive cutoffs and levels listed in Table 7.2 (see Figure 7.1). This approach produced a binary variable that labels trials as highly restrictive or permissive, coded as (1/0). As shown in Table 7.3, the second approach was to score each criterion based on cutoffs and levels assigned in the protocol and produce a continuous eligibility criteria score of up to 55 for each trial. These cutoffs and levels were used to assess the restrictiveness of each criteria.

Both approaches were developed with clinical input. A diabetologist, Dr Elaine Butterly and my project supervisors, Professor David McAllister and Dr Peter Hanlon, contributed their clinical knowledge to help formulate clinically informed approaches. Determining the pragmatism of cutoffs and levels with clinical insight contributed to the robustness and reliability of these approaches. Eligibility criteria included for this analysis were age, BMI, HbA1C, T2DM duration, SBP, eGFR, CrCl, treatment stability, ancillary regimen, HF, use of other medications, non-cardiovascular comorbidities, alcohol and drugs, and investigator discretion. Excluded criteria were omitted because they either showed a linear correlation with other criteria or were completely uniform. Excluded criteria included FBG, DBP, carers, mental disability, physical disability, and investigations.

Table 7.2 Binary restrictiveness score based on highly restrictive cutoffs or levels

Eligibility criteria	Define trial as highly restrictive if it excludes
Age	Trial maximum age is ≤ 60 years old
BMI	Trial maximum BMI is ≤ 30 kg/m2 and/or trial minimum BMI ≥ 45 kg/m^2
HbA1C	Trial maximum HbA1C ≤ 8% (64 mmol/mol) and/or trial minimum HbA1C ≥11% (97 mmol/mol)
T2DM Duration	Trial minimum duration of the disease is ≥ 120 months, and trial maximum duration of the disease is ≤ 12 months
SBP	Trial maximum SBP is ≤ 140 mmHg
eGFR	Trial minimum eGFR is ≥ 60 ml/min/1.73m^2
CrCl	Trial minimum CrCl is ≥ 100 ml/min
Heart Failure	"Class I-IV" (i.e., exclude all HF classes)
Alcohol/Drug Use	"yes"
Other medicines	"other"
Non-CV Comorbidities	"comprehensive"
Investigator Discretion	"yes"
Ancillary regimen category	"mono"
Treatment stability	"stable"

Abbreviations: BMI, Body Mass Index, HbA1C, Glycated Haemoglobin, T2DM, Type 2 Diabetes Miletus, SBP, systolic blood pressure, eGFR, estimated Glomerular Filtration Rate, CrCl, Creatinine Clearance, CV, Cardiovascular.

Figure 7.1 Example of scoring eligibility criteria using the binary restrictiveness score

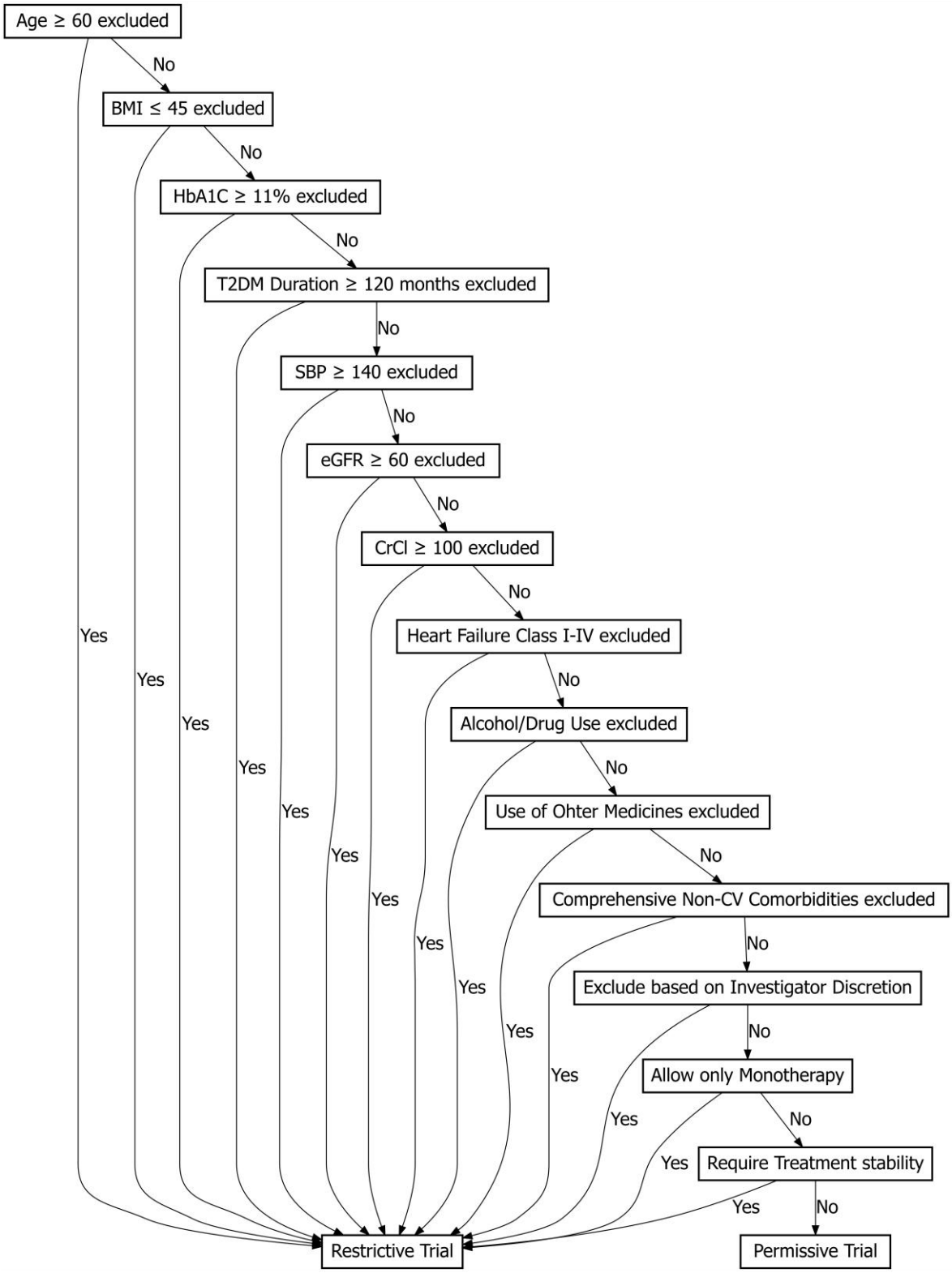


Table 7.3 Scores for each variable based on specific cutoffs

Eligibility criteria	level	Level position	Level position string
Minimum BMI	lower BMI limit ≥45 kg/m^2 (i.e. at least morbidly obese)	1	least permissive
Minimum BMI	lower BMI limit ≥30 kg/m^2 but ≤40 (i.e. at least obese)	2	middle permissive
Minimum BMI	lower BMI limit ≥25 kg/m^2 but ≤30 kg/m^2 (i.e. at least overweight)	3	middle 2 permissive
Minimum BMI	no lower BMI limit or lower limit ≥18.5 kg/m^2 but ≤25 (i.e. normal weight)	4	most permissive
Maximum BMI	upper BMI limit ≤35 kg/m^2	1	least permissive
Maximum BMI	upper BMI limit ≥35 kg/m^2 but ≤40 kg/m^2	2	middle permissive
Maximum BMI	no upper BMI limit or limit ≥40 kg/m^2	3	most permissive
Minimum age	lower age limit 40-60 years	1	least permissive
Minimum age	lower age limit 30-40 years	2	middle permissive
Minimum age	lower age limit 21-30 years	3	middle 2 permissive
Minimum age	no lower age limit or lower age limit 18-21 years	4	middle 3 permissive
Maximum age	upper age limit ≤60 years	1	least permissive
Maximum age	upper age limit ≥60 but ≤65 years	2	middle permissive
Maximum age	upper age limit ≥65 but ≤75 years	3	middle 2 permissive
Maximum age	upper age limit ≥75 but ≤85 years	4	middle 3 permissive
Maximum age	no upper age limit or upper age limit ≥85 years	5	most permissive

Minimum HbA1C	Lower limit on HbA1C ≥80 mmol/mol (9.5%)	1	least permissive
Minimum HbA1C	Lower limit on HbA1C ≥69.5 mmol/mol (8.5%) and ≤80 mmol/mol (9.5%)	2	middle permissive
Minimum HbA1C	lower limit on HbA1C ≥58.5 mmol/mol (7.5%) and ≤69.5 mmol/mol (8.5%)	3	middle 2 permissive
Minimum HbA1C	no lower limit on baseline HbA1C or lower limit on HbA1C ≥53 mmol/mol (7%) and ≤58.5 mmol/mol (7.5%)	4	most permissive
Maximum HbA1C	upper limit on HbA1C ≤75 mmol/mol (9%)	1	least permissive
Maximum HbA1C	upper limit on HbA1C ≥75 mmol/mol (9%) but <86 mmol/mol (10%)	2	middle permissive
Maximum HbA1C	upper limit on HbA1C ≤97 mmol/mol (11%) and ≥86 mmol/mol (10%)	3	middle 2 permissive
Maximum HbA1C	no upper limit on baseline HbA1C or upper limit on HbA1C ≥97 mmol/mol (11%)	4	most permissive
Minimum renal function eGFR	lower limit on eGFR ≥60 ml/min/1.73 m²	1	least permissive
Minimum renal function eGFR	lower limit on eGFR ≥45 ml/min/1.73 m²	2	middle permissive
Minimum renal function eGFR	no lower limit on eGFR or lower limit on eGFR ≥30 ml/min/1.73 m²	3	most permissive
Minimum renal function CrCl	lower limit on eGFR ≥100 ml/min	1	least permissive
Minimum renal function CrCl	lower limit on eGFR ≥60 ml/min	2	middle permissive
Minimum renal function CrCl	no lower limit on CrCl or ≤60 ml/min	3	most permissive

Minimum SBP	Lower limit ≥120 mmHg	1	Least permissive
Minimum SBP	Lower limit ≥100 and ≤120 mmHg	2	middle permissive
Minimum SBP	No lower limit or lower limit ≤ 100 mmHg	3	Most permissive
Maximum SBP	Upper limit ≤140 mmHg	1	Least permissive
Maximum SBP	Upper limit ≥140 and ≤180 mmHg	2	middle permissive
Maximum SBP	No upper limit or upper limit ≥180 mmHg	3	Most permissive
Heart failure	Excludes NYHA HF Class I-IV	1	least permissive
Heart failure	Excludes NYHA HF Class II-IV or III-IV	2	middle permissive
Heart failure	Excludes only NYHA HF Class IV	3	most permissive
Heart failure	Does not exclude based on any NYHA HF Class	3	most permissive
Alcohol and drugs	Excludes based on the history of drinking alcohol or using drugs	1	least permissive
Alcohol and drugs	Does not exclude based on the history of drinking alcohol or using drugs	2	most permissive
Other medicines	Excludes based on other non-antidiabetic medications, e.g. beta-blockers or ACEIs	1	least permissive
Other medicines	Does not exclude based on other non-antidiabetic medications	2	most permissive
Ancillary regimen category	Allow the use of single ancillary medication	1	least permissive

Ancillary regimen category	Allow the use of two ancillary medication	2	middle permissive
Ancillary regimen category	Allow the use of three or more ancillary medications	3	most permissive
Treatment stability	Require treatment stability	1	least permissive
Treatment stability	Unspecified	2	most permissive
Non-cardiovascular comorbidities	Excludes based on the presence of all/comprehensive comorbidities	1	Least permissive
Non-cardiovascular comorbidities	Excludes based on the presence of other comorbidities	2	middle permissive
Non-cardiovascular comorbidities	Does not exclude based on the presence of comorbidities or excludes only comorbidities listed in the BNF as cautions or contraindications	3	Most permissive
Investigator discretion	Excludes based on unspecified investigator discretion	1	least permissive
Investigator discretion	Does not exclude based on unspecified investigator discretion	2	most permissive

Abbreviations: BMI, Body Mass Index, HbA1C, Glycated Haemoglobin, T2DM, Type 2 Diabetes Miletus, CV, Cardiovascular, eGFR, estimated Glomerular Filtration Rate, CrCl, Creatinine Clearance, SBP, systolic blood pressure, SU, Sulfonylurea, BNF, British National Formulary, NYHA, New York Heart Association.

7.4.5 Statistical analysis

Descriptive statistics

Descriptive statistics were used to summarise data and results. Frequencies and percentages were used to describe the categorical eligibility criteria of all GLP-1 RA trials. Statistical summaries and cumulative plots were used to show the distributions of continuous eligibility criteria included for analysis.

Regression analysis

As previously explained in 3.4.4.2, overdispersion was examined by comparing the variance to the mean and calculating the ratio of the residual deviance to the residual degrees of freedom. Overdispersion was found with generalised linear regression models using the Poisson distribution.

Therefore, three negative binomial models were fitted to examine the association between eligibility criteria and the SAE rate. First, a negative binomial regression model was fitted to assess the association between highly restrictive eligibility criteria (binary variable) and the SAE rate. Second, another negative binomial regression model was fitted to examine the association between the continuous score of eligibility criteria and the SAE rate. Third, a negative binomial regression model was fitted to examine the association between specific restrictive criteria and the SAE rate. The outcome variable in these models was the SAE count. An offset was included to account for the variation in person-time.

7.5 Results

7.5.1 SAE reporting in GLP-1 RA trials

SAE reporting in GLP-1 RA trials is described in 3.5.2.

7.5.2 Descriptive statistics

As shown in Figure 7.2, age, BMI, HbA1C, SBP, T2DM duration, eGFR, and CrCl eligibility criteria demonstrated reasonable variability between trials; trials were distributed across multiple cutoffs. Table 7.4 showed that the age and BMI criteria of GLP-1 RA trials were broad and included older adults and obese patients. Patients with high SBP or poor glycaemic control were generally eligible for GLP-1 RA trials. Patients with moderate to severe renal impairment were also eligible for GLP-1 RA trials. These criteria may contribute to observing high SAE rates, especially with highly permissible trials.

As illustrated in Table 7.5, eligibility criteria for heart failure, alcohol and drugs, other medicines, non-cardiovascular comorbidities, and investigator discretion showed noticeable differences in frequencies and percentages across their levels. Most trials had no restriction on HF and non-antidiabetic medications. However, they had restrictions on the allowed number of ancillary medications, treatment stability, alcohol and drugs and non-cardiovascular comorbidities. Moreover, they mostly excluded subjects based on investigator discretion. The total continuous score of eligibility criteria was calculated for all trials. The average continuous score of eligibility criteria was 47.3, with a minimum of 37 and a maximum of 55.

