Understanding and improving the applicability of randomised controlled trials: subgroup reporting and the statistical calibration of trials to real-world populations.

Lili Wei

MB, MPH

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

August 2023

Thesis Abstract

Context and objective

Randomised controlled trials (hereafter, trials) are widely regarded as the gold standard for evaluating treatment efficacy in medical interventions. They employ strict study designs, rigorous eligibility criteria, standardised protocols, and close participant monitoring under controlled conditions, contributing to high internal validity. However, these stringent criteria and procedures may limit the generalisability of trial findings to real-world situations, which often involve diverse patient populations such as multimorbidity and frailty patients. Consequently, there is growing interest in the applicability of trials to real-world clinical practice. In this thesis I will 1) evaluate how well major trials report on variation in treatment effects and 2) examine the use of trial calibration methods to test trial applicability.

Methods

1) A comprehensive and consistent subgroup reporting description was presented, which contributes to the exploration of subgroup effects and treatment heterogeneity for informed decision-making in tailored subgroup populations within routine practice. The study evaluated 2,235 trials from clinicaltrial.gov that involve multiple chronic medical conditions, assessing the presence of subgroup reporting in corresponding publications and extracting subgroup terms. These terms were then standardised and summarised using Medical Subject Headings and WHO Anatomical Therapeutic Chemical codes. Logistic and Poisson regression models were employed to identify independent predictors of subgroup reporting patterns.

2) Two calibration models, namely the regression-based model and inverse odds of sampling weights (IOSW) were implemented. These models were utilised to apply the findings from two influential heart failure (HF) trials - COMET and DIG - to a real-world HF registry in Scotland consisting of 8,012 HF patients mainly with reduced ejection fraction, using individual participant data (IPD) from both datasets. Additionally, calibration was conducted within the subgroup population

(lowest and highest risk group) of the real-world Scottish HF registry for exploratory analyses. The study provided comparisons of baseline characteristics and calibrated and uncalibrated results between the trial and registry. Furthermore, it assessed the impact of calibration on the results with the focus on overall effects and precision.

Results

The subgroup reporting study showed that among 2,235 eligible trials, 48% (1,082 trials) reported overall results and 23% (524 trials) reported subgroups. Age (51%), gender (45%), racial group (28%) and geographical locations (17%) were the most frequently reported subgroups among 524 trials. Characteristics related to the index condition (severity/duration/types, etc.) were somewhat commonly reported. However, reporting on metrics of comorbidity or frailty and mental health were rare. Follow-up time, enrolment size, trial starting year and specific index conditions (e.g., hypercholesterolemia, hypertension etc.) were significant predictors for any subgroup reporting after adjusting for enrolment size and index conditions while funding source and number of arms were not associated with subgroup reporting.

The trial calibration study showed that registry patients were, on average, older, had poorer renal function and received higher-doses of loop diuretics than trial participants. The key findings from two HF trials remained consistent after calibration in the registry, with a tolerable decrease in precision (larger confidence intervals) for the effect estimates. Treatment-effect estimates were also similar when trials were calibrated to high-risk and low-risk registry patients, albeit with a greater reduction in precision.

Conclusion

Variations in subgroup reporting among different trials limited the feasibility to evaluate subgroup effects and examine heterogeneity of treatment effects. If IPD or IPD alternative summarised data is available from trials and the registry, trial applicability can be assessed by performing calibration.

Table of Contents

List of Tables

List of Figures

Publications, Working papers and Conference Presentations

Published

Wei L, Shah A, Cleland JG, Lewsey J, McAllister D. Transportability of two heart failure trials to a disease registry using individual patient data. Journal of Clinical Epidemiology, 162, pp. 160-168. (doi: 10.1016/j.jclinepi.2023.08.019) (PMID:37659583)

Working papers (being drafted for submission)

Wei L, Butterly E, Rodríguez Pérez J, Moar K, Chowdhury A, Shemilt R, Hanlon P, McAllister D. A description of subgroup reporting in clinical trials of medical conditions.

Conference presentations

1. **[Published abstract]** Wei L, Shah A, Cleland JG, Lewsey J, McAllister D. Assessing the Applicability of Three Heart Failure Randomised Controlled Trials by Calibration to Scottish Registry. Circulation. Nov 16 2021;144

2. Wei L, Shah A, Cleland JG, Lewsey J, McAllister D. Maximise the generalisability of two heart failure randomised controlled trials by calibrating the trials to a HF register in Scotland. Guidelines International Network, 2021

3. Wei L, Butterly E, Rodríguez Pérez J, Moar K, Chowdhury A, Shemilt R, Hanlon P, McAllister D. Subgroup reporting in clinical trials and its implications on subgroup analyses. The 6th International Clinical Trials Methodology Conference, 2022

Abbreviations

AACT: Access to Aggregate Content of ClinicalTrials.gov

AF: Atrial Fibrillation

AFT: Accelerated Failure Time

ARR: Absolute Risk Reduction

BMI: Body Mass Index

CAPRIE: Clopidogrel versus Aspirin in Patients at Risk of Ischaemic Events

CHC: Chronic Hepatitis C

CI: Confidence Interval

COPD: Chronic Obstructive Pulmonary Disease

COMET: Carvedilol or Metoprolol European Trial

CONSORT: The Consolidated Standards of Reporting Trials

CVD: Cardiovascular Disease

DES: Discrete Event Simulation

DIG: The Digitalis Investigation Group Trial

DAPT: Dual Antiplatelet Therapy

eGFR: Estimated Glomerular Filtration Rate

GPP3: Good Publication Practice

HF: Heart Failure

HFrEF: Heart Failure with Reduced Ejection Fraction

HFpEF: Heart Failure with Preserved Ejection Fraction

HTE: Heterogeneity in Treatment Effects

IQR: Interquartile Range

IPD: Individual Participant Data

IOSW: Inverse Odds of Sampling Weights

IV: Intravenous Injection

JUPITER: Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin

LDL: Low-Density Lipoprotein

LVEF: Left Ventricular Ejection Fraction

MACEs: Major Adverse Cardiovascular Events

MeSH: Medical Subject Heading

MI: Myocardial Infarction

ML-NMR: Multilevel Network Meta-Regression

NCDs: Noncommunicable diseases

NHSGGC: National Health Service Greater Glasgow & Clyde

NMA: Network Meta-analysis

NYHA: New York Heart Association Classification

OR: Odds Ratio

PMID: PubMed ID

PO: Administer Drug Orally

RA: Rheumatoid Arthritis

RR: Rate Ratio

SBP: Systolic Blood Pressure

SE: Standard Error

Acknowledgements

I would like to hugely thank my supervisors - Prof David McAllister and Prof Jim Lewsey - for their continued support and encouragement, immerse knowledge and skills, patient guidance throughout the whole process of my PhD. David is my master supervisor, PhD supervisor and line manager who is a super excellent clinical epidemiologist that gives me so much guidance in all the obstacles I faced and naturally "dragged" my epidemiology and R skills into a higher level. He cares and provides a lot listening and advice to my needs and future career plans. He can always remember and offer opportunities for me to improve my skills such as collaborating with collaborators in Johns Hopkins University and Peking University, recommending and guiding me to review paper for the "European Heart Journal" etc. which are all invaluable experience for me. He and his lovely family are also very supportive to cure my homesickness. I am hugely grateful for his time, patience and supervision. And I am hugely grateful for studying and working with such an all-round great supervisor for more than 5 years.

I am also extremely grateful for Jim's support and advice during my study. Still clearly remember in my master statistics course in 2017, I was so impressed with such a clear, understandable session with many difficult and tough materials he presented. Then it was within his introduction, I met David. Jim is a super expertise in statistics, he can always point out some little problems in my models in his first sight and give the right suggestions. And he is always happy to help with my needs. He is so wise, patient and erudite, it has always been a pleasure to talk with him. Thanks so much for his time, guidance and help.

I would like to extend my heartfelt gratitude to Mr. William Spence and his wonderful family for their ongoing kindness. It is truly remarkable to consider that the small orchid I brought to their Christmas lunch not only survived but flourished in full bloom the following Christmas. Life can hold such magical moments, much like enduring friendships. And glad now we are neighbours!

I want to express my gratitude to my former colleague, my dear friend, and my emergency contact in the UK - Dr Jesús Alberto Rodríguez Pérez - for his unwavering support that feels like that of a brother. Every time I met him and Natalia, it was as if I

was returning home, especially during times when I felt down and incredibly homesick. It is quite surprising that we are also neighbours!

I am grateful for Dr John McClure from School of Cardiovascular & Medical Sciences for offering me my first part-time job as a teaching assistant in statistics after interviews. I gained invaluable experience with great passion in teaching and communicating with students.

I would also like to acknowledge my Canadian team – Prof Eric Benchimol, Dr Ellen Kuenzig, and James Im – who are very supportive and have broadened my career horizons. It is always fantastic to have different working experience, especially in different countries.

Thank Prof Sarah Wild in the University of Edinburgh for introducing an excellent diabetes collaboration team and giving many wonderful suggestions.

Thank my colleagues Elaine, Peter, Ryan, Jamie, Salah, Khalid within the team, it has always been a pleasure to work with them.

I am very grateful to Ms. Katriona Lloyd and Mr. Paul Mallon, who used to live just 3 minutes away from me. I will always remember the great times I spent at their place.

Thank my very helpful collaborators in heart failure area - Prof John Cleland for giving constructive suggestions, Dr Anoop Shah for the encouragement email that “Don’t lose hope” and Dr David Phillippo for excellent statistical skills.

Thank Prof Ewan Macdonald for my BBQ training and lots of supportive conversations, it was great experience to be his cooking assistant.

Thank you to our HEHTA team for being incredibly supportive! A special appreciation goes out to our fantastic director, Prof Olivia Wu, who is the best leader in the world, and to our deputy director, Prof Emma McIntosh, for always being helpful. Their approachability and great personalities made the team spirit invaluable.

My special thanks must go to my supportive friends: Dr Y. Zou for her all-round “sister-like” love; Dr L. Huang for profound spiritual and philosophical exchanges we have had; C. Sun, ZD Zhang, Dr YS. Liu for our more than 10 years “steel” friendship; Dr YT Huang for constructive career advice; Dr S. Zheng, Dr J. Feng for accompanying some tough times. Thanks all for appearing in my life!

I also want to emphasise my profound gratitude to my uncles - Prof M. Wei and Prof Y. Wei - for being exceptional "beacons in the ocean" within our family, guiding me through life's journey. I will always be proud of them. My special thanks must go to my lovely aunts and cousins for often visiting my parents when I was not by their side. Thank my grands for being concerned about me in their 80s and 90s. I also want to take this chance to show my condolences to my grandfather Dr W. Wei: you are always the pride of the family, my cousins and I have been missing you so much! We are seeing the world that you may wish to see. Just want to remember you in this way - in a peaceful and solemn corner of the world together with my thesis - that can last forever.

I am immensely grateful to my parents - Mr. H. Wei and Ms. M. Li - for their boundless love, unwavering support, and everything. Without their steadfast encouragement, I would not have had the chance to do what I want. Their love and kindness have been instrumental in shaping me into the person I am today. I will always love them!

I consider myself very lucky for having all these dream people in my life.

I must also express my love and gratitude to the nature - the loch, the ocean, the country park, the hills, the sky, the sun, the moon, the stars, the breeze, the drizzle, the clouds… - for taking away lots of my tiredness, stress, loneliness, and homesickness, refreshing me and offering me inner peace for thousands of times.

Finally, I also want to thank myself for the hard-working, positivity and resilience throughout the hundreds of days and nights during my PhD.

It is my best time in my 20s.

Author's Declaration

I declare that the contents of this thesis are my own work and has not been submitted for any other degree at the University of Glasgow or any other institution. Where the work of others has been used it has been indicated and appropriately referenced.

Lili Wei

Chapter 1: Context and Introduction

1.1 Introduction

1.1.1 Randomised controlled trials (hereafter, trials)

Trials are prospective studies aimed at examining the impact on outcomes (including adverse effects) caused by an intervention when it is implemented instead of another intervention or lack of intervention(1). Trials employ randomisation to minimise bias and offer a robust approach for exploring cause-effect relationships(2). Through randomisation, trials ensure that participant characteristics, both observed and unobserved, are evenly distributed between the treatment groups, randomising confounding factors that may bias results, enabling one to attribute any disparities in outcomes to the specific intervention; this is not possible with any other study design(2, 3). Trials are therefore regarded as the highest level of evidence in evidence-based medicine for evaluating efficacy in clinical research(3, 4).

1.1.2 Trial applicability

1.1.2.1 Definition

The Consolidated Standards of Reporting Trials (CONSORT) statement contains guidelines for reporting parallel group trials, and use the terms generalisability, external validity or applicability to describe the aspect of the generalisability of the trial findings(5). Some researchers prefer to use applicability(1, 6). According to David et al, applicability is defined as “the extent to which the effects observed in published studies are likely to reflect the expected results when a specific intervention is applied to a broader population of interest under ‘real-world’ settings”, which they believe this perspective aligns more closely with the reviews conducted by the Agency for Healthcare Research and Quality Effective Health Care Program and many other groups such as guideline developers(6). A conceptual review gives the definition of applicability as “the extent to which the magnitude of effectiveness of an intervention for a specific

patient or specific group of patients in clinical practice is similar to the magnitude of effectiveness in the results of a trial or a systematic review of trials"(1).

1.1.2.2 Propositions (principles) for trial applicability

Based on a conceptual review, some important propositions for trial applicability were extracted with explanations listed below(1):

1. **High internal validity:** For the findings of a trial or a systematic review (with or without meta-analysis) to be applicable to clinical practice, it is essential to have high internal validity. This means the risk of biased findings is low, indicating that results likely reflect the true situation within the specific context of the study(5). Otherwise, the findings may be false. Therefore, it is rational to utilise the findings of a study only when the risk of bias is low(5).

2. **Rationale:** Clinicians or decision-makers require knowledge derived from the trial to obtain answers for specific patients of interest(6).

3. **Documentation of trials:** To obtain a reliable estimate of the intervention effect in the real-world clinical practice, it is crucial to have comprehensive documentation of the characteristics of a trial at two levels: the intended study design and what the trial turned out to be with the latter considering the aspects of participants selection, healthcare settings, the baseline characteristics of participants, interventions, outcomes and follow-up. The documentation is a precondition for accurate estimation. Also, apart from variations in outcomes, trials should also provide information on the probabilities of favourable and unfavourable (adverse) outcomes between the different treatment arms.

4. **Documentation of registries:** The representative clinical registries in the real-world with uniform documentation with the trial enables the systematic comparison between trial data and registry data(7).

This conceptual review also displayed other propositions such as applicability needs to base on trials with a plausible mechanism intervention, single intervention trials can provide more applicable evidence et al. It also highlights that applicability is reduced due to heterogeneity in the study population and the multidimensionality of the intervention. Additionally, factors such as human perception, behaviour, environmental considerations, and health economic issues further lesson the applicability(8). In summary, ensuring high internal validity is crucial to enhance the applicability of a trial which indicates the findings are likely to reflect the truth. And based on the nature of the trial this feature is always fulfilled. Additionally, the trial should be designed with rational interventions that address the specific needs of the target population. Moreover, accessibility to both the trial information and the registry data is important to enable meaningful comparisons between them.

1.1.2.3 Trial applicability

By design, trials ensure the high internal validity via randomisation and stringent eligibility criteria, participants selection and allocation methods under ideal conditions with minimised bias (2, 9-11). This can meet the pre-condition for trial applicability as it is likely to reflect the "truth" between the outcomes and interventions(5). However, the rigorous implementing requirements of trials also means that it may not fully capture the complexities and challenges experienced in routine clinical practice. There are multiple factors including trial settings, patients selection and characteristics, differences between trial protocol and routine practice, outcome measures and follow-up, treatment adverse events et al that can affect the applicability of trials(12).

Pragmatic trials offer greater applicability to real-world patient populations by enrolling a broader range of patients, relaxing the inclusion criteria for a better reflection of real-world populations(10, 11, 13). Additional details can be found in Chapter 2. Even for few highly pragmatic trials, it remains uncertain if they have well represented the target population. Therefore, this thesis will not focus on pragmatic trials but the vast majority of trials.

1.1.2.3.1 Trial settings and resources

A comprehensive understanding of the trial context is crucial when assessing its applicability, including any unique characteristics of the healthcare system in different countries. For example, different healthcare systems can influence the speed that patients got diagnosed and investigated, further having impact on time from last symptoms to randomisation and even treatment effects(14). It also remains questionable how trials conducted in developed countries apply to the developing countries. Moreover, differences in disease diagnosis and management methods between countries, which can be substantial, as well as important racial variations in disease pathology and natural history, can also impact the applicability. An illustrative example is the heterogeneity of results in trials of bacilli calmette guerin vaccination for tuberculosis prevention, where efficacy progressively declines with decreasing latitude(15).

The selection process of centres and clinicians participating in trials, although often underreported, can also impact the applicability(12). An example is the Asymptomatic Carotid Artery Study trial, which focused on endarterectomy for asymptomatic carotid stenosis. This trial exclusively admitted surgeons with exceptional safety records, rejecting 40% of initial applicants and subsequently excluding those who experienced adverse surgical outcomes during the trial(16). This is not feasible in real-world clinics. Also, while trials should include centres capable of safely treating patients, the selection process should not be overly exclusive to the extent that the results cannot be generalised to routine clinical practice.

1.1.2.3.2 Patients selection and characteristics

There are also multiple factors regarding patients selection and characteristics such as pathways to recruitment, heterogeneity of patients, volunteer bias, going through pre-randomisation run-in that will influence trial applicability.

1.1.2.3.2.1 Pathways to recruitment

In trials, there are often earlier stages of selection that are commonly overlooked, but present challenges(12). For example, when recruiting participants for a trial of a new blood pressure-lowering drug in a hospital clinic, maybe less than 10% of patients with hypertension are managed in hospital clinics, which may differ from those managed in primary care settings. Additionally, if only one out of the ten physicians who treat hypertensive patients in the hospital is involved in the trial, and this physician primarily sees young patients with resistant hypertension, it creates a situation where the potential recruits are already highly unrepresentative of the local community. Therefore, it is crucial for trials to document and report the recruitment pathways whenever possible(12).

1.1.2.3.2.2 Heterogeneity of patients

Patients in real-world settings can be more heterogeneous, which unavoidably questioning the applicability(8). The heterogeneity and variability of real-world patients comparing with trial participants can reflect on multiple aspects such as demographics (age, gender, race et al), comorbidities and disease severities, treatment adherence, concomitant medication use and polypharmacy et al(17-19). Some studies have consistently demonstrated differences between patients in real-world settings and those enrolled in clinical trials. For instance, research by Tan et al showed that among 43,895 eligible trials they have examined, adolescents experience the highest proportion of exclusions, reaching a peak of 90.3% in cardiovascular trials, and the lowest exclusion rate at 70.7% in ear, nose, and oropharynx trials. However, adolescents represent a median of 28.0% (interquartile range (IQR): 19.3%-30.8%) of the real-world population across different cohorts(19). Additionally, the exclusion proportion increases as patients get older after their 60s. It also showed that trials are very likely to exclude patients with comorbidities which dramatically reduce the eligible participants. However, in the real-world practice, multimorbidity (referred to two or more clinical specialties) is common with a median prevalence at 41·0% (IQR 34·9% – 46·0%)(19). Another example compared 226 hypertension trials and 21 corresponding observational studies and it found that the mean age of participants in trials was 54.46 years which was significantly younger compared to the observational studies with the mean age at 66.35 years ($P<0.05$)(20). It also showed that duration of hypertension and severity in trial participants was significantly lower than those in the real-world (3.89 years vs 12.96

years for duration, 17% grade III hypertensive patients vs 34%). Trials also tended to enrol hypertensive participants with significantly fewer comorbidities such as heart failure, stroke, diabetes, or coronary heart disease compared to patients in the real-world practice(20).

1.1.2.3.2.3 Volunteer bias

Volunteer bias is a potential source of selection bias which is also a systematic error that occurs when there are differences between individuals who choose to participate in studies and those who do not(21, 22). In trials, this bias can arise from the fact that the participants only include individuals who are willing to participate, leading to systematic differences between volunteers and those who decline or do not respond to invitations, which may not reflect the real-world situations(22). For example, certain trials investigating antipsychotic drugs have specifically recruited patients who have previously shown a positive response to antipsychotic treatment, introducing uncertainties regarding the differences compared to those who have not been actively treated(12, 23).

1.1.2.3.2.4 Pre-randomisation run-in periods

Pre-randomisation run-in periods are commonly used in trials to exclude patients based on factors such as poor adherence, adverse effects, or ineffective treatment. In active treatment run-in periods, all eligible patients receive the active drug, and those who experience serious adverse effects or show signs of treatment ineffectiveness are excluded. Although this approach aims to ensure safety and efficacy, a high rate of exclusion can limit the applicability of the trial findings. Furthermore, some studies indicate that the complication rates observed during the subsequent randomised phase are lower than those observed during the run-in period, which may not accurately reflect real-world conditions.

1.1.2.3.3 Subgroup analyses and reporting

Subgroup patients refer to a specific subset or subgroup of patients within a trial. Subgroup analyses involve dividing participants into specific subgroups based on their

specific characteristics and subsequently comparing the relevant research outcomes among these subgroups(30). They are used to examine consistency/differences in treatment effects between groups to help tailor treatment recommendations and provide reassurance that treatments effects are "portable" to groups with different characteristics(24). For example, treatment effects may be bigger for older patients than younger participants within a trial, then older patients in the real-world may benefit more. This phenomenon also raises a question that the overall trial findings would arguably be less applicable to routine clinical settings, where the patients are generally younger. Hence, along with trial design and baseline characteristics, subgroup reporting is one of the most important factors for considering trial applicability if heterogeneity in treatment effects (HTE) is less likely to be applicable. Therefore, understanding the subgroup distribution and subgroup treatment effects is also important to assess applicability. However, individual trials are rarely sufficiently large to estimate subgroup effects with adequate precision, making subgroup effect estimates difficult to interpret and frequently misleading(25).

To help address this problem, subgroup analyses of similar trials can be combined in meta-analyses(26). This requires that the subgroups of interest are reported consistently across multiple trials(26). However, some studies show that the subgroup reporting is inconsistent overall.

The majority of studies of subgroup reporting have focused on its overall aspects, such as the incidence and factors influencing subgroup reporting, as well as the adherence to reporting guidelines(27-30). Less attention has been given to identifying the specific subgroups that are commonly reported. Additionally, the emphasis has mainly been on individual papers, particularly those published in major general medical journals, rather than considering the comprehensive reporting of subgroups across all papers for a given trial. In short, they have focused on trial reporting from the perspective of single trials, not meta-analysis.

A number of previous studies have examined the reporting of subgroup analyses and the impact of study characteristics on subgroup reporting. For instance, a systematic review investigated 467 trials across 118 core medical journals in 2007, revealing that 44% (n = 207) of them reported subgroups. Higher-impact journals and larger sample sizes were associated with more frequent subgroup reporting, while industry funding sources were more likely to report subgroups in trials without statistically significant primary

outcomes(31). Another study randomly selected 437 trials from five high-impact journals across three time periods, finding that 62% (n = 270) of them reported subgroups. The study also highlighted that disease severity was commonly reported (69%), while age (87%) and sex (73%) were the most frequently reported subgroups (28). Similarly, a systematic review focusing on cardiovascular disease (CVD) trials examined 130 publications from three high-impact journals between 2015 and 2016. The review showed that 68% (n = 89) of these publications presented subgroup analyses, and trials with larger sample sizes were more likely to report subgroups(32). Additionally, another study assessed 97 trials conducted between 2005 and 2006, revealing that 61% (n = 59) of them reported subgroups, and trials with larger sample sizes were more inclined to report subgroups(33). However, it is important to note that these studies often concentrated solely on high-impact journals, limited time periods, specific conditions, and/or specific subgroups, potentially limiting the generalisability of their conclusions (27-30, 32, 34).

Moreover, the denominator in each of these estimates is trials identified from searching the published literature. It is therefore not clear what proportion of trials that are registered go on to publish subgroup effects. Furthermore, it is not clear what subgroups are reported, in which trials, and with what frequency. If existing trial data is to be harnessed to provide more reliable estimation of subgroup effects, it is necessary to understand what proportion of registered trials report subgroups, as well as what subgroups these trials report.

1.1.2.4 The focus of this thesis

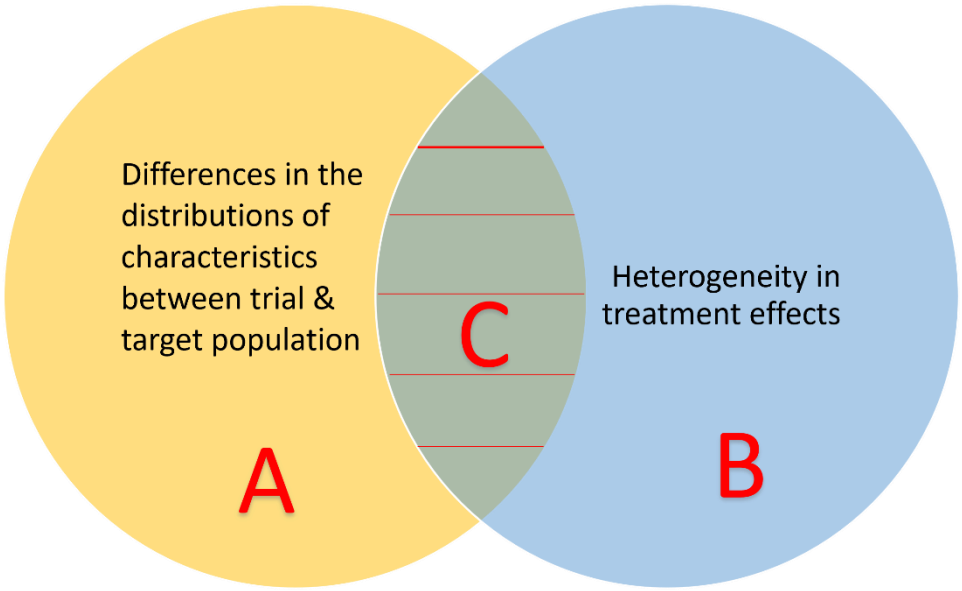
Concerns over trial applicability can arise in various aspects of a trial, including trial settings and resources, patients selection etc.. However, this thesis specifically concentrates on the distribution of patient characteristics in the trial and real-world target population. Real-world data refer to information routinely collected from various sources that pertains to the health status of patients and/or the delivery of health care(35). The real-world target population in this thesis is patients encountered and recorded in clinical practice, which is under the observational and noninterventional setting that is different from the controlled, interventional trial setting. It is assumed that patients from this real-world target population could have some chance of being included in the trial, but it is not necessarily the population from which trial participants are derived. Figure 1 refers to the necessary conditions for applicability.

In scenario A, if differences exist only in the distributions of characteristics between the trial and target population, without heterogeneity in treatment effects (HTE), trial findings can still be applicable.

In scenario B, even with HTE, if there are no differences in patient characteristics between the trial and target population, trial findings can remain applicable.

In scenario C, assuming other factors like trial settings and resources are transferable, trial findings are inapplicable when both differences in patient characteristics and HTE are present. This thesis will centre on this scenario, investigating 1) the assessment of HTE and applicability, and 2) methods to enhance trial applicability in the presence of both HTE and differences in patient characteristics.

Figure 1. Necessary conditions for applicability.



1.1.3 Strengths and weaknesses of trials compared with population-based observational studies.

Trials continue to be the gold standard for evaluating the efficacy of interventions. They are indispensable for testing the effects of new treatments. Meanwhile, observational data derived from real-world clinical practice can offer valuable insights

into treatment characteristics and safety, uncovering aspects that may have been previously overlooked(36). The summarised comparisons between trials and observational studies are displayed below(36, 37).

Strengths of trials:

1) **High internal validity.**
2) **Evaluates efficacy.**
3) **Best for studying an intervention.**
4) **Unbiased distribution of confounders:** randomisation can balance confounding across arms.

Weaknesses of trials:

1) **Resources consuming.**
2) **Relatively short follow-up:** Due to the high costs associated with trials, short trial durations are often implemented, which may not be sufficient to detect rare or delayed side effects and assess long-term efficacy(36).
3) **Volunteer bias:** In trials, it may consist only individuals who are willing to participate due to the strict study design, systematic differences can occur between those who volunteer and those who decline or do not respond to invitations(22).
4) **Limited applicability.**

Strengths of observational studies:

1) **Good external validity.**
2) **Evaluate effectiveness:** The effectiveness of a treatment, which refers to its performance in real-world clinical practice, may not have been adequately studied prior to marketing approval. Observational studies and real-world data are valuable in assessing treatment effectiveness, as they allow for broader inclusion criteria and provide evidence on how the treatment performs in realistic clinical conditions. Observational studies provide comprehensive data on treatment outcomes in the complex environment of routine care, ideally including all cases treated with the intervention.

3) **Good to inform risk factors on an outcome:** it can reflect the natural progression of diseases.
4) **Participants can be matched.**
5) **Can detect rare adverse events, rare diseases or minority patients.**

Weaknesses of observational studies:

1) **Limited internal validity:** It can be challenging to distinguish the effects of a new treatment from other confounding. And it lacks randomisation and control for confounders.
2) **Channelling bias:** Channelling is a selection bias commonly observed in observational studies that compare older and newer drugs within the same therapeutic class. It occurs when drugs with similar indications are prescribed to groups of patients with different baseline prognoses. This bias can be influenced by the timing of drug launches, as medications introduced later to the market may be more likely to be prescribed to patients who have not responded well to existing medications (38).
3) **Difficult to blind.**
4) **Lack details such as disease severity.**

Understanding these complementary approaches, namely trials and observational studies, can help bridge the gap between internal validity and external validity, efficacy and effectiveness and contribute to a comprehensive understanding of healthcare interventions. This thesis aims to examine methods to enhance trial applicability by integrating both trials and a population-based observational study (a disease registry).

1.2 Objectives of the thesis and justification

Continuing from scenario C in Figure 1, my thesis will encompass the following objectives:

1. Explore the conventional availability of trial data for assessing HTE and applicability.

2. Examine methods for enhancing applicability to a wider real-world population encountered in clinical practice by utilising individual participant data from both trials and population-based real-world disease registry.

Initially, for the exploration of assessing HTE and applicability, conducting subgroup analyses with the same index conditions and interventions to assess was considered. However, due to inconsistent reporting of trial subgroups, this approach became unfeasible. Understanding the consistency of subgroup reporting across different index conditions and intervention types is essential to establish a standardised subgroup set for various index conditions and interventions. Additionally, this understanding would be valuable for investigating HTE.

1.3 Structure of the thesis

Chapter 1 defines trial applicability and outlines its prerequisites, including high internal validity, rationality, and clear documentation etc.. It also discusses factors like trial settings and patient selection that influence trial applicability. The chapter then focuses on the specific scenario of differences in patient characteristics distribution between trials and target populations, as well as HTE. By comparing the strengths and limitations of trials and observational studies, it aims to combine both data sources to enhance trial applicability. The chapter finally justifies the thesis objectives: 1) investigating the conventional use of trial data for assessing HTE and applicability, and 2) exploring methods to improve applicability by combining individual participant data from trials and population-based observational studies.

Chapter 2 describes and critiques the literature on existing methods aimed at enhancing the applicability of trials to real-world populations. It presents three types of literature:

1) Descriptive comparisons of characteristics between trial participants and real-world patients.
2) Modification of trial design to include more representative samples.
3) Statistical strategies to apply trial findings into the real-world population.

Chapter 3 aims to address the first objective that is about the subgroup reporting situations in clinical trials with different chronic medical conditions. It assesses over

2,000 trials from clinicaltrial.gov, keeping eligible trials defined by the criteria, screening their related publications, extracting every subgroup reported, harmonising and summarising those subgroups. It provides the descriptive picture of the presence and numbers of subgroups reported in different index conditions. It highlights the commonest subgroups and the uncommonest subgroups across different trials. It also analyses the relationship between trial characteristics and subgroup reporting. It further provides implications for further reporting guidance.

Chapter 4 aims to apply the findings from two historical heart failure (HF) trials to a real-world HF registry in Scotland using two statistical methods. One method involves reweighting by utilising individual participant data (IPD) obtained from the literature, which is considered the gold standard method. The other method employs a parametric survival model. Both methods demonstrate that when applying the results of these two HF trials to the Scottish HF registry, patients in the real-world setting experience similar treatment effects as observed in the trials. Furthermore, this chapter discusses the necessary data format for calibrating trial findings and provides additional recommendations to both trialists and routine data managers.

Finally, chapter 5 summarises the main findings and contributions of the thesis by revisiting two research questions and corresponding case studies trying to answer those questions. The challenges and recommendations arising from the practical case studies and methodologies to enhance the trial applicability are also discussed and summarised. It also discusses the strengths and weaknesses of the research in this thesis. It ultimately presents the overall conclusions derived from the conducted research and outlines the potential areas for future research.

Chapter 2: Literature review of methods to improve the applicability of trials in the real-world population.

2.1 Aim

To stay updated on the latest scope, breadth, and characteristics of research pertaining to the applicability of trials involving pharmaceutical drug therapy in real-world populations, a literature review was conducted. The primary objective was to survey the existing literature and provide an overview of the current methodologies utilised to assess and enhance the applicability of trials.

2.2 Methods

2.2.1 Search strategy

The search strategy was reviewed by the librarian and is shown in Table 1. Searches were run in EMBASE (1947 to present) and MEDLINE (1946 to present) on 18th February 2021 for the first time including published studies from 1946/1947 to 31/12/2020. The key words used are "representativeness", "randomised controlled trial", "pharmaceutical drug therapy", "real-world population" and their synonyms. Searches was run on May 2023 for the second time for updates from 01/01/2021 to 16/05/2023 by the same search strategy.

Table 1. Search strategy in Embase and Medline.

1	external validity.tw.
2	(generalisab* or generalizab*).tw.
3	representat*.tw.
4	applicab*.tw.
5	***or/1-4***
6	Clinical Trial/
7	Randomised Controlled Trial/
8	controlled clinical trial/

9	multicenter study/
10	Phase 3 clinical trial/
11	Phase 4 clinical trial/
12	exp RANDOMISATION/
13	Single Blind Procedure/
14	Double Blind Procedure/
15	Crossover Procedure/
16	PLACEBO/
17	randomi?ed controlled trial$.tw.
18	rct.tw.
19	(random$ adj2 allocat$).tw.
20	single blind$.tw.
21	double blind$.tw.
22	((treble or triple) adj blind$).tw.
23	placebo$.tw.
24	Prospective Study/
25	***or/6-24***
26	(real-world population or real-world).tw.
27	(real world or real life or real patient$ or real practice$ or real clinical$ or realpopulation$).tw.
28	(actual world or actual life or actual patient$ or actual practice$ or actual clinical$ or actual population$).tw.
29	***or/26-28***
30	exp Drug Therapy/
31	exp Pharmaceutical Preparations/
32	exp Drug Interactions/
33	(drug or drugs or pharmaceutical$1 or pharmacotherap$ or pharmaco-therap$ or chemotherap$ or chemo-therap$ or pharmacolog$ or medicin$ or medicat$ or agent$1 or dose$1 or dosage$1 or dosing).tw.
34	***or/30-33***
35	***5 and 25 and 29 and 34***
36	limit 35 to yr="1946 - 2020"
37	limit 36 to (english language and humans)

2.2.2 Review process

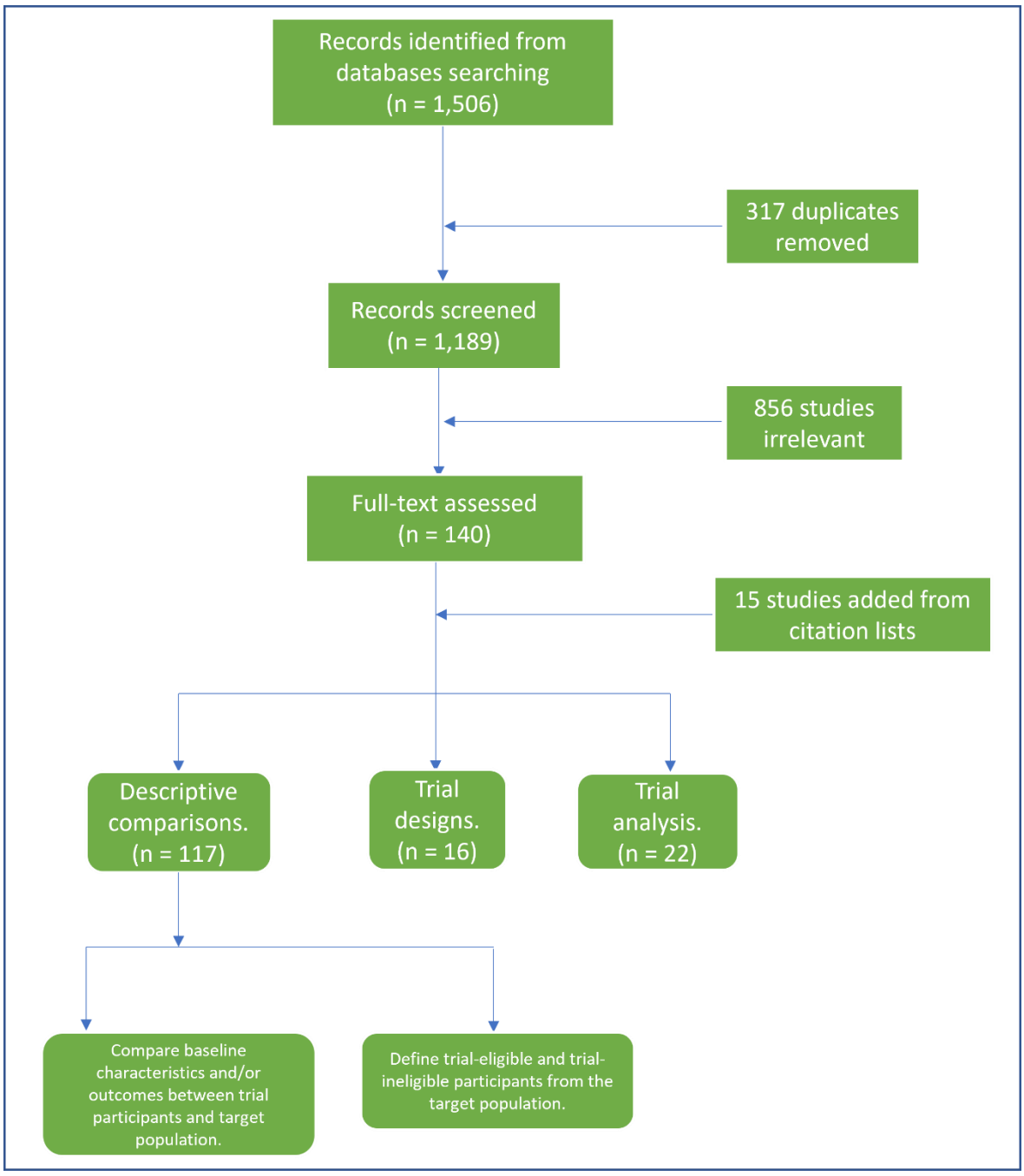
Inclusion was limited to literature published in English related to humans from 1946 to 08/2021. Studies were considered eligible if their methodology involved exploring strategies to enhance the real-world representativeness of trials. The review included both the review articles of different methods and individual research paper. Conferences paper mentioning above were also included. The review specifically focused on trials involving one or more pharmaceutical drugs, excluding surgical trials. Studies were also excluded if they solely analysed trial data or routine data without attempting to apply trial results in real-world contexts.

2.3 Results

The PRISMA diagram is as Figure 2. Title and abstract were screened for studies. 140 studies were retained for full-text screening. After screening the full text, additional 15 references were detected from the citation list and were added.

After screening all of them, they were mainly classified as 3 types of references.

Figure 2. PRISMA diagram of the literature review.



2.3.1 Descriptive comparisons.

This type of study accounted for a major proportion of the literature which includes two sub-types. The first sub-type compared the baseline characteristics and/or outcome between the target population and trial participants to describe the representativeness. For example, researchers compared the baseline characteristics and outcomes between the Acute Study of Nesiritide in Decompensated Heart Failure trial and an eligible complementary registry. Patients in the observational registry were more elderly, more

likely to be female, had more chronic respiratory disease, higher SBP and ejection fraction level, and had less diabetes. They had a similar incidence of ischemic HF and atrial fibrillation. For the endpoint, in-hospital mortality was significantly higher for the registry than the trial patients(39). This study showed that the patients in the trial differed significantly with those in the registry according to the baseline characteristics and outcomes, which reflected the under-representativeness issue and emphasised the need for improving the generalisability of the findings of trials(39).

Palmowski et al conducted a systematic review and meta-analysis to compare the characteristics between participants in rheumatoid arthritis (RA) trials with glucocorticoids and real-world patients(40). They pulled out 56 trials with a total of 7,053 participants and 10 cohorts with a total of 14,688 patients. 12 characteristics were reported with sufficient frequency to enable comparative analysis. Trial participants were found to be younger (-4.7 years [95% CI -7.2 to -2.1]; $p < 0.001$) and had higher erythrocyte sedimentation rates (11.8 mm/h [5.7 to 17.8]; $p < 0.001$) compared to real-world patients. There were no statistically significant differences observed in any of the other analysed characteristics between trials and the routine data, including proportion of females, body mass index (BMI), proportion of current or previous smokers, disease duration, disease activity score, proportion of individuals positive for rheumatoid factor or nticitrullinated peptide antibody, health assessment questionnaire, pain, and patient's global assessment of disease activity. Sensitivity analyses were also conducted and gave similar statements. In this study, comorbidities were unable to be assessed due to insufficient reporting. Researchers found that the study populations in those glucocorticoids trials were generally representative of current real-world patients, with the exception of elderly patients who were underrepresented which was consistent with the trend observed in RA trials in general(41). Also, the representation of patients with comorbidities remains unclear as they were unable to assess this characteristic adequately.

Although 12 characteristics have been compared between glucocorticoid trials for RA and real-world patients, with 10 out of 12 showing no significant differences, the descriptive comparison of baseline characteristics and the application of trial findings to real-world RA patients are still subject to questioning. It shows the representativeness issue and justifies the "what" question while is not adequate to answer the "how" question which is how to improve the trial representativeness.

Evans et al conducted a systematic review to identify published studies that examined the treatment of patients with atrial fibrillation (AF) using warfarin in real-world clinical practice(42). The data from these studies were then compared with pooled data from five AF trials using warfarin(42, 43). Three studies were identified from the systematic review that met the predefined criteria which is in line with trials, conducted in different healthcare settings (the US, UK and Canada) and involving a total of 410 patients. Compared to participants in clinical trials, patients in real-world clinical practice were 6 years older and had a higher proportion of women compared to the trial population. Additionally, a significantly higher proportion of patients from routine practice had a history of cerebrovascular disease. However, the rate (expressed as events per 100 patient-years of exposure) of ischemic stroke was similar between clinical practice and randomised studies, with rates of 1.8% (95% [CI], 0.9%-2.7%) and 1.4% (95% CI, 0.9%-2.0%), respectively. Rates of intracranial hemorrhage (0.1% [0%-0.3%] vs 0.3% [0.06%-0.5%]) and major bleeding (1.1% [0.4%-1.8%] vs 1.3% [0.8%-1.8%]) were also similar. However, the rate of minor bleeding was higher in clinical practice compared to trials, with rates of 12.0% (9.7%-14.3%) and 7.9% (6.6%-9.2%), respectively(42).

This study only extracted and analysed published data for the real-world population, it is unavoidable to introduce bias caused by the non-publication of negative results(44).

The second sub-type illustrated the proportion of the target population that would meet the inclusion and exclusion criteria of the trial, that is, defining the target population as trial-eligible and trial-ineligible participants and comparing the baseline characteristics and/or outcome across them. For example, heart protection study trial is a typical statins use trial in diabetes patients(45, 46) as statins were broadly used to reduce cardiovascular risk among patients with diabetes(47). The endpoint was composite CVD events, and the treatment arms were statins verse placebo. When this trial was applied to the real-world patients in Finland diabetes database (N=56,593), only 57% (N=32,582) patients were eligible for the trial and patients who were ineligible had a higher cumulative risk for CVD events (48). This study also indicated that this trial was under-representative for the female patients. This was a simple example to assess the representativeness of a trial and it provided a general picture of representativeness instead of quantifying it.

Another example is in chronic hepatitis C (CHC) patients. Berden et al conducted a retrospective cohort study involving CHC patients treated in real-world settings nation-widely to compare the effectiveness and safety of patients eligible and ineligible for registration trials(49). They identified registration trials through a systematic search for telaprevir and boceprevir in CHC patients. From the published protocols, the eligibility criteria was extracted and a general set of criteria based on the least stringent criteria across all studies was developed. This general set was then applied to the real-world population to determine eligibility. They compared the outcomes between eligible and ineligible patients and performed sensitivity analyses using strict criteria. Among the cohort of 467 patients, 47% would have been ineligible for registration trials. The main exclusion criteria were related to hepatic decompensation and comorbidities such as cardiac disease, anaemia, malignancy, and neutropenia. These criteria were associated with an increased risk of serious adverse events (relative risk 1.45-2.31). Ineligible patients experienced significantly more serious adverse events compared to eligible patients (27% vs. 11%, $p<0.001$). The effectiveness of treatment was decreased when strict criteria were applied for sensitivity analyses.

Approximately half of the patients with CHC undergoing treatment in real-world clinical practice would not meet the eligibility criteria for registration trials. This finding therefore highlights the limitation of generalisability of results obtained from trials is to real-world patients who would not meet the eligibility criteria. The strengths of this study lie in its nationwide and multicenter nature, which is from a large and representative real-world cohort. However, the retrospective design of the study resulted in the presence of missing values for some variables.

A literature review conducted to assess the external validity of trials also supported these two sub-types of comparisons between the participants included in the trials and patients from everyday clinical practice(50). That are analysing the baseline characteristics of trial-enrolled patients, comparing them to a real-world population; and assessing the proportion of real-world patients who would have been eligible for trial inclusion and comparing characteristics between trial-eligible and trial-ineligible patients. The findings of the included studies consistently indicated that trial participants were highly selected and had a lower risk profile than real-world populations. Elderly patients and those with comorbidities were frequently excluded. The ineligibility proportion calculated from individual studies revealed that a significant proportion of the general disease population was often excluded from trials. The majority of studies (37 out of 52 retained studies) explicitly concluded that trial

participants were not broadly representative of real-world patients, raising concerns about the external validity of trials.

2.3.2 Trial designs.

When it came to the second type of references, it was about modifying trial design to enrol more representative patents in the real-world. This thesis focuses on maximising the applicability of explanatory trials which is different with pragmatic trials. The brief comparison between them is showed in Table 2. Explanatory trials aim to test the efficacy of an intervention under ideal or controlled conditions that typically include highly selected participants with fewer comorbidities and adhering more closely to the study protocol, strict inclusion and exclusion criteria, standardised protocols, and close monitoring of participants, which aim to minimise confounding factors and maximise internal validity. For the outcome measures, explanatory trials primarily focus on measuring surrogate endpoints or clinical outcomes that directly assess the efficacy of the intervention. The findings of explanatory trials, therefore, may have limited generalisability to broader populations in the real-world clinical settings. They provide evidence of efficacy under specific strict conditions that may not reflect the complexities of routine care (9-11).

Unlike explanatory trials which enabled homogeneity by controlling known bias and confounders in strict settings to evaluate the causal effects of the intervention, pragmatic trials aim to assess the effectiveness of interventions under real-world conditions, reflecting routine clinical practice(11). They enrol a broad range of patients who are more representative of real-world populations, including those with

Table 2. Key characteristics comparison between explanatory trials and pragmatic trials.

	Explanatory trials	Pragmatic trials
Research objective	Assessing efficacy of interventions under ideal, controlled conditions	Assessing effectiveness of interventions in real-world clinical practice
Eligibility criteria and Participants enrolled	Strict and selective, enrolling a homogeneous population	Inclusive and representative of the target patient population

		encountered in routine care
Treatment protocols	Highly standardised and protocol-driven, limiting variations in care delivery	Flexible and adaptable to reflect diverse real-world clinical practice
Outcome measures	Usually focused on specific clinical endpoints to measure the efficacy of intervention	Broader range of outcomes, including patient-centred outcomes, healthcare resource utilization, cost-effectiveness et al
Internal validity	High internal validity through strict control of measured variables	Compromised internal validity through less control over confounding factors
External validity	Limited	Higher external validity to real-world populations
Implications to clinical decision-making	Provides insights into the efficacy of interventions in controlled settings	Offers evidence that directly informs clinical decision-making in routine care settings
Resource intensity	Resource requirements can vary depending on several factors, including the study design, sample size, data collection methods, and the specific research question being addressed et al.	

comorbidities and varying levels of adherence. Pragmatic trials have more inclusive eligibility criteria, flexible treatment protocols, and capture the diversity of patients and practices encountered in routine care. These trials provide valuable data on clinically relevant considerations such as different treatments, patient-friendly treatment algorithms, cost-effectiveness, and outcomes that are meaningful to patients(51). They also account for real-world treatment adherence and compliance, offering insights into the direct impact of medications or treatment regimens on patients(52). The findings from pragmatic trials, therefore, have higher generalisability to real-world patient populations (10, 11, 13). It tried to increase the heterogeneity in

all aspects such as patient enrolment and clinical settings etc. to maximise the generalisability. However, this heterogeneity can introduce additional sources of variability and limit the translatability to different settings and locations, which could compromise the ability to draw causal inferences so it supposed to be large enough to increase power and simple enough for performing and follow-up purpose(10, 53). Also, pragmatic trials conducted in real-world settings often require larger sample sizes, longer follow-up periods, and greater resource allocation compared to explanatory trials. This increased complexity and resource requirements can sometimes pose challenges in terms of time, cost, and logistical considerations(13).

One example that illustrates the value of pragmatic trial design is the investigation of patient-driven insulin titration protocols(54). A total of 244 insulin-naive subjects with type 2 diabetes and HbA1c levels between 7.0% and 9.0% on oral antidiabetic treatment were enrolled in the Treat to target with once-daily Insulin Therapy: Reduce A1C by Titrating Effectively study. The subjects were randomly assigned in a 1:1 ratio to one of two treatment arms with 3.9–5.0 or 4.4–6.1 mmol/l fasting plasma glucose as titration targets(54). This study provides valuable insights into the effectiveness and feasibility of patient-directed insulin titration by focusing on real-world conditions, aligning with the practical needs of patients in their everyday experiences rather than reflecting the needs of a highly controlled, well-motivated population in an explanatory trial setting. This trial design bridges the gap between clinical research and routine clinical practice, offering evidence that is more applicable and relevant to routine patient care.

Generally, trials require substantial resources with complex design, burdensome administrative procedures, staff training, participants recruitment, data collection, safety reporting and substantial funding etc.(55), which already posed a great challenge for investigators to initiate it. Pragmatic trial could probably improve the external validity by relaxing the inclusion criteria and enrolling a wider range of more representative participants based on the real-world situation, but it initiates a new trial with the expense of costly resources rather than make more use of the existing trials which is not a sustainable way and is not always feasible. Some researchers also believe that it may not be accurate to assume that pragmatic trials inherently have higher applicability than explanatory trials(12). Although pragmatic trials offer several advantages such as broad eligibility criteria, and inclusion of diverse centres with varying expertise and more representative patient populations, these factors can also present challenges in implementation when attempting to generalise the overall average treatment effect to a specific clinical setting(12).

2.3.3 Trial analysis - statistical methods.

It is of increasing scientific interest whether there are methods that can assess the real-world treatment effects based on the existing trials and the baseline characteristics of real-world target population without costing more resources. Researchers did try exploring the “bridge” between this gap. This third type of references described the attempts to solve this issue which were statistical strategies. The comprehensive overview of different methods for applying trial findings to a target population is provided in Table 3.

Each method has its own set of key features, strengths, limitations, assumptions, and data requirements with more details described below.

2.3.3.1 Overview

Table 3. Comparison of Methods for Extrapolating Trial Findings to Target Populations: Key Features, Strengths, Limitations, Assumptions, and Data Requirements.

Method	Key Features	Strengths	Assumptions	Limitations	Relative treatment effects come from trial	Data Needed in Trial	Data Needed in Registry
Re-weighting by using individual data	Reweight trial data to resemble the registry cohort	Regarded as gold-standard approach among re-weighting methods	Models to predict the sampling probability were correctly specified; all individuals in the target population had some chance of being included in the trial	Unmeasured factors may introduce disparities	Yes	IPD	IPD
Re-weighting by using simulated data	Simulate individual-level data based on summary statistics from the registry	More widely applicable	No correlations between each covariate from the registry	Unmeasured factors and assumption may introduce disparities	Yes	IPD	Aggregated
Re-weighting by using the	Adjust patient characteristics in the	More widely applicable	-	Unmeasured factors are not considered;	Yes	IPD	Aggregated

method of moments	trial to match aggregated data in the registry			method is a bit sophisticated			
Post-stratification	Reweight effects based on population distributions	Conceptually straightforward	-	Applicability is limited in terms of the number of variables	Yes	IPD	Aggregated
Expected absolute risk reduction	Combine relative risk in the trial and baseline risk in the target population	The absolute effect and number needed to treat can be obtained	Uniform relative risk across trial and target population	The assumption may not always be valid	Yes	IPD	Aggregated
Multilevel Network Meta-Regression model	Establish an IPD level regression model and combine outcomes from aggregate data study while adjusting for differences in effect modifiers	Important for decision making; can be used in any relative target population	Common heterogeneity variance; shared effect modifier; conditional constancy of relative effects	May suffer from low power; the unmeasured effect modifier may differ between studies int the network and the target population	Yes	Both	Both
Extrapolation by using cross-	Integrate findings from randomised and	Can extrapolate trial findings to	No unmeasured correlates of the effect measure; the observed trends in	Subjective judgment in selecting algorithms and	Yes	IPD	Aggregated

design synthesis	non-randomised studies	excluded populations	relationships between the risk factors and the endpoints remain constant.	statistical models; caution is needed for extrapolation			
Extrapolation by using discrete event simulation	Model disease pathways and outcomes over time	Accounts for dynamic risk factors over time; can extrapolate trial findings to excluded populations	Same as cross-design synthesis	Requires validation and careful interpretation	Not clear	IPD	IPD
Maximum entropy weighting	Match trial strata and reweights based on observed characteristics in the target population	Combines benefits of trials and observational data sources	Consistency under parallel studies, strong ignorability of sample assignment	May be complicated to implement	Yes	IPD	Either
Non-parametric Bayesian approach	Model observational data with a Dirichlet process	Incorporates trial data with prior distribution	-	Limited details available for evaluation	Not clear	IPD	Either
	IPD: individual participant data.						

2.3.3.2 Re-weighting by using individual data.

The overall aim of this method is to resemble the registry cohort in terms of all measured variables by reweighting the trial, and then to re-estimate the treatment effects of trial on outcomes in the real-world population using reweighted sample. It combines the trial cohort and registry cohort together and estimates the probability of trial participation for every individual if they were selected from the target population by multivariate logistic regression model which includes all effect modifiers that might have the potential to influence the treatment effects such as sociodemographic factors, medical history et al (56, 57). To calculate the weights (inverse odds) of each patient from the real-world target population included in the trial, logistic regression was first used to model the probability of being included in the trial sample, with the patient characteristics as predictors. The logistic regression can be written as the formular 1 below. $pi=Pr(Si=1|Xi)$ denotes the probability of subject i, with a p-dimensional patient covariates Xi, had membership in trial sample ($Si=1$)(56). This step examines how well those baseline characteristics or effect modifiers capture the differences between the trial samples and target population(57). After getting the probability, the sampling weights (the inverse of the estimated odds of trial participation conditional on baseline covariates) for the trial participants can be obtained by getting the inverse odds of the sampling probability [(1-p)/p] (58). Then the inverse odds of trial participation weighting method were employed to re-estimate the baseline characteristics and cumulative incidence of trial outcomes in both the treatment and placebo arms of the trial cohort in terms of the distribution of characteristics in the registry.

$$log(p_i/1-p_i) = \beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}$$

p_i: the probability of subject i from the target population, with a p-dimensional predictors/patient characteristics from X_{i1} to X_{ip}, of being included in the trial.

i: each subject in the target population.

X: predictors such as age, gender et al.

p: number of predictors.

(Formula 1)

This approach aims to enhance the representation of individuals in the linked trial cohort who share characteristics that are more prevalent in the registry cohort. Conversely, individuals in the linked trial cohort with characteristics that are less

common in the registry cohort are given less weight. This process creates a pseudo-population that mimics the distribution of observed covariates in the registry cohort(56). Within the weighted trial, multiple models such as logistic regression or cox proportional hazard model can be applied to obtain the risk of getting the endpoint on the treatment and effect modifiers(59).

It's important to highlight that inverse odds of sampling weights (IOSW) are typically utilised in situations where the effects are estimated in a population entirely external to the study sample. This concept is often referred to as "transportability". This differs from another scenario where study results are extended from a sample to the population it was drawn from, and in this case, inverse probability of sampling weights (IPSW) are used, commonly referred to as "generalisability"(58, 60, 61).

This method has been applied to Dual Antiplatelet Therapy (DAPT) trial and a contemporary real-world population of 568,540 patients undergoing percutaneous coronary intervention with drug-eluting stent(56, 62). DAPT was a large pragmatic trial where patients who had undergone a coronary stent procedure with a drug-eluting stent were enrolled. Following 12 months of treatment with a thienopyridine drug (clopidogrel or prasugrel) and aspirin, patients were randomly assigned to either continue thienopyridine treatment or receive a placebo for an additional 18 months. Throughout the study, all patients continued to receive aspirin. The co-primary endpoints were stent thrombosis and major adverse cardiovascular and cerebrovascular events (a composite of death, myocardial infarction (MI), or stroke) occurring between months 12 and 30. The primary safety outcome was moderate or severe bleeding. DAPT was observed to reduce stent thrombosis (hazard ratio [95% confidence interval (CI)] 0.29 [0.17 to 0.48]), major adverse cardiovascular and cerebrovascular events (0.71 [0.59 to 0.85]) and MI (0.47 [0.37 to 0.61]), but at the cost of increased bleeding (1.61 [1.21 to 2.16]). These findings led to current guidelines recommending the continuation of DAPT for patients with an acceptable risk of bleeding beyond the initial one-year period(62). In comparison to the trial population, the registry patients exhibited more comorbidities and were more likely to present with myocardial infarction and receive 2nd-generation drug-eluting stents. After applying reweighting method to represent the registry, there was no longer a statistically significant effect of prolonged DAPT in reducing stent thrombosis (reweighted treatment effect [95% CI] -0.40, [-0.99% to 0.15%]), major adverse cardiac and cerebrovascular events (-0.52 [-2.62% to 1.03%]), or myocardial infarction (-0.97% [-2.75% to 0.18%]). However, the observed increase in bleeding associated with prolonged DAPT remained significant (2.42% [0.79% to 3.91%].

This study assessed the applicability of the DAPT to a more contemporary registry population receiving PCI and indicated the benefit of prolonged DAPT was attenuated in the real-world population. It showed findings from trials can be applied to the real-world population based on the observed effect modifiers and this method could be used in a wider context in other cardiovascular trials. This method based on the assumptions that logistic regression models used to predict the sampling probability were correctly specified and all individuals in the target population had some chance of being included in the trial(59). Unmeasured factors that could not be addressed during the reweighting process might introduce disparities between the extrapolated registry treatment effect derived from the registry data and the actual treatment effect that would be observed if a contemporary trial were conducted.

2.3.3.3 Re-weighting by using simulated individual data in the target population.

In this method, IPD was used from trials and the aggregated data was used from the target population. Based on the summary statistics of each effect modifiers and the total number of the target population, IPD can be simulated under the assumption that there are no correlations between each covariate. All variables were simulated independently. There are two ways of simulation. Continuous variables such as age, BMI can be simulated as continuous variables based on means and standard deviations. They can also be simulated as categorical variables (e.g., categorise age into <65 and >= 65-year-old). Categorical variables were simulated based on the proportions. In order to achieve a more stable sampling distribution, researchers in this study increased the sample size by a factor of 100 compared to the size of the target population. Then this data was combined with the trial data and the probability of inclusion and sampling weights can be obtained and the treatment effects can be estimated as the above method. This study only involves the main effects without treatment interactions as interaction terms might worsen the covariates balance(59).

The availability of IPD in the target population sometimes is limited such as being stored in the restricted platforms and cannot be accessed. This method can be more widely adapted to deal with the real-world questions although IPD from the trial is still needed. Also, unmeasured covariates were not considered and the assumption that no correlations between each covariate may cause some disparities between the treatment effects from the estimated and observed real-world situations.

2.3.3.4 Re-weighting by using the method of moments.

Signorovitch et al. introduced this approach to enable indirect comparisons between trials when IPD are limited(63). In situations where only IPD is accessible for trials of one treatment (such as a novel drug), while only aggregated data is available for trials of the comparator treatment, it is possible to utilise all available data by adjusting the average patient characteristics in IPD trials (hereafter, T1) to match the characteristics reported for trials without IPD (hereafter, T0). After matching, all available baseline characteristics can be well balanced across trials. Treatment outcomes can then be compared across balanced trial populations.

The principle behind this method is still re-weighing in which participants in T1 are re-weighted to match the distribution of participants in T0. To address the under-representation of participants who are more likely to have received treatment in T0 compared to treatment in T1, their weights will be increased (up-weighted) accordingly in the T1 sample. Participants less likely to have received T0 versus T1 will be down-weighted to compensate for their over-representation in the T1 sample(63). The weight assigned to the i-th participant receiving treatment in T1 is the odds that the i-th participant receives treatment in T0 versus T1 (being enrolled in T0 vs T1) based on the i-th participant's baseline characteristics. The weights can be estimated using the logistic regression as the equation below, where x_i is the covariate vector for the i-th participant(61). However, the regression parameters are not estimable using standard methods due to the lack of IPD in T0. Then Signorovitch et al. proposed the use of a method of moments estimate for ß and the details was described in their paper(63). They also showed the weights balance the mean covariate values between the weighted T1 population and T0 population.

Similarly, in the scenario that only aggregated data is available for the target population instead of T0, participants in the trial with IPD can also be re-weighted to have average values of variables that match baseline characteristics in the target population with aggregated data. After getting the weights, the treatment effects can still be estimated as the above methods(59).

This method estimates weights for target population with only aggregated data by estimating the regression parameter through the method of moments. Compared with

using IPD in the target population, it can also add more flexibility for data acquisition. Also, it is said that the variables can be well balanced between trials and target population. However, it may require a sophisticated understanding of the method of moments and programming skills in practice.

2.3.3.5 Post-stratification

Post-stratification, originally used in sample surveys to align survey samples with population distributions(64), has been extended to estimate population-level effects(57). It re-weights the effects according to population distributions, which is the common way to apply trial effects to the target population. For example, if a trial contains a 20% female and 80% male composition, whereas the target population is evenly split between genders. In this scenario, post-stratification would involve taking an equally weighted average of the gender-specific effect estimates from the trial to estimate the effect in the population. The 95% CI can be obtained based on the pooled standard deviation across strata(59, 65). Hong et al suggested that it can only be used on binary or categorical effect modifiers and better for one-at-a-time variable(59). The calculation was performed iteratively for each effect modifier individually. Subsequently, the poststratification estimates of treatment effect, pertaining to all effect modifiers, were aggregated by computing the unweighted mean of the estimated treatment effects(59). This method can work effectively when there are only a small number of variables which is also binary or categorical, it might not be suitable for many or continuous variables(59).

While this method is conceptually straightforward, its applicability is limited in terms of the number of variables that can be adjusted. When attempting to post-stratify on basic demographic factors such as gender, race, ethnicity, and age groups, the resulting post-stratification cells may become very small(57).

Li et al proposed post-stratification can also be used for generalisation with discrete effect modifiers and continuous effect modifiers with detailed steps and equations described(66). In situations involving multiple continuous effect modifiers, discrete pseudo-strata can be constructed for each point by selecting nearest neighbours based on a multivariate distance measure. However, when both continuous and discrete effect modifiers are present, it is advisable to initially select pseudo-strata based on

continuous predictors and subsequently stratify within those strata using discrete stratifiers. This approach helps avoid the issue of empty strata, where for instance, there are no non-smokers aged 34 (66).

Tipton introduces a method that integrates post-stratification and propensity scores to account for a broader range of variables. This approach exhibits close similarities to the re-weighting methods above(67).

2.3.3.6 Expected absolute risk reduction.

The expected absolute risk reduction can be obtained by the difference of the actual observed risk of unexposed patients in the target population and the expected risk of the exposed patients if they were treated as the trial, assuming the relative risk calculated from the trial is also true for the target population. For example, Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER) evaluates the effectiveness of rosuvastatin vs placebo among patients with low levels of low-density lipoprotein (LDL) cholesterol and elevated levels of C-reactive protein. The primary outcome is the major cardiac event(68). JUPITER - eligible patients naive to statins were defined in the target population, the actual cardiovascular risk was treated as the baseline risk. If the 1-year risk ratio is 0.55 for rosuvastatin in JUPITER and the 1-year cardiovascular risk is 1.5% in the target population, the expected risk in the target population if they were treated with rosuvastatin can be calculated as 0.55 * 1.5% = 0.825% assuming the relative risk is also uniform for the target population. Then the expected absolute risk reduction at 1-year is 0.825% - 1.5% = -0.675 percentage points. The 95% confidence intervals can be obtained based on standard deviations of estimates from 200 bootstraps of the JUPITER data(59). This method based on the assumption that the relative risk is uniform for the trial and target population which may not always be valid.

Another example using this method was applied in Clopidogrel versus Aspirin in Patients at Risk of Ischaemic Events (CAPRIE) trial to Saskatchewan routine health population (69). CAPRIE trial evaluated the relative efficacy of clopidogrel compared with aspirin in reducing the risk of a composite ischemic events including ischaemic stroke, myocardial infarction, or vascular death (70). 12,931 patients from Saskatchewan population who fulfilled the CAPRIE eligible criteria were selected and the data was linked with hospital

admission records, physician visits and prescriptions et al. In order to determine the potential absolute risk reduction associated with clopidogrel for the prevention of ischemic events in real-world settings, the relative risk reduction observed in the CAPRIE trial (8.7%) was multiplied by the reference risk (per year) estimated for patients receiving aspirin in the Saskatchewan population. Chi-square tests were then employed to compare the event proportions between those trial-eligible patients from Saskatchewan populations and the participants randomly assigned to aspirin treatment in the CAPRIE trial. For the results, patients in the real-world were slightly older and there were more females than those in the trial. The rates of subsequent outcomes were higher in real-world practice compared to the CAPRIE controls. In Saskatchewan, patients experienced outcomes at a rate of 159 per 1,000 person-years, whereas in CAPRIE, the rate was only 69 per 1,000 person-years. This indicates that patients in Saskatchewan had an event rate slightly more than twice as high (relative risk 2.3, 95% CI: 2.2 to 2.5, $P < 0.0001$) as that of CAPRIE. Based on the data from patients receiving aspirin in the CAPRIE trial, it was estimated that treatment with clopidogrel instead of aspirin would prevent 5 events per 1,000 person-years which corresponds to a number needed to treat of 200 per year. By utilising the event rate in Saskatchewan (159 per 1,000 person-years) as an approximation of the reference risk, and applying the same relative risk reduction of 8.7%, the calculation yields 14 adverse events prevented per 1,000 person-years. Consequently, the number needed to treat is 70 per year.

When assessing the cost-effectiveness and health benefits of a therapy within the context of healthcare assessment, it is important to take into account the absolute effect and the number needed to treat(71). This method has implications on determining the cost effectiveness of new therapies. It is commonly assumed that the relative risk reduction observed in CAPRIE, is applicable to the real-world Saskatchewan population. If this assumption were invalid, the utility of randomised trials would be compromised. Nevertheless, it is important to critically examine this assumption. The key consideration is whether the factors influencing the reference risk also impact the relative risk reduction. For instance, this could happen if the causes of adverse outcomes vary among patients in a population compared to those in a clinical trial. Despite the pragmatic and representative nature of clinical trials, there is a possibility of underestimating the reference risk and, consequently, the absolute effects(69). Also, this method did not take the multiple baseline characteristics into account which can only make inferences in the population level rather than an individual level.

2.3.3.7 Multilevel Network Meta-Regression (ML-NMR) model

ML-NMR model is a powerful method for conducting population-adjusted indirect comparisons and it is an extension of the network meta-analysis (NMA) framework designed to incorporate both individual and aggregate data from a connected network formed by any number of studies and treatments(72).

It follows specific steps:

1. Data Integration: ML-NMR integrates both IPD and aggregate data from a connected network formed by any number of studies and treatments.
2. Individual-Level Regression Model: It establishes an individual-level regression model directly fitted to participants with IPD(73).
3. Incorporation of Aggregate Data: It incorporates summarised outcomes from studies with aggregate data by integrating it across the covariate distribution within each aggregate data study. ML-NMR adeptly combines networks of IPD and aggregate data studies of varying sizes, adjusting for differences in effect modifiers and avoiding aggregation bias and noncollapsibility bias.
4. Population-Adjusted Estimates: ML-NMR can produce population-adjusted estimates of quantities of interest in any target population for which covariate information is available, such as average treatment effects or absolute event probabilities(72).

This approach relies on several assumptions. Firstly, all effect modifiers should be appropriately accounted for to maintain the validity of the conditional constancy of relative effects assumption. Additionally, it assumes the consistency of relative treatment effects, extending this consistency assumption to the interactions involving effect modifiers. The shared effect modifier assumption, where interaction parameters of effect modifiers are assumed to be common for treatments, is also employed. However, this assumption can be challenging to hold when data is insufficient. Furthermore, it may suffer from low power when data is lacking. Unmeasured effect modifiers may also be omitted between studies in the network and the target population(72).

The method can been implemented in the “multinma” R package(74) and Phillippo et al provides more details on fitting ML-NMR models(72).

2.3.3.8 Extrapolation by using cross-design synthesis

A broader category of methodologies, known as research synthesis or cross-design synthesis, incorporates similar principles to meta-analysis but offers greater potential to explicitly address the question of generalisability (57). Research synthesis allows for the integration of findings from both randomised and nonrandomised studies, enabling the combination of information on program effects from diverse sources. For instance, it can merge results from a trial with those from an observational study, which can provide different, complementary strengths and weaknesses (i.e., trials are usually used to explore the causal relationship with restricted subjects and study design and observational studies might contain more representative samples with less restrictive inclusion criteria)(57, 75, 76) to extrapolate findings of trial to the target population (77-79). This approach involves modelling multiple parameters from each study and incorporating study characteristics into the analysis(80). It assumes that there are no unmeasured correlates of the effect measure of interest by inclusion criteria for the trial and the observed trends in relationships between the risk factors and the endpoints remain constant. However, further investigation is required to fully explore the applicability of research synthesis in addressing the specific question of generalisability, as the explicit goal of these methods may not always be to estimate population effects(57).

Wang et al proposed cross-design synthesis could be used to extrapolate trial findings to estimate treatment effects in excluded populations(77). They conducted a fictional case study to extrapolate trial results for fantastistatin compared to normostatin to a target population that includes older patients with a longer lag between MI and treatment initiation than those in the trial. The algorithm they used utilises observed trends in the rate of major adverse cardiovascular events (MACEs) with increasing age among normostatin initiators in the observational data to extrapolate rates to older patients who were excluded from the trial. They then checked the consistency by assessing if the rate of MACEs with increasing age among normostatin initiators under 65 years in the observational data aligns with the trends observed in trial participants under 65 years who initiated normostatin. If there were significant deviations, cross-design synthesis would not be applied in this case. If not, the rates of MACEs in both arms of the trial

would be extrapolated and the expected rates for each group in the excluded target population would be estimated. To estimate the average efficacy in the defined target population, the observed and extrapolated rates of events can be combined by reweighting them according to the distribution of age and/or interval between MI and treatment initiation in the target population.

Cross-design synthesis lacks formal mechanisms to assess the similarity or difference between subjects in trials and individuals in the target population. Subjective judgment in some parts of the process such as the selection of the algorithms and the complex statistical adjusting models with the possibility of pooling inappropriate data still remains questioning(76). It still needs to be cautious when extrapolation being carried out and interpreted.

2.3.3.9 Extrapolation by using discrete event simulation

Discrete event simulation (DES) is a method used to model disease pathways and outcomes over time, taking into account treatment and individual patient-level variables. By tracking patient-level characteristics and incorporating changes in risk factors, DES can estimate event rates, absolute risks, and treatment effects(81-84). For instance, DES can consider the increased risk of MACEs as patients age or develop comorbidities. DES can incorporate these “dynamic” risk factors as parameters when estimating individual risks in the simulation. It can, therefore, extrapolate over time as patients’ risk varying based on different transition pathways and health status according to the characteristics such as more elderly age or more comorbidity burden. Similar to reweighting and cross-design synthesis, DES can generalise results to target populations with different characteristics and extrapolate evidence to populations excluded from trials. Like cross-design synthesis, it also assumes that there are no unmeasured correlates of the effect measure of interest and the observed associations between the risk factors and the endpoints remain constant(77). To implement DES, the following steps can be followed:

1. Develop and validate outcome prediction models: Create models that describe how patient characteristics relate to outcomes for each exposure group using trial data and external information.

2. Design the DES model: Construct a model that represents various health states and pathways, incorporating the prediction model from step 1 to define probabilities of outcomes based on changing patient characteristics.

3. Simulate the trial participants: Generate a cohort of patients with characteristics matching those in the original trial population.

4. Validate the DES model: Compare the simulated event rates and effect measures from the DES model with the observed rates and measures from the trial.

5. Simulate the target population: Use the DES model to simulate a cohort of patients with characteristics reflecting the target population in routine care. Obtain the relevant covariate distributions from literature or healthcare data sources.

6. Run DES: Obtain predicted absolute event rates and effect measures for the target population. Consider uncertainty by incorporating model estimates and standard errors at each transition.

By following these steps, DES can provide estimates of outcomes and treatment effects for a target population by accounting for patient characteristics and their impact on health outcomes over time.

Outcome prediction models can be derived from various sources, such as published literature or regression models fitted with individual-level data from trials or observational studies. It is important to validate newly developed models to assess their predictive performance in out-of-sample data. In a recent study, published outcome models from the Randomised Evaluation of Long-Term Anticoagulation Therapy trial were combined with baseline characteristics from two previously published observational studies comparing dabigatran to warfarin in atrial fibrillation patients to develop a DES model that accurately replicated the rates of ischemic stroke and major bleeding observed in the trial. The well-fitted DES model was then utilised to predict trial outcomes in populations similar to those encountered in routine care settings(81).

By using this method, researchers need to get access to IPD in both trial and routine data to be able to develop and validate outcome prediction models and examine the effects and "dynamic" risk over time. DES model can extrapolate trial findings to excluded populations, but it also needs to be cautious when validating the model and interpreting it.

2.3.3.10 Maximum entropy weighting

The principle of maximum entropy states that, in situations where information is incomplete, the preferred probability distribution is the one that maximises entropy or a form of probabilistic uncertainty while satisfying the given constraints (53). The natural constraint is that the sum of all probabilities must equal one. Maximum entropy weighting does not assume the propensity score is rightly specified and does not make additional assumptions about the distribution of weights.

It uses an automatic matching approach to create matched strata within the trial and then by using maximum entropy weighting to reweight the individual trial strata according to the observed characteristics in the target population(85). More details and equations can be found in the appendix(85). It ensures the weights of the matched pairs sum to 1, but simultaneously satisfy the constraints based on the characteristics. This method can be used when either aggregated or individual data of the target population are available and it has some in common with the method of moments(86, 87). It is suggested to incorporate covariates in the model that are expected to have an impact not only on the outcomes but also on the selection of patients into the trial.

It combines the benefits of trials with those of large observational data sources and retains the advantages of both types of data. However, it requires sufficient assumptions such as the consistency under parallel studies, strong ignorability of sample assignment for treated and controls(85). It may also be complicated to implement.

2.3.3.11 Non-parametric Bayesian approach

According to Yovanna et al in 2017, a non-parametric Bayesian approach was applied that models long-term observational data with a Dirichlet process. The fitted Dirichlet process serves as the prior distribution, while the Kaplan Meier estimate from the trial

data acts as the likelihood function. Trial data was then incorporated with the prior distribution from the observational data, resulting in the non-parametric Bayesian estimator (88). As this approach was published as the meeting abstract, more details are to be explored.

2.3.3.12 Discussion

Re-weighting by using individual data, aiming to adjust the trial cohort to resemble the target population in terms of measured variables, is regarded as the gold-standard approach among re-weighting methods(59). It based on three assumptions that 1) it captured most of the potential factors that may modify the effect, 2) logistic regression models utilised to estimate the sampling probability were accurately specified and 3) every individual in the target population had a non-zero probability of being selected for the trial.

Weighting methods that rely on aggregate data from the target population have inherent limitations when it comes to matching multidimensional distributions of effect modifiers between the weighted trial and target population. This is due to the lack of IPD or joint distribution data. Modelling-based weighting methods require the assumption of correct model specification but often neglect to assess covariate balance in joint distributions for confounding control(89). However, achieving covariate balance in joint distributions is crucial as it involves reweighting trial participants to the target population based on all effect modifiers, including those specific to certain covariate patterns. A study evaluated variable balance in subgroups stratified by sex after reweighting trial participants to match the target population's marginal distributions of variables(59). It showed that while the gold-standard method maintained balanced covariates after stratification by sex, re-weighting by using simulated data and the method of moments showed a deterioration in covariate balance. To address this limitation, IPD or data on joint distributions of relevant effect modifiers are necessary when the access to IPD is not always possible due to data sharing agreement and regulatory approvals. The flexible application of re-weighting methods therefore depends on the availability of the data from the real-world. Re-weighing by using IPD is recommended where possible.