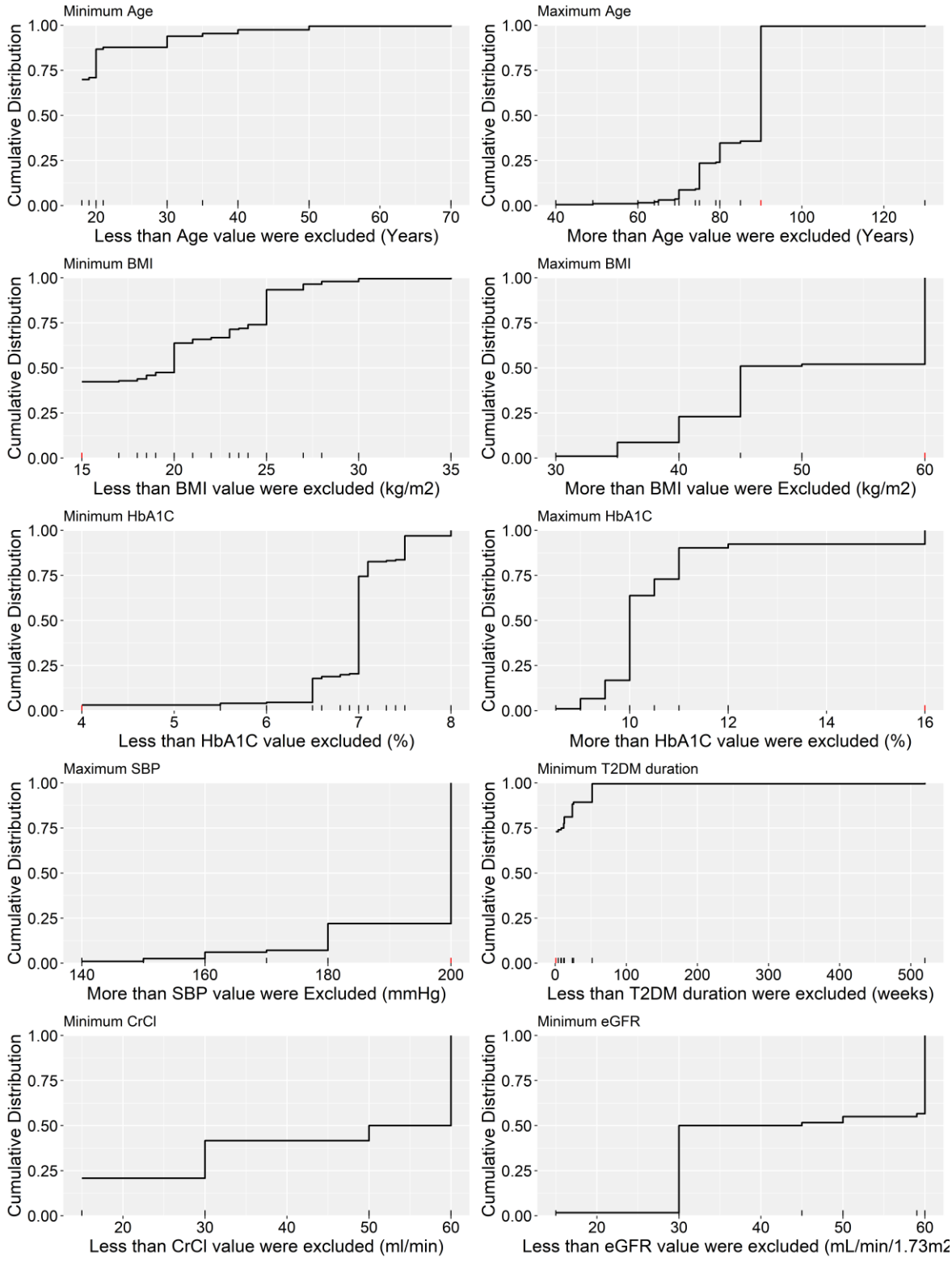
Table 7.4 Descriptive statistics of continuous criteria

Eligibility criteria	Min	Max	Mean	Mode	Not reported
Min age (years)	18	70	20.8	18	2
Max age (years)	40	90	77.1	75	121
Min BMI (kg/m2)	17	35	23	25	83
Max BMI (kg/m2)	30	50	42	45	94
Max SBP (mmHg)	140	180	172.3	180	153
Min HbA1C (%)	5	8	7	7	6
Max HbA1C (%)	8.5	12	10.2	10	15
Min T2DM duration (months)	1.1	13	10	130	143
Min eGFR (ml/min/1.73 m²)	15	60	44.2	30	136
Min CrCl (ml/min)	15	60	43.5	60	172

Table 7.5 The distributions of categorical criteria of GLP-1 RA trials

Variable	Trials (%)
Heart failure	196
Class I-IV	4 (2%)
Class II-IV	5 (2.6%)
Class III-IV	31 (15.8%)
Class IV	31 (15.8%)
Do not exclude subjects even if they have HF	125 (63.8%)
Alcohol and drugs	
Yes	37 (18.9%)
Ancillary regimen category	
Monotherapy	25 (12.8%)
Dual therapy	50 (25.5%)
Triple+ therapy	115 (58.7%)
Treatment stability	
Stable	111 (56.6%)
Unspecified	85 (43.4%)
Other medicines	
Other	82 (41.8%)
none	114 (58.2%)
Non-cardiovascular comorbidities	
Any other	46 (23.5%)
Comprehensive	82 (41.8%)
None	68 (34.7%)
Investigator discretion	
Yes	52 (26.5%)

Figure 7.2 The distributions of continuous eligibility criteria of GLP-1 RA trials



7.5.3 The association between eligibility criteria and the SAE rate

As shown in Figure 7.3, the association between highly restrictive eligibility criteria and the SAE rate was analysed using a generalised linear model with Poisson likelihood and found that permissive trials were associated with increased SAE rates by 29%; CI did not include the null (IRR=1.29, CI 1.26 – 1.32). Furthermore, the association between the continuous score of eligibility criteria and the SAE rate was analysed using a generalised linear model with Poisson likelihood and found that for each increase in the score, the SAE rate increased by 15%, CI did not include the null (IRR= 1.15, CI 1.14 – 1.15). The results from the generalised linear models with Poisson likelihood indicated an association between eligibility criteria and the SAE rate. However, these models did not account for overdispersion. Further models were fitted, which accommodated overdispersion. As illustrated in Table 7.6 and Table 7.7, the results of negative binomial models suggested that the association between eligibility criteria and the SAE rate is also apparent.

As shown in Table 7.8, trials with highly restrictive eGFR were associated with lower SAE rates by 46%; CI did not include the null (IRR= 0.54, CI 0.38 – 0.77). Moreover, trials that excluded patients based on non-CV comorbidities, investigator discretion, use of other medications, allowed number of ancillary medications and treatment stability were associated with decreased SAE rates by 44%, 37%, 47%, 68% and 52%, respectively; all CIs did not include the null. However, other trials with specific highly restrictive eligibility criteria were not associated with SAE rates; all CIs included the null. Moreover, trials that did not have specific highly restrictive criteria were not associated with SAE rates; CI included the null (IRR= 0.84, 95% CI 0.57 – 1.24).

Figure 7.3 The association between eligibility criteria and the SAE rate

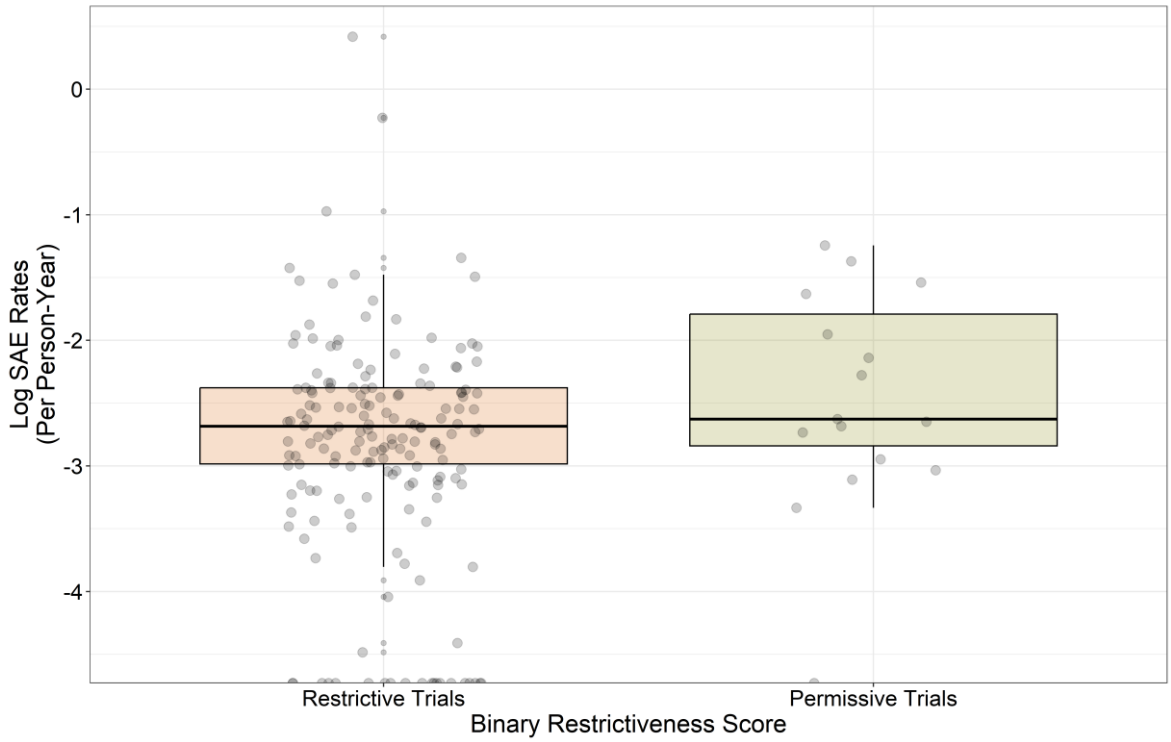


Table 7.6 The association between highly restrictive eligibility criteria and the SAE rate

	Restrictive trials	**Permissive trials**
Number of RCTs	178 (90.8%)	18 (9.2%)
Number of subjects at risk	115670	40425
Number of SAEs	13659	12218
Person years	202288.4	140292.7
IRR GLM (Poisson)	1.29 (95% CI, 1.26 – 1.32)	
IRR negative binomial	1.52 (95% CI, 1.11 – 2.14)	

Table 7.7 The association between continuous eligibility criteria score and the SAE rate

Models	**IRR Estimates**	**95% CI**
GLM (Poisson)	1.15	1.14 – 1.15
Negative binomial	1.04	1.02 – 1.07

Table 7.8 The associations between specific highly restrictive criteria and the SAE rate

Restrictive criteria	IRR Estimates	95% CI
Maximum age	0.47	0.16 – 1.34
Maximum SBP	0.76	0.15 – 3.82
eGFR	0.54	0.38 – 0.77
Non-CV Comorbidities: comprehensive	0.56	0.34 – 0.92
Heart failure: Class I-V	0.51	0.22 – 1.18
Investigator discretion: yes	0.63	0.41 – 0.97
Other medications: other	0.53	0.39 – 0.74
Ancillary regimen: mono	0.32	0.22 – 0.47
Treatment stability: stable	0.48	0.34 – 0.67
Permissive trials	0.84	0.57 – 1.24

7.6 Discussion

In the attempt to explore whether the SAE rate can reflect trial representativeness, this chapter explored the association between the SAE rate and eligibility criteria by conducting two separate approaches. First, it examined the association between highly restrictive eligibility criteria and the SAE rate. Second, it studied the association between the continuous score of eligibility criteria and the SAE rate.

7.6.1 Summary of findings

I found that trials with permissive eligibility criteria were associated with increased SAE rates. Moreover, I found that trials with increased continuous scores of eligibility criteria were associated with increased SAE rates.

7.6.2 Interpretation

Permissive trials had higher SAE rates than highly restrictive trials, and trials with increased continuous eligibility criteria scores were associated with increased SAE rates. These findings suggested that eligibility criteria as a driver of trial representativeness are linked to SAE rates. As expected, permissive trials were associated with higher SAE rates, which could be attributed to the inclusion of older adults and patients with comorbidities in these trials.

Furthermore, the restrictiveness of specific criteria was associated with decreased SAE rates. Trials with restrictive eGFR excluded patients who may be candidates for the intervention in routine care, resulting in SAE rates that do not reflect the general population. Additionally, trials that exclude patients based on non-CV comorbidities, investigator discretion, use of other medications, use of antidiabetics and treatment stability excluded elderly and patients with comorbidities who are typically seen in routine care, thus reporting SAE rates that do not reflect SAE rates in routine care. Consequently, restrictive trials are likely to under-represent older adults and patients with long-term conditions and have lower SAE rates. These observations may support that SAE rates may reflect trial representativeness, though further research is needed to confirm this hypothesis.

7.6.3 Comparison with previous literature

The association between eligibility criteria and the SAE rate has not been widely studied. Most of the literature is focused on quantifying the effect of eligibility criteria on trial representativeness. A study examined the association between ineligibility for psoriasis trials and SAE rates and found that ineligible patients had higher SAE rates than eligible patients (IRR= 2.7, 95% CI 1.5 – 4.7), which was consistent with this study (Garcia-Doval et al., 2012). Furthermore, the findings of an observational study compared SAE rates between standard trials and trials of older people in the RCTs of RAAS drugs for hypertension, which were consistent with this study. They found that older people trials

had higher SAE rates than standard trials (IRR= 1.70, CI 1.07 – 2.77) (Hanlon et al., 2021). Additionally, a pilot study examined the correlation between eligibility criteria and SAEs in 16 sepsis trials, which was consistent with the findings of this study. They found that trials with restrictive eligibility criteria were less likely to have more SAEs (Sen et al., 2017).

7.6.4 Strengths and limitations

This study included trials from different registries, which provided a variety of RCTs. Moreover, this study conducted two separate investigations on the association between eligibility criteria and the SAE rate. Additionally, this study protocolised the cutoffs of assessment and blindly conducted the analysis. One limitation is that this study extracted eligibility criteria mainly from ClinicalTrials.gov for registered trials and journal publications for unregistered trials; other sources were only checked when spars of information was available on ClinicalTrials.gov or journal publications. Another limitation is that this study focuses on GLP-1 RA trials, which may limit the generalisability of the findings to trials of other drug classes or diseases, especially trials of toxic interventions.

7.6.5 Implications

Permissive trials had higher SAE rates than restrictive trials. Moreover, the increased continuous score of eligibility criteria was associated with increased SAE rates. Therefore, the SAE rate can be used as a quick-to-measure proxy of the restrictiveness of eligibility criteria, as it can differentiate between permissive and restrictive trials. Furthermore, they can be cautiously used in conjunction with other measures of trial representativeness. The SAE rate provides a quantitative, tangible, and time-efficient measure that can minimise the subjectivity and complexity associated with other methods. Although the findings of this chapter may support that SAEs may reflect trial representativeness, additional research is needed to validate this hypothesis. Future research may study the association between eligibility criteria and the SAE rate in routine care to examine whether SAE rates differ between trial and general populations. Moreover, investigating the similarities between

eligibility criteria and baseline characteristics may further elucidate the effect of differences in eligibility criteria on trial participation.

7.7 Conclusion

The SAE rate was associated with eligibility criteria. As the SAE rate can differentiate between permissive and restrictive trials, they can be used to measure the restrictiveness of eligibility criteria. Furthermore, they can be cautiously used to measure trial representativeness in combination with other metrics.

Chapter 8 Discussion

8.1 Review of background and main findings

8.1.1 Knowledge gap

RCTs are considered the gold standard for estimating the efficacy of interventions (Akobeng, 2005). However, RCTs with restrictive eligibility criteria usually exclude a notable proportion of the general population, especially older adults and patients with comorbidities (Herrera et al., 2010). Nevertheless, trial representativeness may not always be determined by eligibility criteria. For example, eligible patients may be excluded earlier during the informal recruitment in routine care or refuse to participate in trials (Kemeny et al., 2003; Go et al., 2006; Nipp, Hong and Paskett, 2019). Consequently, these factors may limit trial representativeness, contributing to the uncertainty in decision-making in usual care (Øvretveit, Leviton and Parry, 2011; Blonde et al., 2018).

Although there are several measures of trial representativeness, they are either subjective or challenging to implement (Burchett et al., 2020). The SAE rate has been proposed as a potential measure of trial representativeness (Hanlon et al., 2021, 2022). The use of the SAE rate as a metric can overcome the challenges of other measures by offering a reasonably objective, quantifiable, and time-efficient metric that is likely to be reported for most trials. However, the ability of this metric to compare trial representativeness between trials and capture the under-representation of subgroups such as older adults and patients with comorbidities is still unclear. Therefore, there is a knowledge gap on whether the SAE rate can be used as a measure of trial representativeness.

8.1.2 Approach

In this thesis, I investigated whether the SAE rate can be used as a metric of trial representativeness, using GLP-1 RA trials as an exemplar. **In Chapters 2, 3 and 4,** I explored the feasibility of using the SAE rate as a metric of trial representativeness. **In Chapter 2,** I examined the literature on SAE reporting in RCTs to examine how frequently SAE data are reported. **In Chapter 3,** I assessed SAE capturing in GLP-1 RA trials to enable the calculation of SAE rates. **In Chapter 4,** I compared SAE rates between trial arms to examine if combining SAE rates of the intervention and control arms is feasible, as this will increase the statistical power. **In Chapters 5, 6 and 7,** I used a triangulation of evidence to examine whether the SAE rate can reflect trial representativeness. **In Chapter 5,** I assessed GLP-1 RA trials retrospectively using the PRECIS-2 tool to study the challenges of using this tool and enable a fair comparison with SAE rates. **In Chapter 6,** I examined the association between the SAE rate and the PRECIS-2 tool. Moreover, I compared between the PRECIS-2 tool and the SAE rate based on differences in their associations with several markers of trial representativeness, which serve as fair umpires. **In Chapter 7,** I examined the association between eligibility criteria and the SAE rate.

8.1.3 Main findings and interpretation

8.1.4 The feasibility of using the SAE rate as a metric of trial representativeness

The literature review concluded that SAEs were sufficiently reported, especially on CSRs and trial registries. However, SAE timeframes were frequently underreported. Moreover, journal publications had inconsistent SAE reporting with trial registries and CSRs.

SAEs were reported for most GLP-1 RA trials, indicating the feasibility of calculating SAE rates for most trials. However, less than half of these trials reported timeframes of SAE in the reported event totals AACT table. Therefore, SAE timeframes of these trials were

inferred from timeframes reported in the result group (event type) AACT table and primary outcome AACT table or journal publications. However, this may lead to biased results if inferred timeframes are shorter than actual SAE timeframes. Furthermore, trials had inconsistent reporting of MACE counts in the total SAEs, which was considered while calculating SAE rates.