In re-weighting by using simulated data, it is assumed that there is no correlation between effect modifiers. The impact of correlation on the generalisability of trial results is not well understood. While correlation between variables may not affect certain methods of confounding control, it can potentially introduce interactions in logistic regression models used to predict sampling probabilities, thereby influencing the estimates in the generalisability of trial results(90). Therefore, further research is needed to explore the incorporation of correlation between covariates in simulations(59).

Poststratification can be viewed as another weighting approach that assesses the generalisability by reweighting subgroup-specific treatment effects to align with the distribution of those subgroups in the target population. However, this method is limited by the number of variables it can account for. Additionally, it only standardizes for the distribution of one effect modifier at a time. This method can be particularly valuable when there is a strong effect modifier or when identifying effect modifiers that have a substantial impact on the generalisability of trial results(59).

Extrapolation by using cross-design synthesis or DES both need IPD from the target population to model multiple parameters or measure the dynamic risk factors over time. They both based on the assumption that there are no unmeasured correlates of the effect measure and the observed trends in relationships between the risk factors and the endpoints remain constant. And the models need to be validated before it can be applied. The strength is that they can extrapolate trial findings to excluded population that other methods cannot achieve. In theory they are feasible while researchers just used the C cases studies to implement the methods. It remains unsure if they are practical in the real-world situations and it requires a more comprehensive understanding and interpretation of the methods.

While re-weighting by simulated data and method of moments that are based on aggregated data are theoretically straightforward to implement, practical challenges often arise. The ability to align multidimensional distributions of effect modifiers between the weighted trial and target population may be constrained by the absence of IPD or comprehensive joint distributions. Therefore, matching existing aggregated data that precisely with the trial's inclusion and exclusion criteria can be difficult unless common effect modifiers were pre-specified before aggregation. On the other side, comparing with getting access to IPD in the target population through data sharing

agreement and regulatory approvals, weighting methods using simulated individual data and the method of moments are preferable with much less administrative burden but require more advanced programming techniques. However, study highlights the limitation of using continuous variables in weighting methods based on aggregate data due to the challenge of understanding actual distributions without IPD(59). In certain situations where a strong effect modifier is suspected or complex programming is not feasible, alternative approaches such as poststratification and expected absolute risk reduction should be considered.

2.4 Chapter discussion

Austin Hill said in 1984, “At its best a trial shows what can be accomplished with a medicine under careful observation and certain restricted conditions. The same results will not invariably or necessarily be observed when the medicine passes into general use”(23, 91). The disparities between trials and real-world practice reflect a phenomenon known as the "development paradox." In the drug development process, phase II-III trials typically focus on enrolling patients who are relatively easy to treat, whereas in real-world clinical settings, priority is given to treating patients with more challenging conditions(92-94). The sequential approach of initially studying drugs in easy-to-treat patients is generally deemed appropriate. However, the final step of conducting trials specifically targeting difficult-to-treat patients is frequently bypassed or postponed until after market authorisation(49). There are always uncertainties when the findings from trials are applying to patients in the real-world practice especially for those with different characteristics. This literature review focuses on methods to improve the applicability of trials to the real-world practice and there are 3 types of studies.

The descriptive comparisons mainly describe two types of comparisons to assess the representativeness of the trial. Comparing the baseline characteristics and/or outcome between the target population and trial participants is the first type. The second type of assessment is defining the target population as trial-eligible and trial-ineligible participants and comparing the baseline characteristics and/or outcome across them. The issue of representativeness is well described by those comparisons with some examples provided, emphasising the importance of addressing the "what" question. However, it falls short in providing a solution to the "how" question, which pertains to improving the trial representativeness.

The second type of study refers to conducting pragmatic trials rather than explanatory trials to improve the trial representativeness by enrolling more representative samples from the routine practice to better reflect the real-world situation. Pragmatic trials have a higher external validity than explanatory trials from the design, but the cost of substantial resources poses a great challenge due to the complex design, burdensome administrative procedures, staff training, participant recruitment, data collection, safety reporting, and the need for substantial funding, among other factors (55). These resource-intensive requirements can create barriers and limitations in conducting research studies and clinical trials. In the words of Theodore Roosevelt, "Do what you can, with what you have, where you are." Nowadays conducting randomised effectiveness trials that involve significant financial investments of tens or hundreds of millions of dollars may not be feasible(95). Also, even for highly pragmatic trial, it remains uncertain if they have well represented the target population. However, with the emergence of registries and powerful digital platforms, there is an opportunity to leverage the available resources, such as bigger data and smaller budgets, to design and execute megatrials. These megatrials can still provide valuable insights and contribute to advancing scientific knowledge within the constraints of current resources(95).

The third type of literature is about re-analysing the available trials and routine registry with statistical methods to maximising the generalisability of explanatory trials. These methods are all based on the statistical strategy to apply findings from trials to the target population in the real-world in terms of the existing trials without costing substantial resources to modify trial design to involve more representative samples. Each method has unique features, strengths, limitations, assumptions, and data requirements. Re-weighting methods are used to adjust trial cohorts to resemble the target population. Re-weighting using individual data is considered the gold-standard approach, but it relies on the model is correctly specified for the trial inclusion probability and individuals in the target population had some chance of being included in the trial. Although obtaining IPD is preferred, it is not always possible. Re-weighting using aggregated data have limitations in matching effect modifiers and covariate balance due to the lack of IPD or joint distribution data. Poststratification is useful for strong effect modifiers but limited in accounting for multiple variables. Extrapolation methods require IPD and assumptions of no unmeasured factors. They can apply trial findings to excluded population which is a strength over re-weighting methods. However, they normally require comprehensive understanding and programming skills, and the models always need validation for practical application in real-world situations.

Using statistical strategies to maximise the applicability of trials is more sustainable based on existing data sources compared to modifying trial designs to enrol more respective samples, which is also the focus of this thesis. Understanding those aspects of different methods is crucial for choosing the appropriate method. It depends on the data availability and type, conditions to meet the assumptions, the understanding to the statistical methods and programming skills et al. Also cautions are needed when justifying the purpose and interpreting the results.

There is no existing literature comparing whether the importance of trial representativeness varies by the type of condition being investigated (e.g. non-communicable diseases, infectious diseases, cancers, screening interventions etc). The swiftness with which individuals can spread infections around the globe makes infectious diseases a tremendous challenge for governments, the public, and primary healthcare systems(96). The screening interventions require careful consideration of both the prevalence and severity of the disease being screened for(97). Noncommunicable diseases (NCDs, also called chronic diseases), including cardiovascular disease, cancer, chronic respiratory ailments, and diabetes et al represent 74% of global deaths(98). Population growth, rising global average age, and considerable declines in age-, sex-, and cause-specific mortality rates drive a shift from infectious, maternal, neonatal, and nutritional causes to non-communicable diseases(99). Among NCDs, cardiovascular diseases, causing predominant deaths about 17.9 million yearly, followed by cancers, resulting in 9.3 million deaths(98). Therefore, my thesis will firstly focus on NCDs, then specifically using cardiovascular diseases as an example to explore trial applicability.

Chapter 3: Case study 1 - A description of subgroup reporting in clinical trials of chronic medical conditions.

3.1 Chapter summary

This study examined a wide range of clinical trials for subgroup reporting obtained from clinicaltrial.gov and screened each trial-corresponding publication. Every reported subgroup was extracted. Using MeSH terms and WHOATC code, all reported subgroups have been standardised and categorised, and summarised according to trial index condition with their frequencies. It described the association between trial characteristics (such as trial starting year, enrolment size, follow-up time, trial sponsors, number of arms, index conditions) and subgroup reporting. It also provided implications for future trial reporting.

3.2 Abstract

3.2.1 Introduction

In trials, subgroup analyses are used to examine whether treatment effects differ by important patient characteristics. However, which subgroups are most commonly reported have not been comprehensively described. Therefore, using a set of trials identified from the US clinical trials register (ClinicalTrials.gov), every reported subgroup for a range of conditions and drug classes (PROSPERO CRD42018048202) was described.

3.2.2 Methods

Trial characteristics from ClinicalTrials.gov via the Aggregate Analysis of ClinicalTrials.gov database was obtained. Subsequently all corresponding PubMed indexed papers were also obtained and screened for subgroup reporting. Tables and text for reported subgroups were extracted and standardised using Medical Subject

Headings and WHO Anatomical Therapeutic Chemical codes. Via logistic and Poisson regression models, independent predictors of result reporting (any vs none) and subgroup reporting (any vs none and counts) were identified. Next, subgroup reporting by index condition was summarised and all subgroups were presented for all trials via a web-based interactive heatmap (https://ihwph-hehta.shinyapps.io/subgroup_reporting_app/).

3.2.3 Results

Among 2,235 eligible trials, 48% (1,082 trials) reported overall results and 23% (524 trials) reported subgroups. For any subgroup reporting, adjusting for enrolment size and index conditions, the predictive characteristics were follow-up time (odds ratio (OR), 95%CI: 1.13, 1.04-1.24), enrolment (per 10-fold increment, 3.48, 2.25-5.47)), trial starting year (1.07, 1.03-1.11) and specific index conditions (e.g., hypercholesterolemia, hypertension etc., OR ranged from 2.48 to 10.44). Funding source and number of arms were not associated with subgroup reporting. Results were similar on modelling any result reporting (except number of arms, 1.42, 1.15-1.74) and the total number of subgroups.

Age (51%), gender (45%), racial group (28%) and geographical locations (17%) were the most frequently reported subgroups. Characteristics related to the index condition (severity/duration/types etc.) were somewhat commonly reported (e.g., 69% of MI trials reported on MI severity/duration/types). Also, 16% cardiovascular trials reported on diabetes. However, reporting on metrics of comorbidity or frailty (5 trials) and mental health (4 trials) were rare.

3.2.4 Conclusion

Other than age, sex, race ethnicity, geographic location and characteristics related to the index condition, information on variation in treatment effects is sparse in trial reporting.

3.3 Introduction

3.3.1 Heterogeneity of treatment effect (HTE)

Trials provide estimations of average treatment effects, yet they are less suitable for comprehending the variability, known as HTE, that exists among individuals. Observing the difference in outcome caused by a treatment in an individual is infeasible, given that an individual cannot experience both the treatment and non-treatment simultaneously(100).

The average treatment effect derived from a trial is the difference in average outcomes between the intervention group and the comparator group (such as placebo). However, applying these average outcomes to a particular individual necessitates the assumption of homogeneity, implying that the average treatment effect observed within the study population reflects the impact on any individual within it, which is usually not true. A trial can indicate an overall average benefit for an intervention by showing a large benefit in a small subset of individuals, even if there is no benefit or potential harm observed for the majority(101). Therefore the results of that trial, an average benefit, would be greatly misleading for nearly all individuals(100).

The most common method to investigate whether the treatment effect varies among individuals is to estimate the benefit separately in subgroups of patients. This is based on the assumption that a subgroup is more homogeneous than the entire study population. Consequently, the average effect within the subpopulation may provide a better prediction of the benefit for any individual within that subgroup(100).

3.3.2 What are subgroup analyses?

Subgroup analyses normally split participants into subgroups according to their specific characteristics and make comparisons across them based on the research of interest. The subgroup analyses may be conducted for subset of subjects such as males and females, patients that are over 65-year-old and less than 65-year-old, or for subset of studies in different locations et al. Subgroup analysis is most commonly used to explore HTE(100). It is generally the evaluation of a treatment effect in a specific subset of

subjects as the response to interventions may vary by the subjects' baseline characteristics(102). For example, low-risk participants may fail to show the benefits of intervention in a prevention trial compared with high-risk participants(102). Subgroup analysis is usually run as a post-hoc analysis after individual patient data being collected and obtaining the treatment effects based on the whole participants. It is not a re-test for the null hypothesis in any subgroup as it can be misleading by chance and sample size, instead, it aims to explore if the treatment effects obtained from the whole participants from the trial is heterogeneous across subgroups. For example, male patients who are over 65-year-old in the trial might get a higher benefit than the female patients under 65-year-old. That is usually conducted by running a treatment by subgroup interaction to assess whether the treatment effect is significantly different in the subgroups. Furthermore, it is hopefully expected to detect information about the tailored subgroup patients who can gain most or least benefit from the treatment and generate hypothesis for further research(103, 104).

3.3.3 Challenges of subgroup analyses.

When assessing HTE between subgroups, an interaction test is appropriate(33).

The most informative approach will implement both estimating the treatment effect in that subgroup and an interaction test(105). However, interpreting and reporting subgroup analyses can pose challenges(34), often leading to misleading outcomes due to increased risks of false positives (caused by unadjusted multiple comparisons) and false negatives (resulting from inadequate statistical power)(106, 107).

3.3.4 How to evaluate the heterogeneity of treatment effects across subgroups?

The most widely used methods to assess the heterogeneity is to test the hypothesis of treatment-by-subgroup interaction. If the hypothesis is rejected or, in other words, there is a significant interaction, it will indicate a substantial heterogeneity in treatment effects across subgroups. This case usually indicates further analysis and interpretation(102, 108). Alternatively, if not, then the difference in the effects across subgroups might be subtle and the overall mean effects might be able to capture the treatment effects(102). The interaction testing can decrease the risk of finding false-positive subgroups while its power in detecting true subgroups remains low(109). It is

also worth noting that statistical significance of the findings within each subgroup analysis should not be compared(110).

Trials are generally not powered to detect subgroup effects, and the sample size in subgroup analyses is frequently insufficient to detect clinically significant differences in treatment effects even if these were to exist.(25) Conversely, by testing multiple subgroups, the likelihood of chance findings (i.e. false positives) is increased(25, 111).

To increase statistical power to detect differences in effectiveness, and to confirm or refute apparent subgroup effects from single studies, subgroup analyses of similar trials can be combined in meta-analyses to pool information across multiple trials(26). For such meta-analyses to be possible and reliable, however, studies of comparable agents must assess similar subgroups. Furthermore, these analyses need to be reported consistently. However, not all trials report subgroup analyses(28, 29). Those trials that do report subgroups vary widely in their reporting and adherence to published guidelines(28, 29, 112).

Well-conducted subgroup analyses can sometimes inform health policy recommendations. A meta-analysis of six large diabetes trials with cardiovascular and kidney outcomes revealed that non-White participants had higher rates of cardiovascular and other comorbidities compared to the White participants(113). However, non-White participants accounted for only approximately 21% of the overall enrolled trial populations, which is under-represented. The American Diabetes Association and the European Association for the Study of Diabetes recommended in their consensus report that the increased burden of complications in underrepresented populations with diabetes should be considered in tailoring personalized treatment plans. Ongoing and future trials should aim to recruit participants who are more representative of the entire population with diabetes. This approach will facilitate a more accurate assessment of the effects of interventions within less studied subgroups(114, 115).

Another meta-analysis of all major beta-blocker trials in heart failure with reduced ejection fraction (HFrEF) has revealed no benefit concerning hospitalisation and mortality in the subgroup of HFrEF patients with atrial fibrillation(116). However, as this analysis is a retrospective subgroup analysis and considering that beta-blockers did

not increase the risk, the guideline committee decided not to make a separate recommendation based on heart rhythm(117).

3.3.5 Research questions and rationale

Cochrane Handbook had advised that the investigation of heterogeneity which may be sought by conducting subgroup analyses or meta-regression is less likely to provide useful findings unless there is a considerable number of studies such as at least ten studies/trials for each characteristic of interest, although even ten studies may be too few when the covariates are distributed unevenly(118).

I originally intended to conduct subgroup analysis for trials with the same index conditions and interventions to assess HTE. However, the inconsistent subgroup reporting across trials made this objective unfeasible. Instead, to understand the consistency of subgroup reporting across different index conditions and intervention types to further enable the exploration of HTE, this study will assess a set of trials with multiple chronic medical conditions for subgroup reporting. It aims to address questions regarding:

1) Which subgroups are reported, in which trials, and at what frequency?
2) What are the predictors of subgroup reporting among trial characteristics?

3.4 Methods

3.4.1 Identifying trials registered in ClinicalTrials.gov

The trials selection has been described previously(119) (PROSPERO (CRD42018048202)). The eligible trials were identified through the US Clinical Trials register (clincialtrials.gov). This study was restricted to this database because it allowed efficiently obtaining a large trial-level (rather than paper-level) denominator. Trials were sought between January 1990 (since initial scoping indicated that trials with accessible IPD generally initiated after this date) and November 2016 using the Access to Aggregate Content of ClinicalTrials.gov (AACT) database. This is a copy of ClinicalTrials.gov in a relational database format(120). Additionally, the selected trials

were of phase 2/3, 3, or 4, with a minimum recruitment of ⩾ 300 participants. These trials encompassed participants aged ⩾ 60 years (or without an upper age limit) and focused on evaluating drugs for specific chronic conditions. The chosen conditions were those that necessitate long-term pharmacological therapy. The selection criteria are shown in Supplementary I Table S1. A range of cardiovascular, musculoskeletal, gastrointestinal, respiratory, neurological, urological, metabolic and autoimmune disorders were included. A full list of included conditions, Medical Subject Heading (MeSH) terms and MeSH code are provided in Table S2.

3.4.2 Identifying publications relating to registered trials

All PubMed indexed publications related to the identified trials from the ClinicalTrials.gov database were searched using two approaches. First, the ClinicalTrials.gov database was searched for PubMed IDs (PMIDs) of all relevant registered trials. Trial sponsors are obligated to update the ClinicalTrials.gov database with PMIDs of publications associated with registered trials. Secondly, PubMed was searched using the trial registration number for each relevant trial to identify publications that were not yet added to the database. This search was performed using the R Eutils package(121). This was last updated in April 2019.

3.4.3 Screening of publications

All papers were screened manually and via automatic text searches, as depicted in Figure 3. Initially, an automatic full-text search was conducted using specific strings such as "subgroup," "sub-group," "strata," "by baseline," "subpopulation," or "sub-population." In cases where the automatic screening did not identify any of these terms in the manuscript text, articles (including supplementary appendices) were manually reviewed once to confirm the absence of relevant results. Otherwise, the studies were independently screened by two reviewers.

3.4.4 Data extraction

Trial-level data for all trials identified from ClinicalTrials.gov, regardless of publication status and the presence of subgroup analyses, were extracted from AACT. Extracted

data included ClinicalTrials.gov identifier, index condition, interventions and comparators, number of participants, phase of trial, number of arms, trial sponsor, start date, completion date, countries included, and eligibility criteria. Tabulated subgroup data were extracted using an interactive web-app (https://tabletidier.org/), subgroup results in the manuscript text or in figures without accompanying tables were extracted manually. The text used to describe subgroups were extracted verbatim. To allow comparison of subgroups across different studies, this text was assigned to standard terms using the MeSH and/or WHO Anatomical Therapeutic Chemical (ATC) vocabularies. MeSH is created by National Library of Medicine for indexing journal articles and books in life sciences which is widely used by PubMed and ClinicalTrials.gov registry. I initially extracted the tabulated data via TableTidier with the help of trained clinicians or medical student. Then text or figure data was extracted. I harmonised the collected data, and subsequently allocated appropriate MeSH and/or WHO codes to each to correspond with the relevant subgroups. For potentially ambiguous subgroups, like those in abbreviations, I cross-referenced the original paper to ensure they have been captured correctly. All assigned MeSH and/or WHO code were then reviewed by a clinically qualified investigator for accuracy. I also added additional qualifiers to assigned MeSH terms for subgroups such disease severity or duration to capture more information (i.e., duration of diabetes is one of the subgroups in the diabetes trials).

3.4.5 Statistical analysis

Via an interactive heatmap, all original subgroup terms as well as MeSH terms at the level of individual subgroup for all trials were summarised. The heatmap allows users to examine subgroup reporting according to the type of subgroup as well as the index condition, drug class and other trial characteristics (https://ihwph-hehta.cognishinyapps.io/subgroup_reporting_app/), and where possible, links directly to the extracted tables. It should be clarified that this R shiny app was designed by my supervisor – Prof David McAllister, while the prepared data needed is my work. In this chapter, a concise overview is provided by presenting simple summary statistics such as ranks, counts, and percentages. Additionally, specific terms of interest are presented, and certain terms are collapsed into broader categories using the MeSH hierarchy. For example, heart failure and myocardial infarction are collapsed into the category of CVD (122).

Two sets of logistic regression models for two sets of binary outcomes were fitted, i) any results reported and ii) any subgroups reported (taking those with any results reported as the denominator). For both outcomes, variables included were the year the trial started, number of arms ((>2 arms vs ≤2 arms), number of participants enrolled, sponsor type (industry versus other), duration of follow-up and the index condition (see Table S2 for list of conditions). Among trials with any subgroup reporting, we examined the total number of subgroups using quasi-Poisson models, again including the same covariates. This latter model was confined to trials with one or more subgroup. Data analysis was performed using R version 4.2.

3.5 Results

As reported previously(119), 2,235 registered clinical trials with a pre-specified set of conditions and treatment comparisons were identified. Of the 1,082 trials with published results, 524 (48.43%) trials reported findings from subgroup analyses (907 manuscripts, of which 681 presented these results in tabular form) (Figure 3). Over 2,000 unique strings were reduced to 345 unique MeSH terms. Of these MeSH terms, 182 were further described using qualifiers (eg severity, duration).

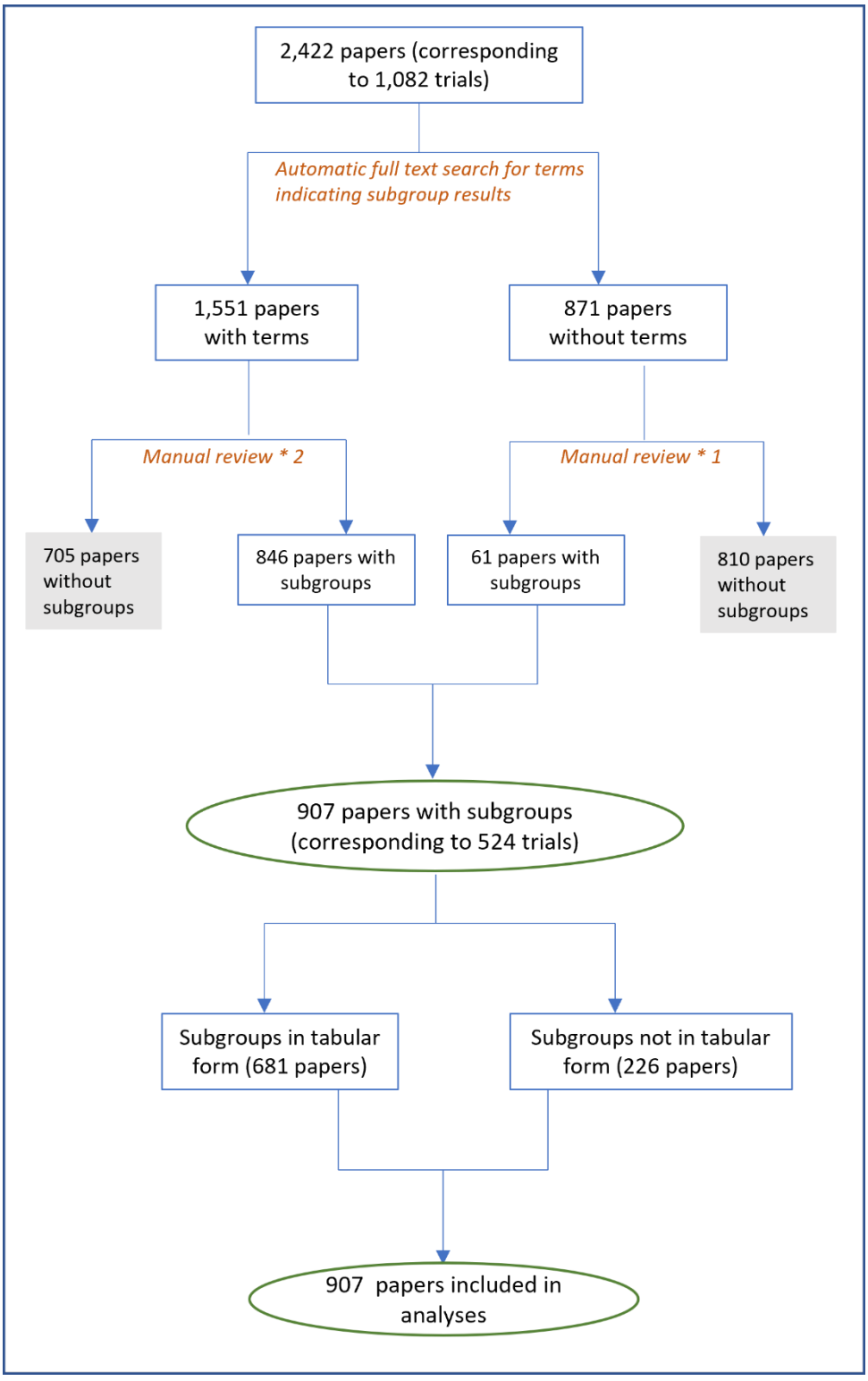
3.5.1 Presence and numbers of subgroups reported

Of the 524 trials reporting subgroups, 156 (30%) reported a single subgroup, 90 (17%) reported 2-3 subgroups, 73 (14%) reported 4-5 subgroups and 205 (39%) reported 6 or more subgroups. Compared to trials without subgroup reporting, trials reporting subgroups were generally larger (median 827 participants, interquartile range (IQR 499 to 1912) versus 610 participants enrolled (IQR 418 to 1000), had longer follow-up (median 2 years (IQR 2 to 4 years) versus 2 (IQR 1 to 3 years)), a higher percentage of non-industry sponsorship (14% versus 9%) and a higher percentage with more than 2 arms (39% versus 35%).

Figure 4 shows associations for any result reporting (yes/no), any subgroup reporting (yes/no among those trials reporting results), and total number of subgroups reported (among those trials reporting >=1 subgroup), using logistic regression models and a Poisson model, respectively. All of the covariates shown were included in the models.

Of the trial characteristics, the number of participants enrolled was the most important predictor of any result reporting OR per 10-fold increase in number enrolled 1.63; 95% CI 1.22 - 2.19, Figure 4), any subgroup reporting (OR per 10-fold increase in number

Figure 3. Screening of subgroups analyses from eligible papers.



enrolled 3.48; 95% CI 2.25 – 5.47, Figure 4) and the total number of subgroups reported (rate ratio (RR) per 10-fold increase 1.69; 95% CI 1.65 – 1.73). Duration of follow-up also predicted any result reporting (OR 1.10 per year of follow-up; 95% CI 1.03 - 1.18), subgroup reporting (OR 1.13; 95% CI 1.04 - 1.24) and the total number of subgroups (RR 1.03; 95% CI 1.02 - 1.03). More recent trials were similar to older trials (OR 0.97, 95% CI 0.95 - 0.99, OR 1.07, 95% CI 1.03 - 1.11 and RR 1.02, 95% CI 1.02 - 1.02 for result reporting, subgroup reporting and number of subgroups respectively). Trials with 3 or more arms were more likely to report results (OR 1.42, 95% CI 1.15 - 1.74) but were not associated with increased subgroup reporting (OR 1.00, 95% CI 0.73 - 1.37) or a higher total number of subgroups (RR 1.01, 95% CI 0.99 - 1.04). Industry funding was not associated with any of the three outcomes (OR 1.03, 0.73 - 1.45; OR 1.58, 0.94 - 2.69 and RR 1.00, 0.97 - 1.03 respectively).

Taking asthma trials as a reference (asthma was chosen to make the ratios easier to interpret as it was an index conditions with lower odds of reporting), subgroup reporting was more likely within trials of cardiovascular, metabolic, thromboembolic index conditions (overall index conditions ORs ranged from 2.48 to 10.44, see Supplementary I Table S3). These trials were also more likely to report larger numbers of subgroups. Results for other indications were more mixed (Figure 4).

3.5.2 Commonest subgroups reported

There was substantial variation in subgroups across index conditions. Across 49 index conditions there were a total of 345 subgroup terms, with a median of 11 terms per index condition ranging from 1 to 97 (interquartile range 6 to 29). Nonetheless, some subgroups were common across all index conditions. Age (268 out of 524 trials, 51%) and gender (in 235 trials, 45%) were the commonest reported subgroups. Despite being the most common, these subgroups were only reported in approximately 50% of trials with documented subgroups, which accounts for roughly 25% (268 out of 1,082 trials) of the trials with reported results. It was followed by comorbid diabetes (154 out of 524 trials, 29%), racial group (in 146 trials, 28%), BMI (in 125 trials, 24%), geographical locations (in 88 trials, 17%), Glycated Haemoglobin A (in 72 trials, 14%) and cigarette smoking (in 63 trials, 12%). Most of the BMI subgroup reporting appeared in the context of type 2 diabetes trials (out of 125 trials reporting BMI, 44 of them are in type 2 diabetes trials, 35%), followed by hypercholesterolemia trials (12%, n = 15) and Chronic Obstructive Pulmonary Disease (COPD) and hypertension trials (both 7%, n = 9). Among trials

reporting cigarette smoking subgroups, most were COPD trials (18 out 63 trials, 29%) followed by coronary artery disease (13%, n = 8) and type 2 diabetes trials (11%, n = 7).

For many trials, subgroups relating to the index condition (e.g., duration or severity) were commonly reported which meant that treatment effects were stratified by the type, duration or severity of the index condition. For example, among 26 myocardial infarction trials with subgroup reporting, 69% reported severity/history/type of myocardial infarction as a subgroup, for type 2 diabetes trials, 29 of 120 trials reported diabetes characteristics (mainly duration) as a subgroup and for COPD trials 30 of 40 trials (75%) reported severity of COPD as a subgroup, while 88% stroke trials which reported subgroup analyses reported previous/severity/type of stroke as a subgroup (Table 4).

Figure 4. Predictors of subgroup reporting and total number of subgroups.

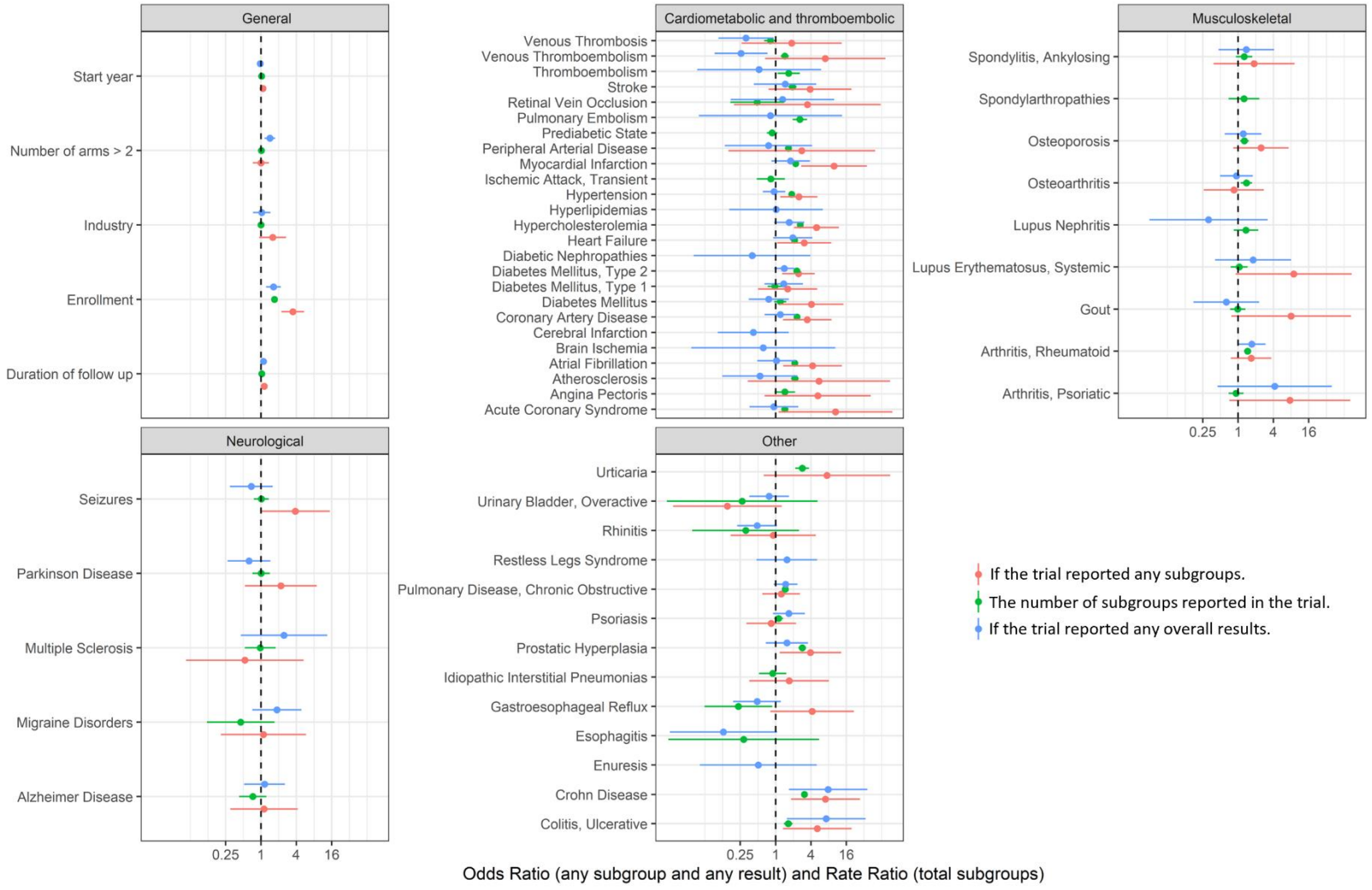


Table 4. The proportion of subgroup reporting and commonest subgroups in each index condition.