Interestingly, SAE rates did not differ between placebo and intervention arms nor between intervention and active comparator arms in GLP-1 RA trials, suggesting that SAEs were not mainly attributed to trial interventions and that SAE rates of trial arms can be combined to overcome low event numbers and increase the statistical power. Therefore, SAEs from the intervention and control arms were combined in subsequent analyses.

8.1.5 Triangulation of evidence to determine whether the SAE rate can reflect trial representativeness

Correlations between domains of the PRECIS-2 tool were generally negligible, indicating that the domains are independent, and each domain scored different aspects of trial design as expected for a well-formed score. The assessment using the PRECIS-2 tool was challenging, especially because of the notable subjectivity, the lack of time efficiency and the high rates of missing data in some domains. Despite these challenges, the PRECIS-2 was assessed as objectively as possible to ensure a fair comparison with the SAE rate. The mean PRECIS-2 score was calculated for each trial and used in the comparison with the SAE rate.

No association was found between the normalised mean PRECIS-2 score and the normalised SAE rate. Moreover, no association was found between PRECIS-2 domains and the normalised SAE rate, suggesting that these measures are independent, likely because both reflect representativeness but measure different aspects, particularly, the PRECIS-2 tool may be more focused on the trial design and settings, whereas SAEs are mainly a reflection of the baseline health condition of the recruited patients.

In the absence of a gold standard measure, the use of fair umpires enables the comparison between the SAE rate and the PRECIS-2 score in a balanced way. Most fair umpires were not constantly strongly associated with the normalised SAE rate. Only baseline sample size, baseline mean age, and baseline mean T2DM duration were positively associated with the normalised SAE rate. However, longer trial duration had a negative association with the normalised SAE rate. The analysis of the difference between metrics showed that all fair umpires did not strongly favour the normalised SAE rate. However, the direction of the differences of baseline mean age, baseline male percentage, baseline mean T2DM duration, industry-sponsored trials and phase IV trials favoured the normalised SAE rate as expected, except for phase IV trials, the favouring was unexpected. Other fair umpires showed null differences. Therefore, these findings do not support the use of the SAE rate as a measure of trial representativeness.

Trials with permissive eligibility criteria were associated with increased SAE rates. Moreover, trials with higher eligibility criteria scores were associated with higher SAE rates. These findings indicate that the SAE rate can be used as a quick-to-measure proxy of the restrictiveness of eligibility criteria. Measuring the restrictiveness of eligibility criteria is complicated and time-consuming. Furthermore, these findings suggest that the SAE rate may reflect trial representativeness.

Assessing trial representativeness is complex and may be influenced by different factors. Furthermore, some characteristics that reflect trial participation, such as frailty and multimorbidity, were unavailable. Moreover, there has been no gold standard measure of trial representativeness to compare the SAE rate with. Therefore, I used a triangulation approach to examine whether the SAE rate can reflect trial representativeness. I found no association between the SAE rate and the PRECIS-2 score. Moreover, all fair umpires did not strongly favour the SAE rate. However, the direction of half of the fair umpires slightly favoured the SAE rate. Furthermore, eligibility criteria were associated with the SAE rate. Consequently, the findings of the triangulation of evidence are inconsistent in supporting the use of the SAE rate as a standalone metric of trial representativeness. However, the SAE rate may have potential value when used cautiously in conjunction with other metrics of trial representativeness.

8.2 Contribution to the literature

First, I found that most GLP-1 RA trials reported SAE counts and relevant timeframes. This finding indicated an important methodological aspect for using the SAE rate as a metric of trial representativeness, particularly the calculation of SAE rates for most trials. Moreover, trials registered on ClinicalTrials.gov had more complete SAE reporting than unregistered trials, highlighting the importance of database registration and the value of obtaining data from ClinicalTrials.gov. However, I identified a critical gap in SAE reporting in RCTs, where trials with serious hard outcomes, such as MACE, may not necessarily include their counts in the SAE reporting, resulting in inconsistent serious hard outcome reporting in total SAEs. This inconsistency may significantly impact the implication of the SAE rate as a metric of trial representativeness or any research focused on total SAEs. Furthermore, SAE timeframes were not explicitly reported for nearly half of the trials, indicating another gap in SAE reporting. Therefore, rigorous and consistent reporting of serious hard outcomes and SAE timeframes in RCTs is advised.

Furthermore, I found that SAE rates did not differ between intervention and control arms of GLP-1 RA trials, indicating an important methodological aspect of using the SAE rate as a metric of trial representativeness in future research, particularly the ability to combine SAE rates and increase the statistical power. However, this may not apply to trials of toxic interventions as SAE rates may differ between interventions and placebo arms.

Additionally, I protocolised and operationalised the PRECIS-2 tool to assess GLP-1 RA trials, contributing to a more objective and standardised assessment. The process involved modifying the tool to suit the unique characteristics and requirements of GLP-1 RA trials, providing a standardised framework to help researchers use this tool for GLP-1 RA trials. Furthermore, tailoring the PRECIS-2 tool for GLP-1 RA trials would help improve the translatability of research findings into clinical practice. Furthermore, the retrospective assessment of the pragmatism of a wide range of GLP-1 RA trials using the PRECIS-2 tool may help understand the applicability of GLP-1 RA trials to the real-world and consequently improve the quality of decision-making in clinical practice. Moreover, it identified challenges researchers may face when retrospectively using the PRECIS-2 tool.

Finally, I explored whether the SAE rate can reflect trial representativeness, showing the factors that may or may not be associated with the SAE rate. I found that half of the umpires slightly favoured the SAE rate over the PRECIS-2 score. Moreover, eligibility criteria were associated with the SAE rate. Therefore, the SAE rate can be used as a quick-to-measure proxy of the restrictiveness of eligibility criteria. Measuring the restrictiveness of eligibility criteria is a complex and time-consuming process. Although the evidence is still insufficient to suggest that the SAE rate can be solely used as a measure of trial representativeness, incorporating the SAE rate alongside other metrics of trial representativeness may enhance their utility.

8.3 Implications

The findings of this thesis contributed to a better understanding of the use of the SAE rate as a metric of trial representativeness. There are several areas where this thesis can help clinical practice, researchers, and policymakers. Firstly, sufficient SAE reporting enables the calculation of SAE rates for most trials, accounting for the differences in SAE timeframes when comparing SAEs between trials. Furthermore, registering trials and posting results on ClinicalTrials.gov can enhance SAE reporting due to its mandatory SAE reporting, standardised reporting format, and publicly accessible platform. However, the inconsistent reporting of SAE timeframes in the reported event totals AACT table highlights the need for standardised and adequate reporting practices.

Additionally, the lack of the difference in SAE rates between intervention and control arms of GLP-1 RA trials indicates the feasibility of combining SAE rates from all arms. Using total SAE rates to assess trial representativeness is more powerful, as it deals with low event numbers by increasing the statistical power.

Furthermore, the association between the SAE rate and eligibility criteria indicates that the SAE rate can be used as a quick-to-measure marker of the restrictiveness of eligibility criteria. Adopting the SAE rate as a proxy measure of the restrictiveness of eligibility

criteria is efficient as it addresses the complexity and time-consuming nature of measuring the restrictiveness of eligibility criteria.

Finally, although the triangulation of evidence did not fully support using the SAE rate as a standalone metric of trial representativeness, it can be carefully used in combination with other metrics. The SAE rate provides a relatively objective, quantitative, tangible, and time-efficient metric.

8.4 Strengths and limitations

This thesis has several strengths that have allowed it to get novel and robust insight into whether the SAE rate can be used to measure trial representativeness. Firstly, data used in this thesis were obtained from different sources, including trial registries, CSRs, study protocols and journal publications. This approach contributed to minimising data missingness, which may enhance the credibility and robustness of this thesis. Furthermore, there is no gold standard measure of trial representativeness. Therefore, the SAE rate was compared to an existing proxy measure of trial representativeness (the PRECIS-2 score) and a potential driver of trial representativeness (eligibility criteria). Finally, the high dimensionality of eligibility criteria of GLP-1 RA trials was reduced into a single variable to simplify the analysis of their association with the SAE rate.

However, several limitations may have affected the findings of this thesis. Firstly, this thesis used GLP-1 RA trials as an exemplar, which could limit the generalisability of the findings to trials of other diseases and drug classes. Researchers should be cautious when applying thesis results to other trials, especially trials of toxic interventions, where SAE of trial interventions may complicate the analysis. Secondly, the information needed to score some domains of the PRECIS-2 tool was unavailable. However, the PRECIS-2 mean scores were calculated without the missing domains to minimise the impact of missingness. Still, the analysis could have been affected by missing data. Thirdly, although the PRECIS-2 tool was operationalised and protocolised to reduce the subjectivity of assessment, this thesis was still limited by the inherent subjectivity of the PRECIS-2 score,

which may introduced bias to the comparison with the SAE rate. Fourth, baseline and trial characteristics were missing in some trials. However, missing data were imputed to minimise the impact of missingness on the results. Fifth, trials had heterogeneous reporting of their eligibility criteria. It was unclear whether trials reported all eligibility criteria on ClinicalTrials.gov or publications and whether these criteria are sometimes copied from similar studies. This uncertainty is problematic as it may have led to biased results. Finally, there is a lack of information to better assess trial participation, such as multimorbidity and frailty. This information was not available across data sources used in this thesis. Factors such as multimorbidity and frailty are important to understand whether the SAE rate can reflect trial representativeness.

8.5 Future research

Based on the findings of this thesis, several areas of research could be explored in the future. Here are some potential directions for future research. Firstly, future studies may examine the similarities between trial baseline characteristics and trial eligibility criteria, as this may further explain the effect of the differences in eligibility criteria on trial representativeness. Secondly, future research may investigate that vulnerable eligible subjects may be disproportionately excluded during the informal referral in usual care. Thirdly, future research may study the association between eligibility criteria and the SAE rate in routine care to examine whether SAE rates differ between trial and general populations. Finally, future research may use Individual participant data (IPD) to explore the differences in SAEs between subgroups in large pragmatic trials. This exploration may provide insight into which subgroup may experience higher SAEs. Moreover, it may explore the prevalence of specific SAEs, such as cardiovascular events.

8.6 Conclusion

In trials for GLP-1 receptor agonists for type 2 diabetes, SAE counts and relevant timeframes were reported for most trials. Moreover, no differences in SAE rates were found between intervention and control arms. Therefore, it is feasible to calculate SAE rates for most trials and combine SAE rates from intervention and control arms.

The SAE rate was not associated with the PRECIS-2 score. Moreover, none of the umpires strongly favoured the SAE rate over the PRECIS-2 score. However, half of these umpires slightly favoured the SAE rate. Furthermore, the SAE rate was associated with eligibility criteria. Although this triangulation of evidence is inconsistent in indicating that the SAE rate can be used as a standalone measure of trial representativeness. However, there may still be value in examining the SAE rate when assessing trial representativeness when considered carefully in conjunction with other measures.

Future studies may examine the association between eligibility criteria and the SAE rate in routine care. Future research may also examine the association between baseline characteristics and eligibility criteria. Moreover, it may investigate the premature exclusion of eligible patients during informal referral in usual care. Finally, this thesis contributed to the literature by improving the understanding of the complexities surrounding the use of the SAE rate to measure trial representativeness. It also highlighted the challenges and possible solutions to improve SAE reporting in RCTs.

References

Abdel Shaheed, C. et al. (2022) ‘Strengthening the reporting of harms of all interventions in clinical trials.’, The Medical journal of Australia, 217(10), pp. 502–504. doi: 10.5694/mja2.51755.

Adam, G. P. et al. (2018) ‘Does information from ClinicalTrials.gov increase transparency and reduce bias? Results from a five-report case series.’, Systematic reviews, 7(1), p. 59. doi: 10.1186/s13643-018-0726-5.

Akobeng, A. K. (2005) ‘Understanding randomised controlled trials.’, Archives of disease in childhood, 90(8), pp. 840–4. doi: 10.1136/adc.2004.058222.

Anderson, R. M. et al. (2024) ‘Reporting of Harms in Randomized Controlled Trials Published in Urology Journals: An Updated Analysis’, Journal of Urology, 211(1), pp. 48–54. doi: 10.1097/JU.0000000000003740.

Aslam, A. et al. (2013) ‘Registration of clinical trials: Is it really needed?’, North American Journal of Medical Sciences, pp. 713–715. doi: 10.4103/1947-2714.123266.

Asmar, R. and Hosseini, H. (2009) ‘Endpoints in clinical trials: Does evidence only originate from hard’ or mortality endpoints?’, Journal of Hypertension, 27(SUPPL. 2), pp. 45–50. doi: 10.1097/01.hjh.0000354521.75074.67.

Averitt, A. J. et al. (2020) ‘Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations’, npj Digital Medicine. Springer US, 3(1). doi: 10.1038/s41746-020-0277-8.

Aves, T. et al. (2017) ‘The role of pragmatism in explaining heterogeneity in meta-analyses of randomised trials: A protocol for a cross-sectional methodological review’, BMJ Open, 7(9), pp. 1–6. doi: 10.1136/bmjopen-2017-017887.

Aws. Rutjes et al. (2007) ‘Evaluation of diagnostic tests when there is no gold standard’, Health Technology Assessment, 11(50), pp. 1–86.

Belknap, S. M. et al. (2013) ‘Reporting of serious adverse events during cancer clinical trials to the institutional review board: An evaluation by the research on adverse drug events and reports (RADAR) project’, Journal of Clinical Pharmacology, 53(12), pp. 1334–1340. doi:

10.1002/jcph.177.

Bernhardt, J. et al. (2015) ‘Exploring threats to generalisability in a large international rehabilitation trial (AVERT).’, BMJ open, 5(8), p. e008378. doi: 10.1136/bmjopen-2015-008378.

Bethel, M. A. et al. (2017) ‘Assessing the Safety of Sitagliptin in Older Participants in the Trial Evaluating Cardiovascular Outcomes with Sitagliptin (TECOS).’, Diabetes care, 40(4), pp. 494–501. doi: 10.2337/dc16-1135.

Bhaskar, S. B. (2018) ‘Clinical trial registration: A practical perspective.’, Indian journal of anaesthesia, 62(1), pp. 10–15. doi: 10.4103/ija.IJA_761_17.

Bhide, A., Shah, P. S. and Acharya, G. (2018) ‘A simplified guide to randomized controlled trials.’, Acta obstetricia et gynecologica Scandinavica, 97(4), pp. 380–387. doi: 10.1111/aogs.13309.

Blonde, L. et al. (2018) ‘Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician’, Advances in Therapy. Springer Healthcare, 35(11), pp. 1763–1774. doi: 10.1007/s12325-018-0805-y.

Bolland, M. J. et al. (2013) ‘Differences between self-reported and verified adverse cardiovascular events in a randomised clinical trial’, BMJ Open, 3(3), pp. 1–6. doi: 10.1136/bmjopen-2012-002334.

Bolton, M. et al. (2019) ‘Serious adverse events reported in placebo randomised controlled trials of oral naltrexone : a systematic review and meta- analysis’, pp. 1–13.

Braslow, J. T. et al. (2005) ‘Generalizability of studies on mental health treatment and outcomes, 1981 to 1996.’, Psychiatric services (Washington, D.C.), 56(10), pp. 1261–8. doi: 10.1176/appi.ps.56.10.1261.

Burchett, H. E. D. et al. (2020) ‘When assessing generalisability, focusing on differences in population or setting alone is insufficient.’, Trials, 21(1), p. 286. doi: 10.1186/s13063-020-4178-6.

Butterly, E. et al. (2022) ‘Calibrating a network meta-analysis of diabetes trials of sodium glucose cotransporter 2 inhibitors, glucagon-like peptide-1 receptor analogues and dipeptidyl peptidase-4 inhibitors to a representative routine population: a systematic review

protocol.', BMJ open, 12(10), p. e066491. doi: 10.1136/bmjopen-2022-066491.

Cader, M. Z. et al. (2011) 'Performance of the 2010 ACR/EULAR criteria for rheumatoid arthritis: comparison with 1987 ACR criteria in a very early synovitis cohort.', Annals of the rheumatic diseases, 70(6), pp. 949–55. doi: 10.1136/ard.2010.143560.

Cahan, A., Cahan, S. and Cimino, J. J. (2017) 'Computer-aided assessment of the generalizability of clinical trial results', International Journal of Medical Informatics, 99, pp. 60–66. doi: 10.1016/j.ijmedinf.2016.12.008.

Cardoso, J. R. et al. (2014) 'What is gold standard and what is ground truth?', Dental Press Journal of Orthodontics, 19(5), pp. 27–30. doi: 10.1590/2176-9451.19.5.027-030.ebo.