Conditions	Total subgroups	The proportion of subgroup reporting among 2,235 trials n_T/N (%)	The proportion of subgroup reporting among 1,082 trials with results reporting n_R/N_R (%)	Five commonest subgroups in each condition
Myocardial Infarction	99	26/47 (55%)	25/30 (83%)	Age Factors (25); Diabetes Mellitus (23); Gender Identity (23); **Myocardial Infarction (18)**; Hypertension (8)
Diabetes Mellitus, Type 2	89	120/460 (26%)	117/235 (50%)	Age Factors (59); Glycated Hemoglobin A (58); Gender Identity (47); Body Mass Index (44); Racial Groups (44)
Coronary Artery Disease	77	27/80 (34%)	27/46 (59%)	Diabetes Mellitus (23); Age Factors (20); Gender Identity (20); Myocardial Infarction (10); Hypertension (9)
Hypertension	64	44/247 (18%)	44/98 (45%)	Age Factors (26); Gender Identity (23); Diabetes Mellitus (17); Racial Groups (16); Blood Pressure (12)
Heart Failure	51	17/40 (42%)	17/27 (63%)	Age Factors (12); Diabetes Mellitus (11); Gender Identity (11); Stroke Volume (10); **Heart Failure (9)**
Hypercholesterolemia	48	28/72 (39%)	28/43 (65%)	Lipoproteins (20); Diabetes Mellitus (19); Age Factors (18); Gender Identity (17); Body Mass Index (15)
Atrial Fibrillation	46	13/39 (33%)	13/20 (65%)	Age Factors (8); Gender Identity (7); Heart Failure (7); **Atrial Fibrillation (6)**; Hypertension (5)

Pulmonary Disease, Chronic Obstructive	40	40/186 (22%)	39/96 (41%)	**Pulmonary Disease, Chronic Obstructive (30)**; Age Factors (20); Cigarette Smoking (18); Gender Identity (17); Steroids (16)
Acute Coronary Syndrome	37	9/22 (41%)	9/10 (90%)	Age Factors (8); Gender Identity (7); Diabetes Mellitus (6); Myocardial Infarction (5); Percutaneous Coronary Intervention (5)
Arthritis, Rheumatoid	35	28/106 (26%)	28/65 (43%)	**Arthritis, Rheumatoid (13)**; Age Factors (7); Gender Identity (6); Immunosuppressive Agents (6); C-Reactive Protein (5)
Stroke	35	8/20 (40%)	8/13 (62%)	**Stroke (7)**; Age Factors (5); Gender Identity (5); Diabetes Mellitus (3); Hypertension (3)
Atherosclerosis	30	2/9 (22%)	2/3 (67%)	Age Factors (2); Body Mass Index (2); Cigarette Smoking (2); Diabetes Mellitus (2); Gender Identity (2)
Crohn Disease	29	11/18 (61%)	11/16 (69%)	Immunosuppressive Agents (7); Tumor Necrosis Factor Inhibitors (7); C-Reactive Protein (6); **Crohn Disease (5)**; Steroids (5)
Osteoporosis	29	11/44 (25%)	11/23 (48%)	Age Factors (6); Fractures, Bone (6); **Osteoporosis (5)**; Body Mass Index (3); Geographic Locations (3)
Prostatic Hyperplasia	28	9/30 (30%)	9/15 (60%)	Body Mass Index (4); Age Factors (3); Erectile Dysfunction (3); Adrenergic alpha-Antagonists (2); Antihypertensive Agents (2)
Peripheral Arterial Disease	24	3/8 (38%)	3/4 (75%)	Diabetes Mellitus (2); Age Factors (1); Ankle Brachial Index (1); Blood Pressure (1); Body Weight (1)
Venous Thromboembolism	23	7/36 (19%)	7/8 (88%)	Age Factors (6); Gender Identity (6); **Venous Thromboembolism (4)**; Anticoagulants (3); Body Weight (3)

Asthma	22	19/147 (13%)	19/62 (31%)	**Asthma (6)**; Eosinophilia (6); Steroids (5); Age Factors (4); Gender Identity (4)
Colitis, Ulcerative	21	8/14 (57%)	8/12 (67%)	Steroids (5); Tumor Necrosis Factor Inhibitors (5); C-Reactive Protein (3); Gender Identity (3); Age Factors (2)
Psoriasis	19	13/62 (21%)	13/37 (35%)	Immunosuppressive Agents (5); **Psoriasis (5)**; Tumor Necrosis Factor Inhibitors (4); Biological Therapy (2); Cyclosporins (2)
Diabetes Mellitus	16	8/36 (22%)	8/15 (53%)	Age Factors (6); Body Mass Index (6); Gender Identity (6); Racial Groups (5); Glycated Hemoglobin A (3)
Osteoarthritis	14	6/64 (9%)	6/26 (23%)	Age Factors (3); Arthritis, Rheumatoid (3); Diabetes Mellitus (2); Gender Identity (2); Pain (2)
Urticaria	12	2/3 (67%)	2/3 (67%)	Age Factors (1); Angioedema (1); Autoantibodies (1); Body Weight (1); Gender Identity (1)
Diabetes Mellitus, Type 1	11	7/35 (20%)	7/17 (41%)	Glycated Hemoglobin A (4); Insulin (3); Age Factors (2); Body Mass Index (2); Glucose (2)
Hyperlipidemias	11	1/7 (14%)	1/4 (25%)	Age Factors (1); C-Reactive Protein (1); Diabetes Mellitus (1); Gender Identity (1); Geographic Locations (1)
Pulmonary Embolism	11	1/2 (50%)	1/1 (100%)	Age Factors (1); Body Mass Index (1); Fibrin Fibrinogen Degradation Products (1); Gender Identity (1); Neoplasms (1)
Lupus Erythematosus, Systemic	10	4/8 (50%)	4/5 (80%)	Autoantibodies (2); Racial Groups (2); Steroids (2); Albuminuria (1); Antimalarials (1)
Arthritis, Psoriatic	9	3/5 (60%)	3/4 (75%)	Immunosuppressive Agents (2); Antirheumatic Agents (1); Arthritis, Juvenile (1); **Arthritis,**

				Psoriatic (1); Arthritis, Rheumatoid (1)
Gastroesophageal Reflux	9	5/29 (17%)	5/8 (62%)	Body Mass Index (2); Age Factors (1); Gastrointestinal Diseases (1); Gender Identity (1); Heartburn (1)
Seizures	9	6/31 (19%)	6/12 (50%)	Anticonvulsants (5); Age Factors (3); other antiepileptics (3); Racial Groups (2); Gender Identity (1)
Spondylitis, Ankylosing	9	3/15 (20%)	3/8 (38%)	C-Reactive Protein (2); Tumor Necrosis Factor Inhibitors (2); Arthritis (1); Cigarette Smoking (1); Gender Identity (1)
Angina Pectoris	8	2/4 (50%)	2/4 (50%)	Age Factors (2); Gender Identity (2); Body Weight (1); Diabetes Mellitus (1); Electrocardiography (1)
Gout	8	5/11 (45%)	3/4 (75%)	Glomerular Filtration Rate (3); Renal Insufficiency (3); Age Factors (2); Comorbidity (2); Diuretics (2)
Parkinson Disease	8	4/38 (11%)	4/12 (33%)	**Parkinson Disease (4)**; Age Factors (3); Gender Identity (3); Body Weight (1); Depression (1)
Idiopathic Interstitial Pneumonias	7	3/8 (38%)	3/8 (38%)	Vital Capacity (2); Age Factors (1); Cigarette Smoking (1); Geographic Locations (1); Hydroxymethylglutaryl-CoA Reductase Inhibitors (1)
Thromboembolism	7	1/4 (25%)	1/1 (100%)	Age Factors (1); Embolism and Thrombosis (1); Gender Identity (1); Obesity (1); Specialties, Surgical (1)
Alzheimer Disease	6	4/31 (13%)	4/16 (25%)	**Alzheimer Disease (2)**; Dementia (2); Apolipoprotein A-I (1); Gender Identity (1); Genetic Profile (1)
Multiple Sclerosis	6	2/8 (25%)	2/6 (33%)	Age Factors (2); Coronary Artery Disease (1); Gadolinium (1);

				Gender Identity (1); Multiple Sclerosis (1)
Prediabetic State	6	1/1 (100%)	1/1 (100%)	Body Mass Index (1); Body Weight (1); Diabetes Mellitus (1); Gender Identity (1); Racial Groups (1)
Venous Thrombosis	6	2/21 (10%)	2/5 (40%)	Age Factors (1); Body Weight (1); Gender Identity (1); Neoplasms (1); Renal Insufficiency (1)
Ischemic Attack, Transient	5	1/1 (100%)	1/1 (100%)	Age Factors (1); Coronary Artery Disease (1); Gender Identity (1); Racial Groups (1); Stroke (1)
Lupus Nephritis	5	1/4 (25%)	1/1 (100%)	Cyclophosphamide (1); Gender Identity (1); Geographic Locations (1); Racial Groups (1); unclassifiable (1)
Spondylarthropathies	5	1/1 (100%)	1/1 (100%)	Age Factors (1); Antirheumatic Agents (1); Axial Spondyloarthritis (1); Gender Identity (1); Tumor Necrosis Factor Inhibitors (1)
Migraine Disorders	3	2/22 (9%)	2/11 (18%)	Adrenergic beta-Antagonists (1); **Migraine Disorders (1)**; sumatriptan (1)
Raynaud Disease	3	1/1 (100%)	1/1 (100%)	Blood Pressure (1); Gender Identity (1); unclassifiable (1)
Retinal Vein Occlusion	2	1/4 (25%)	1/2 (50%)	Macular Edema (1); unclassifiable (1)
Rhinitis	2	2/41 (5%)	2/11 (18%)	Geographic Locations (1); unclassifiable (1)
Esophagitis	1	1/10 (10%)	1/1 (100%)	unclassifiable (1)
Urinary Bladder, Overactive	1	1/39 (3%)	1/14 (7%)	Urinary Bladder Diseases (1)

Some trials might correspond to multiple index conditions, the commonest condition among 2,235 trials were kept for simplicity; the number for some subgroups is the same in the 5th place and only one was kept based on the alphabetical order; the subgroup in bold is the subgroup same as the condition term with additional information such as type, severity, duration et; $n_{T:}$ number of trials with subgroup reporting among 2,235 trials; $n_{R:}$ number of trials with subgroup reporting among 1,082 trials with results reporting; $N_{R:}$ trials with results reporting and N_R = 1,082.

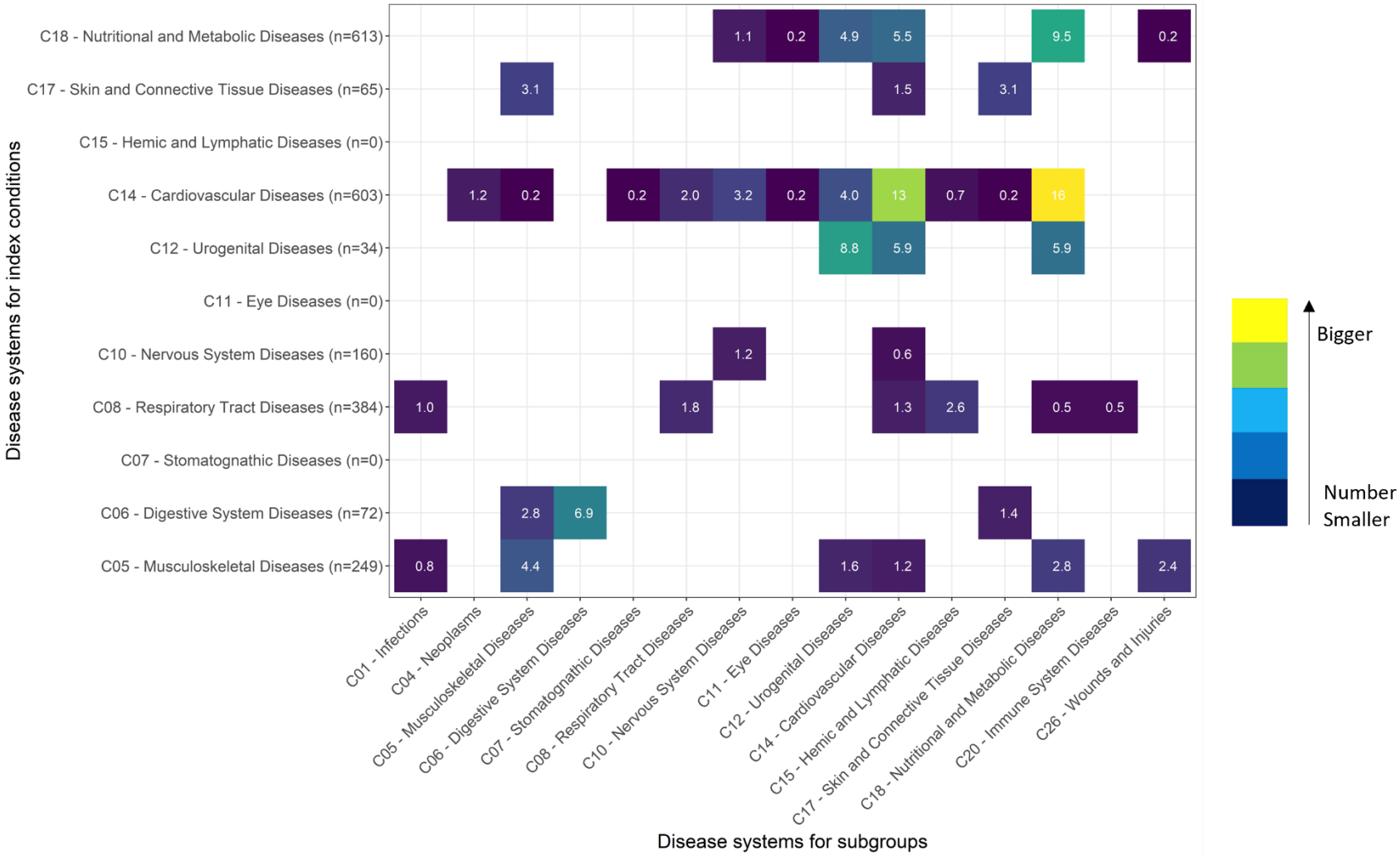
3.5.3 Comorbidity subgroup reporting

Where conditions other than the index conditions were reported as subgroups, this was largely confined to diseases within the same body system as the index condition. Figure 5 illustrates this - the organ system for each index condition and each subgroup are shown on the y and x-axis respectively and the % of subgroups reported per organ system are shown on each cell - frequencies above 5% were generally seen on the diagonals (e.g., 13% trials of cardiovascular diseases reported a non-index condition cardiovascular disease subgroup – e.g., stroke trials reported hypertension as a subgroup which are both cardiovascular disease). Where there were high percentages off the diagonal (i.e., where the index condition and subgroup pertained to different organ systems), the subgroup conditions were either known causes or known sequelae of the index condition such as nutritional and metabolic disease (predominantly diabetes) in cardiovascular disease trials (16%), or cardiovascular diseases (5.5%) and renal disease (4.9% urogenital diseases) in diabetes trials. In contrast, only 1.3% of respiratory tract diseases trials reported subgroup results according to presence/characteristics of cardiovascular diseases.

3.5.4 Comorbidity, multimorbidity, frailty and mental health

Trials rarely included metrics of comorbidity, multimorbidity or frailty (5 trials). 78 trials (15%) reported estimated glomerular filtration rate or renal insufficiency as renal impairment measures and the majority were either type 2 diabetes trials (n = 28) or heart failure (n = 8) trials. Subgroups related to mental health were particularly rarely reported with only 4 of the 524 trials (1%) including MeSH terms within these categories.

Figure 5. Comorbidities reported in each disease system.



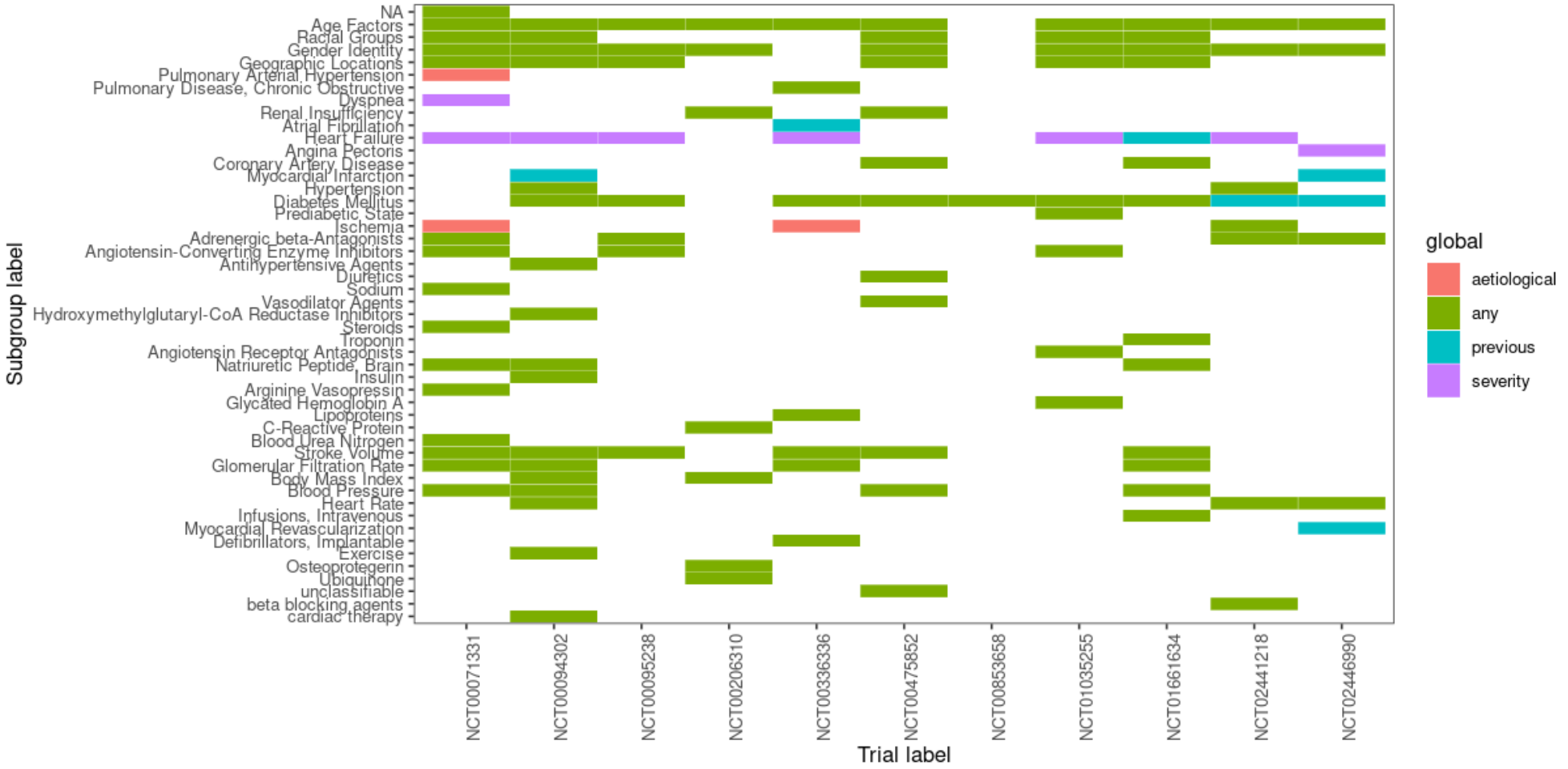
3.5.5 Demonstration of the heatmap for large heart failure trials.

Figure 6 serves as an illustrative example of the heatmap. Specifically, I opted for large HF trials with an enrolment size exceeding 2,000 for the purpose of this demonstration. The interactive heatmap can be accessed through the following link: https://ihwph-hehta.shinyapps.io/subgroup_reporting_app/. Multiple heatmaps can be created, which can be customised based on various criteria such as trial index conditions, drug class, subgroups, and trial-specific details like sample size. Comprehensive instructions can be found on the webpage.

Before drawing the heatmap, after the selection of HF trials with a sample size of over 2,000, the shinyapp produced summary statistics for eligible HF trials. 11 trials were eligible with 33 published papers and there were 46 MeSH and ATC code to categorise a total of 280 subgroup strings captured from the papers.

As showed in Figure 6, age (10 out of 11 trials) and gender (9 trials) were most commonly reported. Diabetes (9 trials) was also very commonly reported with 2 trials reported as history of diabetes, which is in line with the overall finding that diabetes were commonly reported in CVD trials.

Figure 6. Heatmap for large heart failure trials.



3.6 Discussion

On reviewing more than 2,000 trials registered on ClinicalTrials.gov, I made a number of observations about subgroup reporting. First, only around a quarter of clinical trials report subgroup effects. Secondly, of those that report subgroup effects, just under half (47%) report on 3 or fewer subgroups. Thirdly, the number of participants enrolled, the duration of trial follow-up and trial starting year predict subgroup reporting. Fourthly, after accounting for participants enrolled, industry funded trials are not more likely to report subgroup effects. Fifthly, some trials with conditions of cardiovascular, metabolic and thromboembolic disease are the most likely to report on subgroups.

Finally, this study showed that even where trials do report subgroups, this is largely confined to "general" subgroups such as age, sex, race/ethnicity, geographic variation or to features of the index condition. Few trials report on comorbidities related to other body systems. Mental health disorders or metrics of comorbidity, multimorbidity or frailty were rarely covered. Together these findings suggest that – with the exception of cardiometabolic and thromboembolic diseases, and especially for subgroups not closely related to the index condition - the published literature contains only sparse information on how treatment effects differ within clinical trials.

Certain variables, like age and gender with their trial index conditions, may possess adequate information for the examination of HTE and the inclusion in the pre-trial protocol. However, for variables such as multimorbidity and frailty, there is minimal available information, requiring the use of IPD to investigate HTE and enhance the applicability of the trial.

3.6.1 Strengths and weaknesses of this study

A strength of this study is that, unlike most previous studies(32, 123), registered trials were included regardless of where they were published. Secondly, this study was the largest, to my knowledge, to assess subgroup reporting among trials of chronic medical conditions. Thirdly, this was the only study to assign terms to standard terminologies allowing comparison across multiple conditions and drug classes. However, there were a number of limitations. First, where papers were neither notified to the

ClinicalTrials.gov register nor included a trial registration identifier in PubMed, they were not obtained. However, the number of papers missed is likely to be small because the trial registration number is required by the International Committee of Medical Journal Editors(124). Secondly, subgroup results in non-indexed sources (e.g., clinical study reports) will have been missed, although many of these are only accessible after a formal application process which may require a data sharing agreement. Thirdly, a small number of terms could not be assigned to MeSH or ATC codes due to their complexity. Finally, the results are confined to chronic medical conditions, and exclude trials in infectious diseases, oncology and (other than dementia) psychiatric disorders.

3.6.2 Strengths and weaknesses in relation to other studies

The majority of previous subgroup studies were concerned with the reliability of subgroup findings in the context of a single paper. As such, since higher impact journal publications are likely to be the most influential, most confined their analysis to papers published in one or more high impact medical journal or in the case of Sun et al, on core medical journals (as defined by the national library of medicine)(27). Only one study, Kasenda et al examined all papers regardless of the journal type, but this was confined to a set of trials which had been approved by one of six research ethnics committees in Switzerland(125). This difference in papers included could account for the fact that previous studies found an association between industry funding and subgroup reporting while I did not. Alternatively, the null association for type of funding could be due to heterogeneity in this association according to the statistical significance of the primary outcome(27, 126). Sun et al found that the association between industry funding and subgroup reporting was only present when the primary analysis was not statistically significant, and data on this variable was not collected in this study.

Nevertheless, there were a number of findings common to this study and previous findings; particularly that larger studies were more likely to report subgroup effects(28). It is understandable that larger trials tend to have more subgroup analyses since detecting differences in effects between subgroups typically requires larger sample sizes compared to assessing the overall treatment effect(127). Small sample sizes may have little power for subgroup analysis, which can cause false negatives(128). Pre-specification of a subgroup is crucial for making the findings from subgroups more convincing(106). It is recommended to pre-specify subgroups during the study design or implementation stage. This helps collect the appropriate data to identify subgroup

members and ensures the study has sufficient statistical power to detect relevant subgroup differences in intervention effects, which can provide further evidence for policy decisions tailored to specific subgroup populations(129). Previous literature holds differing views regarding the stratified randomisation by the factors defining the subgroup on the balance of prognostic factors within the subgroup(130). A pre-specified subgroup analysis is generally regarded as more credible and valid, and this credibility is not affected by whether the randomisation is stratified based on subgroups (128, 131, 132). The association between the stratification factors is not well demonstrated in the literature to my best knowledge. Sun et al, also found that trials of surgical interventions were less likely to report subgroups than non-surgical trials(27). This study extended this finding by showing considerable variances among non-surgical trials; even after adjusting for trial size and other variables, cardiovascular and metabolic trials were considerably more likely to report subgroup effects.

Only one study found reported detailed information on which subgroups were reported(28). In the study appendix, Gabler et al reported the percentage of the 1,042 reported variables which were allocated to specific categories. These included centre or site (3%), anthropomorphics (4%), demographics (25%), comorbidities (10%), disease severity marker (32%), medical history (6%), medications at baseline (9%), temporal features (4%), multivariable risk scores (3%) or others (5%). Several of these were further sub-categorised. For example, comorbidity was categorised diabetes (31%), cardiovascular disease (35%) and demographics into age, sex, race/ethnicity, smoking status and other. These percentages appear consistent with the observations in this chapter as to which subgroups were commonest, although treating the variables examined as the denominator meant that it cannot be directly compared with my findings.

The reporting of subgroup analyses in trials varies, with some trials reporting only the numbers, proportions, and event rates of patients in the subgroups(133), while others provide detailed subgroup-specific results(134). This variability can pose challenges for conducting a meta-analysis that aims to estimate subgroup effects across different studies. Proctor et al conducted a NMA based on direct and indirect evidence integrating to estimate the treatment effect for a patient subgroup(135). This model could potentially mitigate the challenge of integrating subgroup-specific results with those that only provide the indirect information of patients in the subgroups. However, it may introduce bias when using less patient level data and suffer from low power

when the study or covariate effect is small(135). Therefore, caution is needed when using this model to integrate information and results have to be interpreted with care.

3.6.3 Meaning of the study

According to the Cochrane Handbook for Systematic Reviews of Interventions, subgroup analyses are "uncommon in systematic reviews based on published literature because sufficient details to extract data about separate participant types are seldom published in reports"(118). This study showed that considerable variation in reporting between trials even within the same index condition and drug class was one reason for this lack of detail. Nonetheless, common variables did emerge such as age, sex, geographic region, race/ethnicity and features of the index condition.

In contrast I found there was very little information contained in the publicly available literature about comorbidity and multimorbidity. Given that multimorbidity is common, increasing in prevalence, and is known to complicate clinical decision making, the lack of such information is a challenge for decision-makers(136). Hanlon et al previously showed that, while under-represented, multimorbidity is not absent from clinical trials(119). Despite this, very few trials have reported treatment effects according to comorbidity, multimorbidity or frailty scores. Moreover, for individual comorbidities, the majority of reporting was for conditions in the same body system as the index disease (e.g., a history of ischaemic heart disease in a trial of an antihypertensive), so there was little information about "discordant" comorbidities (e.g., coexisting prostate disease and heart failure), which are the most complex and difficult to treat. Nonetheless, given the large number of ways in which multimorbidity can be defined and measured, standards are needed if these are to be incorporated into clinical trial reporting.

An interesting contrast between this study and most previous reports was the focus; it was concerned with all subgroup reports for trials regardless of whether the subgroup was reported in a high impact journal. Underlying this difference is a difference in the consumer of the subgroups - the person looking at a single trial, versus the secondary researcher. For the reader of a single trial, to avoid dangers of over-interpretation, individual papers should be very cautious in reporting subgroup effects. However, this is

the opposite of what is desirable for meta-analyses across multiple trials, where completeness and consistency would be helpful.

At present, neither audience is well served. As this study shows, trials are highly variable in what subgroups are reported, while as others have shown papers rarely meet the published standards for pre-specification(34). It would seem that in the digital age, both audiences could be served. Trial reports could continue to limit subgroup reporting in line with current recommendations, while a wider common set of subgroup effect standard terminologies for use by secondary researchers. This is an exactly opposite strategy to reduce bias in subgroup reporting from that normally advocated – confining subgroups reporting to a small set of pre-specified variables – instead rather this study reduce bias through completeness. This would of course require an agreement as to what should constitute such a wider common set of subgroup effects. I hope that these findings, showing dramatic and unhelpful variation across trials, and a paucity of information on the impact of health states important for decision-making (such as comorbidities and frailty), help demonstrate a need for such a consensus.

3.7 Conclusion

Approximately 1 in 4 trials report results for one or more subgroups. Age, sex, race/ethnicity and features of the index condition were the most common subgroups. Where subgroup effects for other conditions were reported, these were largely confined to the same body system as the index condition. Outside these areas information on variation in treatment effects was sparse.

Chapter 4: Case study 2 - Transportability of two heart failure trials to a disease registry using individual patient data.

4.1 Chapter summary

This chapter described the use of a parametric survival model and an inverse odds of sampling weights model to calibrate trial findings from two landmark HF trials. The objective was to assess whether the trial findings remain applicable in the real-world settings by using individual patient data from both the trials and the real-world HF register. It included baseline characteristics selection, model building, exploratory analyses, results comparisons and interpretation, and implications for data availability.

4.2 Abstract

4.2.1 Background

Trials are the gold-standard for determining therapeutic efficacy and safety, but the characteristics of patients participating in trials differ from those encountered in clinical practice. Calibration can partially account for these differences, improving the applicability of trial findings, without breaking randomisation. I calibrated characteristics of patients from two HF trials to those enrolled in a HF registry.

4.2.2 Method

Individual-patient-level data from two trials (COMET, comparing carvedilol and metoprolol, and DIG, comparing digoxin and placebo) and a Scottish HF registry with 8,012 HF patients were obtained. The primary endpoint for both trials was all-cause mortality; secondary composite outcomes were all-cause mortality or hospitalisation for COMET and worsening HF culminating in death or hospitalisation for DIG. I performed

regression-based and inverse odds of sampling weights (IOSW) transportation approaches.

4.2.3 Results

Registry patients were older and received higher-doses of loop-diuretics than trial participants. For each trial, point estimates were similar for uncalibrated and IOSW (e.g., DIG composite outcome: OR, with placebo as reference, 0.75 95% CI (0.69, 0.82) versus 0.73 (0.64, 0.83)). Treatment effect estimates were also similar when calibrated to high-risk OR 95% CI (0.64 (0.46, 0.89)) and low-risk registry patients (0.73 (0.61, 0.86)). Similar results were obtained using regression-based transportation.

4.2.4 Conclusion

Regression-based or IOSW approaches can be used to calibrate trial effect estimates to patients administrative/registry data, with only moderate reductions in precision.

4.3 Introduction

Trials are the gold-standard for determining the efficacy and safety of treatments (137, 138). However, participants in heart failure trials are generally younger, more likely to be men, and have fewer comorbidities such as chronic respiratory or kidney disease than those encountered in clinical practice (39, 139). If the patient characteristics that are under-represented are also associated with differences in treatment efficacy (e.g., if efficacy is lower in older people), the applicability of trial findings to clinical practice is attenuated. Partly for this reason, trials sometimes report baseline characteristics (such as age, sex, and disease severity) as well as treatment effects stratified by subgroups. However, individual patients may have many co-existing characteristics (for instance anaemia and renal dysfunction) which are not represented in trial analysis with one-variable-at-a-time subgroup reporting(24).

Statistical trial transportation, also called calibration or population adjustment in other contexts(58, 60, 61), addresses these difficulties by weighting trial results to reflect the

characteristics of target populations more closely and, importantly, without breaking randomisation. Briefly, transportation apportions greater weight to randomised participants that were under-represented in the trial compared to the target population and less weight to participants who were over-represented in the trial, compared to the target population(58, 140). Calibration has been used in other conditions such as HIV (141) and lung cancer (142) and employed in a dual antiplatelet therapy (DAPT) study(143) but to my knowledge there was no previous attempt to transport HF trials to a clinical practice registry for patients with HF in order to estimate effects in clinical practice. trials require considerable resources, in terms of research staff, finances and patient commitment (144); it is important to maximise their utility for clinical practice.

Accordingly, I examined the effect of transporting two landmark HF trials to patients from a Scottish clinical practice HF registry using two different methods with differing assumptions – inverse odds of sampling weights (IOSW), and regression modelling.

4.3.1 Research questions and rationale

This chapter aims to answer three questions:

(1) Is the HF population encountered in clinical practice in Scotland different from the participants included in two HF trials?

(2) Can we calibrate trial data using disease registry?

(3) How does the treatment effect estimate change when performing calibration?

This chapter can offer valuable insights into the applicability of the primary results of two HF trials to the real-world population. It can also serve as an example for evaluating the applicability of other HF trials to broader populations using real-world data.

4.4 Methods

4.4.1 Data sources and governance

4.4.1.1 Data sources and study population

4.4.1.1.1 Carvedilol or Metoprolol European Trial (COMET)

COMET was a multicentre, randomised, double-blind, parallel-group comparison of carvedilol and metoprolol in participants with a left ventricular ejection fraction (LVEF) of 35% or less. Conducted in 15 European countries, 1,511 participants were randomly assigned to carvedilol and 1,518 to metoprolol tartrate. The mean trial duration was 58 months. The primary endpoints were all-cause mortality, and a composite of all-cause mortality and all-cause hospitalisation (11, 12).

4.4.1.1.2 The Digitalis Investigation Group Trial (DIG)

DIG was a randomised, double-blind trial of the effect of digoxin on all-cause mortality compared to placebo among people with chronic HF. Studied in 302 centres, 7,788 participants were involved. The main trial was conducted for 6800 participants which had a LVEF of 45% or less. The average follow-up time was 37 months. The primary outcome of the trial was all-cause mortality. Worsening HF culminating in death or hospitalisation was reported as a composite secondary outcome (13, 14).

4.4.1.1.3 Heart Failure Registry

A clinical practice registry of individuals with HF (predominantly with HFrEF) was obtained from the largest regional health authority in Scotland (National Health Service Greater Glasgow & Clyde, NHSGGC), which covers 1.14 million people (almost a quarter of the Scottish population)(15). People with HF in the region who were assessed by community HF nurses or HF clinics were included in the registry. Each patient's clinical features (diabetes, ischaemic heart disease etc.), therapy, vital signs (heart rate, systolic blood pressure (SBP)), results of blood tests (serum sodium, potassium,

creatinine etc.) were routinely recorded in an electronic health record to support clinical care. Missing values in the HF registry were assumed to be missing at random, which means the probability that a value is missing only depends on observed values and not on unobserved values(145). It also means that after all available data (i.e., the variables included in the imputation model) have been accounted for, any remaining missingness can be regarded as random(146). In this chapter, missing data were imputed by a predictive mean matching algorithm with one imputed dataset being generated for simplicity(147). Briefly, the variable with missing values is regarded as the dependent variable in a regression model while all other variables are independent variables. Then the missing values in this dependent variable are replaced with imputations from this regression model. And this imputed dependent variable with both observed and imputed values can subsequently be used as an independent variable in the regression models for other variables. It then repeated for each variable that has missing values(148). The imputation includes diabetes (12%), SBP (14%), heart rate (12%), serum sodium (24%), estimated glomerular filtration rate (eGFR, 24%) and loop diuretics (17%) with dose expressed in furosemide equivalents (e.g., 1mg of bumetanide = 40mg of furosemide).

4.4.1.2 Data storage

Data was stored in the Safe Haven platforms via Robertson Centre for Biostatistics. Storing in a Safe Haven environment ensures the secure handling of sensitive data, minimising the risk of unauthorized disclosure. This is achieved by strict control over access, data analysis, and output dissemination. It provides a safeguarded setting for the linkage, storage, and analysis of personal data. Access to the Safe Haven is granted exclusively to authorized researchers listed on the study's application form and possessing valid information governance training certification. Remote access to the Safe Haven is enabled via a virtual private network, which ensures a secure connection to the network. This approach essentially allows working on a restricted terminal where data cannot be copied, removed, or stored. The user agreement outlines user responsibilities, along with sanctions and penalties for any breaches. All analyses were conducted exclusively within the Safe Haven, and outputs were meticulously reviewed to prevent any possibility of identifying individual data before release. Only controlled and non-disclosing outputs were transferred to me as the researcher.

Trial data and the real-world HF register data were stored in different Safe Havens.

4.4.1.3 Variables extraction and data cleaning

4.4.1.3.1 Variables selection

Variables were included as model covariates if they were available in both the trial and the registry and if they were regarded to have the potential to modify treatment effectiveness by cardiologists.

4.4.1.3.2 Variables extraction among trials and the registry

4.4.1.3.2.1 Baseline characteristics

For baseline characteristics, age (years), systolic blood pressure (mm Hg), heart rate (beats per minute), sodium (mmol/l), eGFR (mL/min/1.73m^2), loop diuretics (mg, referred to frusemide), male (%) and history of diabetes (%) were selected from both the trials and the registry. In the registry, age was calculated by using the date of referral to subtract the date of birth. Creatinine was extracted from both trials and the registry to calculate eGFR by using formula 2 along with age, gender and race information. Different datasets used various loop diuretics, including furosemide, bumetanide, ethacrynic acid, and torsemide. These could be administered orally (PO), via intramuscular injection, slow intravenous injection (IV), or intravenous infusion(149). During data cleaning, patients taking bumetanide, ethacrynic acid, or torsemide were converted to equivalent doses of furosemide. Since patients in the HF registry visited outpatient clinics and some units of the drug were recorded as tablets, it implies oral administration. Loop diuretics (specifically furosemide) were then calculated using formula 3(150-153). This harmonized terminology between trials and registry data, making them comparable. Other covariates were straightforwardly extracted from the datasets.

eGFR = 186 x (Creatinine/88.4)-1.154 x (Age)-0.203 x (0.742 if Female) x (1.210 if in Black race)

(Formula 2)

40 mg PO furosemide ≈ 1 mg PO bumetanide ≈ 20 mg PO torsemide ≈ 100 mg PO ethacrynic acid

(Formula 3)

For variables like heart rate and creatinine, which were measured multiple times during the follow-up, the initial measurement was selected to represent the baseline situation. After collecting all the variables, outliers in numerical variables were addressed. For instance, some recorded SBP measurements were extremely high, like 15,080, 13,900, 1,140 mmHg, which were clear outliers. These values were corrected to the median value of 120 mmHg. Similarly, certain heart rate values were recorded as 6, 7, or 694 beats/min, and these were corrected to the median of 72 beats/min.

4.4.1.3.2.2 Endpoint variables

The endpoint variables were extracted accordingly, based on the primary outcomes of the trials. In COMET they were all-cause mortality, and a composite of all-cause mortality and all-cause hospitalisation. In DIG the primary outcome was all-cause mortality. Worsening HF culminating in death or hospitalisation was reported as a composite secondary outcome.

The extraction of endpoint variables in the registry was as below:

- All-cause death: Patients with recorded death dates were coded as having experienced death, while those without such records were coded as having no outcome or being censored. For most patients, referral dates were available. In cases where the referral date was missing, the first contact date was used to calculate the time to death by subtracting the date of referral or first contact from the date of death.

- All-cause death and all-cause hospitalisation: The initial step involved capturing all-cause hospitalizations. A patient management variable in the register indicated whether a patient was treated as a "day case" (not retained overnight) or an "inpatient." Those designated as inpatients were identified as being hospitalized. Some patients experienced multiple admissions, including instances before their referral date, suggesting prior hospitalization. To determine the time of admission, the first instance of admission after the referral date was selected, and the time to admission was calculated by subtracting the referral or first contact date from the admission date. In cases where patients experienced both admission and death, the time of the first occurrence was selected as the endpoint time, typically the time of admission. For the identifier, a value of 1 (1 = yes) was assigned if either admission or death occurred. Conversely, it was set to 0 (0 = no) when neither endpoint event took place.

4.4.2 Statistical methods

Summary

Each trial was analysed separately. For the primary endpoint (all-cause mortality in both trials) and the composite endpoint (all-cause mortality or all-cause hospitalisation in COMET and worsening HF culminating in death or hospitalisation in DIG), each trial was calibrated to the 8,012 patients in the HF registry, first using a regression-based method (Figure 7) and then using inverse odds of sampling weights (IOSW, Figure 9). All analyses were conducted in R (R 3.4.0 for trial and R 3.5 for the registry). The parametric survival models were fitted using the “flexsurv” package(17) and the weighted logistic regression models were fitted using the “survey” package(18). Each method is described below, with detailed steps and selected R code provided in the supplementary appendix.

Regression-based calibration

Method description

The overview of this method is showed in Figure 7. Variables were included as model covariates based on their availability in both the trial and registry and their potential to modify treatment effectiveness (as Table 7). A model based on the trial data was first constructed. Parametric survival models of the primary and composite outcome on the treatment effect and these covariates were fitted using a range of distributions ("Weibull", "Generalised gamma", "Exponential", "Log-logistic", "Log-normal", "Gompertz"). For subsequent analyses I selected the distribution which had the best fit based on visual inspection of diagnostic plots and the Akaike Information Criterion (AIC). For all covariates, main effects and 2-way interactions with the treatment variable were included in the final model. Where there was evidence of non-linearity for continuous covariates, the covariates were transformed (SBP, estimated glomerular filtration rate (eGFR) in DIG).

A regression model using the same distribution for the outcome variables, covariates and transformations was fitted to the HF registry, except that treatment main effects and interaction were not included (since the treatment effects are estimated solely using the trial data). This was the registry model.