Chan, A.-W. and Hróbjartsson, A. (2018) 'Promoting public access to clinical trial protocols: challenges and recommendations.', Trials, 19(1), p. 116. doi: 10.1186/s13063-018-2510-1.

Chen, T. W. et al. (2015) 'A systematic review of immune-related adverse event reporting in clinical trials of immune checkpoint inhibitors', Annals of Oncology, 26(9), pp. 1824–1829. doi: 10.1093/annonc/mdv182.

Choi, M. Y. et al. (2019) 'Pragmaticism of Randomized Controlled Trials of Biologic Treatment With Methotrexate in Rheumatoid Arthritis: A Systematic Review', Arthritis Care and Research, 71(5), pp. 620–628. doi: 10.1002/acr.23620.

Ciani, O. et al. (2022) 'Surrogate endpoints in trials: a call for better reporting', Trials, 23(1), pp. 1–4. doi: 10.1186/s13063-022-06904-7.

Ciolino, J. D., Kaizer, A. M. and Bonner, L. B. (2023) 'Guidance on interim analysis methods in clinical trials.', Journal of clinical and translational science, 7(1), p. e124. doi: 10.1017/cts.2023.552.

Clapp, J. T. et al. (2023) 'Clinical reasoning in pragmatic trial randomization: a qualitative interview study.', Trials, 24(1), p. 431. doi: 10.1186/s13063-023-07445-3.

Clark, L. T. et al. (2019) 'Increasing Diversity in Clinical Trials: Overcoming Critical Barriers', Current Problems in Cardiology, 44(5), pp. 148–172. doi: 10.1016/j.cpcardiol.2018.11.002.

Clemens, K. K. et al. (2020) 'Sex disparities in cardiovascular outcome trials of populations with diabetes: A Systematic Review and Meta-analysis', Diabetes Care, 43(5), pp. 1157–

1163. doi: 10.2337/dc19-2257.

Crépin, S., Villeneuve, C. and Merle, L. (2016) ‘Quality of serious adverse events reporting to academic sponsors of clinical trials: far from optimal’, Pharmacoepidemiology and Drug Safety, 25(6), pp. 719–724. doi: 10.1002/pds.3982.

Dagenais, S. et al. (2022) ‘Use of Real-World Evidence to Drive Drug Development Strategy and Inform Clinical Trial Design.’, Clinical pharmacology and therapeutics, 111(1), pp. 77–89. doi: 10.1002/cpt.2480.

Dal-Ré, R. (2020) ‘Articles provided insufficient information to conduct an appropriate retrospective assessment of the pragmatic/explanatory features of medicine trials with the PRECIS-2 tool’, European Journal of Clinical Pharmacology, 76(8), pp. 1093–1102. doi: 10.1007/s00228-020-02901-4.

Dal-Ré, R., Janiaud, P. and Ioannidis, J. P. A. (2018) ‘Real-world evidence: How pragmatic are randomized controlled trials labeled as pragmatic?’, BMC Medicine. BMC Medicine, 16(1), pp. 1–6. doi: 10.1186/s12916-018-1038-2.

Davis, A. L. and Miller, J. D. (2017) ‘The European Medicines Agency and Publication of Clinical Study Reports: A Challenge for the US FDA.’, JAMA, 317(9), pp. 905–906. doi: 10.1001/jama.2017.0918.

De Angelis, C. et al. (2004) ‘Clinical trial registration: a statement from the International Committee of Medical Journal Editors.’, The New England journal of medicine, 351(12), pp. 1250–1. doi: 10.1056/NEJMe048225.

de Vries, Y. A. et al. (2016) ‘Bias in the reporting of harms in clinical trials of second-generation antidepressants for depression and anxiety: A meta-analysis’, European Neuropsychopharmacology. Elsevier, 26(11), pp. 1752–1759. doi: 10.1016/j.euroneuro.2016.09.370.

Dekkers, O. M. et al. (2010) ‘How to assess the external validity of therapeutic trials: A conceptual approach’, International Journal of Epidemiology, 39(1), pp. 89–94. doi: 10.1093/ije/dyp174.

Del Olmo-Garcia, M. I. and Merino-Torres, J. F. (2018) ‘GLP-1 Receptor Agonists and Cardiovascular Disease in Patients with Type 2 Diabetes.’, Journal of diabetes research, 2018, p. 4020492. doi: 10.1155/2018/4020492.

Dennis, J. M. (2020) ‘Precision Medicine in Type 2 Diabetes: Using Individualized Prediction Models to Optimize Selection of Treatment.’, Diabetes, 69(10), pp. 2075–2085. doi: 10.2337/dbi20-0002.

DeVito, N. J., Bacon, S. and Goldacre, B. (2020) ‘Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study.’, Lancet (London, England), 395(10221), pp. 361–369. doi: 10.1016/S0140-6736(19)33220-9.

Devos, F. et al. (2019) ‘Study characteristics impacted the pragmatism of randomized controlled trial published in nursing: a meta-epidemiological study’, Journal of Clinical Epidemiology, 116, pp. 18–25. doi: 10.1016/j.jclinepi.2019.07.017.

Dimairo, M. et al. (2018) ‘Development process of a consensus-driven CONSORT extension for randomised trials using an adaptive design’, BMC Medicine. BMC Medicine, 16(1), pp. 1–20. doi: 10.1186/s12916-018-1196-2.

Doshi, P. and Jefferson, T. (2013) ‘Clinical study reports of randomised controlled trials: an exploratory review of previously confidential industry reports.’, BMJ open, 3(2). doi: 10.1136/bmjopen-2012-002496.

Emdin, C. et al. (2015) ‘Association of cardiovascular trial registration with positive study findings: Epidemiological study of randomized trials (ESORT)’, JAMA Internal Medicine, 175(2), pp. 304–307. doi: 10.1001/jamainternmed.2014.6924.

Ettori-Ajasse, I. et al. (2020) ‘How pragmatic are the randomised trials used in recommendations for control of glycosylated haemoglobin levels in type 2 diabetic patients in general practice: An application of the PRECIS II tool’, Trials, 21(1), pp. 20–26. doi: 10.1186/s13063-020-4215-5.

Evans, S. R. (2010) ‘Fundamentals of clinical trial design.’, Journal of experimental stroke & translational medicine, 3(1), pp. 19–27. doi: 10.6030/1939-067x-3.1.19.

Ferreira-González, I. et al. (2007) ‘Problems with use of composite end points in cardiovascular trials: Systematic review of randomised controlled trials’, British Medical Journal, 334(7597), pp. 786–788. doi: 10.1136/bmj.39136.682083.AE.

Fitzpatrick, C. et al. (2020) ‘Effect of pragmatic versus explanatory interventions on medication adherence in people with cardiometabolic conditions: A systematic review and meta-analysis’, BMJ Open. doi: 10.1136/bmjopen-2019-036575.

Forbes, G. et al. (2017) ‘Understanding the applicability of results from primary care trials: lessons learned from applying PRECIS-2’, Journal of Clinical Epidemiology. Elsevier Inc, 90, pp. 119–126. doi: 10.1016/j.jclinepi.2017.06.007.

Ford, I. and Norrie, J. (2016) ‘Pragmatic Trials’, New England Journal of Medicine. Edited by J. M. Drazen et al., 375(5), pp. 454–463. doi: 10.1056/NEJMra1510059.

Fox, C. S. et al. (2004) ‘The Significant Effect of Diabetes Duration on Coronary Heart Disease Mortality: The Framingham Heart Study’, Diabetes Care, 27(3), pp. 704–708. doi: 10.2337/diacare.27.3.704.

Garcia-Doval, I. et al. (2012) ‘Risk of serious adverse events associated with biologic and nonbiologic psoriasis systemic therapy: patients ineligible vs eligible for randomized controlled trials.’, Archives of dermatology, 148(4), pp. 463–70. doi: 10.1001/archdermatol.2011.2768.

Gartlehner, G. et al. (2006) ‘A simple and valid tool distinguished efficacy from effectiveness studies.’, Journal of clinical epidemiology, 59(10), pp. 1040–8. doi: 10.1016/j.jclinepi.2006.01.011.

Gewandter, J. S. et al. (2015) ‘Reporting of adverse events and statistical details of efficacy estimates in randomized clinical trials of pain in temporomandibular disorders: Analgesic, anesthetic, and addiction clinical trial translations, innovations, opportunities, and networks syst’, Journal of the American Dental Association. Elsevier Inc, 146(4), pp. 246-254.e6. doi: 10.1016/j.adaj.2014.12.023.

Ghouse, J. et al. (2020) ‘Effect of diabetes duration on the relationship between glycaemic control and risk of death in older adults with type 2 diabetes’, Diabetes, Obesity and Metabolism, 22(2), pp. 231–242. doi: 10.1111/dom.13891.

Girbes, A. R. J. and de Grooth, H.-J. (2020) ‘Time to stop randomized and large pragmatic trials for intensive care medicine syndromes: the case of sepsis and acute respiratory distress syndrome’, Journal of Thoracic Disease, 12(S1), pp. S101–S109. doi: 10.21037/jtd.2019.10.36.

Glasgow, R. E. et al. (2012) ‘Applying the PRECIS criteria to describe three effectiveness trials of weight loss in obese patients with comorbid conditions’, Health Services Research, 47(3 PART 1), pp. 1051–1067. doi: 10.1111/j.1475-6773.2011.01347.x.

Glasziou, P., Irwig, L. and Deeks, J. J. (2008) ‘When should a new test become the current reference standard?’, Annals of Internal Medicine, 149(11), pp. 816–821. doi: 10.7326/0003-4819-149-11-200812020-00009.

Go, R. S. et al. (2006) ‘Clinical trial accrual among new cancer patients at a community-based cancer center.’, Cancer, 106(2), pp. 426–33. doi: 10.1002/cncr.21597.

Grayek, E. N. et al. (2022) ‘The value of adherence information during clinical pharmaceutical trials.’, Clinical trials (London, England), 19(3), pp. 326–336. doi: 10.1177/17407745221084127.

Gresham, G. et al. (2022) ‘Update on the clinical trial landscape: analysis of ClinicalTrials.gov registration data, 2000–2020’, Trials, 23(1), p. 858. doi: 10.1186/s13063-022-06569-2.

Gumber, L. et al. (2024) ‘Operational complexities in international clinical trials: a systematic review of challenges and proposed solutions.’, BMJ open, 14(4), p. e077132. doi: 10.1136/bmjopen-2023-077132.

Gupta, S. (2011) ‘Intention-to-treat concept: A review’, Perspectives in Clinical Research, 2(3), p. 109. doi: 10.4103/2229-3485.83221.

Gupta, U. and Verma, M. (2013) ‘Placebo in clinical trials’, Perspectives in Clinical Research, 4(1), p. 49. doi: 10.4103/2229-3485.106383.

Haleem, R. M. et al. (2015) ‘Quality in the pharmaceutical industry - A literature review’, Saudi Pharmaceutical Journal. King Saud University, 23(5), pp. 463–469. doi: 10.1016/j.jsps.2013.11.004.

Hamaker, M. E., Stauder, R. and van Munster, B. C. (2014) ‘Exclusion of older patients from ongoing clinical trials for hematological malignancies: an evaluation of the National Institutes of Health Clinical Trial Registry.’, The oncologist, 19(10), pp. 1069–75. doi: 10.1634/theoncologist.2014-0093.

Hanlon, P. et al. (2019) ‘Representation of people with comorbidity and multimorbidity in clinical trials of novel drug therapies: an individual-level participant data analysis.’, BMC medicine, 17(1), p. 201. doi: 10.1186/s12916-019-1427-1.

Hanlon, P. et al. (2021) ‘Observed and expected serious adverse event rates in randomised

clinical trials for hypertension: an observational study comparing trials that do and do not focus on older people', The Lancet Healthy Longevity. The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license, 2(7), pp. e398–e406. doi: 10.1016/S2666-7568(21)00092-1.

Hanlon, P. et al. (2022) 'Assessing trial representativeness using serious adverse events : an observational analysis using aggregate and individual - level data from clinical trials and routine healthcare data', pp. 1–15.

Hao, T. et al. (2014) 'Clustering clinical trials with similar eligibility criteria features.', Journal of biomedical informatics, 52, pp. 112–20. doi: 10.1016/j.jbi.2014.01.009.

Harrington, J. et al. (2023) 'Medication-Attributable Adverse Events in Heart Failure Trials.', JACC. Heart failure, 11(4), pp. 425–436. doi: 10.1016/j.jchf.2022.11.026.

Harris, K. B. and McCarty, D. J. (2015) 'Efficacy and tolerability of glucagon-like peptide-1 receptor agonists in patients with type 2 diabetes mellitus.', Therapeutic advances in endocrinology and metabolism, 6(1), pp. 3–18. doi: 10.1177/2042018814558242.

Hartung, D. M. et al. (2014) 'Reporting discrepancies between the ClinicalTrials.gov results database and peer-reviewed publications.', Annals of internal medicine, 160(7), pp. 477–83. doi: 10.7326/M13-0480.

He, Z. et al. (2020) 'Clinical Trial Generalizability Assessment in the Big Data Era: A Review', Clinical and Translational Science, 13(4), pp. 675–684. doi: 10.1111/cts.12764.

Hernán, M. A. and Robins, J. M. (2017) 'Per-Protocol Analyses of Pragmatic Trials', New England Journal of Medicine, 377(14), pp. 1391–1398. doi: 10.1056/NEJMsm1605385.

Herrera, A. P. et al. (2010) 'Disparate inclusion of older adults in clinical trials: priorities and opportunities for policy and practice change', American Journal of Public Health, 100(SUPPL. 1), pp. 105–112. doi: 10.2105/AJPH.2009.162982.

Hindiyeh, N. A. et al. (2022) 'Development and validation of a novel model for characterizing migraine outcomes within real-world data', Journal of Headache and Pain. Springer Milan, 23(1), pp. 1–7. doi: 10.1186/s10194-022-01493-x.

Hirsch, B. R. et al. (2013) 'Characteristics of oncology clinical trials: insights from a systematic analysis of ClinicalTrials.gov.', JAMA internal medicine, 173(11), pp. 972–9.

doi: 10.1001/jamainternmed.2013.627.

Hodkinson, A. et al. (2013) 'Reporting of harms data in RCTs: A systematic review of empirical assessments against the CONSORT harms extension', BMJ Open, 3(9), pp. 1–8. doi: 10.1136/bmjopen-2013-003436.

Hodkinson, A. et al. (2021) 'Benefits and harms of Risperidone and Paliperidone for treatment of patients with schizophrenia or bipolar disorder: a meta-analysis involving individual participant data and clinical study reports.', BMC medicine, 19(1), p. 195. doi: 10.1186/s12916-021-02062-w.

Hodkinson, A., Gamble, C. and Smith, C. T. (2016) 'Reporting of harms outcomes: A comparison of journal publications with unpublished clinical study reports of orlistat trials', Trials. Trials, 17(1), pp. 1–11. doi: 10.1186/s13063-016-1327-z.

Howard, L. et al. (2009) 'Why is recruitment to trials difficult? An investigation into recruitment difficulties in an RCT of supported employment in patients with severe mental illness', Contemporary Clinical Trials, 30(1), pp. 40–46. doi: 10.1016/j.cct.2008.07.007.

Hughes, S., Cohen, D. and Jaggi, R. (2014) 'Differences in reporting serious adverse events in industry sponsored clinical trial registries and journal articles on antidepressant and antipsychotic drugs: A cross-sectional study', BMJ Open, 4(7). doi: 10.1136/bmjopen-2014-005535.

Hunsinger, M. et al. (2014) 'Adverse event reporting in nonpharmacologic, noninterventional pain clinical trials: ACTTION systematic review', Pain. International Association for the Study of Pain, 155(11), pp. 2253–2262. doi: 10.1016/j.pain.2014.08.004.

Hurria, A. et al. (2015) 'Improving the Evidence Base for Treating Older Adults With Cancer: American Society of Clinical Oncology Statement.', Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 33(32), pp. 3826–33. doi: 10.1200/JCO.2015.63.0319.

Ioannidis, J. P. A. et al. (2004) 'Better reporting of harms in randomized trials: An extension of the CONSORT statement', Annals of Internal Medicine, 141(10), pp. 781–788. doi: 10.7326/0003-4819-141-10-200411160-00009.

Ioannidis, J. P. A. and Lau, J. (2001) 'Completeness of safety reporting in randomized trials: An evaluation of 7 medical areas', Jama, 285(4), pp. 437–443. doi: 10.1001/jama.285.4.437.

Jakobsen, J. C. et al. (2017) ‘When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts.’, BMC medical research methodology, 17(1), p. 162. doi: 10.1186/s12874-017-0442-1.

James, E. C. et al. (2020) ‘Overlap between adverse events (AEs) and serious adverse events (SAEs): a case study of a phase III cancer clinical trial.’, Trials, 21(1), p. 802. doi: 10.1186/s13063-020-04718-z.

Johansson, M. et al. (2018) ‘Benefits and harms of screening men for abdominal aortic aneurysm in Sweden: a registry-based cohort study.’, Lancet (London, England), 391(10138), pp. 2441–2447. doi: 10.1016/S0140-6736(18)31031-6.

Johnson, K. E. et al. (2016) ‘Use of PRECIS ratings in the National Institutes of Health (NIH) Health Care Systems Research Collaboratory’, Trials. Trials, 17(1). doi: 10.1186/s13063-016-1158-y.

Jones, C. W. et al. (2021) ‘Delays in reporting and publishing trial results during pandemics: cross sectional analysis of 2009 H1N1, 2014 Ebola, and 2016 Zika clinical trials.’, BMC medical research methodology, 21(1), p. 120. doi: 10.1186/s12874-021-01324-8.

Kabisch, M. et al. (2011) ‘Randomisierte kontrollierte studien: Teil 17 der serie zur bewertung wissenschaftlicher publikationen’, Deutsches Arzteblatt, 108(39), pp. 663–668. doi: 10.3238/arztebl.2011.0663.

Kaushal, P. et al. (2022) ‘Evaluation of the design, conduct and reporting of randomised controlled trials in the haemodialysis population: a scoping review and interview study.’, BMJ open, 12(3), p. e058368. doi: 10.1136/bmjopen-2021-058368.

Kavalci, E. and Hartshorn, A. (2023) ‘Improving clinical trial design using interpretable machine learning based prediction of early trial termination.’, Scientific reports, 13(1), p. 121. doi: 10.1038/s41598-023-27416-7.

Kemeny, M. M. et al. (2003) ‘Barriers to clinical trial participation by older women with breast cancer.’, Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 21(12), pp. 2268–75. doi: 10.1200/JCO.2003.09.124.

Kendall, J. M. (2003) ‘Designing a research project: randomised controlled trials and their principles’, Emergency Medicine Journal, 20(2), pp. 164–168. doi: 10.1136/emj.20.2.164.

Kennedy-Martin, T. et al. (2015) ‘A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results’, Trials. Trials, 16(1), pp. 1–14. doi: 10.1186/s13063-015-1023-4.

Kim, E. S. et al. (2021) ‘Continuing to Broaden Eligibility Criteria to Make Clinical Trials More Representative and Inclusive: ASCO-Friends of Cancer Research Joint Research Statement.’, Clinical cancer research : an official journal of the American Association for Cancer Research, 27(9), pp. 2394–2399. doi: 10.1158/1078-0432.CCR-20-3852.

Kirkman, M. S. et al. (2012) ‘Diabetes in Older Adults’, Diabetes Care, 35(12), pp. 2650–2664. doi: 10.2337/dc12-1801.

Kishimoto, I. et al. (2014) ‘Hemoglobin A1c predicts heart failure hospitalization independent of baseline cardiac function or B-type natriuretic peptide level’, Diabetes Research and Clinical Practice. Elsevier Ireland Ltd, 104(2), pp. 257–265. doi: 10.1016/j.diabres.2014.02.009.

Knox, J. et al. (2019) ‘Prevention, screening, and treatment for heavy drinking and alcohol use disorder.’, The lancet. Psychiatry, 6(12), pp. 1054–1067. doi: 10.1016/S2215-0366(19)30213-5.

Koppenaal, T. et al. (2011) ‘Pragmatic vs. explanatory: An adaptation of the PRECIS tool helps to judge the applicability of systematic reviews for daily practice’, Journal of Clinical Epidemiology, 64(10), pp. 1095–1101. doi: 10.1016/j.jclinepi.2010.11.020.

Kurki, S. et al. (2024) ‘A comparative study of clinical trial and real-world data in patients with diabetic kidney disease.’, Scientific reports, 14(1), p. 1731. doi: 10.1038/s41598-024-51938-3.

Kwasi, S. et al. (2022) ‘PLOS ONE Vaccines to prevent COVID-19 : A living systematic review with Trial Sequential Analysis and network meta-analysis of randomized clinical trials’, pp. 1–23. doi: 10.1371/journal.pone.0260733.

Lakey, W. C. et al. (2013) ‘Are current clinical trials in diabetes addressing important issues in diabetes care?’, Diabetologia, 56(6), pp. 1226–35. doi: 10.1007/s00125-013-2890-4.

Li, G. et al. (2018) ‘Enhancing primary reports of randomized controlled trials: Three most common challenges and suggested solutions.’, Proceedings of the National Academy of Sciences of the United States of America, 115(11), pp. 2595–2599. doi:

10.1073/pnas.1708286114.

Li, Q. et al. (2020) ‘Using Real-World Data to Rationalize Clinical Trials Eligibility Criteria Design: A Case Study of Alzheimer’s Disease Trials.’, AMIA ... Annual Symposium proceedings. AMIA Symposium, 2020, pp. 717–726. Available at: http://www.ncbi.nlm.nih.gov/pubmed/33936446.

Lichtman, S. M. et al. (2017) ‘Modernizing Clinical Trial Eligibility Criteria: Recommendations of the American Society of Clinical Oncology-Friends of Cancer Research Organ Dysfunction, Prior or Concurrent Malignancy, and Comorbidities Working Group.’, Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 35(33), pp. 3753–3759. doi: 10.1200/JCO.2017.74.4102.

Lipman, P. D. et al. (2017) ‘Framing the conversation: use of PRECIS-2 ratings to advance understanding of pragmatic trial design domains.’, Trials, 18(1), p. 532. doi: 10.1186/s13063-017-2267-y.

London, J. W. et al. (2009) ‘The automation of clinical trial serious adverse event reporting workflow’, Clinical Trials, 6(5), pp. 446–454. doi: 10.1177/1740774509344778.

Loudon, K. et al. (2013) ‘Making clinical trials more relevant: improving and validating the PRECIS tool for matching trial design decisions to trial purpose.’, Trials, 14, pp. 1–9. doi: 10.1186/1745-6215-14-115.

Loudon, K. et al. (2015) ‘The PRECIS-2 tool: Designing trials that are fit for purpose’, BMJ (Online), 350. doi: 10.1136/bmj.h2147.

Loudon, K. et al. (2017) ‘The PRECIS-2 tool has good interrater reliability and modest discriminant validity’, Journal of Clinical Epidemiology, 88, pp. 113–121. doi: 10.1016/j.jclinepi.2017.06.001.

Lu, L. et al. (2017) ‘The application of PRECIS-2 ratings in randomized controlled trials of Chinese herbal medicine’, Oncotarget, 8(63), pp. 107002–107010. doi: 10.18632/oncotarget.22204.

Luoma, K. A. et al. (2017) ‘How can clinical practices pragmatically increase physical activity for patients with type 2 diabetes? A systematic review’, Translational Behavioral Medicine. Translational Behavioral Medicine, 7(4), pp. 751–772. doi: 10.1007/s13142-017-0502-4.

Madi, K. et al. (2023) ‘Quality of reporting of adverse events in clinical trials of covid-19 drugs: systematic review.’, BMJ medicine, 2(1), p. e000352. doi: 10.1136/bmjmed-2022-000352.

Madsbad, S. (2016) ‘Review of head-to-head comparisons of glucagon-like peptide-1 receptor agonists.’, Diabetes, obesity & metabolism, 18(4), pp. 317–32. doi: 10.1111/dom.12596.

Mahinbakht, A., Lavasani, S. M. and Guirguis, M. (2014) ‘The quality of reporting harms-related data in clinical trials of adjuvant trastuzumab in early-stage breast cancer treatment’, Therapeutic Innovation and Regulatory Science, 48(3), pp. 299–304. doi: 10.1177/2168479013509383.

Maillet, D. et al. (2016) ‘The reporting of adverse events in oncology phase III trials: A comparison of the current status versus the expectations of the EORTC members’, Annals of Oncology, 27(1), pp. 192–198. doi: 10.1093/annonc/mdv485.

Malay, S. and Chung, K. C. (2012) ‘The choice of controls for providing validity and evidence in clinical research’, Plastic and Reconstructive Surgery, 130(4), pp. 959–965. doi: 10.1097/PRS.0b013e318262f4c8.

Malmivaara, A. (2021) ‘Applicability of evidence from randomized controlled trials and systematic reviews to clinical practice: A conceptual review.’, Journal of rehabilitation medicine, 53(6), p. jrm00202. doi: 10.2340/16501977-2843.

Martin, S. S. et al. (2013) ‘Patient- and trial-specific barriers to participation in cardiovascular randomized clinical trials.’, Journal of the American College of Cardiology, 61(7), pp. 762–9. doi: 10.1016/j.jacc.2012.10.046.

Maund, E. et al. (2014) ‘Benefits and harms in clinical trials of duloxetine for treatment of major depressive disorder: Comparison of clinical study reports, trial registries, and publications’, BMJ (Online), 348(June), pp. 1–9. doi: 10.1136/bmj.g3510.

Mayo-Wilson, E. et al. (2017) ‘Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy.’, Journal of clinical epidemiology, 91, pp. 95–110. doi: 10.1016/j.jclinepi.2017.07.014.

Mayo-Wilson, E. et al. (2019) ‘Opportunities for selective reporting of harms in randomized clinical trials: Selection criteria for non-systematic adverse events’, Trials. Trials, 20(1), pp.

1–11. doi: 10.1186/s13063-019-3581-3.

Mc Cord, K. A. et al. (2022) ‘Reporting transparency and completeness in Trials: Paper 2 - reporting of randomised trials using registries was often inadequate and hindered the interpretation of results.’, Journal of clinical epidemiology, 141, pp. 175–186. doi: 10.1016/j.jclinepi.2021.09.012.

Medeiros, F. A. (2017) ‘Biomarkers and Surrogate Endpoints: Lessons Learned From Glaucoma.’, Investigative ophthalmology & visual science, 58(6), pp. BIO20–BIO26. doi: 10.1167/iovs.17-21987.

Ménard, T. et al. (2019) ‘Enabling Data-Driven Clinical Quality Assurance: Predicting Adverse Event Reporting in Clinical Trials Using Machine Learning’, Drug Safety. Springer International Publishing, 42(9), pp. 1045–1053. doi: 10.1007/s40264-019-00831-4.

Michos, E. D., Lopez-Jimenez, F. and Gulati, M. (2023) ‘Role of Glucagon-Like Peptide-1 Receptor Agonists in Achieving Weight Loss and Improving Cardiovascular Outcomes in People With Overweight and Obesity.’, Journal of the American Heart Association, 12(11), p. e029282. doi: 10.1161/JAHA.122.029282.

Miklavcic, J. J. et al. (2020) ‘Effectiveness of a community program for older adults with type 2 diabetes and multimorbidity: a pragmatic randomized controlled trial.’, BMC geriatrics, 20(1), p. 174. doi: 10.1186/s12877-020-01557-0.

Misra, S. (2012) ‘Randomized double blind placebo control studies, the “gold Standard” in intervention based studies’, Indian Journal of Sexually Transmitted Diseases, 33(2), pp. 131–134. doi: 10.4103/0253-7184.102130.

Moberg, C. A. and Humphreys, K. (2017) ‘Exclusion criteria in treatment research on alcohol, tobacco and illicit drug use disorders: A review and critical analysis.’, Drug and alcohol review, 36(3), pp. 378–388. doi: 10.1111/dar.12438.

Moher, D. et al. (2010) ‘CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials.’, BMJ (Clinical research ed.), 340. doi: 10.1136/bmj.c869.

Moher, D. and Olkin, I. (1995) ‘Meta-analysis of randomized controlled trials. A concern for standards’;, JAMA, 274(24), pp. 1962–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8568993.

Monti, S. et al. (2018) ‘Randomized controlled trials and real-world data: differences and similarities to untangle literature data.’, Rheumatology (Oxford, England), 57(57 Suppl 7), pp. vii54–vii58. doi: 10.1093/rheumatology/key109.

Moore, R. A. et al. (2005) ‘Research article Tolerability and adverse events in clinical trials of celecoxib in osteoarthritis and rheumatoid arthritis : systematic review and meta-analysis of information from company clinical trial reports’, 7(3). doi: 10.1186/ar1704.

Moustgaard, H. et al. (2020) ‘Impact of blinding on estimated treatment effects in randomised clinical trials: Meta-epidemiological study’, The BMJ, 368, pp. 1–4. doi: 10.1136/bmj.l6802.

Nardini, C. (2014) ‘The ethics of clinical trials.’, Ecancermedicalscience, 8, p. 387. doi: 10.3332/ecancer.2014.387.

Navarro-pérez, J. et al. (2018) ‘Mortality and cardiovascular disease burden of uncontrolled diabetes in a registry-based cohort. BMC Cardiovasc Disor. 2018;18(180):1-9.’ BMC Cardiovascular Disorders, pp. 1–9.

Negoro, T. et al. (2023) ‘Poor Result Reporting Rate in Cell Therapy Trials Registered at ClinicalTrials.gov.’, Tissue engineering. Part B, Reviews, 29(6), pp. 623–633. doi: 10.1089/ten.TEB.2023.0053.

Niemeijer, A. et al. (2020) ‘Adverse events of exercise therapy in randomised controlled trials: A systematic review and meta-Analysis’, British Journal of Sports Medicine, 54(18), pp. 1073–1080. doi: 10.1136/bjsports-2018-100461.

Nijsten, T. et al. (2009) ‘The Misperception That Clinical Trial Data Reflect Long-term Drug Safety’, Archives of Dermatology, 145(9). doi: 10.1001/archdermatol.2009.188.

Nipp, R. D., Hong, K. and Paskett, E. D. (2019) ‘Overcoming Barriers to Clinical Trial Enrollment.’, American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting, 39, pp. 105–114. doi: 10.1200/EDBK_243729.

Norton, W. E. et al. (2021) ‘Designing provider-focused implementation trials with purpose and intent: introducing the PRECIS-2-PS tool’, Implementation Science. Implementation Science, 16(1), pp. 1–11. doi: 10.1186/s13012-020-01075-y.

O’Shea, M., Teeling, M. and Bennett, K. (2013) ‘The prevalence and ingredient cost of

chronic comorbidity in the Irish elderly population with medication treated type 2 diabetes: a retrospective cross-sectional study using a national pharmacy claims database.', BMC health services research, 13, p. 23. doi: 10.1186/1472-6963-13-23.

Oude Rengerink, K. et al. (2017) 'Series: Pragmatic trials and real world evidence: Paper 3. Patient selection challenges and consequences.', Journal of clinical epidemiology, 89, pp. 173–180. doi: 10.1016/j.jclinepi.2016.12.021.

Øvretveit, J., Leviton, L. and Parry, G. (2011) 'Increasing the generalisability of improvement research with an improvement replication programme.', BMJ quality & safety, 20 Suppl 1(Suppl_1), pp. i87-91. doi: 10.1136/bmjqs.2010.046342.

Pablos-Méndez, A., Barr, R. G. and Shea, S. (1998) 'Run-in periods in randomized trials: implications for the application of results in clinical practice.', JAMA, 279(3), pp. 222–5. doi: 10.1001/jama.279.3.222.

Palese, A., Bevilacqua, M. G. and Dante, A. (2014) 'Do randomized controlled nursing trials have a pragmatic or explanatory attitude? Findings from the Pragmatic-Explanatory Continuum Indicator Summary (PRECIS) tool exercise', Journal of Nursing Research, 22(3), pp. 216–220. doi: 10.1097/jnr.0000000000000045.

Paludan-Müller, A. S., Créquit, P. and Boutron, I. (2021) 'Reporting of harms in oncological clinical study reports submitted to the European Medicines Agency compared to trial registries and publications-a methodological review.', BMC medicine, 19(1), p. 88. doi: 10.1186/s12916-021-01955-0.

Pearson-Stuttard, J. et al. (2022) 'Variations in comorbidity burden in people with type 2 diabetes over disease duration: A population-based analysis of real world evidence.', EClinicalMedicine, 52, p. 101584. doi: 10.1016/j.eclinm.2022.101584.

Péron, J. et al. (2013) 'Adherence to CONSORT adverse event reporting guidelines in randomized clinical trials evaluating systemic cancer therapy: A systematic review', Journal of Clinical Oncology, 31(31), pp. 3957–3963. doi: 10.1200/JCO.2013.49.3981.

Phillips, R. et al. (2019) 'Analysis and reporting of adverse events in randomised controlled trials: a review', BMJ open, 9(2), p. e024537. doi: 10.1136/bmjopen-2018-024537.