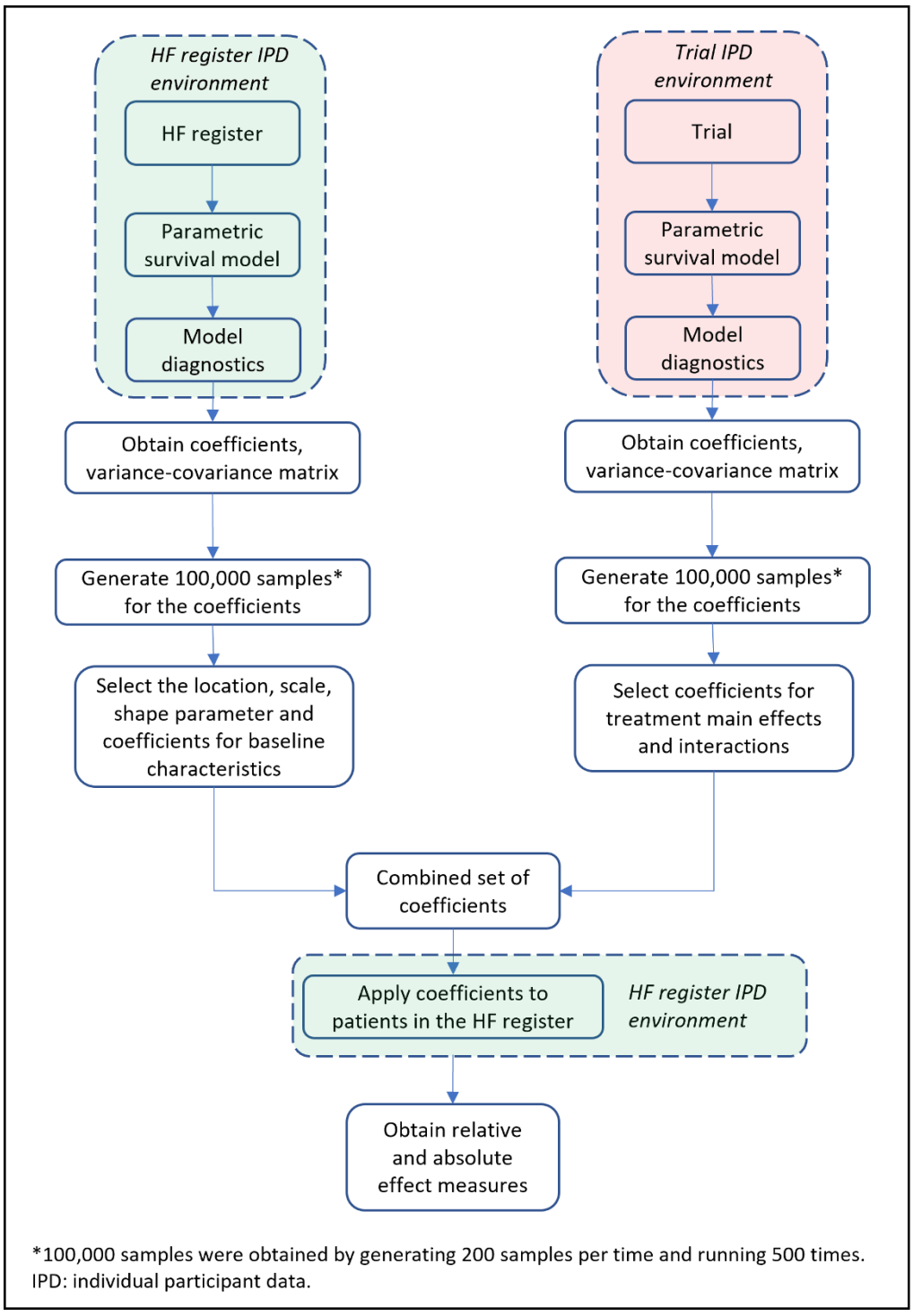
Coefficients from the registry and trial models were applied to patients in the HF registry to estimate the predicted rate of the primary/composite outcome, first assuming that patients in the registry received the trial intervention and then assuming instead that they received the comparator. For this estimation, the coefficients for the covariate main effects (e.g., age, sex) were obtained from the registry model, and the coefficients for treatment effects and treatment covariate interactions (e.g., treatment + age*treatment) were obtained from the trial model. The predicted outcome under each intervention was then summed across individuals in the registry and then compared for the trial and comparator interventions to obtain relative and absolute effect measures.

For this analysis, uncertainty in the coefficients was propagated to the final model via simulation - I obtained 100,000 samples from both the trial and registry models. I then

sampled from multivariate normal distributions where the means and variance-covariance matrices corresponded to the coefficient point estimates and variance-covariance matrices from the relevant models. The correlation between coefficients in the trial and registry model was assumed to be zero. The outcome predictions and treatment effects were calculated for each sample and summarised via the mean (geometric mean for relative measures) with the uncertainty expressed via the 2.5th and 97.5th percentiles.

Detailed implementation of the regression-based transportation with selected R code are described in the supplementary appendix.

Figure 7. Overview of the regression-based calibration.



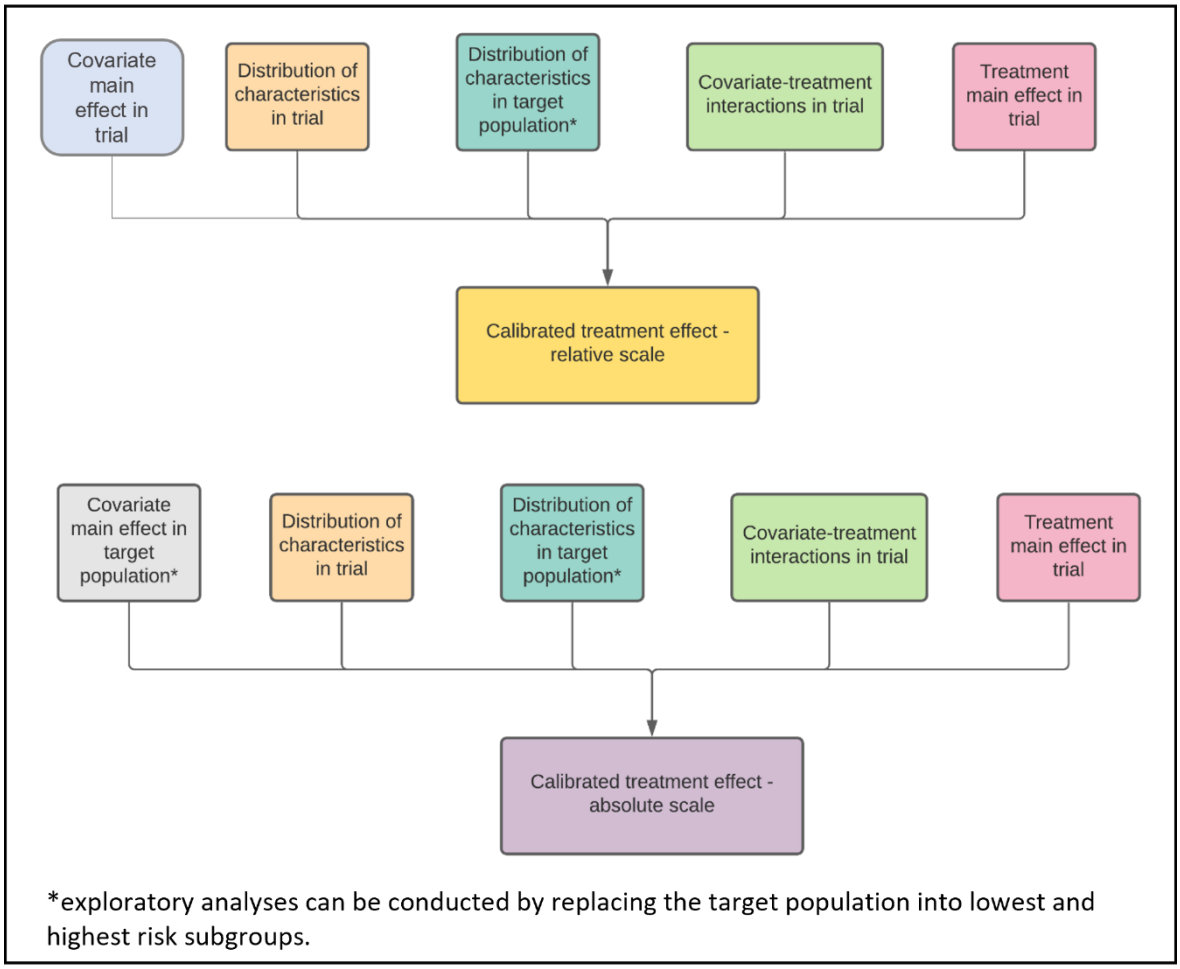
*100,000 samples were obtained by generating 200 samples per time and running 500 times.
IPD: individual participant data.

The principle of regression-based method

The principle of this method is as Figure 8 below. Briefly, the differences of the baseline characteristics between the trial participants and target population had the potential to influence the outcomes. By accounting for these differences, combining treatment-covariate interactions and the treatment main effect from the trial, the

relative scale of the calibrated treatment effect can be determined. To estimate the absolute scale of the calibrated treatment effect, I also consider the covariate main effect in the target population. Different defined target populations might lead to distinct absolute scales of the calibrated treatment effect, such as employing a newly defined sub-population from the registry for exploratory analyses.

Figure 8. The principle of regression-based method.



Inverse Odds of Sampling Weights (IOSW).

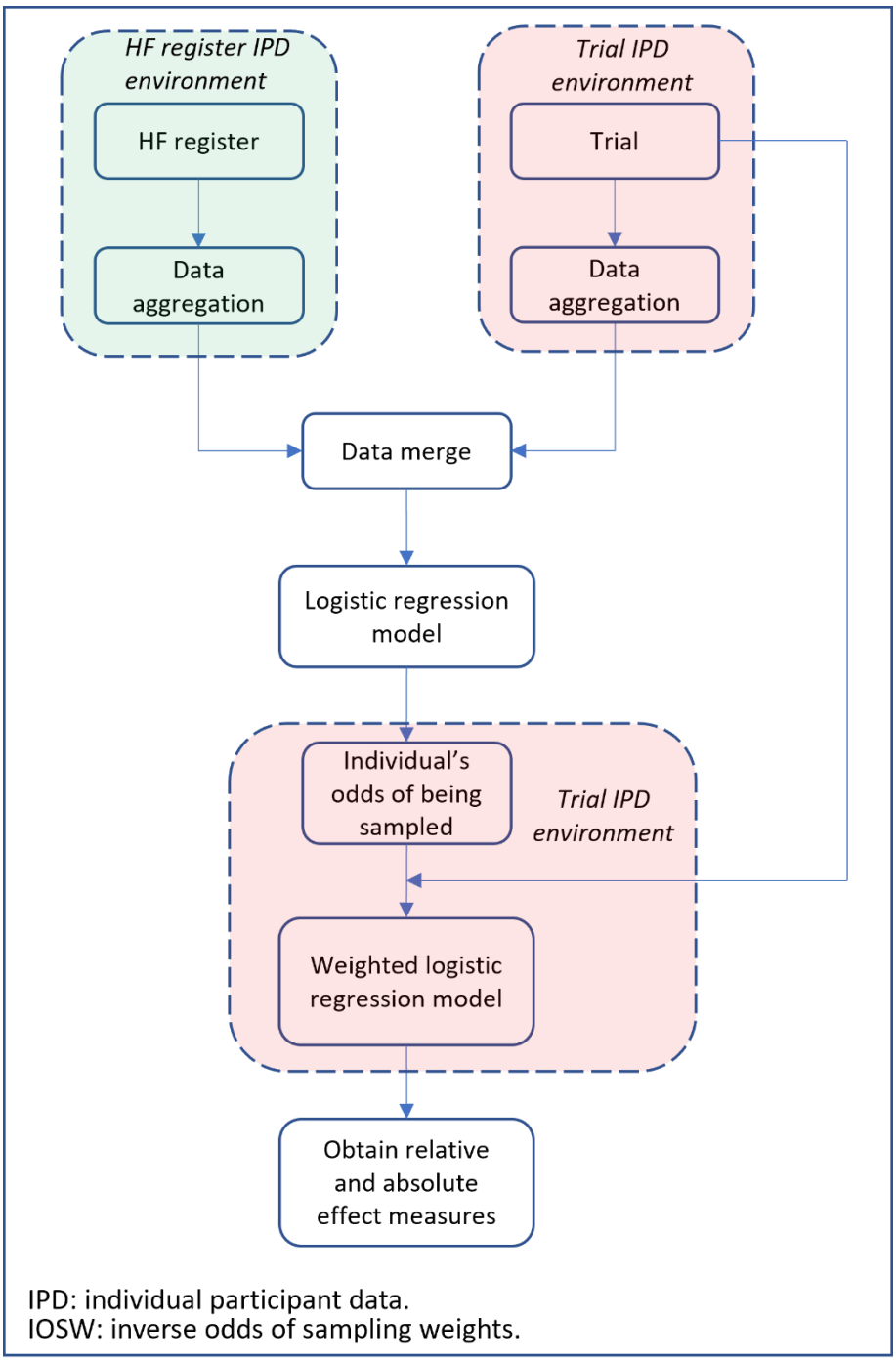
Method description

The overview of this calibration is showed in Figure 9. Briefly, using the same covariates as regression-based method, the trial and registry datasets were aggregated to obtain counts of individuals with each combination of characteristics (Supplemental II Table S4). The probability that an individual from the HF registry is included in the trial

sample, conditional on covariates, divided by the probability of not being in the trial sample - the inclusion odds - was then estimated by comparing these counts.

I then estimated the treatment effects as standard by comparing the odds of the outcome in each treatment arm; except that instead of all participants having the same weight in the analysis, different participants were weighted differently according to their inclusion odds. As an example, if there were 500 individuals with a given set of characteristics in the registry, and 5 participants with that set of characteristics in the trial, the odds of being sampled would be 1% (5/500). This would translate to a raw weighting of 100 (1/odds) for those 5 participants. Final weights for all participants would then be calculated by dividing each participant's weight by the sum of weights for all participants. See supplementary appendix for details on the methods used to calculate the inclusion odds and weightings, and to estimate the treatment effects using the weightings.

Figure 9. Overview of Inverse odds of sampling weights calibration.



Exploratory analyses

Exploratory data analyses are usually utilised to detect mistakes and check the assumptions. It is also used to discover patterns and select the suitable model preliminarily and explore the relationships among predictor variables. Moreover, it can assess the direction and rough size of the associations between the predictor variables

and the endpoints(154). In this chapter, exploratory analyses were conducted for both regression-based and IOSW method.

In additional analyses, I used the natural history model fitted to the registry data to estimate the risk of the covariates (e.g. age, male, SBP) and outcomes (all-cause death et al corresponding with trial outcomes) to estimate the predicted risk for each individual in the register, ranked these, then selected the top 10 percentile highest and top 10 percentile lowest into the highest and lowest risk subgroups (801 patients in each) respectively. two calibration methods were used following the above analyses respectively to calculate the measure of effects again(83, 155). They can also be regarded as the subgroup analyses as the sub-population was chosen and multiple characteristics were considered at the same time.

Assumptions for two calibration methods.

Table 5. Assumptions for two calibration methods.

	Regression-based method	Inverse Odds of Sampling Weights (IOSW)
Assumptions common to both methods for relative treatment effects	• Treatment assignment is random and independent of sample selection. • There are no unmeasured covariates that are related to both trial inclusion and that are treatment effect modifiers. (Note in the *standard* approach of applying relative treatment effect estimates to target populations, the assumption is that there are no (i.e., not just no unmeasured) covariates that are related to both trial inclusion and that are treatment effect modifiers).	
Different assumptions across methods for relative treatment effects	• For the trial model the treatment effect estimates are correctly modelled including all effect modifiers. This means all interactions that are present are	• The trial inclusion logistic regression model includes all characteristics that both 1) differ between trial sample and target population and 2)

	included in the final model and that any departures from linearity are correctly modelled.	demonstrate heterogeneity in the treatment effect.
Assumptions required for absolute effect estimates	• For the Registry model the associations are correctly modelled including all variables which predict the outcome of interest (including all relevant interaction terms and correctly accounting for any departures from linearity). • The right censoring is assumed to be non-informative conditional on the covariates.	• Conditional on the characteristics included in the trial inclusion model, the risk of the outcomes are the same in the trial and registry populations. Note that this could be relaxed by applying the calibrated relative effect estimate to the event rate in the registry. • The right censoring is assumed to be non-informative by assigned treatment.

4.5 Results

4.5.1 Baseline characteristics

4.5.1.1 Missing values

The percentage of missing values either in the register or trials is showed as Table 6 below. There are few missingness in two trials (no more than 0.26%) while in the

register, the percentage is much higher (24.35% for eGFR), therefore imputation was conducted for the HF register only.

Table 6. Percentage of missingness in the HF register and two trials.

	HF register (N=8,012)	COMET (N=3,029)	DIG (N=7,788)
Age (years)	0	0.10%	0
Systolic blood pressure (mm Hg)	13.90%	0.10%	0.04%
Heart rate (beats per minute)	11.80%	0.10%	0.10%
Sodium (mmol/l)	24.10%	0.07%	--
eGFR (mL/min/1.73m2)	24.35%	0.10%	0
Frusemide (mg)	17.40%	0.10%	--
*Male, n (%)	0	0.10%	0
*History of diabetes, n (%)	12.40%	0.26%	0.01%
*Categorical variables, others are and continuous variables			

4.5.1.2 Baseline characteristics in the registry and trials

Patients in the HF register were more elderly (mean (sd): 73 (12) vs 62 (11) and 64 (11) years in COMET and DIG respectively), had a slightly lower SBP (120 (21) vs 126 (19) and 126 (20) mmHg), and eGFR levels (59 (23) vs 67 (21) and 62 (21) mL/min/1.73m^2), and a much higher loop diuretics dosage (62 (31) vs 20 (46) mg/day in COMET) than trial participants. There were more men than women in the register (61% men vs 39% women), COMET (80% vs 20%) and DIG (78% vs 22%). Patients with diabetes history in the register were slightly fewer than those in the trials (23% vs 24% and 28%).

Table 7. Baseline characteristics in each dataset included in calibrations.

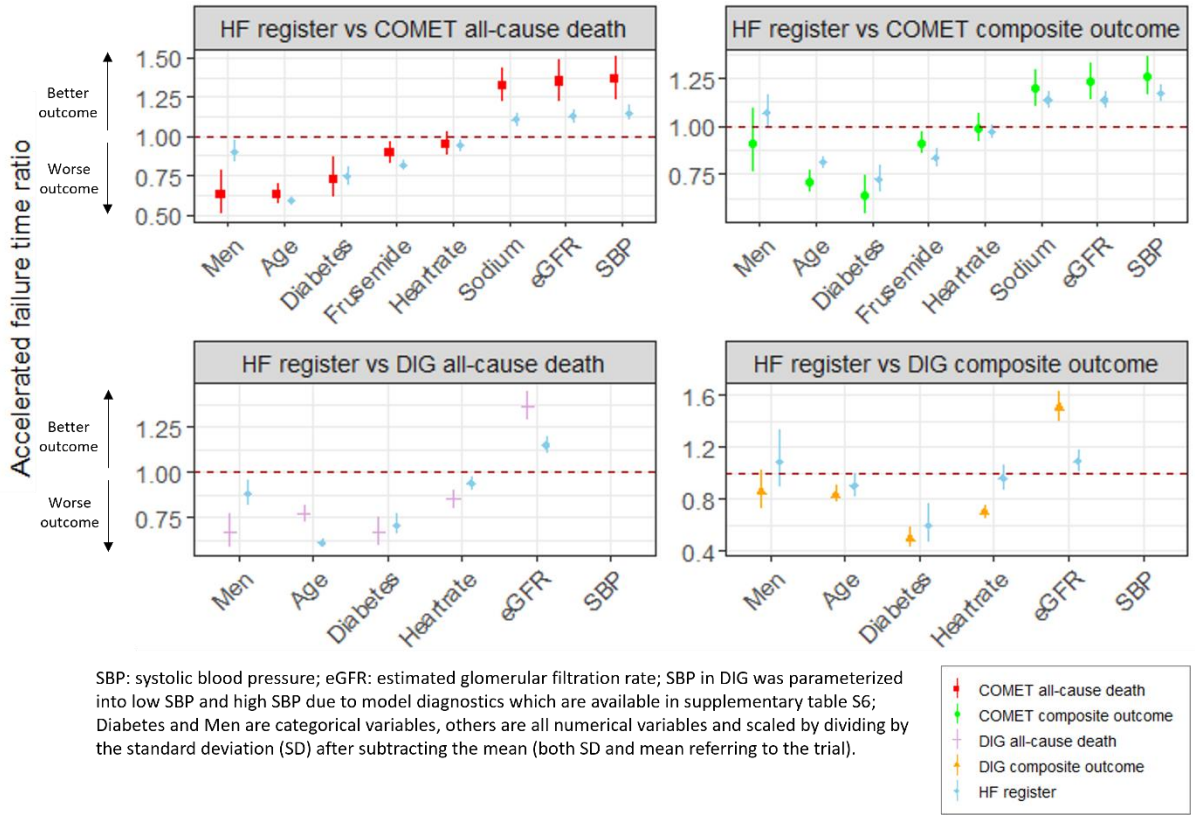
	HF registry (N=8,012)	COMET (N=3,029)	DIG (N=6,800)
Age (years)	73 (12)	62 (11)	64 (11)
Men, n (%)	4906 (61%)	2412 (80%)	5281 (78%)
History of diabetes, n (%)	1863 (23%)	728 (24%)	1933 (28%)
Heart rate (beats per minute)	73 (13)	81 (13)	79 (13)
Systolic blood pressure (mm Hg)	120 (21)	126 (19)	126 (20)
Serum sodium (mmol/l)	138 (4)	140 (3)	--
eGFR (mL/min/1.73m^2)	59 (23)	67 (21)	62 (21)
Loop diuretics (mg/day)	62 (31)	20 (46)	*
Categorical variables are shown as counts (%s) and continuous variables as means (standard deviations); -- not available; *In DIG loop diuretics was recorded as a categorical variable (whether participants had taken it or not or unknown) and the dosage information was not available.			

4.5.2 Effect of baseline characteristics on outcomes

The parametric survival model with a generalised gamma distribution had the best fit and was used for the HF registry and each trial. The coefficient for the covariates for the HF registry and trials are shown in Figure 10 and Supplementary II Table S5. These coefficients are mutually adjusted. In both trials, male sex, older age, history of diabetes predicted a worse prognosis. The accelerated failure time (AFT) ratios with 95% CIs for male sex were 0.65 (0.53, 0.81), 0.91 (0.76, 1.09), 0.67 (0.58, 0.77), and 0.86 (0.72, 1.02) for COMET all-cause death, composite outcome, DIG all-cause death, and composite outcome, respectively. For age, the AFT ratios were 0.63 (0.57, 0.70), 0.71 (0.65, 0.77), 0.77 (0.72, 0.82), and 0.84 (0.77, 0.91) respectively. For a history of

diabetes, the AFT ratios were 0.73 (0.65, 0.87), 0.64 (0.54, 0.75), 0.67 (0.59, 0.76), and 0.50 (0.43, 0.58) respectively. Higher eGFR predicted longer survival in both trials (AFT ratio: 1.35 (1.22, 1.49), 1.23 (1.13, 1.34), 1.37 (1.29, 1.46) and 1.50 (1.39, 1.63) respectively). In COMET, the use of higher dose loop diuretics also predicted a worse outcome (AFT ratio: 0.9 (0.84, 0.96) in all-cause death and 0.91 (0.85, 0.97) in the composite outcome). Conversely, higher serum sodium concentration (AFT ratio: 1.32 (1.22, 1.43) and 1.20 (1.11, 1.29)) and higher SBP (AFT ratio: 1.37 (1.25, 1.49) and 1.26 (1.17, 1.36)) predicted a better prognosis.

Figure 10. Main effects in HF registry and two trials.



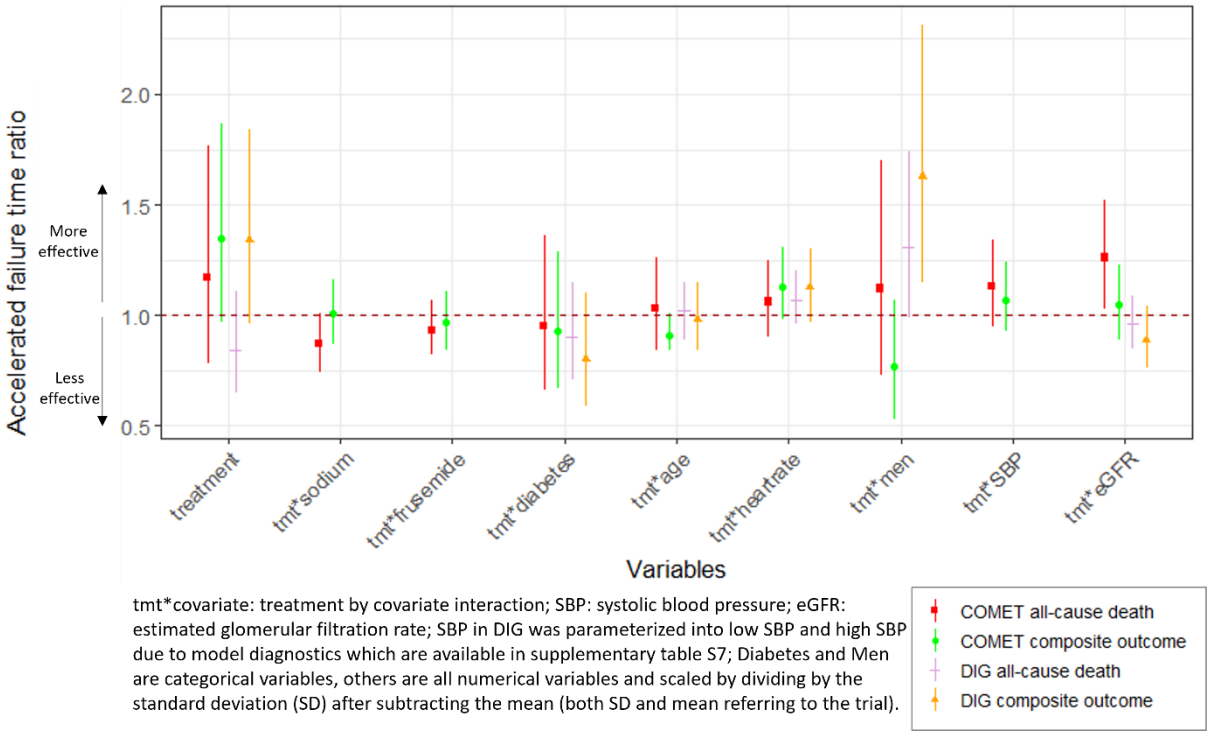
SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate; SBP in DIG was parameterized into low SBP and high SBP due to model diagnostics which are available in supplementary table S6; Diabetes and Men are categorical variables, others are all numerical variables and scaled by dividing by the standard deviation (SD) after subtracting the mean (both SD and mean referring to the trial).

4.5.3 Effect of baseline characteristics on treatment efficacy

The estimates for the treatment effects (at the mean of all the covariate levels) and the treatment-covariate interactions are shown in Figure 11 and Supplementary II Table S6. The treatment-covariate interaction estimates were wide, and for some variables the magnitude and direction of the point estimates varied between trials. For both COMET and DIG, treatment efficacy appeared to be lower for patients with diabetes (accelerated failure time (AFT) ratio: 0.95 (0.66, 1.37) and 0.93 (0.67, 1.29) for COMET all-cause death and composite outcome; 0.90 (0.71, 1.15) and 0.81 (0.59, 1.10) for DIG

all-cause death and composite outcome) and greater for heartrate (AFT ratio: 1.06 (0.90, 1.25) and 1.13 (0.98, 1.31) for COMET and 1.07 (0.96, 1.20) and 1.12 (0.97, 1.30) for DIG), but the CIs almost all included the null.

Figure 11. Treatment and treatment-covariate interactions in two trials.

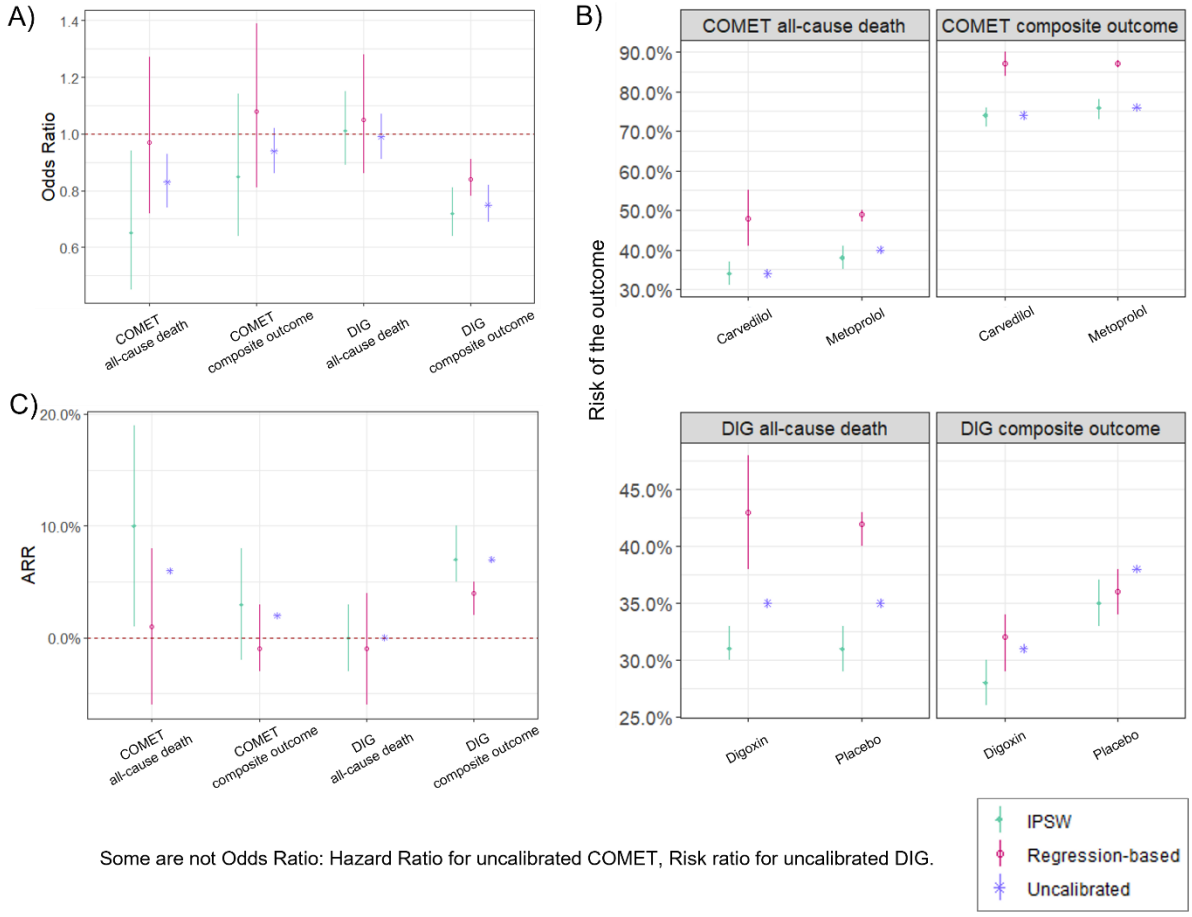


tmt*covariate: treatment by covariate interaction; SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate; SBP in DIG was parameterized into low SBP and high SBP due to model diagnostics which are available in supplementary table S7; Diabetes and Men are categorical variables, others are all numerical variables and scaled by dividing by the standard deviation (SD) after subtracting the mean (both SD and mean referring to the trial).

4.5.4 Effect of transportation on treatment effects

Figure 12 and Supplementary II Table S9 show the calibrated treatment effects. For either primary or composite outcome in DIG over a period of 3 years, the uncalibrated and calibrated effect estimates (odds ratios, ORs) were similar (OR: 0.99 (0.91, 1.07) vs 1.06 (0.92, 1.21) vs 1.05 (0.86, 1.28) for uncalibrated analysis, IOSW and regression-based transportation for all-cause death and 0.75 (0.69, 0.82) vs 0.73 (0.64, 0.83) vs 0.84 (0.78, 0.91) for the composite outcome), indicating similar efficacy in the trial and HF registry. For COMET the efficacy was higher for IOSW (OR: 0.62 (0.39, 0.99) and 0.87 (0.59, 1.30) for all cause death and composite outcome over a period of 4 years) but lower for the regression-based transportation (0.97 (0.72, 1.27) and 1.08 (0.81, 1.39)) although the 95% CIs overlapped those of the uncalibrated estimates (0.83 (0.74, 0.93) and 0.94 (0.86, 1.02)). The impact of transportation was similar where the trials were calibrated to the high-risk and low-risk subgroups (Supplementary II Table S11 and Fig 4).

Figure 12. Measure of effects in uncalibrated and calibrated analyses in two trials. A) Odds Ratio; B) Risk of the outcome; C) Absolute Risk Reduction (ARR).



Where differences exist between calibrated and uncalibrated results, the influence of each covariate on this divergence can be estimated as the covariate-treatment interaction multiplied by the mean difference in the covariate between the registry and trial populations. For instance, in the case of COMET death, the standardised mean for loop diuretics in the registry and trial are 0.9 and 0 respectively, and the treatment-loop diuretics interaction is -0.07. The contribution to the discrepancy is thus calculated as -0.07 * 0.9 = -0.06. Consequently, for COMET (death), eGFR and loop diuretics dose were the primary influencers; for COMET (composite), age and heart rate were the key influencers; and for DIG, male sex and heart rate were the main contributors (see Supplementary II Table S7).

Compared to the uncalibrated and IOSW models, the estimated risk of the outcome within each treatment arm (except for carvedilol arm in COMET all-cause death) was larger for the regression-based model (Figure 12B), e.g., the estimated mortality in the digoxin arm of DIG was 35%, 33% and 43% in the uncalibrated, IOSW and regression-

based analysis respectively. However, these differences of risk in each treatment arm did not translate to large differences in the absolute risk reductions (ARRs, see Figure 12C).

4.5.5 Effect of calibration on precision of treatment efficacy

The ratio of the standard error (SE) from calibrated and uncalibrated analysis was used to denote the precision of treatment efficacy as showed in Table 9. As expected, compared to the standard analysis, the standard errors (SEs) were generally larger for the calibrated effect estimates (Table 8). For transportation to the overall target population, this ranged from no increase to 4.6-fold wider SEs (e.g., SEs are 0.06, 0.15 and 0.24 for uncalibrated analysis, regression-based and IOSW transportation for COMET all-cause mortality). Where the results were calibrated to the highest and lowest risk subgroups of the registry, which by design were more different from the trial populations based on baseline characteristics than was the overall population, the SEs ranged from 1.6-fold to 12.7-fold wider.

4.5.6 Influence of highest weights on the precision of treatment effects

In additional analyses (Supplementary II Table S11), the exclusion of individuals with lowest 1% odds (largest 1% weights) in the IOSW transportation slightly changed the point estimates, SEs and narrowed the CI. After excluding the 1% patients with the lowest odds (highest weights) of trial inclusion the SE ranged from no increase to 2.6-fold wider for overall target population, and it ranged from 1.2-fold to 5-fold wider for highest and lowest risk subgroups ((e.g., SEs are 0.06, 0.15 and 0.11 for uncalibrated analysis, regression-based and IOSW transportation for COMET all-cause mortality)).

This 1% extreme large weights (low odds of inclusion) were characterised by older age, higher loop diuretics doses and lower eGFR (Figure 13).

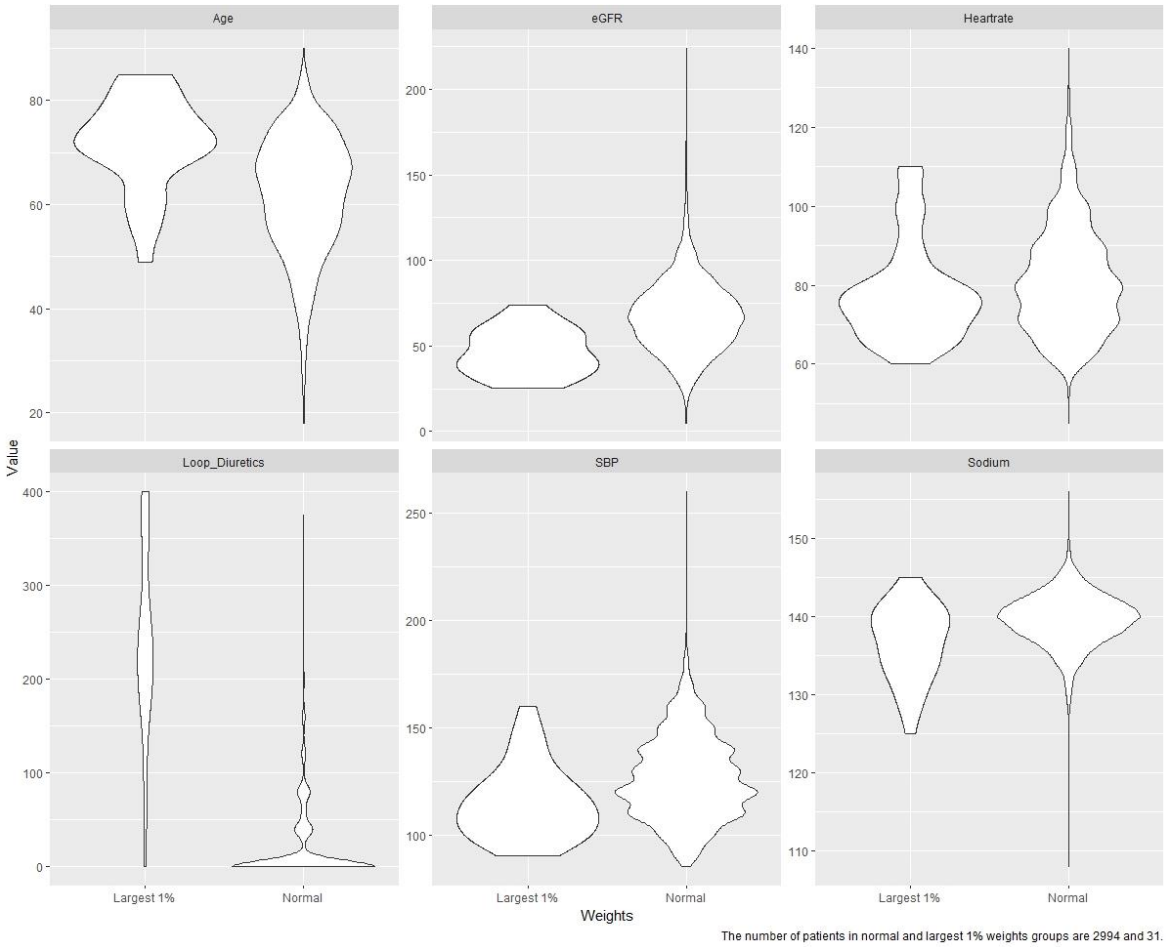
Table 8. Precision of estimates in uncalibrated and calibrated analyses.