Pitkala, K. H. and Strandberg, T. E. (2022) 'Clinical trials in older people.', Age and ageing, 51(5). doi: 10.1093/ageing/afab282.

Pitrou, I. et al. (2009) ‘Reporting of safety results in published reports of randomized controlled trials.’, Archives of internal medicine, 169(19), pp. 1756–61. doi: 10.1001/archinternmed.2009.306.

Qi, M. et al. (2021) ‘Quantifying representativeness in randomized clinical trials using machine learning fairness metrics.’, JAMIA open, 4(3), p. ooab077. doi: 10.1093/jamiaopen/ooab077.

Ramirez, A. G. et al. (2012) ‘Early Phase Clinical Trials: Referral Barriers and Promoters among Physicians.’, Journal of community medicine & health education, 2(8). Available at: http://www.ncbi.nlm.nih.gov/pubmed/24073358.

Ray, P. et al. (2010) ‘Statistical evaluation of a biomarker.’, Anesthesiology, 112(4), pp. 1023–40. doi: 10.1097/ALN.0b013e3181d47604.

Riveros, C. et al. (2013) ‘Timing and Completeness of Trial Results Posted at ClinicalTrials.gov and Published in Journals’, PLoS Medicine, 10(12), pp. 1–12. doi: 10.1371/journal.pmed.1001566.

Romera, I. et al. (2019) ‘A Review of Practical Issues on the Use of Glucagon-Like Peptide-1 Receptor Agonists for the Management of Type 2 Diabetes’, Diabetes Therapy, 10(1), pp. 5–19. doi: 10.1007/s13300-018-0535-9.

Rosas, L. G. et al. (2015) ‘Applying the Pragmatic-Explanatory Continuum Indicator Summary Model in a Primary Care-Based Lifestyle Intervention Trial’, American Journal of Preventive Medicine. Elsevier, 49(3), pp. S208–S214. doi: 10.1016/j.amepre.2015.05.011.

Rosati, P. et al. (2016) ‘Major discrepancies between what clinical trial registries record and paediatric randomised controlled trials publish.’, Trials, 17(1), p. 430. doi: 10.1186/s13063-016-1551-6.

Ross, J. et al. (2010) ‘Analysis of eligibility criteria complexity in clinical trials.’, Summit on translational bioinformatics, 2010, pp. 46–50. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21347148.

Rothwell, P. M. (2005) ‘External validity of randomised controlled trials: “to whom do the results of this trial apply?”.’, The Lancet, 365(9453), pp. 82–93. doi: 10.1016/S0140-6736(04)17670-8.

Rothwell, P. M. (2006) ‘Factors That Can Affect the External Validity of Randomised Controlled Trials’, PLoS Clinical Trials, 1(1), p. e9. doi: 10.1371/journal.pctr.0010009.

Saesen, R. et al. (2023) ‘Analysis of the characteristics and the degree of pragmatism exhibited by pragmatic-labelled trials of antineoplastic treatments.’, BMC medical research methodology, 23(1), p. 148. doi: 10.1186/s12874-023-01975-9.

Scharf, O. and Colevas, A. D. (2006) ‘Adverse event reporting in publications compared with sponsor database for cancer clinical trials’, Journal of Clinical Oncology, 24(24), pp. 3933–3938. doi: 10.1200/JCO.2005.05.3959.

Schroll, J. B., Penninga, E. I. and Gøtzsche, P. C. (2016) ‘Assessment of Adverse Events in Protocols, Clinical Study Reports, and Published Papers of Trials of Orlistat: A Document Analysis’, PLoS Medicine, 13(8), pp. 1–14. doi: 10.1371/journal.pmed.1002101.

Sedrak, M. S. et al. (2021) ‘Older adult participation in cancer clinical trials: A systematic review of barriers and interventions.’, CA: a cancer journal for clinicians, 71(1), pp. 78–92. doi: 10.3322/caac.21638.

Sen, A. et al. (2017) ‘Correlating eligibility criteria generalizability and adverse events using Big Data for patients and clinical trials.’, Annals of the New York Academy of Sciences, 1387(1), pp. 34–43. doi: 10.1111/nyas.13195.

Sen, A. et al. (2018) ‘The representativeness of eligible patients in type 2 diabetes trials: A case study using GIST 2.0’, Journal of the American Medical Informatics Association, 25(3), pp. 239–247. doi: 10.1093/jamia/ocx091.

Sepehrvand, N. et al. (2019) ‘Trends in the Explanatory or Pragmatic Nature of Cardiovascular Clinical Trials over 2 Decades’, JAMA Cardiology, 4(11), pp. 1122–1128. doi: 10.1001/jamacardio.2019.3604.

Seruga, B. et al. (2011) ‘Reporting of serious adverse drug reactions of targeted anticancer agents in pivotal phase III clinical trials’, Journal of Clinical Oncology, 29(2), pp. 174–185. doi: 10.1200/JCO.2010.31.9624.

Shields, B. M. et al. (2023) ‘Patient stratification for determining optimal second-line and third-line therapy for type 2 diabetes: the TriMaster study.’, Nature medicine, 29(2), pp. 376–383. doi: 10.1038/s41591-022-02120-7.

Sil, A. et al. (2019) ‘Selection of Control, Randomization, Blinding, and Allocation Concealment.’, Indian dermatology online journal, 10(5), pp. 601–605. doi: 10.4103/idoj.IDOJ_149_19.

Singh, S. and Loke, Y. K. (2012) ‘Drug safety assessment in clinical trials: methodological challenges and opportunities’, Trials. Trials, 13(1), p. 1. doi: 10.1186/1745-6215-13-138.

Sivendran, S. et al. (2014) ‘Adverse event reporting in cancer clinical trial publications’, Journal of Clinical Oncology, 32(2), pp. 83–89. doi: 10.1200/JCO.2013.52.2219.

Smith, S. M. et al. (2012) ‘Adherence to CONSORT harms-reporting recommendations in publications of recent analgesic clinical trials: An ACTTION systematic review’, Pain. International Association for the Study of Pain, 153(12), pp. 2415–2421. doi: 10.1016/j.pain.2012.08.009.

Smith, S. M. et al. (2013) ‘Adverse event assessment, analysis, and reporting in recent published analgesic clinical trials: ACTTION systematic review and recommendations’, Pain. International Association for the Study of Pain, 154(7), pp. 997–1008. doi: 10.1016/j.pain.2013.03.003.

Sørensen, H. T., Lash, T. L. and Rothman, K. J. (2006) ‘Beyond randomized controlled trials: A critical comparison of trials with nonrandomized studies’, Hepatology, 44(5), pp. 1075–1082. doi: 10.1002/hep.21404.

Spieth, P. M. et al. (2016) ‘NDT-101938-randomized-clinical-trials--a-matter-of-design’, 12, pp. 1341–1349. doi: 10.2147/NDT.S101938.

Stewart, S. et al. (2020) ‘PRAGMATISM OF RANDOMIZED CLINICAL TRIALS ON RANIBIZUMAB FOR THE TREATMENT OF DIABETIC MACULAR EDEMA: Impact on Clinical Outcomes’, Retina (Philadelphia, Pa.), 40(5), pp. 919–927. doi: 10.1097/IAE.0000000000002476.

Stuart, E. A., Ackerman, B. and Westreich, D. (2018) ‘Generalizability of Randomized Trial Results to Target Populations’, Research on Social Work Practice, 28(5), pp. 532–537. doi: 10.1177/1049731517720730.

Stuart, E. A., Bradshaw, C. P. and Leaf, P. J. (2015) ‘Assessing the generalizability of randomized trial results to target populations.’, Prevention science : the official journal of the Society for Prevention Research, 16(3), pp. 475–85. doi: 10.1007/s11121-014-0513-z.

Sun, Y. et al. (2021) ‘A Framework for Systematic Assessment of Clinical Trial Population Representativeness Using Electronic Health Records Data.’, Applied clinical informatics, 12(4), pp. 816–825. doi: 10.1055/s-0041-1733846.

Taillefer de Laportalière, T. et al. (2023) ‘Reporting of harms in clinical trials of esketamine in depression: a systematic review.’, Psychological medicine, 53(10), pp. 4305–4315. doi: 10.1017/S0033291723001058.

Talbot, J. C. C. and Nilsson, B. S. (1998) ‘MAT8_RF_001_06.pdf’, pp. 427–431. doi: 10.1046/j.1365-2125.1998.00713.x.

Tan, Y. Y. et al. (2022) ‘Comparing clinical trial population representativeness to real-world populations: an external validity analysis encompassing 43 895 trials and 5 685 738 individuals across 989 unique drugs and 286 conditions in England’, The Lancet Healthy Longevity, 3(10), pp. e674–e689. doi: 10.1016/S2666-7568(22)00186-6.

Tang, E. et al. (2015) ‘Comparison of serious adverse events posted at ClinicalTrials.gov and published in corresponding journal articles’, BMC Medicine. BMC Medicine, 13(1), pp. 1–8. doi: 10.1186/s12916-015-0430-4.

Tasneem, A. et al. (2012) ‘The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty.’, PloS one, 7(3), p. e33677. doi: 10.1371/journal.pone.0033677.

Tfelt-Hansen, P., Lindqvist, J. K. and Do, T. P. (2018) ‘Evaluating the reporting of adverse events in controlled clinical trials conducted in 2010–2015 on migraine drug treatments’, Cephalalgia, 38(12), pp. 1885–1895. doi: 10.1177/0333102418759785.

Thorpe, K. E. et al. (2009) ‘A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers’, Journal of Clinical Epidemiology, 62(5), pp. 464–475. doi: 10.1016/j.jclinepi.2008.12.011.

Tongtong Wang, Jean-Paul Collet, Stan Shapiro, M. A. W. (2008) ‘Adverse effects of medical cannabinoids: a systematic review’, Cmaj, 178(13), pp. 1669–1678. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2413308/.

Tosh, G. (2011) ‘Pragmatic vs explanatory trials: The pragmascope tool to help measure differences in protocols of mental health randomized controlled trials’, Dialogues in Clinical Neuroscience, 13(2), pp. 209–215. doi: 10.31887/dcns.2011.13.2/gtosh.

Treweek, S. and Zwarenstein, M. (2009) ‘Making trials matter: Pragmatic and explanatory trials and the problem of applicability’, Trials, 10, pp. 1–9. doi: 10.1186/1745-6215-10-37.

Tripathi, B. K. and Srivastava, A. K. (2006) ‘Diabetes mellitus: complications and therapeutics.’, Medical science monitor : international medical journal of experimental and clinical research, 12(7), pp. RA130-47. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16810145.

United States Food and Drug Administration Authority. (2023) What is a Serious Adverse Event? Available at: https://www.fda.gov/safety/reporting-serious-problems-fda/what-serious-adverse-event.

Unkel, S. et al. (2019) ‘On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies.’, Pharmaceutical statistics, 18(2), pp. 166–183. doi: 10.1002/pst.1915.

Usman, M. S. et al. (2022) ‘The need for increased pragmatism in cardiovascular clinical trials’, Nature Reviews Cardiology, 19(11), pp. 737–750. doi: 10.1038/s41569-022-00705-w.

Van Lieshout, E. M. M. and Wijffels, M. M. E. (2020) ‘Patient-reported outcomes: Which ones are most relevant?’, Injury, 51 Suppl 2, pp. S37–S42. doi: 10.1016/j.injury.2019.10.073.

Van Spall, H. G. C. et al. (2007) ‘Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review.’, JAMA, 297(11), pp. 1233–40. doi: 10.1001/jama.297.11.1233.

Varga, Z., Sabzwari, S. R. A. and Vargova, V. (2017) ‘Cardiovascular Risk of Nonsteroidal Anti-Inflammatory Drugs: An Under-Recognized Public Health Issue.’, Cureus, 9(4), p. e1144. doi: 10.7759/cureus.1144.

Varma, T. et al. (2021) ‘Reporting of Study Participant Demographic Characteristics and Demographic Representation in Premarketing and Postmarketing Studies of Novel Cancer Therapeutics’, JAMA Network Open, 4(4), pp. 1–15. doi: 10.1001/jamanetworkopen.2021.7063.

Verghis, R. et al. (2023) ‘Heterogeneity of surrogate outcome measures used in critical care studies: A systematic review’, Clinical Trials. doi: 10.1177/17407745231151842.

Vestergaard Kvist, A. et al. (2021) ‘Cardiovascular Safety Profile of Romosozumab: A Pharmacovigilance Analysis of the US Food and Drug Administration Adverse Event Reporting System (FAERS).’, Journal of clinical medicine, 10(8). doi: 10.3390/jcm10081660.

Wang, J. et al. (2022) ‘Biomarker-based risk model to predict cardiovascular events in patients with acute coronary syndromes - Results from BIPass registry.’, The Lancet regional health. Western Pacific, 25, p. 100479. doi: 10.1016/j.lanwpc.2022.100479.

Ware, J. H. and Hamel, M. B. (2011) ‘Pragmatic Trials — Guides to Better Patient Care?’, New England Journal of Medicine, 364(18), pp. 1685–1687. doi: 10.1056/nejmp1103502.

Wertli, M. M. et al. (2013) ‘Incomplete reporting of baseline characteristics in clinical trials: an analysis of randomized controlled trials and systematic reviews involving patients with chronic low back pain.’, PloS one, 8(3), p. e58512. doi: 10.1371/journal.pone.0058512.

Westergren, T., Narum, S. and Klemp, M. (2018) ‘Characterization of gastrointestinal adverse effects reported in clinical studies of corticosteroid therapy’, Journal of Clinical Epidemiology. Elsevier Inc, 94, pp. 19–26. doi: 10.1016/j.jclinepi.2017.10.018.

Whiting, P. F. et al. (2013) ‘A systematic review classifies sources of bias and variation in diagnostic test accuracy studies’, Journal of Clinical Epidemiology, 66(10), pp. 1093–1104. doi: 10.1016/j.jclinepi.2013.05.014.

Williams, M. R. et al. (2016) ‘Adverse Event Reporting in Clinical Trials of Intravenous and Invasive Pain Treatments: An ACTTION Systematic Review’, Journal of Pain. Elsevier Inc, 17(11), pp. 1137–1149. doi: 10.1016/j.jpain.2016.07.006.

Wise, R. A. et al. (2019) ‘Effect of Aclidinium Bromide on Major Cardiovascular Events and Exacerbations in High-Risk Patients With Chronic Obstructive Pulmonary Disease: The ASCENT-COPD Randomized Clinical Trial.’, JAMA, 321(17), pp. 1693–1701. doi: 10.1001/jama.2019.4973.

Wisniewski, S. R. et al. (2009) ‘Can phase III trial results of antidepressant medications be generalized to clinical practice? A STAR*D report’, American Journal of Psychiatry, 166(5), pp. 599–607. doi: 10.1176/appi.ajp.2008.08071027.

Witt, C. M. et al. (2012) ‘How well do randomized trials inform decision making: Systematic review using comparative effectiveness research measures on acupuncture for back pain’,

PLoS ONE, 7(2). doi: 10.1371/journal.pone.0032399.

Wolff, J. E. et al. (2022) ‘Adverse Events in Placebo Arms of Cancer Trials’, Anticancer Research, pp. 2805–2812. doi: 10.21873/anticanres.15761.

Wu, S. et al. (2022) ‘Association of glucagon-like peptide-1 receptor agonists with cardiac arrhythmias in patients with type 2 diabetes or obesity: a systematic review and meta-analysis of randomized controlled trials.’, Diabetology & metabolic syndrome, 14(1), p. 195. doi: 10.1186/s13098-022-00970-2.

Yao, Y. et al. (2021) ‘Serious Adverse Events Reporting in Phase III Randomized Clinical Trials of Colorectal Cancer Treatments: A Systematic Analysis.’, Frontiers in pharmacology, 12, p. 754858. doi: 10.3389/fphar.2021.754858.

Yuniar, C. T. et al. (2022) ‘Adverse Events Reporting Quality of Randomized Controlled Trials of COVID-19 Vaccine Using the CONSORT Criteria for Reporting Harms: A Systematic Review’, Vaccines, 10(2), pp. 1–15. doi: 10.3390/vaccines10020313.

Zarin, D. A. et al. (2011) ‘The ClinicalTrials.gov results database--update and key issues.’, The New England journal of medicine, 364(9), pp. 852–60. doi: 10.1056/NEJMsa1012065.

Zwarenstein, M. et al. (2008) ‘Improving the reporting of pragmatic trials: an extension of the CONSORT statement.’, BMJ (Clinical research ed.), 337, p. a2390. doi: 10.1136/bmj.a2390.

Zwarenstein, M. et al. (2020) ‘PRECIS-2 for retrospective assessment of RCTs in systematic reviews’, Journal of Clinical Epidemiology. Elsevier Inc, 126, pp. 202–206. doi: 10.1016/j.jclinepi.2020.06.023.