	Uncalibrated$_{(U)}$	Regression-based$_{(R)}$	IOSW$_{(IA)}$ (all included)	IOSW$_{(IT)}$ (trimming the largest 1% of weights)
	SE$_{(U)}$	SE$_{(R)}$ (SE$_{(R)}$/SE$_{(U)}$)	SE$_{(IA)}$ (SE$_{(IA)}$/SE$_{(U)}$)	SE$_{(IT)}$ (SE$_{(IT)}$/SE$_{(U)}$)
Overall				
COMET all-cause death	0.058	0.145 (2.500)	0.240 (4.138)	0.111 (1.914)
COMET all-cause death or hospitalisation	0.044	0.137 (3.114)	0.203 (4.614)	0.113 (2.568)
DIG all-cause death	0.053	0.102 (1.925)	0.069 (1.302)	0.066 (1.245)
DIG death or hospitalisation due to worsening heart failure	0.053	0.038 (0.717)	0.069 (1.302)	0.066 (1.245)
Low risk decile				
COMET all-cause death	0.058	0.165 (2.845)	0.165 (2.845)	0.119 (2.052)
COMET all-cause death or hospitalisation	0.044	0.121 (2.750)	0.129 (2.932)	0.1 (2.273)
DIG all-cause death	0.053	0.106 (2.000)	0.105 (1.981)	0.088 (1.660)
DIG death or hospitalisation due to worsening heart failure	0.053	0.061 (1.151)	0.086 (1.623)	0.076 (1.434)
High risk decile				
COMET all-cause death	0.058	0.338 (5.828)	0.739 (12.741)	0.234 (4.034)
COMET all-cause death or hospitalisation	0.044	0.344 (7.818)	0.296 (6.727)	0.222 (5.045)
DIG all-cause death	0.053	0.197 (3.717)	0.248 (4.679)	0.134 (2.538)
DIG death or hospitalisation due to worsening heart failure	0.053	0.076 (1.434)	0.171 (3.226)	0.114 (2.151)

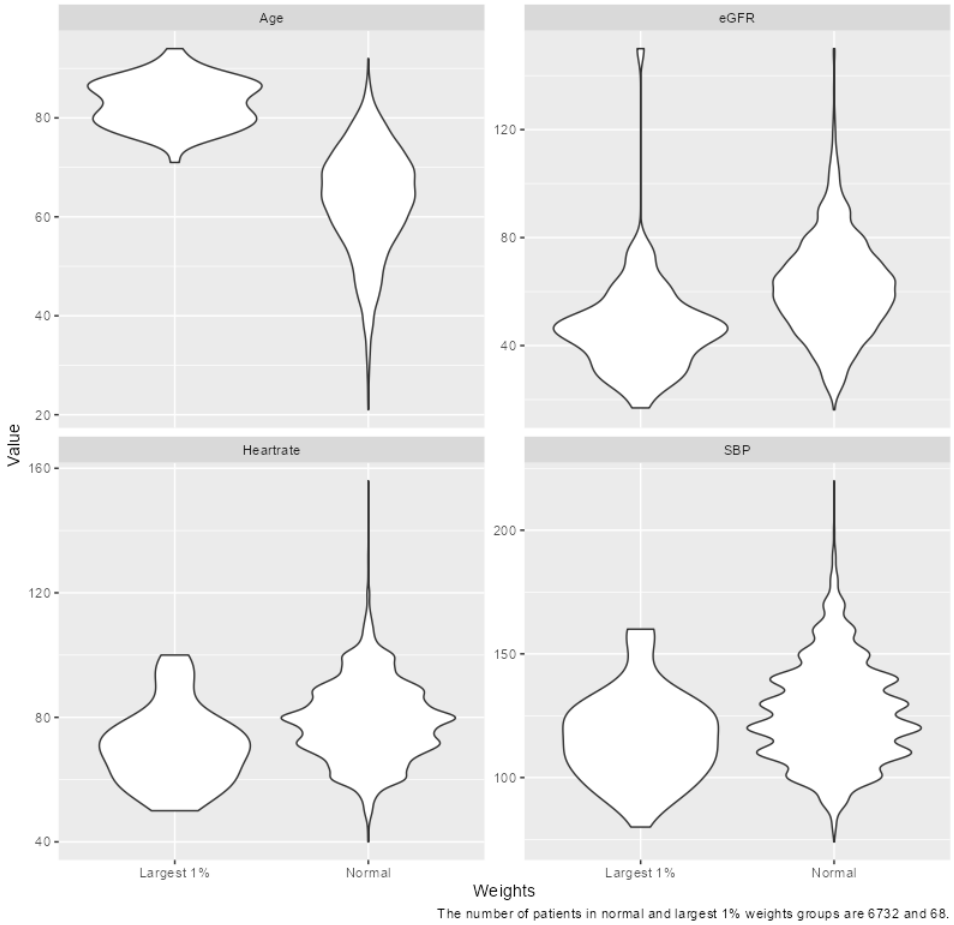
SE: standard error; $_{(U)}$: Uncalibrated analysis; $_{(R)}$: Regression-based method; $_{(IA)}$: Inverse Odds of Sampling Weights including all patients; $_{(IT)}$: Inverse Odds of Sampling Weights trimming the largest 1% of weights; SEs and ratios of SEs are in 3 decimal places.

Figure 13. Comparison of distribution of individual characteristics between patients with normal and extreme large 1% weights in 1) COMET and 2) DIG.

1)



2)



4.6 Discussion

4.6.1 Summary

Two landmark HF trials were calibrated to a Scottish "real-world" population using two approaches, regression-based and IPSW. Both were straightforward to perform, with only moderate loss of precision manifested as larger SEs. This suggests that trials can be calibrated to registry data, maximising representativeness and applicability while preserving the benefits of randomisation.

4.6.2 Previous literature and what this study adds

Previous studies have employed calibration using IOSW or generalisation via inverse probability of sampling weights (IPSW)(141-143). In DAPT study, IOSW was used to

account for patient and procedural factors between trial and the registry and estimated the real-world treatment effect comparing 30 months to 12 months of DAPT after coronary stent procedures. The weighted analyses no longer showed a significant effect of prolonged DAPT on reducing stent thrombosis, major adverse cardiac and cerebrovascular events, or myocardial infarction but the increase in bleeding persisted(143). In the PARADIGM-HF trial, IPSW was used to re-analyse the treatment effects accounting for participants excluded during the run-in period by adding additional weight for participants completing run-in most closely resembling those excluded. It showed that the weighted analysis of key outcomes taking into account drop-outs during the run-in phase did not change the benefit of sacubitril/valsartan compared to enalapril (156). Cole and Stewart used IPSW to calibrate a major HIV trial, using counts of people with HIV in the US stratified by age, sex and CD4 count to define the target population(141). The GetReal project calibrated a trial of chemotherapy for non-small cell lung cancer to a cohort study using IPSW with 15 baseline characteristics and IPSW showed a similar hazard ratio for pemetrexed compared with gemcitabine with greater uncertainty (a wider CI)(142). I add to this literature by showing that HF trials can be calibrated to the more complex populations encountered in clinical practice with only moderate loss in precision, yielding similar results for both IOSW and a regression-based approach. Furthermore, HF trials can be calibrated to different risk subgroups based on multiple characteristics. Unlike conventional subgroup analyses this approach simultaneously accounts for the impact of all measured characteristics which differ between the trial and real-world settings.

In these analyses the calibration was performed to improve transportability rather than generalisability. When re-weighting for generalisability, the technique is identical, except that the inverse of the probability of trial inclusion is used rather than the inverse odds.

4.6.3 Assumptions

Both IOSW and regression make assumptions (Table 5). It is essential that the main effects and interactions are correctly modelled in the regression-based approach, and that all variables that predict both heterogeneity in participation and the outcome have been included in the trial inclusion odds model for the IOSW-approach. For both approaches, I also assume that there are no treatment-covariate interactions for unmeasured variables; although it is worth noting that the current standard approach of

applying the relative treatment effect from trials to target populations makes the more extreme assumption that there are no covariate-treatment interactions of any kind (measured or unmeasured). Importantly, while transportation helps address under-representation, caution is needed when extrapolating trial results to patients who could not have been included in the trial (thus violating the positivity assumption), in this case for example children or people living in Africa. From a purely technical point of view, there are differences between the different approaches when extrapolating the trial findings to patients with combinations of characteristics beyond the range of the trial data. Using the IOSW approach, it is technically impossible to re-weight the estimates for levels of characteristics beyond the range of the trial data (eg if no trial participants were aged over 65 years one cannot estimate relative effects in a population over the age of 65). In contrast, in the regression-based approach, so where covariates are modelled as continuous variables (eg linear terms, polynomials etc) it is technically straightforward to extrapolate beyond the data. Nonetheless, whether applying regression or IOSW it is important to consider whether the applicability of the predicted effect estimates, on the required scale, are genuinely transportable to the desired target population. In other words, whether the relevant assumptions are met. Furthermore, participant/patient characteristics are only one way in which the circumstances of the trial may differ from the target population, for example there may be differences in clinical settings or time periods of enrolment. Differences in diagnosis, treatment delivery and monitoring may lead to differential efficacy (eg due to improved adherence, better tailoring of dosages etc)(157).These also need to be carefully considered when assessing the transportability of effect estimates, and are generally less amenable to the kind of adjustments described in this chapter.

Differences in the assumptions of IOSW and regression approaches alone, provides justification for performing both. However, they also provide different information. For example, the IOSW approach involves calculating the trial inclusion odds, and this then provides an overall single summary measure for all trial participants and registry patients. This allows comparisons within and between these populations, in order to determine, for example, whether the trial and registry populations are sufficiently similar to undertake calibration. This is analogous to an advantage of propensity score weighting in pharmacoepidemiologic analyses (e.g., control for measured confounding, identify barriers for treatment such as age) (158). In contrast, an advantage of the regression approach is that I can explore which differences between trial participants and registry patients are driving any observed discrepancies between calibrated and uncalibrated treatment effect estimates. This can be done by examining the magnitude

of covariate-treatment interactions and comparing levels of these covariates between trial participants and registry patients.

The regression-based calibration approach builds on the standard evidence synthesis modelling process for producing absolute treatment effects in a target population, recommended in NICE Technical Support Document 5(159), wherein:- i) a standard care model for absolute outcomes is fitted to data representative of the target population, ii) a relative treatment effect model is fitted to trial data and iii) the two models are combined (usually using Monte Carlo methods or bootstrapping) to estimate absolute treatment effects. My model differs in two ways. First, homogeneity of relative treatment effects is not assumed but allow these to differ according to individual participant characteristics. Secondly, rather than having a single estimate for the natural history model or having two or more estimates stratified by some important characteristics (e.g. disease severity), the rates to differ according to individual patient characteristics are allowed. Importantly, this approach works on the assumption that relative treatment effects are transportable between trial and target populations conditional on the covariates included in the relative effects model in the trial data, with the standard care model fitted in the target population providing the baseline absolute rates to which the transported relative effects are applied. This is in contrast to alternative standardisation/g-computation approaches (e.g. as described by Dahabreh et al(160).) which solely use the trial data-derived model to produce absolute predictions in the target population (i.e. the standard care model is estimated within the trial), and thus are based on the assumption that absolute effects are transportable between trial and target populations. This is a much more stringent assumption to meet, since differences in all prognostic factors and effect modifiers between trial and target population must be accounted for instead of just the effect modifiers and is generally considered far less plausible. When non-collapsible relative effects measures are used (e.g. odds ratios or hazard ratios), we must additionally take care to ensure that the parameters from the standard care model are compatible with the parameters from the relative treatment effects model; that is, that they are conditioned in the same manner. This is not necessarily true in this analysis as some of the individuals in the register were taking digoxin and/or carvedilol. However, in many applications where a standard care population can be readily defined (e.g. because a new treatment is being considered), this condition is likely to be true; because the standard care population is restricted to (i.e. conditioned on) a common standard treatment and so the parameters of the standard care model have the same interpretation as their counterparts in the relative treatment effects model. This condition is trivially met by

alternative standardisation/g-computation approaches that use only the trial data to produce absolute predictions from one single model, although as noted above these approaches make much stronger assumptions to transport absolute effects. Since they exhibit different assumptions, some researchers may wish to explore the use of both approaches as a triangulation exercise.

4.6.4 Doubly robust estimation

Doubly robust estimation combines an outcome regression form with an exposure model (e.g., propensity score) to estimate the causal estimate of an exposure on an outcome(161). When used independently, both outcome regression and propensity score approaches are unbiased only when the statistical model is correctly specified. The doubly robust estimator combines these two approaches, ensuring that the effect estimator remains robust even when one (but not both) of these models is mis-specified. This implies that only one of the two models needs to be correctly specified to obtain an unbiased effect estimator(161). I focused on comparing two methods of calibration (regression-based and IOSW-based). However, it is also possible to combine both using what are termed doubly robust approaches where both regression and inverse-weighting are used together. See Li et al for an example(162).

4.6.5 Influence of highest weights on the precision of treatment effects

Butala et al suggested to trim the extremely large weights which may be caused by small sample size to ensure stable estimates(143). This can be achieved by truncating the top weights (such as 1%) or normalising weights. In both the main and exploratory analyses (Supplementary II Table S11), the largest 1% weights were truncated. After truncation, for the overall target population, the standard errors were reduced to 2.6-fold wider from 4.6-fold wider before truncation. For the highest and lowest risk subgroups of the registry, the standard errors ranged were reduced to 5-fold wider from 12.7-fold wider, resulting in increased precision with the exclusion of just 1% extreme large weights.

4.6.6 Challenges and implications

A challenge of calibration is the need to access IPD for both the trial and target populations. This is complex (e.g., data sharing agreements and regulatory approvals) and requires considerable analyst time. However, as illustrated in Figure 7, changes in trial reporting could improve this situation. Were trialists to provide the coefficients and the variance-covariance matrix for a treatment effect model including all non-negligible treatment-covariate interactions, secondary researchers (with access to registry IPD) could produce calibrated treatment effect estimates for specific target settings. To enable such an approach would also require trialists to select the relevant covariates and to correctly specify the treatment covariate analysis. To be widely practiced, it would likely also require consensus among trialists and guidance from regulatory agencies. Similarly, it may also be possible in the future for estimates to be produced by trialists if those managing disease registries (such as NHSGGC) were able to provide adequate summary data to reconstruct the joint distribution of patient characteristics. I illustrate some of the information that would be need in Supplementary I Table S12 for HF clinical trials (age, sex, SBP etc). As has previously been shown, joint covariate distributions may be reconstructed from routinely collected data given published marginal summary statistics (eg means, standard deviations) and correlation matrices if we are willing to make assumptions about the functional form of the marginal distributions and the correlation structure, for example by using a multivariate normal a copula to capture the correlation structure(72). Moreover, simulation studies have shown that the results are likely to be robust to the assumptions used to reconstruct the joint distribution(163). However, for such an approach to be adopted, additional methodological work is first needed however in order to i) reassure those holding routinely collected data that the risk of re-identifying individuals is sufficiently low and ii) reassure analysts that this parametric summary of the data is generally adequate for trial calibration. For the widespread adoption of transportation, the reporting of such summaries would need to be standardised(164). Clinical trials are already highly standardised and sophisticated with mature ontologies and reporting standards (165). These would need to be expanded to cover reporting of treatment-covariate interactions from multivariable models. Current proposals to standardise and harmonise HF registries would also need to incorporate reporting standards for population summaries. Considerable efforts by the HF research community would be required to implement such changes in both trial and registry settings. This observation that calibration yielded more applicable estimates with only a moderate loss of precision suggests that this effort is worthwhile. After incorporating patient

characteristics from both the trial and the real world using two statistical calibration methods, it shows similar point estimates, albeit with moderately wider confidence intervals when considering the enrolment of real-world patients receiving the same interventions as in the trial. 'Calibration yielded more applicable estimates' implies greater trustworthiness due to the calibration process being performed.

Not every patient encountered in real-world clinical practice would be enrolled in the trial, but clinicians need to treat these real-world patients in their daily practice. Additionally, policymakers require more evidence to create guidelines. Pragmatic trials tend to enrol more representative samples from the real world, but the increased complexity and resource requirements pose challenges in implementation in terms of time, cost, and logistical considerations(13). Statistical calibration can yield trustworthy real-world estimates without the need for implementing new trials, thereby providing more evidence without incurring additional trial-related costs. In comparison to the efforts involved in initiating a new trial to inform treatments for real-world patients, performing calibration is a more straightforward and sustainable approach.

Another challenge in analysing IPD pertains to working within secure data storage platforms (Safe Havens). In terms of data protection, Safe Havens offer a highly secure environment characterized by restricted access and meticulous management of output exports. However, this level of security can also introduce complexities when it comes to tasks like downloading packages, conducting data analyses, transferring coefficients, and exporting results. This challenge is particularly pronounced when dealing with datasets stored in different Safe Havens, as is the case with the HF registry data and trial data.

The process of transferring coefficients or aggregated data from one Safe Haven to another involves multiple steps and approvals. This necessitates the review of contents to be transferred, seeking necessary approvals, and eventually executing the file transfer. These procedural intricacies can pose significant challenges and may require a considerable amount of time and effort to navigate effectively.

4.6.7 Strengths and limitations

These two calibration methods keep the advantages of the randomisation with high internal validity and combines multiple characteristics of routine data to inform the situations of patients in the clinical practice. Another advantage of this study is that the HF register includes all HF patients who had visited the specialist in NHSGGC that covers a wide area in west central Scotland and almost a quarter of Scottish population, which is representative for Glasgow HF patients. Furthermore, this study is from the perspective of methodology with each step being transparent and reproducible (see more details in the appendix) so it can also be used in other scenarios. Trials normally require substantial resources while it remains questionable on the treatment effectiveness in patients encountered in clinical practice and it is unfeasible to enrol every patient into trials. These two calibration methods, therefore, can maximise the generalisability of existing trials in a sustainable way without costing extra resources.

There are several important limitations in this analysis. I used routine data to define the target population because it was highly representative of patients encountered in clinical practice. However, some important variables were incompletely recorded, such as the New York Heart Association Classification (NYHA, 77.25% missing) and LVEF (84.87% missing) and therefore could not be included in the calibration. Although a numerical value for LVEF was available for only 15% of patients, a semi-quantitative measure of left ventricular function was available for 88% and indicated a reduced LVEF in 85% of cases, indicating that patients in the registry are predominantly HFrEF. This case calls for a better quality of data during the data collection procedure in routine clinical practice in the future especially for those important variables.

4.7 Conclusion

Calibration of HF trials to HF registry data is feasible and may be used, without breaking randomisation, to help address concerns about the representativeness of trials to patient population encountered in clinical practice. Consideration should be given to trial reporting standards and harmonisation of HF registry data to facilitate trial calibration and translation of clinical trials into clinical practice(164).

Chapter 5: Discussion and Conclusion.

5.1 Summary of the findings

5.1.1 Subgroup reporting

5.1.1.1 Overview

Chapter 3 involved the identification and examination of 2,235 trials from clinicaltrial.gov. Following the eligibility criteria outlined in Supplementary I Table S1, a total of 1,082 trials, corresponding to 2,422 publications, were included for further analysis. Each publication underwent at least one manual review to determine if subgroups were reported or not. Out of the reviewed papers, 907 were found to report subgroups. For each reported subgroup, the term was extracted, and standardization was performed using MeSH terms to ensure consistency in subgroup comparison across different trials.

Subsequently, logistic and Poisson regression models were constructed to explore the relationship between various trial characteristics (such as trial starting year, follow-up time, index conditions, sample size, number of arms, and industry sponsorship) and the reporting of results (whether any results were reported or not) as well as subgroup reporting (whether any subgroups were reported or not, and the count of subgroups reported). These models aimed to provide insights into the factors influencing result reporting and subgroup reporting in trials.

5.1.1.2 Main findings

Among 1,082 trials with reported results, 524 trials reported subgroup. Trials reporting subgroups tended to have larger sample sizes, longer follow-up durations, higher percentage of non-industry sponsorship, and more arms compared to trials without subgroup reporting.

The number of participants enrolled was the most significant predictor for any result reporting, any subgroup reporting, and the total number of subgroups reported. Follow-up duration is also a significant predictor. More recent trials were similar to older trials in terms of reporting patterns. Trials with 3 or more arms were more likely to report results but not necessarily associated with increased subgroup reporting or a higher total number of subgroups. Industry sponsorship did not significantly impact the reporting patterns. In comparison to asthma trials, cardiovascular, metabolic, and thromboembolic trials are more likely to report subgroups, along with a tendency to report a larger number of subgroups.

The analysis of 524 trials across 49 index conditions revealed variations in the number and types of subgroups reported. There were 345 subgroup terms in total, with a median of 11 terms per index condition. Some subgroups were commonly reported across all index conditions, including age, gender, comorbid diabetes, racial group, BMI, geographical locations, Glycated Hemoglobin A, and cigarette smoking.

In many trials, subgroups related to the index condition, such as duration, severity, or type, were frequently reported. For example, in MI trials, severity/history/type of MI was commonly reported as a subgroup. In type 2 diabetes trials, diabetes duration was frequently reported. Where conditions other than the index conditions were reported as subgroups, this was largely limited to diseases within the same body system as the index condition. For example, 13% of cardiovascular disease trials reported a non-index condition cardiovascular disease subgroup, such as hypertension. In these cases, both the index condition and subgroup belonged to the cardiovascular disease category. However, when the subgroup conditions differed from the index condition in terms of the body system, they were typically known causes or sequelae of the index condition. For instance, nutritional and metabolic diseases, predominantly diabetes, were reported as subgroups in cardiovascular disease trials (16%). Similarly, in diabetes trials, cardiovascular diseases (5.5%) and renal disease (4.9% urogenital diseases) were reported as subgroups. This study also found that trials rarely report metrics of comorbidity, multimorbidity or frailty and mental health.

This subgroup reporting description study showed there are variations in the subgroup reporting across different trials and interventions. It also identified specific subgroups that could be of greater interest for different trial index conditions and interventions. Future work could be directed towards improving trial reporting standards to achieve

more consistent and standardised subgroup reporting across various index conditions and interventions.

5.1.2 Trial calibration

5.1.2.1 Overview

Two landmark HF trials – COMET and DIG - were calibrated to a Scottish HF registry by using regression-based method and IPSW. Baseline characteristics were compared between trial participants and real-world patients. And the treatment effects in the real-world registry after calibration were also compared with those in trials.

5.1.2.2 Main findings

8 variables common in both the trials and the registry which were also believed to have potential to modify treatment effectiveness by cardiologists were included in the analyses. The patients in the HF registry differed from the trial participants. The registry patients were older, had slightly lower SBP and eGFR, and higher dosages of loop diuretics compared to the trial participants. There was a higher proportion of men than women in the registry, while the trials had a more balanced gender distribution. Additionally, the percentage of patients with a history of diabetes was slightly lower in the registry compared to the trials.

In both trials, male sex, older age, and history of diabetes were associated with a poorer prognosis and higher eGFR predicted longer survival. In COMET, use of higher dose loop diuretics also predicted a worse outcome, higher serum sodium concentration, and higher SBP predicted better prognosis. For both trials, treatment efficacy appeared to be lower for patients with diabetes and greater for heartrate but the CIs almost all included the null.

In the DIG trial, the uncalibrated and calibrated effect estimates for primary or composite outcomes were similar, indicating comparable efficacy between the trial and the HF registry. For the COMET trial, the efficacy was higher with IPSW calibration but

lower with regression-based calibration compared to uncalibrated estimates. However, the confidence intervals of the calibrated estimates overlapped with those of the uncalibrated estimates. The impact of calibration was consistent when trials were calibrated to high-risk and low-risk subgroups. The SEs of calibrated effect estimates were generally larger than those of standard analysis, indicating a loss of precision. Calibration to the overall target population resulted in SEs that were up to just over 3-fold wider. When calibrated to subgroups with the highest and lowest risk deciles, which differed more from the trial populations in terms of baseline characteristics, the SEs ranged from around 1-fold to around 8-fold wider.

Overall, this study demonstrated that trials can be calibrated to real-world registries while maintaining the strengths of trials and without significant loss of precision, assuming that the models were correctly specified. The findings also indicated that the results and messages derived from the trials remained applicable when applied to the Scottish HF registry. These methodologies have the potential for broader application in other trials and routine datasets, provided that the necessary data information is available.

5.1.2.3 Influence of highest weights on the precision of treatment effects

Extremely large weights which may be caused by small sample size may hurt the stable estimates(143). This can be improved by truncating the top weights (such as 1%) or normalising weights. The additional analyses excluded the lowest 1% odds (largest 1% weights) in the IOSW transportation and slightly changed the point estimates, SEs and narrowed the CI. This 1% extreme large weights (low odds of inclusion) were characterised by older age, higher loop diuretics doses and lower eGFR (Figure 13). In real-world practice, trade-offs may be necessary to achieve better precision by excluding patients with extremely low probabilities when using IPSW. However, caution is advised when implementing and interpreting these findings.

5.2 Strengths and limitations of this thesis

5.2.1 Subgroup reporting

This study has several strengths. Firstly, it detected registered trials from clinicaltrial.gov and then identified their corresponding publications, which means that it included trials from various publication sources, unlike some previous studies that only focused on publications in high impact journals(32, 123). This also potentially reduced the publication bias for subgroup reporting and made the description more credible. Additionally, this study represents the largest assessment of subgroup reporting in trials of multiple chronic medical conditions to date, drawing results from 2,235 trials. Another unique aspect is the assignment of standardised terminologies, enabling comparison across multiple conditions and drug classes. By standardising and harmonising terminologies, this study was able to reduce over 2,000 unique strings to 345 unique MeSH terms. This approach was crucial for facilitating comparisons of subgroups among different trials, as it would have been impractical and challenging to reconcile diverse subgroup terms reported by different researchers worldwide.

However, there were several limitations. Firstly, some papers might have been missed if they were not registered with ClinicalTrials.gov or lacked trial registration identifiers in PubMed. However, the number of such missed papers is expected to be small because the International Committee of Medical Journal Editors requires trial registration numbers (124). Secondly, subgroup results from non-indexed sources such as clinical reports could have been overlooked due to limited accessibility as the access to these reports often requires a formal application process or a data sharing agreement. Thirdly, some terms could not be assigned to standard MeSH or ATC codes due to their complexity but after the best attempts and there were not many terms remain unassigned. Lastly, the findings are specific to chronic medical conditions and do not encompass trials in infectious diseases, oncology, and psychiatric disorders (excluding dementia) for simplicity.

5.2.2 Trial calibration

This study demonstrated the advantages of using calibration methods to combine the strengths of randomised trials with routine data in informing clinical practice. The HF register utilised provides a representative sample of HF patients in west central Scotland and almost a quarter of Scottish population, offering valuable insights into real-world patient situations. Secondly, the transparent and reproducible methodology employed in this study makes it applicable to other scenarios as well. By maximising the applicability of existing trials, these calibration methods offer a sustainable approach without requiring additional resources.

However, there are limitations to consider. Incompleteness of certain variables in the routine data, such as NYHA classification and LVEF, limited their inclusion in the calibration process. This highlights the need for improved data quality in routine clinical practice, particularly for important variables. Efforts should be made to enhance data collection procedures in the future to ensure more comprehensive and reliable information.

5.3 Implications and recommendations of research

5.3.1 Feasibility and likely benefits of subgroup analysis

In order to conduct subgroup analysis for tailoring treatments to patients with specific characteristics, there are two approaches. Firstly, primary researchers can design subgroup analyses aligned with predefined subgroup specifications, collect subgroup data, and present the subgroup results. Secondly, secondary researchers can gather subgroup data (such as numbers, proportions, and events) and conduct post-hoc analyses. Pre-specified subgroup analysis is generally considered more credible and valid than post-hoc analyses. And sometimes it is challenging for secondary researchers to obtain sufficient data from published literature for subgroup analysis, especially if IPD is not available.

In the long term, researchers and the community will benefit more if primary researchers are able to provide more details about subgroup analyses within RCT. This

can be achieved through incorporating subgroup specification in the trial registration checklist and incorporating subgroup reporting in the trial reporting guideline. Chapter 3 can offer further insights into which subgroups should be considered in specific trials. Although this approach demands additional effort from primary researchers, the enhanced credibility of subgroup analyses and their ongoing benefits for policymakers, clinicians, and the targeted subgroup of patients may demonstrate its worthiness.

5.3.2 Improve real-world data quality and incorporate more variables

5.3.2.1 Improve routine data quality

Data quality issues in routine data are commonly described in terms of incomplete registers, inconsistencies between registers and reports, and low levels of data accuracy (166-170). In Chapter 4, the trial calibration study encountered missing data for important variables related to HF severity, such as NYHA classification and LVEF, as well as mislabelling of certain measurements from the Scottish HF registry. To address these issues, the study employed alternative approaches and solutions. These challenges highlight the need for improving data collection, standardization, and quality assurance processes in order to enhance the reliability and validity of routine data sources. The quality of routine data plays a crucial role in the effective functioning of the health system and enables policymakers to evaluate the impact of health system interventions aimed at improving population health(171). Although multiple efforts are being made to improve the quality of routine data, it remains insufficient and requires further time and attention(172). Future research can do more to enhance the data quality such as deploying and training monitoring and evaluation officers, conducting routine data assessments, collecting feedback et al(173, 174).

5.3.2.2 Incorporate variables more consistent with trial documentation

Comprehensive and consistent data on important variables that can influence treatment outcomes are often lacking in routine medical records (1). For instance, routine clinical practice frequently lacks consistently reporting of disease severity using standardised scales employed in trials(37, 175). To bridge the gap between effectiveness research and clinical medicine and to enhance the trial applicability, it is crucial to establish consistent documentation of patient characteristics with trials, including selection

criteria, adherence to interventions (such as off-label use), and outcome measurements et al(1). Although creating such documentation presents significant challenges, partly due to the fact that they often involve large-scale datasets requiring substantial resources (37), there is a pressing need for disease-specific clinical registries that are rigorous planned and designed following the same principles as trials. These registries should adopt similar documentation practices, enabling the collection of evidence on treatment effectiveness. This approach aligns with the proposition of trial applicability from Chapter 1, emphasising that representative clinical registries in the real world should maintain uniform documentation standards comparable to trials. This enables systematic comparisons between trial data and registry data (1, 7).

5.3.3 Improve trial and routine registry reporting

5.3.3.1 Complete, consistent, and unbiased trial reporting

Making the trial design and findings comprehensive, concise and transparent is important to ensure its accurate assessment and enhance the applicability (5). This also aligns with the principles of the open science framework, which might be the future direction for medical research. It facilitates the inclusion of various elements of the research lifecycle such as study design, data storage and analysis, protocol registration, etc..(176). Reporting guidelines and frameworks such as CONSORT and GPP3 were well established to improve trial reporting quality (5, 177). However, in real practice, the trial reporting situations are not adequate with the lack information about patient selection, study setting and patient characteristics such as comorbidities and equity factors(8). Some reporting also suffers from the selection bias with the tendency to less report trials with nonsignificant outcomes (178). These factors can all harm trial applicability and mislead future research and clinicians in routine practice. Therefore, adhering to reporting guidelines and providing thorough descriptions and details of trial design and unbiased findings becomes crucial.

5.3.3.2 Consistent trial subgroup reporting

Subgroup effects are different in a single trial and meta-analysis. From a single trial it is more likely to be over-interpreted and misleading as it may suffer from false positive (from multiple testing) or false negative (due to reduced statistical power). Instead,

considering results from multiple trials or conduct meta-analyses that combine data from different studies would be a better option to increase statistical power and enhance precision(26). For such purpose, ensuring the completeness and consistency of subgroup reporting would be helpful.

Chapter 3 presented findings from an analysis of 2,235 eligible trials, revealing that only a quarter of them (524 trials) reported subgroups. Furthermore, significant variation in subgroup reporting was observed even among trials within the same index condition and drug class, posing challenges for meta-analyses aiming to incorporate subgroup effects. Providing a wider common set of subgroup effect estimates via clinicaltrial.gov or digital repositories in machine readable formats would be very useful to incorporate them and further assess HTE. Another issue identified in subgroup reporting is the inconsistent use of subgroup terms by different researchers, necessitating time-consuming standardisation efforts using MeSH terms and WHOATC codes for comparability. To address these challenges, additional items can be added to existing trial reporting guidelines or frameworks, towards unifying subgroup terminology. Developing a checklist for subgroup reporting, encompassing categories such as demographic subgroups (e.g., age, race, gender), disease severities, and comorbidity subgroups et al could also enhance the completeness and consistency.

5.3.3.3 Make IPD alternatives available in both trials and routine registries

Chapter 4 showed that a key challenge in calibration is the requirement for accessing IPD from both the trial and target populations. This process is complex and time-consuming, involving data sharing agreements, regulatory approvals, data protection training certificate et al. Working with multiple Safe Havens, one for trial data and another for the registry, introduces additional complexity. For example, I could not directly transfer the summarised matrix generated from one Safe Haven to another. These summaries require a thorough review to eliminate any potential identification of IPD before being transferred to another Safe Haven to be included in the model. These factors have increased the complexity of the analyses, adding further challenges and causing extra time in addition to the initial task of obtaining access to IPD. However, small changes in trial reporting can help overcome this challenge. If trialists provide coefficients and the variance-covariance matrix for a treatment effect model that includes all relevant treatment-covariate interactions, secondary researchers with access to routine registry data can generate calibrated treatment effect estimates for

specific target population settings. If disease registries provide summary data on the joint distribution of commonly recorded patient characteristics in their corresponding trials (e.g., age, sex, diabetes), trialists with access to IPD can calibrate trials to these populations using methods like IOSW or regression. These efforts could significantly facilitate the trial transportation to real-world populations.

To facilitate the widespread adoption of calibration, the reporting of such summaries should be standardised. Expanding current trial reporting standards to include reporting of treatment-covariate interactions from multivariable models is necessary. Additionally, ongoing efforts to standardise and harmonise registries should incorporate reporting standards for population summaries. Implementing these changes in both trial and registry settings will require significant efforts from the research community. The calibration study yielding more applicable estimates with only a moderate loss of precision supports the value of undertaking these endeavours.

5.4 Recommendations for future search

5.4.1 For subgroup analysis

5.4.1.1 Define subgroups

Defining and using the consistent definition of subgroup enables comparison of outcomes among similar subgroups across different clinical trials. When subgroups are determined by continuous variables, it is preferable to utilise well-established or published cutoffs (105, 179).

5.4.1.2 Identify important subgroups for trials across different conditions

This can draw upon insights from Chapter 3, considering commonly reported subgroups, or obtain information from published literature regarding important prognostic factors or effect modifiers for different diseases. Expertise from clinicians in the field can also contribute to this identification process.

5.4.1.3 Incorporate subgroup analysis into trial registration platforms

The pre-specification of subgroup is to enhance the reliability of subgroup analysis from the trial design stage.

5.4.1.4 Include subgroup reporting in the guidelines for trial reporting

This will facilitate secondary researchers in obtaining comprehensive information on subgroup reporting, enabling them to conduct meta-analyses and derive treatment effects specific to certain subgroups, thereby informing clinical practice.

5.4.2 For trial calibration

5.4.2.1 Define important effect modifiers for trials across diverse conditions

Drawing insights from published literature and expert clinicians can aid trialists in creating models by using treatment and treatment-main covariates interactions to allow secondary researchers to perform calibration.

5.4.2.2 Examine the magnitude of variation in correlations among potential effect modifying variables between different settings

IPD is often limited. Instead, chapter 4 proposed to use the covariate joint distribution to re-construct pseudo-IPD which requires marginal summary statistics (mean and standard deviation for numerical variables, proportions for categorical variables) and correlations between covariates. However, the information about correlations may also be insufficient while marginal summary statistics may be accessible via the cohort profile description or the annual report of the registry. It is desirable to create pseudo-IPD using single-variable summary statistics for the real-world target population of interest, but taking between-variable statistics from other data sources (eg other registries or trial data). However, this assumes that between-variable characteristics are sufficiently similar across different settings. Further research can focus on examining the variation of correlations between different settings.

5.5 Contribution of the thesis

The thesis makes the following contributions:

1. The subgroup study represents the most extensive evaluation of subgroup reporting across trials involving multiple chronic medical conditions to date. It provides comprehensive and consistent evidence for further investigation into subgroup effects and treatment heterogeneity, addressing questions regarding which subgroups are reported, in which trials, and at what frequency.

2. The subgroup study examines the association between trial characteristics and: 1) the reporting of any results, 2) the reporting of any subgroups, and 3) the number of reported subgroups. With a sample size of 2,235 trials, it compares its findings to existing literature and identifies potential patterns of discrepancy.

3. The calibration study retains the key strength of trials, which is randomisation, and combines it with routine data that captures a broader range of patient information from the real-world. By leveraging the strength from both sources of data, it maximises the applicability of trials to real-world settings.

4. This study is the first to employ a regression-based method for calibrating trials in the real-world HF population, taking into account multiple patient characteristics in both datasets. The reproducible implementation steps enhance the broader utilization of this method in other scenarios.

5. This thesis demonstrates the feasibility of utilising IOSW method to calibrate two landmark HF trials in the real-world population, considering multiple patient characteristics. It also discusses the trade-off between the precision of calibrated estimates and the inclusion of patients with a low odds of being included in the trial.

6. This thesis has implications for expanding trial reporting guidelines to include additional features for subgroup reporting, enabling consistent examination of subgroup effects.

7. It also suggests incorporating coefficients, variance-covariance matrices, and treatment-covariate interactions as IPD alternatives in the trial reporting guidelines to facilitate calibration by researchers.

8. This thesis proposes registry reporting to provide summary data on the joint distribution of patient characteristics, enabling calibration from the trialists' side.

5.6 Conclusions

This thesis showed:

1) Variations in subgroup reporting limited the ability for assessing HTE.
2) Transportation methods were feasible to improve trial applicability to real-world HF populations.
3) Gaps in access to IPD and calibration need to be addressed.

To be able to perform a meta-analysis to examine the subgroup effects of in tailored patients, to further enhance the applicability of trials in these specific populations, this thesis emphasises the need for providing a wider common set of subgroup effect estimates via clinicaltrial.gov or digital repositories and unifying subgroup terms through MeSH terms or WHOATC codes in trial reporting guidelines. This requires future research to define important subgroups for different disease conditions, use consistent subgroups across different trials and pre-specify subgroup analysis in the trial design stage.

Trial calibration methods are novel, practical, and reproducible to enhance the trial applicability. It advocates for the inclusion of IPD alternatives for trial reporting (coefficients and the variance-covariance matrix for a treatment effect model) and routine registry reporting (marginal summary statistics and correlation matrix) to allow

researchers without IPD to perform calibration, which will facilitate wider adoption of calibration methods. This requires future research to identify important effect modifiers for different disease and correctly model the treatment covariate analysis.

References

1. Malmivaara A. APPLICABILITY OF EVIDENCE FROM RANDOMISED CONTROLLED TRIALS AND SYSTEMATIC REVIEWS TO CLINICAL PRACTICE: A CONCEPTUAL REVIEW. Journal of Rehabilitation Medicine. 2021;53(6).
2. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research Study design: randomised controlled trials. Bjog-an International Journal of Obstetrics and Gynaecology. 2018;125(13):1716-.
3. Burns PB, Rohrich RJ, Chung KC. The Levels of Evidence and Their Role in Evidence-Based Medicine. Plastic and Reconstructive Surgery. 2011;128(1):305-10.
4. Spieth PM, Kubasch AS, Penzlin AI, Illigens BM-W, Barlinn K, Siepmann T. Randomised controlled trials - a matter of design. Neuropsychiatric Disease and Treatment. 2016;12:1341-9.
5. Schulz KF, Altman DG, Moher D, Grp CONSORT. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Bmc Medicine. 2010;8.
6. Atkins D, Chang SM, Gartlehner G, Buckley DI, Whitlock EP, Berliner E, et al. Assessing applicability when comparing medical interventions: AHRQ and the Effective Health Care Program. Journal of Clinical Epidemiology. 2011;64(11):1198-207.
7. Stuart EA, Rhodes A. Generalizing Treatment Effect Estimates From Sample to Population: A Case Study in the Difficulties of Finding Sufficient Data. Evaluation Review. 2017;41(4):357-88.
8. Malmivaara A. Generalizability of findings from randomised controlled trials is limited in the leading general medical journals. Journal of Clinical Epidemiology. 2019;107:36-41.
9. Zwarenstein M, Treweek S. What kind of randomised trials do we need? Journal of Clinical Epidemiology. 2009;62(5):461-3.
10. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. Canadian Medical Association Journal. 2009;180(10):E47-E57.
11. Schwartz D, Lellouch J. EXPLANATORY AND PRAGMATIC ATTITUDES IN THERAPEUTICAL TRIALS. Journal of Chronic Diseases. 1967;20(8):637-&.