Appendices

Appendix A: Details of literature review on AE reporting in RCTs

SAE reporting in RCTs

Scharf and Colevas (2006) compared AE data from a sponsor's database with published trial data to assess their differences. They searched the National Cancer Institute (NCI) Clinical Data Update System (CDUS) for studies that used the Common Toxicity Criteria version 2.0. From this, they identified 22 studies. Low-grade AEs were underreported in publications; only (58%) of grade 1 and 2 AEs were reported in articles. Inconsistency was found between published data and CDUS. Indeed, 28% of reported high-grade events in publications could not be linked to the AEs of the agent of interest in CDUS. A total of grade ≥ 3 AEs cited on CDUS was 611, while it was 413 in publications. The authors also found that the recorded number of high-grade events differed between published data and CDUS by 20% or more. Scharf and Colevas concluded that low-grade events were underreported, frequencies were underreported, and reporting of high-grade events was insufficient and inconsistent.

Seruga et al. (2011) conducted a study to compare serious adverse drug reactions (SADRs) that were reported in post-marketing updated labels and publications of targeted anticancer RCTs. They searched the FDA website for authorised, targeted anticancer drugs with updated safety labels and RCTs of these agents. They identified 12 drugs with 36 relevant RCTs. Of these, 76 SADRs were found in the updated labels, 50% of which were considered potentially fatal. Seruga et al. found that 30 of 76 SADRs and 15 of 38 potentially fatal SADRs were not posted in reports of RCTs. Conversely, 37 of 76 SADRs and 22 of 38 potentially fatal SADRs were not mentioned in early labels. They thus concluded that RCTs may not be able to detect SADRs in some populations because they are rare and the duration of RCTs is too short to capture these reactions.

Smith et al. (2013) developed the ACTTION (Analgesic, Anaesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks) AE reporting checklist to assess reporting of AEs in RCTs with a more comprehensive approach. They conducted a systematic review that compares AE reporting in recent RCTs of healthy participants and volunteers with multiple pain conditions using ACTTION. The authors searched three major pain journals for pain treatment RCTs and identified 77 eligible trial publications. The frequency of SAEs was not reported for only 33 (41.2%) of trials. AE severity categories were only described in 5% of trials, while 17.5% merely described the severity scale. Only 12.5% of the trials did not report information about withdrawals. However, 30% did not or only partially mentioned the reason for withdrawal. Moreover, withdrawals due to AEs were not reported in 17.5% of RCTs. According to the authors, 37.5% of trials reported all AEs or clearly mentioned no occurrence of AEs, while 48.7% did not mention the severity of AEs. Concerning SAEs, 41.2% of RCTs did not report any information. The authors subsequently concluded that industry trials exhibited better reporting practices than non-industry trials (22.7% vs 1.7%). However, the frequency of SAE reporting did not significantly differ between Industry-funded and non-industry-funded trials (6.9% vs 4.5%). Furthermore, Smith et al. conclude that the quality of AE reporting tends to be low to moderate. However, this study was limited, as publications were the only source of AE data, and they selected only three journals. Moreover, they selected non-invasive treatment trials, which may affect the rates of SAEs in selected trials.

Riveros et al. (2013) conducted a study to compare the completeness and timing of RCT results posted on ClinicalTrials.gov and journal publications. They systematically searched ClinicalTrials.gov for trials that posted their results and searched PubMed for relevant publications. They found that AE and SAE reporting was complete for trials with results posted on ClinicalTrials.gov compared to those posted on publications (73% versus 45%) and (99% versus 63%), respectively.

Belknap et al. (2013) conducted a study to evaluate the current reporting practice of SAEs in RCTs of cancer treatments to an Institutional Review Board (IRB). They retrieved the SAEs of six RCTs reported to an IRB between 2001 and 2008. Belknap et al. found that 182 (75%) of 205 AEs were not reported to an IRB, and six out of 30 SAEs were not

reported. They concluded that IRB reports were inadequate due to defective methodologies and insufficient investigator training. Two of the RCTs included in this study were Phase II, which is not sufficiently competent to detect rare SAEs due to its small sample size.

Maund et al. (2014) assessed reports of clinical studies and publications of nine RCTs for duloxetine, looking for discrepancies in reporting major harm. The literature search identified 1,578 articles; of nine RCTs, only one trial was registered on ClinicalTrials.gov. The occurrence or non-occurrence of SAEs was not reported in 3 out of 9 trials in their relevant publications. The authors found that SAEs, withdrawals due to AEs, and suicide attempts were adequately recorded in CSRs; 8 of 9 CSRs reported SAEs. However, they found discrepancies between reports of clinical studies and publications in seven SAEs, and eight AEs led to withdrawals without significant bias. Moreover, treatment-related events were not reported between a median of 406 (range 177–645) in publications and 166 (100–241) in the registry of trials. Maund et al. concluded that publications did not include adequate safety data compared to reports of clinical studies.

Hughes, Cohen, and Jaggi (2014) conducted a cross-sectional study of trial summaries to evaluate the consistency between registries of RCTs for antidepressant and antipsychotic therapy and their corresponding publications. They used a bibliography of 244 RCTs to search for related publications and found that 142 RCTs had correspondent publications. Conversely, 102 RCTs did not have correspondent publications. According to Hughes, Cohen, and Jaggi, the frequencies of SAEs were reported in 125 (88 %) of trial summaries, 85 (59.9%) of journal articles and 95 (93.1%) of unpublished trial summaries. They also found that 694 SAEs, out of 1,608 treatment-related events, were not posted in relevant publications. Moreover, no description of counted SAEs was found in 60% of publications and 41% of RCTs. The authors also found that these publications did not report 62.3% of deaths and 53.3% of suicidal events. Inconsistencies between registries and publications were found in the reported number of SAEs (49.3%) and descriptions of posted SAEs that were not matched (67.6%). Furthermore, 41.8% of RCT registries in this study did not have correspondent publications, which may have impacted the applicability of the outcomes of this study to other drug classes.

Tang et al. (2015) carried out a study to assess the consistency between reported SAEs in trials' publications and ClinicalTrials.gov. From their search on ClinicalTrials.gov, they randomly included 300 RCTs out of 1,580, which included at least one SAE. They also searched for publications relevant to RCTs included on ClinicalTrials.gov and found that 26% of the included RCTs had no publications, while 20% did not match records on ClinicalTrials.gov. Therefore, 202 RCTs and their corresponding publications were examined. The authors found that 13% of RCT publications did not report SAEs, while 2% posted no SAEs. Moreover, the reported SAEs were inconsistent between publications and ClinicalTrials.gov in 32% of trials, while matched reporting was found in only 11% of RCTs. The total number of SAEs was not reported in 33 (16%) publications. However, this study was limited, as it only searched a single database for publications, which affected the study sample and may have impacted its findings.

Hodkinson, Gamble, and Smith (2016) carried out a meta-analysis to compare RCT summaries and RCT publications of Orlistat to assess the consistency and quality of AE reporting. They searched Medline and Cochrane to identify RCT publications and found 31 corresponding publications to 31 RCTs. Of these, only five RCTs were included in this study, as they provided summaries. Hodkinson, Gamble, and Smith found that AEs and SAEs were inadequately reported in publications of RCTs. Indeed, of five RCT publications, only one matched with the RCT summary in SAE reporting, while other publications reported a maximum of one SAE. Moreover, >86% of SAEs and AEs were available in the summaries of RCTs, while only 26% of them were available in publications. In 91% of studies, withdrawals due to the intervention were present in summaries of RCTs and 51% of publications. This study was based on a meta-analysis of five RCTs, which may have provided a biased assessment.

Maillet et al. (2016) carried out a systematic review to assess the reporting of major AEs (such as those Grades 3/4 and 5) in RCTs of anticancer drugs (2007–2011). They searched Medline and identified publications of 325 eligible RCTs. Maillet et al. found that 96% of these publications reported the frequency of Grade 3/4 AEs, while 17% posted their reporting level, which may increase the chance of underreporting less-common AEs. Moreover, the rate and characteristics of Grade 5 AEs were adequately mentioned in 50%

of published RCT reports, while only 19% posted AEs that resulted in the withdrawal of participants. Frequencies of grade 5 AEs were reported in (73%) of reports. However, as this study focused on publications as its source of safety data, this data may inadequately reflect the actual practice of RCTs.

De Vries et al. (2016) conducted a study to evaluate the contemporary bias in RCT reports of second-generation antidepressants and their effect on safety results. They searched for FDA reviews of registered trials of second-generation antidepressants and identified 74 trials. Subsequently, they searched for publications of these registered trials and found 97 publications. For both sources, the odds ratio for withdrawals due to AEs was 2.4. Of these publications, 79% reported inadequate data about SAEs, while 63% reported no information about SAEs. Only 29% of the 21 publications that were comparable to FDA reviews were fully consistent, while 43% reported an inconsistent number of SAEs, and 6% did not mention SAEs. De Vries et al. (2016) concluded that inconsistencies in reported SAEs are linked to the elimination of post-therapy SAEs or to the omission of events that were assumed to be unrelated to treatment. Moreover, since SAEs are usually found to be infrequent and can occur coincidentally, authors tend to eliminate infrequent SAEs. However, this study was limited, as the FDA database may miss important SAE data, which may have affected the outcomes of this comparison. Moreover, this study was performed on a single drug class, so its results may not apply to other pharmacological interventions.

Tfelt-Hansen, Lindqvist, and Do (2018) conducted a review to assess the adherence of migraine RCTs to the guidelines of AE reporting published by the International Headache Society (HIS). They searched Medline and PubMed between 2010 and 2015 for publications on migraine drug RCTs. They ultimately included 73 articles: 51 acute attack management RCTs and 22 prophylactic trials. Tfelt-Hansen, Lindqvist, and Do found that four of the HIS recommendation items were reported in 37% of acute attack management RCTs, whereas all items were reported in 23% of prophylactic RCTs. Moreover, 74% of acute attack management RCTs reported the number of participants that experienced AEs, as did 86% of prophylactic trials. They also found that SAEs were reported in 22% of acute attack therapy RCTs, while they were reported in 45% of prophylactic RCTs.

Adequate reporting of all parameters was in only (33%) of RCTs. However, the number of patients with any SAE was adequately reported in (69%) of studies. The study period may result in missing important RCTs concerning acute attack management.

Phillips et al. (2019) conducted a systematic review to assess the current practice of AEs that are reported in RCTs. Having searched the BMJ, the JAMA, the Lancet, and NEJM for publications of Phase III and IV RCTs of pharmacological interventions between 2015 and 2016, they identified 148 RCT reports eligible for this study. Of these, 13% reported only the number of events without reporting the event, while nine RCTs in this 13% provided the exact events in the appendix. Phillips et al. also found that 89% of RCTs reported the number of collected AEs. Moreover, 80% of RCTs reported withdrawals, and 35% of these withdrawals were due to AEs. Additionally, 41% of RCT reports noted the seriousness of events, while SAEs were adequately reported in 73% of RCTs. The authors thus concluded that the reporting practice is partially inadequate and inconsistent. Moreover, adherence to the CONSORT recommendation is insufficient. This study was limited, as it included high-impact journals, which may have led to better outcomes and increased the likelihood of bias. The authors also limited their search period to one year, which may not reflect contemporary practice.

Hodkinson et al. (2021) conducted a systematic review to assess the efficacy and safety of several treatments for schizophrenia or bipolar disorder. They searched MEDLINE, Central, EMBASE and PsycINFO until 2020 for the RCTs of Risperidone, Paliperidone or Paliperidone palmitate and their relevant IPD and clinical study reports (CSRs). They searched for more studies on trial registers such as ClinicalTrials.gov, the WHO ICTRP portal and OpenTrials.net. They found that CSRs reported approximately two times more AEs and eight times more SAEs than journal publications. They also found that AEs and SAE reporting was complete on CSRs 35 (100%). However, they were underreported on trial registries and journal publications, 10 (29%) and 17 (49%), respectively.

Yao et al. (2021) systematically reviewed SAE reporting in phase III trials of colorectal cancer treatments. They searched the PubMed, Embase, Medline, and New England Journal of Medicine databases from 1993 to 2018 to identify RCT publications. Yao et al.

(2021) mentioned that 25.5% of publications reported SAEs. They found that industry-sponsored trials reported higher SAEs than trials with other funding sources (57.6 vs 20.7%). They also found that trials published in high-impact journals such as NEJM, the Lancet and JAMA reported higher SAEs than those published in other journals (31.9% vs 16.7%).

Paludan-Müller, Créquit and Boutron (2021) carried out a systematic review to examine the completeness and discrepancies of harm reporting in the RCTs of cancer treatments. They searched the European Medicines Agency (EMA) for oncology trials and retrieved their relevant CSRs. They also searched for related data in trial registries such as ClinicalTrials.gov and journal publications. They found that CSRs had more complete harm reporting compared to other sources. Moreover, SAE reporting was complete on CSRs (100%) and trial registries (95%). However, 50% of journal publications did not report SAEs.

Taillefer de Laportalière et al. (2023) conducted a systematic review to examine the quality of harm reporting in the RCTs of esketamine. They searched Medline and ClinicalTrials.gov to identify trials and their relevant journal publications. They found that SAEs were reported in 90% of included trials. Moreover, they found that journal publications reported only 41.5% of SAEs reported on ClinicalTrials.gov.

Madi et al. (2023) carried out a systematic review to assess the quality of AE reporting in the RCTs of Covid-19 treatments. They searched PubMed and ClinicalTrials.gov from 2019 to 2022 for trials and their relevant publications. They found that journal publications reported 51% of SAEs compared to trial summaries posted on ClinicalTrials.gov.

Harm reporting in journal publications

Smith et al. (2012) conducted a systematic review to identify the modifications of AE reporting in RCTs after the publication of the ten items of the 2004 CONSORT harm extension. They searched three pain treatment journals and classified them into two periods

to assess discrepancies in reporting patterns between epochs. Subsequently, they identified 101 trials that were eligible publications. 88% of trials fulfilled items concerned with reporting SAE counts and timeframes. Additionally, the fulfilment of this item improved over time from 74% to 97%. Items 4, 7, 8, and 10 of the CONSORT harm recommendation were fulfilled by more than 75% of RCTs, whereas 50–70% met CONSORT harm recommendation items 1, 3, and 6. Moreover, items 2 and 5 were fulfilled by less than 50% of the included trials. Less than 2% of RCTs met CONSORT recommendation item 9, which was not surprising given that this item deals with subgroups and harm analysis, which are not typical for pain treatments. According to Smith et al., the fulfilment of these CONSORT harm items was significantly higher in the second period than in the first, as it improved from 44.8% in the first epoch to 61.8% in the second epoch. Moreover, Industry-funded trials showed higher CONSORT harms total score than other fund resources, 0.22 CI (0.09-0.36). Smith et al. found that trials of pain volunteers have a significant tendency to report harm compared to trials of healthy participants.

Péron et al. (2013) conducted a systematic review to assess the quality of AE reporting in RCTs. Using Medline via PubMed, they identified all trials' publications about systemic solid tumours therapy between January 2007 and December 2011, which were published at least three years after the release of the CONSORT extension. They identified 739 RCTs, 325 of which were included. The authors developed the AE reporting quality score (AERQS) based on the CONSORT harm recommendation to assess reporting quality. The mean AERQS for all trials was 10.1, and the range was between 9.8 and 10.4. the average AERQS of RCTs that received industrial funding was 1.14 points higher than that of RCTs with other fund resources. SAEs were adequately reported for 296 (91%) trials according to the fulfilment of CONSORT harm extension items. Reporting was mostly limited to SAEs due to space limitations. Withdrawal frequencies were described in 70% of RCTs. However, the AEs responsible for withdrawal were not adequately reported in 15% of trials. Moreover, details about AEs that caused death were not sufficiently reported in 40% of trials.

Hodkinson et al. (2013) carried out a systematic review of studies that examine the adherence of RCTs to CONSORT harm recommendations. Seven studies that assess the

quality of reporting in 800 RCTs were reviewed in this study. The authors found considerable variation in studies' adherence to CONSORT harm recommendations. Indeed, six of these seven studies assessed harm reporting in the title and abstract; three studies found that 70% of RCTs adhered to the recommendations, and less than 30% of RCTs followed these CONSORT harm items in three other studies. Hodkinson et al. subsequently concluded that these studies were inconsistent in their results regarding the quality of reporting. AE reporting was inadequate; (50%) of the CONSORT harm checklist was not adhered to in 6 out of 7 studies. However, this study was limited, as it did not include studies of other AE reporting guidelines. Moreover, the number of included studies was limited to only seven studies.