12. Rothwell PM. Factors that can affect the external validity of Rando controlled trials. Plos Clinical Trials. 2006;1(1):5.
13. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials - Increasing the value of clinical research for decision making in clinical and health policy. Jama-Journal of the American Medical Association. 2003;290(12):1624-32.
14. Farrell B, Fraser A, Sandercock P, Slattery J, Warlow CP, Europ Carotid Surg Trial Collaborat G. Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European carotid surgery trial (ECST). Lancet. 1998;351(9113):1379-87.
15. Fine PEM. VARIATION IN PROTECTION BY BCG - IMPLICATIONS OF AND FOR HETEROLOGOUS IMMUNITY. Lancet. 1995;346(8986):1339-45.
16. CAROTID ENDARTERECTOMY FOR PATIENTS WITH ASYMPTOMATIC INTERNAL CAROTID-ARTERY STENOSIS. Journal of the Neurological Sciences. 1995;129(1):76-7.
17. Nouvini R, Parker PA, Malling CD, Godwin K, Costas-Muniz R. Interventions to increase racial and ethnic minority accrual into cancer clinical trials: A systematic review. Cancer. 2022;128(21):3860-9.
18. Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. Archives of Internal Medicine. 2002;162(15):1682-8.
19. Tan YY, Papez V, Chang WH, Mueller SH, Denaxas S, Lai AG. 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. Lancet Healthy Longevity. 2022;3(10):E674-E89.
20. Zhang X, Wu Y, Kang D, Wang J, Hong Q, Le P. The external validity of randomised controlled trials of hypertension within China: from the perspective of sample representation. PLoS ONE [Electronic Resource]. 2013;8(12):e82324.
21. Martinson BC, Crain AL, Sherwood NE, Hayes MG, Pronk NP, O'Connor PJ. Population Reach and Recruitment Bias in a Maintenance RCT in Physically Active Older Adults. Journal of Physical Activity & Health. 2010;7(1):127-35.
22. Jordan S, Watkins A, Storey M, Allen SJ, Brooks CJ, Garaiova I, et al. Volunteer Bias in Recruitment, Retention, and Blood Sample Donation in a Randomised Controlled Trial Involving Mothers and Their Children at Six Months and Two Years: A Longitudinal Analysis. Plos One. 2013;8(7).
23. Rothwell PM. Treating Individuals 1 - External validity of randomised controlled trials: "To whom do the results of this trial apply?". Lancet. 2005;365(9453):82-93.
24. Kent DM, Hayward RA. Limitations of Applying Summary Results of Clinical Trials to Individual PatientsThe Need for Risk Stratification. JAMA. 2007;298(10):1209-12.

25. Hernández AV, Boersma E, Murray GD, Habbema JDF, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? American Heart Journal. 2006;151(2):257-64.
26. Wallach JD, Sullivan PG, Trepanowski JF, Steyerberg EW, Ioannidis JPA. Sex based subgroup differences in randomised controlled trials: empirical evidence from Cochrane meta-analyses. Bmj. 2016;355:i5826.
27. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. Bmj. 2011;342:d1569.
28. Gabler NB, Duan N, Raneses E, Suttner L, Ciarametaro M, Cooney E, et al. No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals. Trials. 2016;17(1):320.
29. Gabler NB, Duan N, Liao D, Elmore JG, Ganiats TG, Kravitz RL. Dealing with heterogeneity of treatment effects: is the literature up to the challenge? Trials. 2009;10(1):43.
30. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. The Lancet. 2000;355(9209):1064-9.
31. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. Bmj-British Medical Journal. 2011;342.
32. Brand KJ, Hapfelmeier A, Haller B. A systematic review of subgroup analyses in randomised clinical trials in cardiovascular disease. Clinical Trials. 2021;18(3):351-60.
33. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine - Reporting of subgroup analyses in clinical trials. New England Journal of Medicine. 2007;357(21):2189-94.
34. Fan J, Song F, Bachmann MO. Justification and reporting of subgroup analyses were lacking or inadequate in randomised controlled trials. Journal of Clinical Epidemiology. 2019;108:17-25.
35. Chodankar D. Introduction to real-world evidence studies. Perspectives in clinical research. 2021;12(3):171-4.
36. Monti S, Grosso V, Todoerti M, Caporali R. Randomised controlled trials and real-world data: differences and similarities to untangle literature data. Rheumatology (Oxford, England). 2018;57(57 Supplement 7):vii54-vii8.
37. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. British Journal of Cancer. 2014;110(3):551-5.

38. Lobo FS, Wagner S, Gross CR, Schommer JC. Addressing the issue of channeling bias in observational studies with propensity scores analysis. Research in social & administrative pharmacy : RSAP. 2006;2(1):143-51.
39. Ezekowitz JA, Hu J, Delgado D, Hernandez AF, Kaul P, Leader R, et al. Acute Heart Failure Perspectives From a Randomised Trial and a Simultaneous Registry. Circulation-Heart Failure. 2012;5(6):735-41.
40. Palmowski A, Nielsen SM, Buttgereit T, Palmowski Y, Boers M, Christensen R, et al. Glucocorticoid-trials in rheumatoid arthritis mostly study representative real-world patients: A systematic review and meta-analysis. Seminars in arthritis and rheumatism. 2020.
41. Palmowski A, Buttgereit T, Palmowski Y, Nielsen SM, Boers M, Christensen R, et al. Applicability of trials in rheumatoid arthritis and osteoarthritis: A systematic review and meta-analysis of trial populations showing adequate proportion of women, but underrepresentation of elderly people. Seminars in Arthritis and Rheumatism. 2019;48(6):983-9.
42. Evans A, Kalra L. Are the results of randomised controlled trials on anticoagulation in patients with atrial fibrillation generalizable to clinical practice? Archives of Internal Medicine. 2001;161(11):1443-7.
43. Laupacis A, Boysen G, Connolly S, Ezekowitz M, Hart R, James K, et al. RISK-FACTORS FOR STROKE AND EFFICACY OF ANTITHROMBOTIC THERAPY IN ATRIAL-FIBRILLATION - ANALYSIS OF POOLED DATA FROM 5 RANDOMISED CONTROLLED TRIALS. Archives of Internal Medicine. 1994;154(13):1449-57.
44. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. Bmj-British Medical Journal. 1997;315(7109):640-5.
45. Catapano. 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias (vol 70, pg 115.e1, 2017). Revista Espanola De Cardiologia. 2018;71(8):691-2.
46. Amer Diabet A. Standards of Medical Care in Diabetes-2013. Diabetes Care. 2013;36:S11-S66.
47. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. International Journal of Behavioral Medicine. 2017;24(3):321-419.
48. Ruokoniemi P, Sund R, Arffman M, Helin-Salmivaara A, Huupponen R, Keskimaki I, et al. Are statin trials in diabetes representative of real-world diabetes care: a population-based study on statin initiators in Finland. Bmj Open. 2014;4(6).
49. Berden FA, de Knegt RJ, Blokzijl H, Kuiken SD, van Erpecum KJ, Willemse SB, et al. Limited Generalizability of Registration Trials in Hepatitis C: A Nationwide Cohort Study. PLoS ONE [Electronic Resource]. 2016;11(9):e0161821.

50. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomised controlled trial samples and implications for the external validity of trial results. Trials. 2015;16(1).
51. Sox HC, Lewis RJ. Pragmatic Trials Practical Answers to "Real World" Questions. Jama-Journal of the American Medical Association. 2016;316(11):1205-6.
52. Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. Advances in Therapy. 2018;35(11):1763-74.
53. Patsopoulos NA. A pragmatic view on pragmatic trials. Dialogues in clinical neuroscience. 2011;13(2):217-24.
54. Blonde L, Merilainen M, Karwe V, Raskin P, Grp TS. Patient-directed titration for achieving glycaemic goals using a once-daily basal insulin analogue: an assessment of two different fasting plasma glucose targets - the TITRATE(TM) study. Diabetes Obesity & Metabolism. 2009;11(6):623-31.
55. Duley L, Antman K, Arena J, Avezum A, Blumenthal M, Bosch J, et al. Specific barriers to the conduct of randomised trials. Clinical Trials. 2008;5(1):40-8.
56. Butala NM, Faridi K, Tamez H, Strom JB, Song Y, Shen C, et al. Estimation of DAPT Study Treatment Effects in Contemporary Clinical Practice: Findings from the EXTEND-DAPT Study. Circulation. 2021;144.
57. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the Generalizability of Randomised Trial Results to Target Populations. Prevention Science. 2015;16(3):475-85.
58. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of Trial Results Using Inverse Odds of Sampling Weights. American Journal of Epidemiology. 2017;186(8):1010-4.
59. Hong JL, Webster-Clark M, Funk MJ, Sturmer T, Dempster SE, Cole SR, et al. Comparison of Methods to Generalize Randomised Clinical Trial Results Without Individual-Level Data for the Target Population. American Journal of Epidemiology. 2019;188(2):426-37.
60. Frangakis C. The calibration of treatment effects from clinical trials to target populations. Clinical Trials. 2009;6(2):136-40.
61. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. Medical Decision Making. 2018;38(2):200-11.
62. Mauri L, Kereiakes DJ, Yeh RW, Driscoll-Shempp P, Cutlip DE, Steg PG, et al. Twelve or 30 Months of Dual Antiplatelet Therapy after Drug-Eluting Stents. New England Journal of Medicine. 2014;371(23):2155-66.
63. Signorovitch JE, Wu EQ, Yu AP, Gerrits CM, Kantor E, Bao Y, et al. Comparative Effectiveness Without Head-to-Head Trials A Method for Matching-Adjusted Indirect

Comparisons Applied to Psoriasis Treatment with Adalimumab or Etanercept. Pharmacoeconomics. 2010;28(10):935-45.
64. Holt D, Smith TMF. POST STRATIFICATION. Journal of the Royal Statistical Society Series a-Statistics in Society. 1979;142:33-46.
65. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomised trials. Journal of the Royal Statistical Society Series a-Statistics in Society. 2011;174:369-86.
66. Li S, Heitjan DF. Generalizing Clinical Trial Results to a Target Population. Statistics in Biopharmaceutical Research. 2023;15(1):125-32.
67. Tipton E. Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. Journal of Educational and Behavioral Statistics. 2013;38(3):239-66.
68. Ridker PM, Danielson E, Fonseca FAH, Genest J, Gotto AM, Jr., Kastelein JJP, et al. Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein. New England Journal of Medicine. 2008;359(21):2195-207.
69. Caro JJ, Migliaccio-Walle K, Rate CCAP. Generalizing the results of clinical trials to actual practice: The example of Clopidogrel therapy for the prevention of vascular events. American Journal of Medicine. 1999;107(6):568-72.
70. Gent M, Beaumont D, Blanchard J, Bousser MG, Coffman J, Easton JD, et al. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). Lancet. 1996;348(9038):1329-39.
71. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, et al. Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR task force on good research practices-modeling studies. Value in Health. 2003;6(1):9-17.
72. Phillippo DM, Dias S, Ades AE, Belger M, Brnabic A, Schacht A, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. Journal of the Royal Statistical Society Series a-Statistics in Society. 2020;183(3):1189-210.
73. Phillippo DM, Dias S, Ades AE, Belger M, Brnabic A, Saure D, et al. Validating the Assumptions of Population Adjustment: Application of Multilevel Network Meta-regression to a Network of Treatments for Plaque Psoriasis. Medical Decision Making. 2023;43(1):53-67.
74. Phillippo DM. multinma: Bayesian Network Meta-Analysis of Individual and Aggregate Data [Available from: https://cran.r-project.org/web/packages/multinma/index.html.
75. Sacristan JA, Soto J, Galende I, Hylan TR. Randomised database studies: A new method to assess drugs' effectiveness? Commentary. Journal of Clinical Epidemiology. 1998;51(9):713-5.

76. CROSS DESIGN SYNTHESIS - A NEW STRATEGY FOR STUDYING MEDICAL OUTCOMES. Lancet. 1992;340(8825):944-6.
77. Wang SV, Schneeweiss S, Gagne JJ, Evers T, Gerlinger C, Desai R, et al. Using Real-World Data to Extrapolate Evidence From Randomised Controlled Trials. Clinical Pharmacology & Therapeutics. 2019;105(5):1156-63.
78. Kaizar EE. Estimating treatment effect via simple cross design synthesis. Statistics in Medicine. 2011;30(25):2986-3009.
79. Verde PE, Ohmann C. Combining randomised and non-randomised evidence in clinical research: a review of methods and applications. Research Synthesis Methods. 2015;6(1):45-62.
80. Brown CH, Wang W, Sandler I. Examining How Context Changes Intervention Impact: The Use of Effect Sizes in Multilevel Mixture Meta-Analysis. Child Development Perspectives. 2008;2(3):198-205.
81. Najafzadeh M, Schneeweiss S, Choudhry NK, Wang SV, Gagne JJ. Simulation for Predicting Effectiveness and Safety of New Cardiovascular Drugs in Routine Care Populations. Clinical Pharmacology & Therapeutics. 2018;104(5):1008-15.
82. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. International Journal of Epidemiology. 2016;45(6):2184-93.
83. Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. International Journal of Epidemiology. 2016;45(6):2075-88.
84. Burke JF, Hayward RA, Nelson JP, Kent DM. Using Internally Developed Risk Models to Assess Heterogeneity in Treatment Effects in Clinical Trials. Circulation-Cardiovascular Quality and Outcomes. 2014;7(1):163-9.
85. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. Journal of the Royal Statistical Society Series a-Statistics in Society. 2015;178(3):757-78.
86. Hansen LP. LARGE SAMPLE PROPERTIES OF GENERALIZED-METHOD OF MOMENTS ESTIMATORS. Econometrica. 1982;50(4):1029-54.
87. Hellerstein JK, Imbens GW. Imposing moment restrictions from auxiliary data by weighting. Review of Economics and Statistics. 1999;81(1):1-14.
88. Castro Y. Long-term outcome estimation combining real world with clinical trial data in metastatic melanoma following a Bayesian approach. Journal of the European Academy of Dermatology and Venereology. 2017;31:65-6.
89. Smith JA, Todd PE. Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics. 2005;125(1-2):305-53.

90. Pingel R, Waernbaum I. Correlation and efficiency of propensity score-based estimators for average causal effects. Communications in Statistics-Simulation and Computation. 2017;46(5):3458-78.
91. Horton R. Common sense and figures: the rhetoric of validity in medicine Bradford Hill Memorial Lecture 1999. Statistics in Medicine. 2000;19(23):3149-64.
92. Pawlotsky J-M, Aghemo A, Back D, Dusheiko G, Forns X, Puoti M, et al. EASL Recommendations on Treatment of Hepatitis C 2015. Journal of Hepatology. 2015;63(1):199-236.
93. Barua S, Greenwald R, Grebely J, Dore GJ, Swan T, Taylor LE. Restrictions for Medicaid Reimbursement of Sofosbuvir for the Treatment of Hepatitis C Virus Infection in the United States. Annals of Internal Medicine. 2015;163(3):215-+.
94. Ferenci P, Dusheiko G. Beyond phase 3 registration trials: defining safety for triple therapy with protease inhibitors in cirrhosis. Gut. 2014;63(7):1033-+.
95. Lauer MS, D'Agostino RB, Sr. The Randomised Registry Trial - The Next Disruptive Technology in Clinical Research? New England Journal of Medicine. 2013;369(17):1579-81.
96. Church DL. Major factors affecting the emergence and re-emergence of infectious diseases. Clinics in Laboratory Medicine. 2004;24(3):559-+.
97. Maxim LD, Niebo R, Utell MJ. Screening tests: a review with examples. Inhalation Toxicology. 2014;26(13):811-28.
98. Noncommunicable diseases [Available from: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.
99. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380(9859):2095-128.
100. Angus DC, Chang C-CH. Heterogeneity of Treatment Effect Estimating How the Effects of Interventions Vary Across Individuals. Jama-Journal of the American Medical Association. 2021;326(22):2312-3.
101. Iwashyna TJ, Burke JF, Sussman JB, Prescott HC, Hayward RA, Angus DC. Implications of Heterogeneity of Treatment Effect for Reporting and Analysis of Randomised Trials in Critical Care. American Journal of Respiratory and Critical Care Medicine. 2015;192(9):1045-51.
102. Alosh M, Huque MF. Multiplicity considerations for subgroup analysis subject to consistency constraint. Biometrical Journal. 2013;55(3):444-62.
103. Sormani MP. Subgroup analysis in MS trials. Multiple Sclerosis Journal. 2017;23(1):34-5.

104. Sormani MP, Bruzzi P. Reporting of subgroup analyses from clinical trials. Lancet Neurology. 2012;11(9):747-.
105. Williamson SF, Grayling MJ, Mander AP, Noor NM, Savage JS, Yap C, et al. Subgroup analyses in randomised controlled trials frequently categorized continuous subgroup information. Journal of Clinical Epidemiology. 2022;150:72-9.
106. Rothwell PM. Treating Individuals 2 - Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005;365(9454):176-86.
107. Wijn SRW, Rovers MM, Le LH, Belias M, Hoogland J, IntHout J, et al. Guidance from key organisations on exploring, confirming and interpreting subgroup effects of medical treatments: a scoping review. Bmj Open. 2019;9(8).
108. Lewis JA. Statistical principles for clinical trials (ICH E9) an introductory note on an international guideline. Statistics in Medicine. 1999;18(15):1903-4.
109. Hernandez AV, Boersma E, Murray GD, Habbema JDF, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? American Heart Journal. 2006;151(2):257-64.
110. Interpretation of subgroup analyses and meta-regressions [Available from: https://handbook-5-1.cochrane.org/chapter_9/9_6_6_interpretation_of_subgroup_analyses_and_meta_regressions.htm.
111. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. Bmj. 2010;340:c117.
112. Schulz KF, Altman DG, Moher D, Group tC. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med. 2010;8(1):18.
113. Bhattarai M, Salih M, Regmi M, Al-Akchar M, Deshpande R, Niaz Z, et al. Association of Sodium-Glucose Cotransporter 2 Inhibitors With Cardiovascular Outcomes in Patients With Type 2 Diabetes and Other Risk Factors for Cardiovascular Disease A Meta-analysis. Jama Network Open. 2022;5(1).
114. Davies MJ, Aroda VR, Collins BS, Gabbay RA, Green J, Maruthur NM, et al. Management of Hyperglycemia in Type 2 Diabetes, 2022. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care. 2022;45(11):2753-86.
115. Mishriky BM, Powell JR, Wittwer JA, Chu JX, Sewell KA, Wu Q, et al. Do GLP-1RAs and SGLT-2is reduce cardiovascular events in black patients with type 2 diabetes? A systematic review and meta-analysis. Diabetes, obesity & metabolism. 2019;21(10):2274-83.
116. Kotecha D, Holmes J, Krum H, Altman DG, Manzano L, Cleland JGF, et al. Efficacy of ß blockers in patients with heart failure plus atrial fibrillation: an individual-patient data meta-analysis. Lancet. 2014;384(9961):2235-43.

117. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. European Journal of Heart Failure. 2016;18(8):891-975.
118. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane Handbook for Systematic Reviews of Interventions Chapter 10: Analysing data and undertaking meta-analyses 2022 [Available from: https://training.cochrane.org/handbook/current/chapter-10.
119. Hanlon P, Hannigan L, Rodriguez-Perez J, Fischbacher C, Welton NJ, Dias S, et al. Representation of people with comorbidity and multimorbidity in clinical trials of novel drug therapies: an individual-level participant data analysis. Bmc Medicine. 2019;17(1).
120. CTTI. What is AACT? [Available from: https://aact.ctti-clinicaltrials.org/.
121. Winter DJ. rentrez: An R package for the NCBI eUtils API. R Journal. 2017;9(2):520-6.
122. Medicine NLo. MeSH Tree View [Available from: https://meshb.nlm.nih.gov/treeView.
123. Khan MS, Khan MAA, Irfan S, Siddiqi TJ, Greene SJ, Anker SD, et al. Reporting and interpretation of subgroup analyses in heart failure randomised controlled trials. Esc Heart Failure. 2021;8(1):26-36.
124. ICMJE. Preparing a Manuscript for Submission to a Medical Journal [Available from: https://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html.
125. Kasenda B, Schandelmaier S, Sun X, von Elm E, You J, Bluemle A, et al. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications (vol 349, g4539, 2014). Bmj-British Medical Journal. 2014;349.
126. Briel M, Grp DS. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. Bmj-British Medical Journal. 2014;349.
127. Song FJ, Bachmann MO. Cumulative subgroup analysis to reduce waste in clinical research for individualised medicine. Bmc Medicine. 2016;14.
128. Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. Bmj-British Medical Journal. 2015;351.
129. Bloom HS, Michalopoulos C. When is the Story in the Subgroups? Prevention Science. 2013;14(2):179-88.
130. Cui L, Hung HMJ, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. Journal of biopharmaceutical statistics. 2002;12(3):347-58.

131. Kaiser LD. Dynamic randomization and a randomization model for clinical trials data. Statistics in Medicine. 2012;31(29):3858-73.
132. Barraclough H, Govindan R. Biostatistics Primer What a Clinician Ought to Know: Subgroup Analyses. Journal of Thoracic Oncology. 2010;5(5):741-6.
133. Margolis KL, Piller LB, Ford CE, Henriquez MA, Cushman WC, Einhorn PT, et al. Blood pressure control in Hispanics in the antihypertensive and lipid-lowering treatment to prevent heart attack trial. Hypertension. 2007;50(5):854-61.
134. Weber MA, Jamerson K, Bakris GL, Weir MR, Zappe D, Zhang Y, et al. Effects of body size and hypertension treatments on cardiovascular event rates: subanalysis of the ACCOMPLISH randomised controlled trial. Lancet. 2013;381(9866):537-45.
135. Proctor T, Jensen K, Kieser M. Integrated evaluation of targeted and non-targeted therapies in a network meta-analysis. Biometrical Journal. 2020;62(3):777-89.
136. Wallace E, Salisbury C, Guthrie B, Lewis C, Fahey T, Smith SM. Managing patients with multimorbidity in primary care. Bmj-British Medical Journal. 2015;350.
137. Concato J, Shah N, Horwitz RI. Randomised , controlled trials, observational studies, and the hierarchy of research designs. New England Journal of Medicine. 2000;342(25):1887-92.
138. Kilcher G, Hummel N, Didden EM, Egger M, Reichenbach S, GetReal Work P. Rheumatoid arthritis patients treated in trial and real world settings: comparison of randomised trials with registries. Rheumatology. 2018;57(2):354-69.
139. Sharma A, Ezekowitz JA. Similarities and differences in patient characteristics between heart failure registries versus clinical trials. Current heart failure reports. 2013;10(4):373-9.
140. Stuart EA, Ackerman B, Westreich D. Generalizability of Randomised Trial Results to Target Populations: Design and Analysis Possibilities. Research on Social Work Practice. 2018;28(5):532-7.
141. Cole SR, Stuart EA. Generalizing Evidence From Randomised Clinical Trials to Target Populations. American Journal of Epidemiology. 2010;172(1):107-15.
142. Happich M, Brnabic A, Faries D, Abrams K, Winfree KB, Girvan A, et al. Reweighting Randomised Controlled Trial Evidence to Better Reflect Real Life - A Case Study of the Innovative Medicines Initiative. Clinical Pharmacology & Therapeutics. 2020;108(4):817-25.
143. Butala NM, Faridi KF, Tamez H, Strom JB, Song Y, Shen CY, et al. Estimation of DAPT Study Treatment Effects in Contemporary Clinical Practice: Findings From the EXTEND-DAPT Study. Circulation. 2022;145(2):97-106.
144. Bentley C, Cressman S, van der Hoek K, Arts K, Dancey J, Peacock S. Conducting clinical trials-costs, impacts, and the value of clinical trials networks: A scoping review. Clinical Trials. 2019;16(2):183-93.

145. Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods. 2002;7(2):147-77.
146. Graham JW. Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology. 2009;60:549-76.
147. Package 'mice' January 27, 2021 [Available from: https://cran.r-project.org/web/packages/mice/mice.pdf.
148. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research. 2011;20(1):40-9.
149. FUROSEMIDE NICE | The National Institute for Health and Care Excellence 2021 [Available from: https://bnf.nice.org.uk/drug/furosemide.html.
150. Khan TM, Patel R, Siddiqui AH. Furosemide. StatPearls. Treasure Island (FL): StatPearls Publishing

151. Guidelines for the Administration of Intravenous and Subcutaneous diuretic for Heart Failure Patients in the Community Setting. NHS North West Coast Strategic Clinical Networks [Available from: https://www.england.nhs.uk/north/wp-content/uploads/sites/5/2019/06/administration-intravenous-subcutaneous-diuretics-heart-failure-patients.pdf.
152. Heart failure: Managing newly diagnosed and decompensated patients admitted to hospital. NHS Wirral University Teaching Hospital [Available from: https://mm.wirral.nhs.uk/document_uploads/guidelines/HeartFailureClinicalGuidance.pdf.
153. Anisman SD, Erickson SB, Morden NE. How to prescribe loop diuretics in oedema. Bmj-British Medical Journal. 2019;364.
154. Seltman HJ. Experimental Design and

Analysis2018.
155. Ioannidis JPA, Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials - A proposed evaluation algorithm. American Journal of Epidemiology. 1998;148(11):1117-26.
156. Desai AS, Solomon S, Claggett B, McMurray JJV, Rouleau J, Swedberg K, et al. Factors Associated With Noncompletion During the Run-In Period Before Randomization and Influence on the Estimated Benefit of LCZ696 in the PARADIGM-HF Trial. Circulation-Heart Failure. 2016;9(6).
157. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. Bmj-British Medical Journal. 2015;350.

158. Stuermer T, Wyss R, Glynn RJ, Brookhart MA. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. Journal of Internal Medicine. 2014;275(6):570-80.
159. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence Synthesis for Decision Making 5: The Baseline Natural History Model. Medical Decision Making. 2013;33(5):657-70.
160. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernan MA. Extending inferences from a randomised trial to a new target population. Statistics in Medicine. 2020;39(14):1999-2014.
161. Funk MJ, Westreich D, Wiesen C, Stuermer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. American Journal of Epidemiology. 2011;173(7):761-7.
162. Li X, Shen C. Doubly Robust Estimation of Causal Effect: Upping the Odds of Getting the Right Answers. Circulation-Cardiovascular Quality and Outcomes. 2020;13(1).
163. Phillippo DM, Dias S, Ades AE, Welton NJ. Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study. Statistics in Medicine. 2020;39(30):4885-911.
164. Aktaa S, Batra G, Cleland JGF, Coats A, Lund LH, McDonagh T, et al. Data standards for heart failure: the European Unified Registries for Heart Care Evaluation and Randomised Trials (EuroHeart). European Heart Journal.
165. CDISC Standards in the Clinical Research Process [Available from: https://www.cdisc.org/standards.
166. Chiba Y, Oguttu MA, Nakayama T. Quantitative and qualitative verification of data quality in the childbirth registers of two rural district hospitals in Western Kenya. Midwifery. 2012;28(3):329-39.
167. Sharma A, Rana SK, Prinja S, Kumar R. Quality of Health Management Information System for Maternal & Child Health Care in Haryana State, India. Plos One. 2016;11(2).
168. O'Hagan R, Marx MA, Finnegan KE, Naphini P, Ng'ambi K, Laija K, et al. National Assessment of Data Quality and Associated Systems-Level Factors in Malawi. Global Health-Science and Practice. 2017;5(3):367-81.
169. Kihuba E, Gathara D, Mwinga S, Mulaku M, Kosgei R, Mogoa W, et al. Assessing the ability of health information systems in hospitals to support evidence-informed decisions in Kenya. Global health action. 2014;7:24859-.
170. Lemma S, Janson A, Persson L-A, Wickremasinghe D, Kallestal C. Improving quality and use of routine health information system data in low- and middle-income countries: A scoping review. Plos One. 2020;15(10).

171. AbouZahr C, Boerma T. Health information systems: the foundations of public health. Bulletin of the World Health Organization. 2005;83(8):578-83.
172. Nisingizwe MP, Iyer HS, Gashayija M, Hirschhorn LR, Amoroso C, Wilson R, et al. Toward utilization of data for program management and evaluation: quality assessment of five years of health management information system data in Rwanda. Global health action. 2014;7:25829-.
173. Mpofu M, Semo B, Grignon J, Lebelonyane R, Ludick S, Matshediso E, et al. Strengthening monitoring and evaluation (M&E) and building sustainable health information systems in resource limited countries: lessons learned from an M&E task-shifting initiative in Botswana. BMC Public Health. 2014;14(1032):(3 October 2014)-(3 October).
174. Gimbel S, Mwanza M, Nisingizwe MP, Michel C, Hirschhorn L, Collaborative APP. Improving data quality across 3 sub-Saharan African countries using the Consolidated Framework for Implementation Research (CFIR): results from the African Health Initiative. Bmc Health Services Research. 2017;17.
175. Kerkhofs TMA, Verhoeven RHA, Van der Zwan JM, Dieleman J, Kerstens MN, Links TP, et al. Adrenocortical carcinoma: A population-based study on incidence and survival in the Netherlands since 1993. European Journal of Cancer. 2013;49(11):2579-86.
176. Open Science Framework (OSF) 2017 [Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5370619/.
177. Battisti WP, Wager E, Baltzer L, Bridges D, Cairns A, Carswell CI, et al. Good Publication Practice for Communicating Company-Sponsored Medical Research: GPP3. Annals of Internal Medicine. 2015;163(6):461-+.
178. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomised trials - Comparison of Protocols to published articles. Jama-Journal of the American Medical Association. 2004;291(20):2457-65.
179. Wang X, Piantadosi S, Le-Rademacher J, Mandrekar SJ. Statistical Considerations for Subgroup Analyses. Journal of Thoracic Oncology. 2021;16(3):375-80.

Supplementary I – Subgroup Reporting

Identifying trials, papers and subgroups.

Identifying eligible trials from clinicaltrials.gov.

Supplementary I Table S1. Pre-specified inclusion criteria for identifying trials from clinicaltrials.gov.

Population	• Adults • Trials must either not exclude based on age or have an upper age limit >60 years
Intervention	"Drug" or "Biological"
Comparator	Comparison with other eligible drug, placebo, usual-care or "standard comparator"
Outcome	Any
Study design	• Randomised controlled trial (search criteria "Factorial assignment", "Parallel assignment" and "allocation random") • Phase 2/3, 3, or 4
Date	• Trials start date after 1st Jan 1990 • Search performed September 2017
Status	'Active, not recruiting', 'Completed' or 'Terminated'
Enrolment	>= 300
Other exclusions	• Upper age limit under 60 years • Topical therapies • Discontinued therapies • Trials with same-drug comparisons

Supplementary I Table S2. Included conditions, Medical Subject Headings (MeSH) terms and MeSH codes.

Category	MeSH term	Code
Musculoskeletal al	Osteoporosis	C05.116.198.579
	Spondyloarthropathies	C05.116.900.853.625.800
	Arthritis	C05.550.114

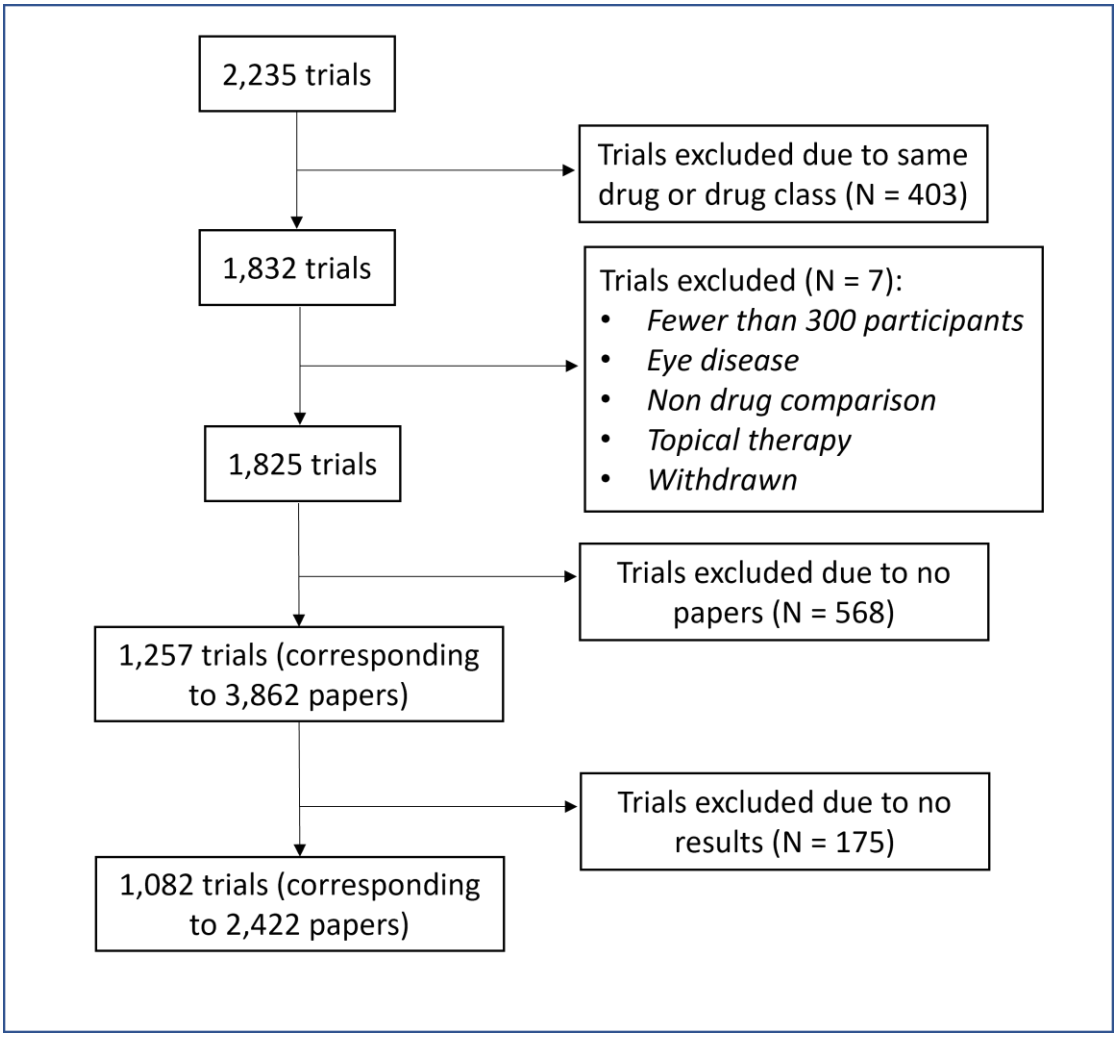
	Arthritis, Rheumatoid Gout	C05.799.114
	Osteoporosis	C05.799.414
Digestive system diseases	CREST Syndrome	C06.405.117.119.500.204
	Oesophageal Achalasia	C06.405.117.119.500.432
	Oesophageal spasm, diffuse	C06.405.117.119.500.450
	Gastro-oesophageal reflux	C06.405.117.119.500.484
	Laryngopharyngeal reflux	C06.405.117.119.500.484.500
	Plummer-Vinson Syndrome	C06.405.117.119.500.742
	Oesophagitis	C06.405.117.620
	Colitis, Ulcerative	C06.405.205.265.231
	Inflammatory Bowel Diseases	C06.405.205.731
	Inflammatory Bowel diseases	C06.405.469.432
	Oesophagitis, peptic	C06.405.608.348
	Duodenogastric reflux	C06.405.748.240
	Gastritis	C06.405.748.398
	Hepatitis, autoimmune	C06.552.380.350.050
Respiratory Tract Diseases	Asthma	C08.127.108
	Bronchiectasis	C08.127.384
	Bronchitis, chronic	C08.127.446.567
	Hypertension, Pulmonary	C08.381.423
	Idiopathic Interstitial Pneumonias	C08.381.483.487
	Idiopathic Pulmonary Fibrosis	C08.381.483.487.500
	Lung Diseases, Obstructive	C08.381.495
	Pulmonary Embolism	C08.381.746
	Pulmonary Fibrosis	C08.381.765
	Rhinitis	C08.460.799
	Asthma	C08.674.095
	Bronchitis, Chronic	C08.730.099.567
Otorhinolaryngologic Diseases	Rhinitis, Allergic	C09.603.799.315
	Multiple Sclerosis	C10.114.375.500
	Parkinsonian Disorders	C10.228.140.079.862
	Brain Ischaemia	C10.228.140.300.150
	Stroke, Lacunar	C10.228.140.300.275.800
	Dementia, Vascular	C10.228.140.300.400
	Infarction, Anterior Cerebral Artery	C10.228.140.300.510.200.325
	Infarction, Middle Cerebral Artery	C10.228.140.300.510.200.387

	Infarction, Posterior Cerebral Artery	C10.228.140.300.510.200.418
	Dementia, Vascular	C10.228.140.300.510.800.500
	Stroke	C10.228.140.300.775
	Alzheimer Disease	C10.228.140.380.100
	Dementia, Vascular	C10.228.140.380.230
	Epilepsy	C10.228.140.490
	Migraine Disorders	C10.228.140.546.399.750
	Parkinsonian Disorders	C10.228.662.600
	Parkinson Disease	C10.574.812
	Alzheimer Disease	C10.574.945.249
	Restless Leg Syndrome	C10.803
Male Urogenital Diseases	Prostatic Hyperplasia	C12.294.565.500
	Diabetic Nephropathies	C12.777.419.192
	Urinary Bladder, Overactive	C12.777.829.866
	Enuresis	C12.777.934.284
	Urinary Incontinence	C12.777.934.852
Female Urogenital Diseases	Urinary Bladder, Overactive	C13.351.968.829.813
	Enuresis	C13.351.968.934.252
	Urinary Incontinence	C13.351.968.934.814
Cardiovascular Diseases	Atrial Fibrillation	C14.280.067.198
	Atrial Flutter	C14.280.067.248
	Heart Failure	C14.280.434
	Myocardial Ischaemia	C14.280.647
	Atherosclerosis	C14.907.137.126.307
	Peripheral Arterial Disease	C14.907.137.126.307.500
	Coronary Artery Disease	C14.907.137.126.339
	Dementia, Vascular	C14.907.137.126.372.500
	Intermittent Claudication	C14.907.137.126.669
	Cerebral Infarction	C14.907.253.092.477.200
	Dementia, Vascular	C14.907.253.560.350.500
	Stroke	C14.907.253.855
	Embolism and Thrombosis	C14.907.355
	Pulmonary Embolism	C14.907.355.350.700
	Thromboembolism	C14.907.355.590
	Thrombosis	C14.907.355.830

	Hypertension	C14.907.489
	Myocardial Ischaemia	C14.907.585
	Peripheral Vascular Diseases	C14.907.617
Skin and Connective Tissue Diseases	Lupus Erythematosus, Systemic	C17.300.480
	Mixed Connective Tissue Disease	C17.300.540
	Rheumatic Diseases	C17.300.775
	Scleroderma, Systemic	C17.300.799
	Scleroderma, Systemic	C17.800.784
	Scleroderma, Diffuse	C17.800.784.602
	Scleroderma, Limited	C17.800.784.801
	CREST Syndrome	C17.800.784.801.500
	Psoriasis	C17.800.859.675
	Urticaria	C17.800.862.945
Nutritional and Metabolic Diseases	Diabetes Mellitus	C18.452.394.750
	Hypercholesterolemia	C18.452.584.500.500.396
	Hyperlipidaemia, Familial Combined	C18.452.584.500.500.438
	Hypertriglyceridemia	C18.452.584.500.500.851
	Hyperlipidaemia, Familial Combined	C18.452.648.398.450
Endocrine System Diseases	Diabetes Mellitus, Type 1	C19.246.267
	Diabetes Mellitus, Type 2	C19.246.300
Immune System Diseases	Anti-Neutrophil Cytoplasmic Antibody-Associated Vasculitis	C20.111.193
	Antiphospholipid Syndrome	C20.111.197
	Arthritis, Juvenile	C20.111.198
	Arthritis, Rheumatoid	C20.111.199
	Multiple Sclerosis	C20.111.258.250.500
	Diabetes Mellitus, Type 1	C20.111.327
	Hepatitis, Autoimmune	C20.111.567
	Asthma	C20.543.480.680.095
	Rhinitis, Allergic	C20.543.480.680.443

Screening eligible trials for reporting results.