Sivendran et al. (2014) conducted a systematic review to describe current AE reporting practices and to assess adherence to CONSORT extension recommendations in oncology trials. After searching PubMed, Embase, and Medline, they found 175 eligible RCT publications. Of these, 96% of trials reported AEs that occurred above a certain level, which may result in the underreporting of rare events, while 22% selectively reported the severity of AEs by limiting them to a specific grade. However, 135 (77%) articles reported AE counts. The authors found that 132 (75%) trials reported death due to AEs. Sivendran et al. concluded that AE reporting in oncology trials was selective and heterogeneous; the median score was 8 out of 14, and the range was 3 to 12. This study was limited by their search time interval of three years, and the probability of poor adherence in the past may have limited the analysis of these results. Moreover, this study was limited to RCT publications as the source of AEs data.

Mahinbakht, Lavasani, and Guirguis (2014) conducted a systematic review to assess adherence to CONSORT harm recommendations in studies of early-phase breast cancer using adjuvant trastuzumab. Five RCTs were included in this review, which found adherence to CONSORT harm recommendations among HERA, NSABP-B31, N9831, PACS-04, and FinHer RCTs to be 70%, 31%, 31%, 49%, and 49.3%, respectively. The authors concluded that these trials did not adequately report AEs, except in one RCT. Moreover, the frequency of SAEs was correctly reported in all trials. The poor adherence of the included RCTs may be because harm is not their primary outcome. However, this

study was limited, as the authors only focused on the early phases of breast cancer. Thus, their results may not apply to all RCT reporting practices. Moreover, this study only evaluated five RCTs.

Chen et al. (2015) carried out a systematic review of RCT publications to assess the quality of reporting of AEs in trials of immune checkpoint inhibitors. The authors searched Medline, Embase, and Cochrane for literature between 2003 and 2013 and thus identified 2,628 articles. Of these, 50 RCTs were included for analysis. The authors adapted the CONSORT harm recommendation to create a quality score of 21 points to assess the quality of reporting. The mean quality score was 11.21 out of 21 (95% CI 10.46–11.96). Grade 3/4 AEs were reported for 96% of studies. Chen et al. found a significant association between high-quality scores and trials that were published in the last five years. They concluded that reporting AEs had improved over time; the mean score increased from 9.09 to 11.81 points, but it was still incomplete. This study scoring tool was adopted from CONSORT harm recommendations instead of using the same items, which may have impacted the quality of their scoring system.

Gewandter et al. (2015) conducted a systematic review to assess the quality of AE reporting in RCTs of temporomandibular disorders (TMDs) therapy. They searched PubMed for publications of RCTs between 1969 and 2013 and identified 90 publications. Gewandter et al. found that 10% of articles complied with CONSORT harm items 2, 3, and 5, while items 1 and 4 were reported in 18–24% of publications. AE reporting was inadequate; most items fulfilment range was between (10%) and (23%), and only one item exceeded (36%). The authors concluded that, although the quality of AE reporting in RCTs of TMDs therapy is inadequate, reporting practices have improved over time. The authors used CONSORT harm items to assess the quality of reporting, which was published in 2004, while they included studies between 1969 and 2013.

Williams et al. (2016) examined AE reporting practices and assessed adherence to the CONSORT harm extension and ACTTION manual. They included RCTs published in three major journals of anaesthesiology and three major journals of intravenous and invasive pain management. They examined these trials during two different periods (2000–2003 and

2008–2012) to assess adherence to the CONSORT harm extension. For the assessment of the ACTTION manual, they included trials from the same six journals but only between 2006 and 2012. Of the 196 identified trials, 165 were used to assess adherence to the CONSORT harm extension, and 132 were used for the ACTTION manual. 140 (85%) fulfilled the reporting of the item in the CONSORT harm extension that concerns SAE reporting. The authors found that CONSORT harm recommendation items 4, 7, and 8 were fulfilled by more than 75% of RCTs, whereas 50–70% met CONSORT harm recommendation items 3, 6, and 10. Moreover, recommendation items 1, 2, and 5 were fulfilled by less than 50% of the included trials, and only less than 2% of RCTs met CONSORT harm recommendation item 9. Comparing the two epochs, the authors found no significant difference in the fulfilment of CONSORT harm recommendations except for items 5 and 10. Thus, the authors found that, of the RCTs assessed according to ACTTION, 8% of trials did not report any information about AEs. Moreover, 61% of trials reported no information about SAEs. However, this study was limited, as publications were the only source of AEs data, and they only selected six journals.

Westergren, Narum, and Klemp (2018) conducted a study of 159 published RCTs, which were included in a meta-analysis of corticosteroids and the related risk of gastrointestinal bleeding. They similarly assessed adherence to CONSORT harm recommendations. According to the authors, the mean score of studies was 5.25 out of 10 with SD of ±2.09. The increase in CONSORT scores related to adequate reporting of GI bleeding (odds ratio [OR] 1.17, 95% CI 1.01-.37, P = 0.042). 130 (81.8%) fulfilled the reporting of the item in the CONSORT harm extension that concerns SAE reporting. Withdrawals of participants were adequately reported in 83.6% of the included studies, while definite GI bleeding was reported in 81.8% of RCTs. Westergren, Narum and Klemp concluded that reporting AEs in RCT publications may be affected by space limits, which means most publications focus on efficacy outcomes. However, this study was limited by its use of publications as the AEs data resource, which may miss important safety data.

Methods of SAE reporting in RCTs

(London et al., 2009) developed an electronic system to report SAEs, instead of paper-based systems, called eSAEy. Thomas Jefferson University applied this system to all RCTs conducted in their institute and by affiliation members in 2007. Subsequently, 588 SAEs were reported using eSAEy, and the median time of the reporting process was <2 days (mean 7±0.2 days). The reporting time was significantly lower than the median reporting time in a paper-based system, which was 24 days (mean of 45 ± 5.7 days). London et al. concluded that eSAEy reduced the reporting period and enhanced reporting precision. This system was designed for RCTs based at Thomas Jefferson University and their affiliation members' institutions. As such, the applicability of this system requires more validation to be deployed in other RCTs.

Bolland et al. (2013) conducted a study to assess the consistency between patient-reported and investigator-verified cardiovascular AEs in a five-year RCT of a 1g calcium supplement. This RCT included 1,471 normal postmenopausal women (mean age 74 years). The authors found that 50% of 64 patient-reported myocardial infarctions (MIs) were verified by investigators, and 58% of 86 self-reported strokes were verified by investigators. Thus, 50% of MIs and 42% of strokes were not reported to investigators. Moreover, 25 of the 58 verified MIs and 13 of the 63 verified strokes were underreported. Bolland et al. concluded that there are discrepancies between patient-reported and investigator-verified AEs. The participants in this RCT were older adults (mean age 74 years) and had comorbidities and cognitive impairment, which may have contributed to underreporting and misleading outcomes.

Crépin, Villeneuve, and Merle (2016) conducted a cross-sectional study to assess the quality of reporting SAEs in RCTs to academic sponsors using forms reported by the investigators of RCTs. From ten RCTs, they received 274 reports of SAEs, 64% of which were pharmacological. The authors found that 3.6% of reports contained no mention of the seriousness of events. Moreover, assessments of causality were not reported in 9.3% of reports. Additionally, 5.7% of reports did not report the date of onset of the SAE, while SAEs were reported during the first 24 hours in only 21% of reports. Crépin, Villeneuve,

and Merle concluded that the quality of SAE reports was low, most forms were not filled, and the fulfilment of items was also low. They attributed these results to the poor uptake of recommendations concerning coding and reporting, as well as insufficient training. However, 36% of the RCTs were not pharmacological, which may have affected the outcomes of this study.

Ménard et al. (2019) developed a tool to predict the frequency of AEs per participant in clinical studies using data from previous RCTs. They hypothesised that this tool would help enhance the reporting of harm in clinical studies. They assessed the ability of this tool to forecast underreporting but not the occurrence of AEs in RCTs. Data were obtained from 104 RCTs funded by Roch-Genentech, and this data was used to train a machine-learning model to forecast the underreporting of the frequency of AEs. In total, 54 characteristics were used to build the final model. Ménard et al. examined the ability of this tool to predict the frequency of AEs in simulated cases and using different scenarios. The authors found that this model scored 0.67 in the area under the curve (AUC) of receiver operating characteristic (ROC) for the statistical scenario and 0.97 for the zero scenario. Moreover, for 25%, 50%, 67%, and 75% scenarios, the AUC was 0.62, 0.79, 0.89, and 0.92, respectively. However, as this model did not examine actual underreported data, its results need more validation. According to Ménard et al., this model is currently applied in a limited number of RCTs that Roch-Genentech conducts, and it will be validated and improved to fit all RCTs.

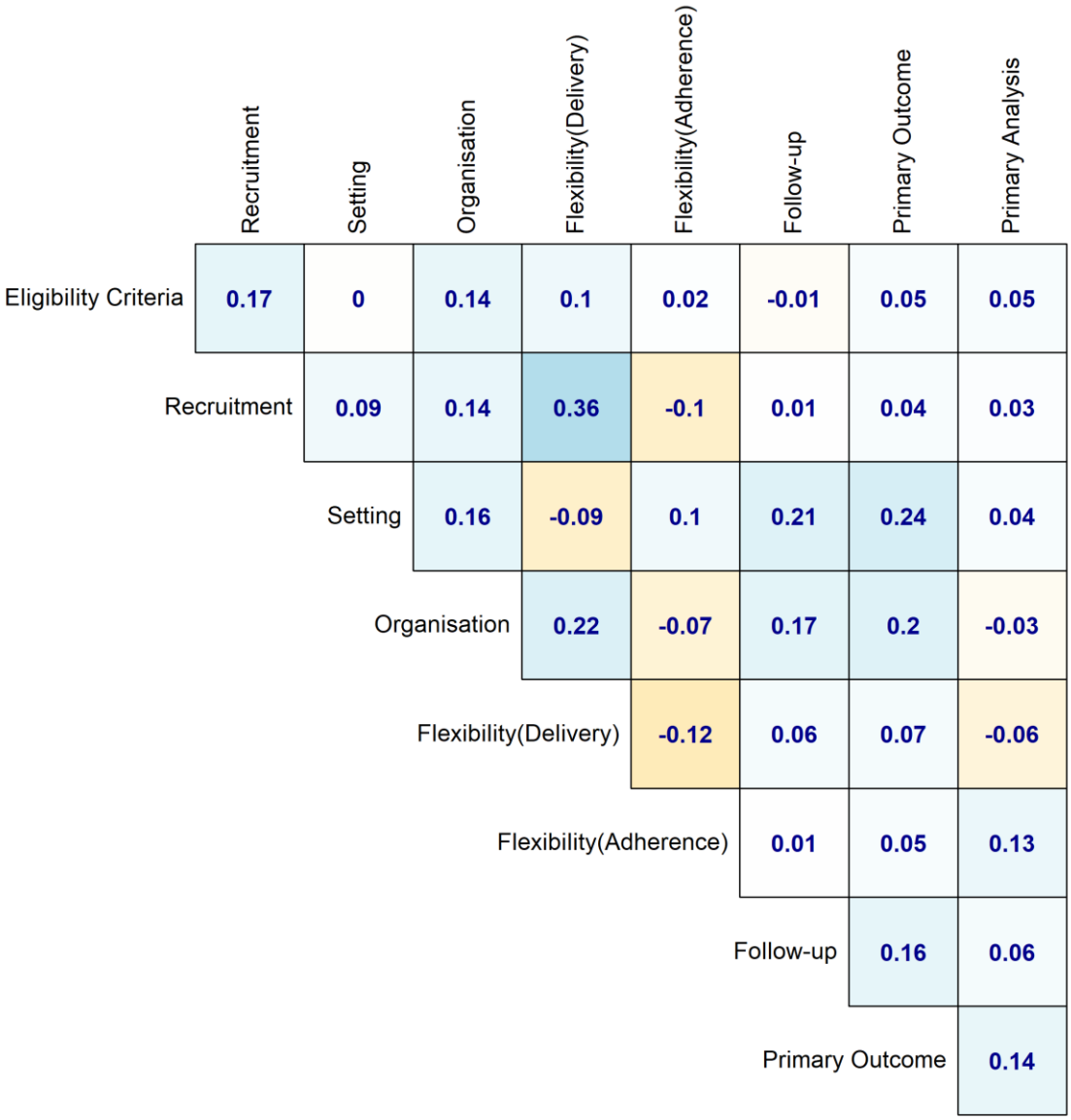
Mayo-Wilson et al. (2019) compared selection criteria for reporting AEs in RCTs and the impact of the differences in criteria of selection on reporting, meta-analysis, and medical practice as part of a methodological study. The authors searched for RCTs of gabapentin for neuropathic pain and quetiapine for bipolar depression and found 21 gabapentin and seven quetiapine RCTs. Besides, they searched for CSRs of identified RCTs and found six gabapentin CSRs and two quetiapine CSRs. Mayo-Wilson et al. compared CSRs with other resources (including CSR synopses and relevant publications) and found that not every CSR applied the selection criteria and reported all AEs, while other resources used the selection criteria but did not report all AEs. Moreover, 22% of gabapentin and 40% of quetiapine RCT publications and CSR synopses reported the application of selection

criteria in their reporting of AEs. Mayo-Wilson et al. concluded that the selection criteria significantly affected the reporting of AEs in RCTs and that most AEs and SAEs were not reported in publications or CSR synopses.

Appendix B: The PRECIS-2 rationale extraction template

Trial ID	**Author**		**Study title**
Domain	**PRECIS-2 score**		**Rationale for assessment**
Eligibility Criteria			
Recruitment			
Setting			
Organisation			
Flexibility (delivery)			
Flexibility (adherence)			
Follow-up			
Primary outcome			
Primary analysis			

Appendix C: Sensitivity analysis for the correlation between the PRECIS-2 domains using MICE imputation



Appendix D: Sensitivity analysis for the association between the normalised SAE rate and the PRECIS-2 domains using MICE imputation

PRECIS-2 domains	β	95% CI	p-value
(Intercept)	0.24	-1.12 – 1.59	0.730
eligibility criteria (unadjusted)	-0.09	-0.24 – 0.05	0.205
eligibility criteria	-0.08	-0.23 – 0.07	0.268
recruitment	-0.01	-0.11 – 0.10	0.893
setting	0.03	-0.11 – 0.18	0.653
organisation	-0.04	-0.15 – 0.07	0.483
flexibility delivery	0.01	-0.24 – 0.27	0.927
flexibility adherence	0.11	-0.22 – 0.43	0.527
follow up	0.04	-0.14 – 0.21	0.680
primary outcome	0.09	-0.07 – 0.26	0.267
primary analysis	-0.11	-0.25 – 0.04	0.141

Appendix E: Complete case analysis of the estimated differences between the SAE rate and the mean PRECIS-2 score

Fair umpires	SAE Rate		Mean PRECIS-2		Difference			SE	2.5%	97.5%	Favouring		
	β	CI (95%)	β	CI (95%)	Null	Unexpected	Expected				Neither	PRECIS	SAE
Intercept	-4.79	-8.96 – -0.64	-3.78	-8.08 – 0.64		-1.04		3.05	-7.12	4.92			
Baseline sample size	0.16	0.03 – 0.29	0.19	0.06 – 0.32		-0.03		0.09	-0.22	0.15	Neither		
Baseline mean age	0.67	0.12 – 1.22	0.08	-0.50 – 0.64		0.60		0.40	-0.17	1.37	Favoured SAE		
Baseline male percentage	-1.06	-2.83 – 0.69	0.03	-1.78 – 1.86		-1.09		1.28	-3.58	1.45	Favoured SAE		
Baseline mean T2DM duration	0.39	-0.24 – 0.99	-0.09	-0.70 – 0.55		0.47		0.45	-0.43	1.32	Favoured SAE		
Baseline mean HbA1C	0.06	-0.25 – 0.38	0.13	-0.20 – 0.45		-0.06		0.23	-0.52	0.39	Neither		

Blinding: Double or more	0.13	-0.18 – 0.44	0.18	-0.14 – 0.48	-0.04	0.22	-0.48	0.40	Neither
Trial Phase: Phase IV	0.27	-0.24 – 0.79	-0.13	-0.66 – 0.37	0.41	0.37	-0.32	1.15	Favoured SAE
Trial duration	-0.28	-0.46 – -0.11	0.11	-0.06 – 0.29	-0.40	0.12	-0.64	-0.16	Favoured PRECIS
Year difference (difference between the year of first register and the year since the intervention was first trialled)	-0.12	-0.60 – 0.34	0.16	-0.35 – 0.67	-0.28	0.35	-0.97	0.38	Favoured PRECIS
Trial Sponsor: Industry	-0.31	-0.98 – 0.34	0.51	-0.17 – 1.16	-0.82	0.48	-1.75	0.11	Favoured PRECIS