Supplementary I Fig 1. The screening of eligible trials with reported results.



Screening eligible trials/papers with reported results for reporting subgroups.

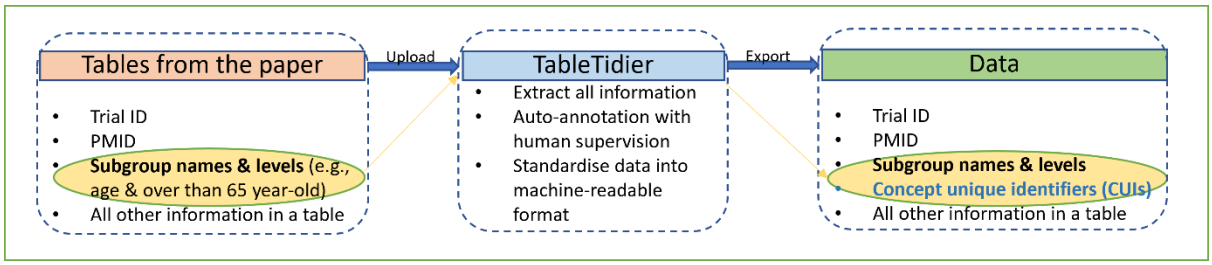
2,422 papers with reported results obtained from the above screening process were then underwent the screening of subgroups analyses showed in Figure 3 in the main paper.

Obtaining standard format for tables obtained from eligible papers.

907 papers contain subgroup reporting after screening as showed in Figure 3 in the main paper. Tables from these 907 papers in a tabular format were uploaded to TableTidier

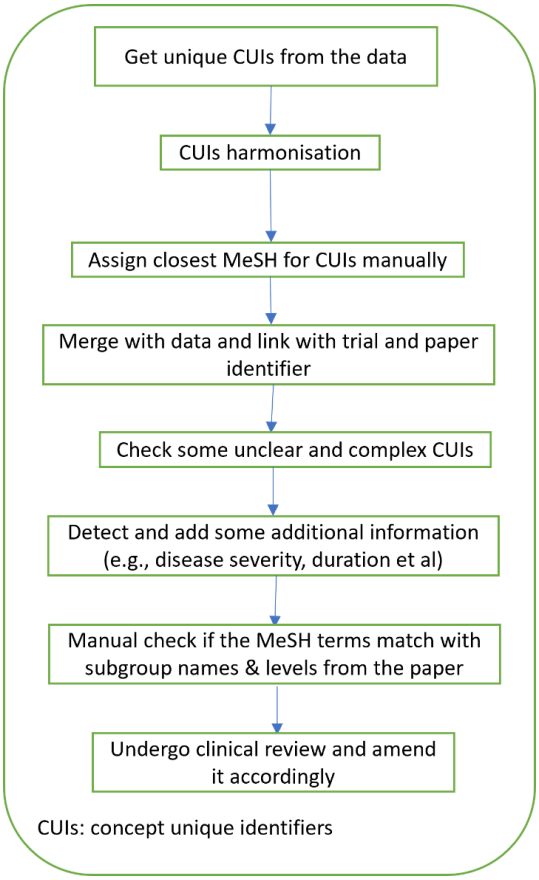
(https://tabletidier.org/) – software that can assist with standardising tables into a machine-readable format. For example, if a table contains sex as a subgroup name and woman as a subgroup level, this is assigned to the Mesh concept identifier (CUI) C0079399. Therefore, synonyms used across the papers are harmonised allowing comparisons across different papers, trials and disease conditions.

Supplementary I Fig 2. Standardisation process.



Assigning MeSH terms

Supplementary I Fig 3. Medical Subject Headings terms assignment.



Model results.

Coefficients from 3 models.

Supplementary I Table S3. Coefficients from subgroup reporting (any vs none) model.

Term	OR 95%CI
Start year	1.07 (1.03, 1.11)
Duration of follow up	1.13 (1.04, 1.24)
Number of arms > 2	1 (0.73, 1.37)
log (enrolment, base = 10)	3.48 (2.25, 5.47)
Industry1	1.58 (0.94, 2.69)
Acute Coronary Syndrome	10.44 (1.57, 210.5)
Alzheimer Disease	1.12 (0.27, 4.05)
Angina Pectoris	5.22 (0.57, 48.21)
Arthritis, Psoriatic	7.62 (0.87, 163.78)
Arthritis, Rheumatoid	1.66 (0.75, 3.73)
Atherosclerosis	5.46 (0.3, 143.72)
Atrial Fibrillation	4.26 (1.37, 14.07)
Colitis, Ulcerative	5.12 (1.39, 21.75)
Coronary Artery Disease	3.44 (1.34, 9.09)
Crohn Disease	7.06 (1.92, 30.14)
Diabetes Mellitus	4.05 (1.14, 14.81)
Diabetes Mellitus, Type 1	1.6 (0.49, 5.1)
Diabetes Mellitus, Type 2	2.44 (1.31, 4.72)
Gastroesophageal Reflux	4.19 (0.81, 24.01)
Gout	8 (0.93, 170.12)
Heart Failure	3.06 (1.08, 9.01)
Hypercholesterolemia	4.96 (2.09, 12.26)
Hypertension	2.48 (1.21, 5.22)
Idiopathic Interstitial Pneumonias	1.7 (0.31, 7.94)
Lupus Erythematosus, Systemic	8.85 (1.18, 181.81)
Migraine Disorders	1.1 (0.15, 5.07)
Multiple Sclerosis	0.53 (0.03, 4.17)
Myocardial Infarction	9.86 (2.94, 40.48)
Osteoarthritis	0.85 (0.24, 2.63)
Osteoporosis	2.45 (0.82, 7.34)
Parkinson Disease	2.18 (0.49, 8.84)
Peripheral Arterial Disease	2.79 (0.15, 77.92)
Prostatic Hyperplasia	3.92 (1.19, 13.71)
Psoriasis	0.84 (0.31, 2.2)
Pulmonary Disease, Chronic Obstructive	1.24 (0.6, 2.61)
Retinal Vein Occlusion	3.48 (0.13, 95.06)
Rhinitis	0.91 (0.13, 4.18)
Seizures	3.85 (0.99, 15.59)
Spondylitis, Ankylosing	1.87 (0.34, 8.97)

Stroke	3.85 (0.77, 22.09)
Urinary Bladder, Overactive	0.15 (0.01, 0.88)
Urticaria	7.49 (0.67, 168.83)
Venous Thromboembolism	6.99 (0.86, 150.52)
Venous Thrombosis	1.87 (0.22, 13.33)

Supplementary I Table S4. Coefficients from number of subgroups model.

Term	OR 95%CI
Start year	1.02 (1.02, 1.02)
Number of arms > 2	1.01 (0.99, 1.04)
log (enrolment, base = 10)	1.69 (1.65, 1.73)
Industry1	1 (0.97, 1.03)
Duration of follow up	1.03 (1.02, 1.03)
Acute Coronary Syndrome	1.43 (1.25, 1.64)
Alzheimer Disease	0.72 (0.4, 1.19)
Angina Pectoris	1.43 (0.94, 2.1)
Arthritis, Psoriatic	0.92 (0.68, 1.23)
Arthritis, Rheumatoid	1.45 (1.26, 1.67)
Atherosclerosis	2.12 (1.83, 2.46)
Atrial Fibrillation	2.11 (1.86, 2.4)
Colitis, Ulcerative	1.64 (1.38, 1.95)
Coronary Artery Disease	2.31 (2.06, 2.6)
Crohn Disease	3.09 (2.72, 3.53)
Diabetes Mellitus	1.2 (0.94, 1.51)
Diabetes Mellitus, Type 1	0.97 (0.72, 1.27)
Diabetes Mellitus, Type 2	2.3 (2.05, 2.58)
Esophagitis	0.29 (0, 2.11)
Gastroesophageal Reflux	0.23 (0.04, 0.69)
Gout	0.99 (0.72, 1.31)
Heart Failure	2.09 (1.85, 2.37)
Hypercholesterolemia	2.63 (2.34, 2.96)
Hypertension	1.87 (1.66, 2.12)
Idiopathic Interstitial Pneumonias	0.89 (0.5, 1.46)
Ischemic Attack, Transient	0.83 (0.45, 1.38)
Lupus Erythematosus, Systemic	1.04 (0.74, 1.43)
Lupus Nephritis	1.36 (0.81, 2.13)
Migraine Disorders	0.45 (0.08, 1.34)
Multiple Sclerosis	0.97 (0.5, 1.68)
Myocardial Infarction	2.19 (1.96, 2.47)
Osteoarthritis	1.39 (1.11, 1.73)
Osteoporosis	1.28 (1.08, 1.51)
Parkinson Disease	1 (0.7, 1.4)
Peripheral Arterial Disease	1.65 (1.44, 1.88)
Prediabetic State	0.87 (0.7, 1.07)
Prostatic Hyperplasia	2.85 (2.45, 3.31)
Psoriasis	1.12 (0.95, 1.33)
Pulmonary Disease, Chronic Obstructive	1.45 (1.29, 1.64)

Pulmonary Embolism	2.58 (1.92, 3.4)
Retinal Vein Occlusion	0.48 (0.14, 1.19)
Rhinitis	0.31 (0.01, 1.47)
Seizures	1.01 (0.75, 1.34)
Spondylarthropathies	1.26 (0.65, 2.18)
Spondylitis, Ankylosing	1.27 (0.91, 1.73)
Stroke	1.93 (1.64, 2.27)
Thromboembolism	1.67 (1.04, 2.52)
Urinary Bladder, Overactive	0.27 (0, 1.99)
Urticaria	2.84 (2.15, 3.7)
Venous Thromboembolism	1.43 (1.23, 1.68)
Venous Thrombosis	0.81 (0.63, 1.03)

Supplementary I Table S5. Coefficients from results reporting (any vs none) model.

Term	OR 95%CI
Start year	0.97 (0.95, 0.99)
Duration of follow up	1.1 (1.03, 1.18)
Number of arms > 2	1.42 (1.15, 1.74)
log (enrolment, base = 10)	1.63 (1.22, 2.19)
Industry1	1.03 (0.73, 1.45)
Acute Coronary Syndrome	0.93 (0.35, 2.43)
Alzheimer Disease	1.15 (0.51, 2.56)
Arthritis, Psoriatic	4.2 (0.58, 84.42)
Arthritis, Rheumatoid	1.71 (1, 2.94)
Atherosclerosis	0.54 (0.11, 2.24)
Atrial Fibrillation	1.03 (0.49, 2.21)
Brain Ischemia	0.61 (0.02, 16.09)
Cerebral Infarction	0.42 (0.09, 1.54)
Colitis, Ulcerative	7.26 (1.87, 48.04)
Coronary Artery Disease	1.2 (0.65, 2.23)
Crohn Disease	7.85 (2.04, 51.75)
Diabetes Mellitus	0.76 (0.34, 1.66)
Diabetes Mellitus, Type 1	1.37 (0.64, 2.93)
Diabetes Mellitus, Type 2	1.41 (0.95, 2.08)
Diabetic Nephropathies	0.4 (0.02, 3.2)
Enuresis	0.51 (0.02, 4.08)
Esophagitis	0.13 (0.01, 0.72)
Gastroesophageal Reflux	0.48 (0.18, 1.18)
Gout	0.63 (0.16, 2.21)
Heart Failure	1.96 (0.92, 4.32)
Hypercholesterolemia	1.7 (0.95, 3.07)
Hyperlipidemias	1.01 (0.13, 6.37)
Hypertension	0.94 (0.6, 1.45)
Lupus Erythematosus, Systemic	1.8 (0.42, 9.18)
Lupus Nephritis	0.31 (0.02, 2.6)
Migraine Disorders	1.87 (0.72, 4.98)
Multiple Sclerosis	2.46 (0.5, 17.89)

Myocardial Infarction	1.8 (0.85, 3.9)
Osteoarthritis	0.94 (0.49, 1.77)
Osteoporosis	1.22 (0.59, 2.52)
Parkinson Disease	0.62 (0.26, 1.4)
Peripheral Arterial Disease	0.75 (0.13, 4.48)
Prostatic Hyperplasia	1.56 (0.68, 3.59)
Psoriasis	1.68 (0.9, 3.17)
Pulmonary Disease, Chronic Obstructive	1.48 (0.93, 2.36)
Pulmonary Embolism	0.82 (0.03, 21.15)
Restless Legs Syndrome	1.55 (0.46, 5.23)
Retinal Vein Occlusion	1.31 (0.15, 11.48)
Rhinitis	0.48 (0.21, 1.03)
Seizures	0.68 (0.29, 1.56)
Spondylitis, Ankylosing	1.38 (0.46, 4.2)
Stroke	1.45 (0.43, 5.26)
Thromboembolism	0.52 (0.02, 5.61)
Urinary Bladder, Overactive	0.78 (0.35, 1.67)
Venous Thromboembolism	0.26 (0.08, 0.68)
Venous Thrombosis	0.31 (0.1, 0.87)

Supplementary II – Trial Calibration

Regression-based method

Regression-based method implementation - COMET and Heart Failure (HF)

The selected code is provided below.

Step 1: Clean and transform variables.

a) Continuous variables standardisation for both COMET and HF registry.

As showed in formula 4 below, age, systolic blood pressure (SBP), heart rate, sodium, eGFR, loop diuretics were scaled by dividing by the standard deviation (SD) after subtracting the mean (both SD and mean referring to the HF registry).

Standardized age in the *trial* = $(X_{age\_Trial} - \overline{X}_{age\_Trial})/\sigma_{age\_Trial.}$

Standardized age in the *register* = $(X_{age\_Register} - \overline{X}_{age\_Trial})/\sigma_{age\_Trial.}$

X_{age}: age values.

$\overline{X}_{age}$: mean age.

σ_{age}: standard deviation of age.

(Formula 4)

Men sex and history of diabetes were categorical variables so did not need to be standarised.

b) Select follow-up time.

The mean trial duration is 58 months. For participants who did not experience death, the time to last contact ranged from 5 to 2175 days with the 1st quantile at 1570 days (4.30 years). Therefore, we selected 4 years as the follow-up time. Set deaths that occurred after 4 years as censored.

Step 2: Build the parametric survival model for COMET and obtain the coefficients.

a) Fit parametric survival model and determine best fitting distribution.

We fit the parametric survival model with treatment interactions by using *"Weibull", "Generalised gamma", "Exponential", "Log-logistic", "Log-normal", "Gompertz"* distribution one at a time. The *"Generalised gamma"* distribution had the best fit based on the visual inspection and Akaike Information Criterion and it was selected.

b) Check linearity assumption for continuous variables in the model.

This is conducted based on visual inspection and clinical judgment and all continuous variables look linear.

R code for step 2a and 2c

```
comet_regression <- flexsurvreg(Surv(time, status) ~ treat + age + men + sbp + heartrate +
loop_diuretics + diabetes + egfr + sodium + treat*age + treat*men + treat*sbp + treat* heartrate +
treat* loop_diuretics + treat*diabetes + treat*egfr + treat*sodium, dist = "gengamma", data =
comet)
coef_comet <- coef(comet_regression)
vcov_comet <- vcov(comet_regression)
```

c) Obtain the coefficients, variance and covariance matrix from the model.

Step 3: Build the parametric survival model for HF registry and obtain the coefficients.

a) Build the same model as COMET in HF Registry.

This model in HF registry also used *"Generalised gamma"* distribution without including treatment and treatment interactions.

b) Obtain the coefficients, variance and covariance matrix from the model.

R code for step 3a and 3b

```
Register_regression <- flexsurvreg(Surv(time, status) ~ age + male + sbp + heartrate +
loop_diuretics + diabetes + egfr + sodium, dist = "gengamma", data = Registry)
coef_Registry <- coef(Registry_regression)
vcov_Registry<- vcov(Registry_regression)
```

Step 4: Generate samples based on the coefficients, variance and covariance matrix from above.

a) Samples generation from HF registry model.

Generate samples from a multivariate normal distribution with the means equal to the coefficients and the variances corresponding to the variance-covariance matrix. This will generate two matrices, each containing 200 rows (can be more, here generate 200 for computational reasons), and the number of columns will match the number of coefficients in the model.

b) Samples generation from COMET model.

Repeat the same as from registry model.

c) Relevant coefficients selection.

Select the scale parameter (mu), parameters of covariate main effects (such as age, SBP) from samples of the HF Registry (see Supplementary II Table S2) and parameters of treatment arm and treatment arm-covariate interactions from samples for the COMET model (Table S1). Combine these 2 sets of parameters into a single matrix with 200 rows and 18 columns. The distribution of each column indicates the uncertainty for each coefficient and the joint distribution of any 2 (or more) columns indicates the joint uncertainty across coefficients.

Step 5: Calculate the risk of death in HF registry.

a) Set the "treatment variable" for HF registry.

For HF Registry individual patient data (IPD), set "treat" equal to 1 to model the allocation of all patients to carvedilol and "treat" equal to 0 to model the allocation of

R code for step 5a

```
Registry_carvedilol <- Registry %>%
  mutate(treat=1)
mtrx_covs_carvedilol <- model.matrix(~ treat + age + male + sbp + heart + loop_diuretics +
diabetes + egfr + sodium + treat*age + treat*male + treat*sbp + treat*heart + treat* loop_diuretics
+ treat*diabetes + treat*egfr + treat*sodium, data = Registry_carvedilol)
Registry_metoprolol <- Registry %>%
  mutate(treat=0)
mtrx_covs_metoprolol <- model.matrix(~ treat + age + male + sbp + heart + loop_diuretics +
diabetes + egfr + sodium + treat*age + treat*male + treat*sbp + treat*heart + treat* loop_diuretics
+ treat*diabetes + treat*egfr + treat*sodium, data = Registry_metoprolol)
```

all patients to metoprolol. Then create a matrix to be in the same format as the simulated parameters in the trial as showed in Table S1. Some rows of this new matrix for HF Registry under the carvedilol allocation (treat = 1) are showed in Table S2 as an example.

b) Calculate the linear predictor by multiplication.

R code for step 5b

```
cell_carvedilol <- mtrx_coef_ac %*% t(mtrx_covs_carvedilol)
cell_metoprolol <- mtrx_coef_ac %*% t(mtrx_covs_metoprolol)
```

Multiply matrix obtained from step 4c and step 5a. Each cell is the sum of the product of each coefficient from step 4c and each covariate level in the HF Registry from step 5a. Each column represents one patient.

c) Calculate the probability of getting the primary endpoint.

The probability of death can be obtained by combining the follow-up time, the cell value above and scale parameters (σ and Q) from the parametric survival model of HF Registry with the cumulative distribution function of the generalised gamma distribution (implemented as pgengamma" in "flexsurv" package). This produces a 200-estimates of the predicted risk on carvedilol for each patient and 200-estimates of the predicted risk on metoprolol for each patient, which means for each patient in each treatment allocation, there are 200 predictions of the risk of death.

R code for step 5c

```
prob_carvedilol <- pgengamma (4, cell_carvedilol, sigma = exp (0.58), Q = 0.67)
prob_metoprolol <- pgengamma (4, cell_metoprolol, sigma = exp (0.58), Q = 0.67)
```

d) repeat these steps 500 times.

Repeat the process from step 4 to step 5c for 500 times to obtain 100,000 samples of the probability of death for each intervention group. Note this was done in batches of 200 for computational reasons. Matrices with more rows could be used if more computer memory is available.

e) Obtain the odds ratio (OR), absolute risk reduction (ARR) and risk in each intervention arm.

For each sample sum the risk across patients to obtain the risks in each arm. Then calculate the absolute risk differences (by subtraction) and the odds ratios (by

transforming these risks to odds and then division). For the resultant estimates take the mean of the 100,000 samples as well as the2.5th and 97.5th centiles as the point estimate and 95% confidence intervals respectively.

Regression-based method implementation - DIG and HF Registry

The process is the same as the process in HF Registry and COMET except:

1. In DIG, there are no sodium and loop_diuretics variables.
2. The follow-up time is *3 years, events that occurred after 3 years are set as censored*.
3. SBP and eGFR showed evidence of departure from linearity, so SBP was parameterised as follows: high SBP is obtained by SBP subtracting 130 if SBP is over than 130 mmHg; low BP is calculated by 120 subtracting SBP if SBP is less than 120. For the values of eGFR which are over than 90, they are set as 90.

Supplementary II Table S1. Example of the combined matrix in the trial.

mu*	treat	age *	Men *	SBP *	heartrate *	Loop_diuretics *	Diabetes *	eGFR *	Sodium *	treat by age	treat by men	treat by SBP	treat by heartrate	treat by loop_diuretics	treat by diabetes	treat by eGFR	treat by sodium
0.02	-0.04	-0.1	-0.16	-0.22	-0.28	0.03	0.34	0.65	0.03	0.34	1.58	0.03	0.34	0.06	-0.05	0.04	0.03
0.26	0.03	-0.2	-0.43	-0.66	-0.89	0.21	1.31	2.41	0.21	0.03	0.34	0.21	1.31	-0.46	-0.94	-0.02	0.03
0.5	0.02	-0.46	-0.94	0.5	0.02	-0.46	-0.94	0.5	0.02	0.21	0.03	0.34	0.01	0.05	0	-0.01	-0.02
0.98	0.26	-0.46	-1.18	-1.9	-2.62	-3.34	0.01	0	0	0.03	0.21	1.31	-0.02	0.06	-0.15	-0.01	-0.01
1.22	0.5	-0.22	-0.94	-1.66	0.5	0.02	-0.46	-0.94	-0.01	0.03	0.03	0.34	0	0.02	-0.05	0.01	0.02
1.46	0.74	0.02	-0.7	-0.7	0.5	0.02	-0.46	-0.94	0	0.02	0.21	1.31	-0.01	0	-0.02	-0.01	-0.01
0.03	0.98	0.05	-0.88	-0.88	4.78	5.73	-0.01	0	-0.01	0.02	0	0.06	0.5	0.02	-0.46	-0.94	-0.02

*these variables are from HF Registry, the rest are from COMET. SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate.

Supplementary II Table S2. Example of the matrix in HF Registry (carvedilol treatment group, treat = 1)

Intercept	treat	age	men	SBP	heartrate	loop_diuretics	diabetes	eGFR	sodium	treat by age	treat by men	treat by SBP	treat by heartrate	treat by loop_diuretics	treat by diabetes	treat by eGFR	treat by sodium
1	1	0.25	1	0.25	1.47	1.471	1	1.42	1.22	1.37	1.17	1.03	0.77	0.69	0.43	0.35	0.09
1	1	0.36	0	0.36	-0.98	-0.977	1	-1.03	-1.23	-1.08	-1.28	-1.42	-1.68	-1.76	-2.02	-2.1	-2.36
1	1	0.05	0	0.05	-1.19	-1.185	1	-1.24	-1.44	-1.29	-1.49	-1.63	-1.89	-1.97	-2.23	-2.31	-2.57
1	1	0.36	1	0.36	0.38	0.375	0	0.33	0.13	0.28	0.08	-0.06	-0.32	-0.4	-0.66	-0.74	-1
1	1	0.68	0	0.68	-0.3	-0.301	1	-0.35	-0.55	-0.4	-0.6	-0.74	-1	-1.08	-1.34	-1.42	-1.68
1	0	0.04	0	0.04	1.05	1.051	0	1	0.8	0.95	0.75	0.61	0.35	0.27	0.01	-0.07	-0.33
1	0	0.25	1	0.25	-1.71	-1.705	0	-1.76	-1.96	-1.81	-2.01	-2.15	-2.41	-2.49	-2.75	-2.83	-3.09

SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate.

Inverse Odds of Sampling Weights (IOSW)

IOSW method description.

The odds of inclusion was estimated using a logistic regression model where the numerator was the number of trial participants and the denominator was the number of registry patients. Based on this model and their individual covariate level, we obtained an inclusion probability for each trial participant. The effect of treatment was then estimated by fitting a weighted logistic regression model of the outcome on the treatment effect within the trial data having re-weighted the contribution of each participant to account for their inclusion probability. Compared to an unweighted model, the IPSW model gives additional weight to randomised participants who were under-represented in the trial compared to the HF registry and give less weight to participants who were over-represented in the trial, compared to the HF registry.

IOSW method implementation.

We have used COMET (all-cause death) and the HF Registry as an example to describe the method. Selected R code is provided below.

Step 1: Data aggregation and merging.

a) Data aggregation.

Aggregate COMET and HF Registry data separately in each safe haven with the same variables chosen as regression-based method. The group intervals are as Table S3.

Supplementary II Table S3. Group intervals.

Age (year)	SBP (mmHg)	Heart rate (beats/min)	Sodium (mmol/L)	eGFR (ml/min/1.73m^2)	Loop_diuretics (mg)
<40	<100	<50	<135	<30	<40
[40,60)	[100,140)	[50,70)	[135,145)	[30,60)	[40,80)
[60,80)	>=140	[70,100)	>=145	>=60	>=80
>=80		>=100			
SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate.					

b) Data merging.

R code for step 1b

```
ath_com <- registry_agg %>%
  left_join(comet_agg) %>%
  mutate(x = if_else(is.na(x), 0L, x))
```

c) Merge two aggregated datasets as Table S4 (only showing the first 8 rows as an example), n and x represent the number of each combination of characteristics in HF Registry and COMET respectively.

Supplementary II Table S4. Example of merged data.

age	men	SBP	heartrate	eGFR	loop_diuretics	diabetes	sodium	n	x
[40,60)	0	[50,100)	[50, 70)	[30, 60)	[20, 40)	0	[90,135)	108	42
[40,60)	1	[50,100)	[50, 70)	[30, 60)	[20, 40)	0	[135,145)	130	78
[40,60)	0	[100,140)	[50, 70)	[30, 60)	[40, 80)	0	[90,135)	52	19
[40,60)	0	[50,100)	[70,100)	[30, 60)	[40, 80)	0	[135,145)	64	54

[40,60)	1	[50,100)	[50, 70)	[30, 60)	[40, 80)	0	[145,170]	37	26
[40,60)	0	[100,140)	[50, 70)	[30, 60)	[40, 80)	1	[90,135)	150	99
[40,60)	0	[50,100)	[70,100)	[30, 60)	[40, 80)	1	[145,170]	88	55
[40,60)	1	[100,140)	[50, 70)	[30, 60)	[40, 80)	0	[145,170]	32	7
SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate.									

d) Set up the mid-point value for the group interval.

For continuous variables, set the mid-point value. E.g., if the age interval is between 40 to 60, the mid-point value is 50.

Step 2: Build logistic regression models.

a) Fit a logistic regression model.

To estimate the inclusion probability in COMET if participants were selected from HF Registry, build a logistic regression model based on the merged data and mid-point value.

R code for step 2a

```
model<- glm(cbind(x,n)~agevalue + male + sbpvalue + heartvalue + sodiumvalue +
        egfrvalue + loop_diureticsvalue + diabetes, data = ath_com, family = binomial)
summary(model)
round (coef (model),2) %>% dput ()
```

b) Estimate the odds of inclusion.

After building the model as above, the coefficients can be derived. Then calculate the sum of the product of each coefficient and each covariate, transform this value into odds, and compute the weight (1/odds).

R code for step 2b

```
coefs <- c(`(Intercept)` = -8.73, agevalue = -0.01, malevalue = 1, sbpvalue = -0.01, heartvalue = 0.06, sodiumvalue = 0.21, egfrvalue = 0.13, loop_diureticsvalue = 0.01, diabetes = -0.01)
comet<- comet %>%
  mutate (sum = coefs["Intercept"] + coefs["agevalue"] * age + coefs["malevalue "] * male + coefs["sbpvalue"] * sbp + coefs["heartvalue"] * heartrate + coefs["sodiumvalue "] * sodium + coefs["egfrvalue"] * egfr + coefs["loop_diureticsvalue "] * fusemide + coefs["diabetes"] * diabetes) %>%
  mutate (odds = exp(sum), weights = 1/odds)
```

c) Build a weighted logistic regression model.

Using the inverse of the odds in step 2b as the weighting variable, perform a weighted logistic regression of death on treatment allocation.

R code for step 2c

```
comet_survey <- svydesign (id=~1, weights = ~weights, data = comet)
model_weighted <- svyglm(status~I(tmt==1), design= comet_survey, family=binomial)
summary(model_weighted)
```

Step 3: Calculate the risk of death in HF registry.

a) Obtain the OR and risk in each intervention arm.

From the weighted model above, the coefficients obtained are on the log scale. Exponentiate these to obtain the odds ratio. For the unweighted and weighted models, the risk in each treatment arm was calculated by applying the inverse link function (logistic) to the linear predictor. The absolute risk reduction was estimated by re-fitting the model using an identity link and gaussian likelihood and obtaining the treatment effect estimate.

The implementation of IOSW in DIG followed this procedure.

Coefficients and plots

Regression-based method

Coefficients for main effects in both registry and trials, and treatment interactions in trials.

Supplementary II Table S5. Main effects in HF Registry and 2 trials.

	HF registry vs COMET (all-cause death)						HF registry vs COMET (all-cause death or all-cause hospitalisation)						HF registry vs DIG (all-cause death)						HF registry vs DIG (death or hospitalisation due to worsening heart failure)					
	HF registry			COMET			HF registry			COMET			HF registry			DIG			HF registry			DIG		
	coefficients	standard error	AFT (95% CI)	coefficients	standard error	AFT (95% CI)	coefficients	standard error	AFT (95% CI)	coefficients	standard error	AFT (95% CI)		coefficients	standard error	AFT (95% CI)	coefficients	standard error	AFT (95% CI)	coefficients	standard error	AFT (95% CI)	coefficients	standard error	AFT (95% CI)
mu	2.78	0.05	--	3.19	0.12	--	0.39	0.06	--	0.9	0.1	--		2.49	0.04	--	2.82	0.08	--	0.06	0.18	--	8.6	0.1	--
sigma	0.25	0.02	--	0.4	0.1	--	0.59	0.02	--	0.59	0.05	--		0.25	0.02	--	0.56	0.08	--	1.46	0.07	--	0.86	0.1	--

Q	0.8	0.04	--	0.58	0.11	--	0.67	0.04	--	0.35	0.07	--		0.8	0.04	--	0.22	0.09	--	-1.92	0.09	--	0.08	0.09	--
History of diabetes	-0.29	0.04	0.75 (0.69, 0.81)	-0.32	0.09	0.73 (0.61, 0.87)	-0.32	0.05	0.72 (0.65, 0.80)	-0.45	0.08	0.64 (0.54, 0.75)		-0.34	0.04	0.71 (0.66, 0.77)	-0.4	0.06	0.67 (0.59, 0.76)	-0.51	0.12	0.60 (0.47, 0.76)	-0.7	0.08	0.50 (0.43, 0.58)
Age*	-0.53	0.02	0.59 (0.57, 0.61)	-0.46	0.05	0.63 (0.57, 0.7)	-0.21	0.02	0.81 (0.77, 0.84)	-0.34	0.04	0.71 (0.65, 0.77)		-0.49	0.02	0.61 (0.59, 0.64)	-0.26	0.03	0.77 (0.72, 0.82)	-0.1	0.05	0.91 (0.82, 1.00)	-0.18	0.04	0.84 (0.77, 0.91)
Loop diuretics*	-0.2	0.02	0.82 (0.78, 0.86)	-0.11	0.04	0.9 (0.84, 0.96)	-0.18	0.03	0.84 (0.79, 0.89)	-0.09	0.03	0.91 (0.85, 0.97)		--	--	--	--	--	--	--	--	--	--	--	--
Men	-0.1	0.04	0.91 (0.84, 0.98)	-0.46	0.11	0.65 (0.53, 0.81)	0.07	0.04	1.07 (0.98, 1.17)	-0.09	0.09	0.91 (0.76, 1.09)		-0.12	0.04	0.89 (0.83, 0.96)	-0.4	0.07	0.67 (0.58, 0.77)	0.09	0.1	1.09 (0.89, 1.34)	-0.15	0.09	0.86 (0.72, 1.02)
Heart rate*	-0.06	0.02	0.94 (0.91, 0.98)	-0.05	0.04	0.95 (0.88, 1.03)	-0.03	0.02	0.97 (0.92, 1.01)	-0.01	0.04	0.99 (0.92, 1.07)		-0.06	0.02	0.94 (0.91, 0.97)	-0.16	0.03	0.86 (0.81, 0.91)	-0.04	0.05	0.96 (0.88, 1.06)	-0.36	0.04	0.70 (0.65, 0.75)
Sodium*	0.1	0.02	1.11 (1.07, 1.14)	0.28	0.04	1.32 (1.22, 1.43)	0.13	0.02	1.14 (1.10, 1.19)	0.18	0.04	1.20 (1.11, 1.29)		--	--	--	--	--	--	--	--	--	--	--	--

eGFR*	0.12	0.02	1.13 (1.09, 1.18)	0.3	0.05	1.35 (1.22, 1.49)	0.13	0.02	1.14 (1.09, 1.19)	0.21	0.04	1.23 (1.13, 1.34)		0.14	0.02	1.15 (1.11, 1.19)	0.31	0.03	1.37 (1.29, 1.46)	0.09	0.04	1.10 (1.00, 1.19)	0.41	0.04	1.50 (1.39, 1.63)
SBP*	0.14	0.02	1.15 (1.11, 1.18)	0.31	0.05	1.37 (1.25, 1.49)	0.16	0.02	1.17 (1.13, 1.22)	0.23	0.04	1.26 (1.17, 1.36)	Low SBP*	-0.1	0.01	0.91 (0.88, 0.93)	-0.25	0.03	0.78 (0.73, 0.82)	-0.23	0.04	0.80 (0.74, 0.86)	-0.34	0.04	0.71 (0.66, 0.77)
													High SBP*	0.08	0.02	1.09 (1.04, 1.13)	0.07	0.03	1.08 (1.01, 1.15)	0.21	0.05	1.23 (1.11, 1.37)	0.06	0.04	1.06 (0.98, 1.14)

SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate; Diabetes and Men are categorical variables; others (with *) are all numerical variables and scaled by dividing by the standard deviation (SD) after subtracting the mean (both SD and mean referring to the trial); AFT: accelerated failure time ratio with 95% confidence interval; --: Not available.

Supplementary II Table S6. Treatment interactions in each trial.

	COMET (all-cause death)		COMET (all-cause death or all-cause hospitalisation)		DIG (all-cause death)			DIG (death or hospitalisation due to worsening heart failure)	
	coefficients	HR (95%CI)	coefficients	HR (95%CI)		coefficients	HR (95%CI)	coefficients	HR (95%CI)
treatment	0.16	1.17 (0.78, 1.76)	0.3	1.35 (0.97, 1.87)		-0.17	0.85 (0.65, 1.11)	0.29	1.33 (0.96, 1.84)

treatment by men	0.11	1.11 (0.73, 1.70)	-0.26	0.76(0.53, 1.07)		0.27	1.31 (0.99, 1.74)	0.49	1.63 (1.15, 2.31)
treatment by age	0.03	1.03 (0.84, 1.26)	-0.09	0.92(0.84, 1.01)		0.02	1.02 (0.89, 1.15)	-0.02	0.98 (0.84, 1.15)
treatment by Diabetes	-0.05	0.95 (0.66, 1.37)	-0.07	0.93(0.67, 1.29)		-0.1	0.90 (0.71, 1.15)	-0.22	0.81 (0.59, 1.10)
treatment by loop diuretics	-0.07	0.94 (0.82, 1.07)	-0.03	0.97(0.84, 1.11)		--	--	--	--
treatment by sodium	-0.14	0.87 (0.74, 1.01)	0.01	1.01(0.87, 1.16)		--	--	--	--
treatment by eGFR	0.23	1.25 (1.03, 1.52)	0.05	1.05(0.89, 1.23)		-0.04	0.97 (0.85, 1.09)	-0.12	0.89 (0.76, 1.04)
treatment by SBP	0.12	1.13 (0.95, 1.34)	0.07	1.07(0.93, 1.24)	treat by low SBP	-0.05	0.95 (0.85, 1.06)	0.08	1.08 (0.94, 1.26)
					treat by high SBP	-0.09	0.92 (0.81, 1.03)	-0.12	0.89 (0.76, 1.03)
treatment by heart	0.06	1.06 (0.90, 1.25)	0.12	1.13(0.98, 1.31)		0.07	1.07 (0.96, 1.20)	0.12	1.12 (0.97, 1.30)

--: Not available; SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate; Diabetes and Men are categorical variables; others are all numerical variables and scaled by dividing by the standard deviation (SD) after subtracting the mean (both SD and mean referring to the trial).

Supplementary II Table S7. Contribution of each covariate to the results differ between uncalibrated and regression-based method.

	COMET (all-cause death)				COMET (all-cause death or hospitalisation)				DIG (all-cause death)					DIG (death or hospitalisation due to worsening heart failure)			
	coefficients	mean of each covariate in HF registry	mean of each covariate in COMET	mean difference * coefficients	coefficients	mean of each covariate in HF registry	mean of each covariate in COMET	mean difference * coefficients		coefficients	mean of each covariate in HF registry	mean of each covariate in DIG	mean difference * coefficients	coefficients	mean of each covariate in HF registry	mean of each covariate in DIG	mean difference * coefficients
treatment	0.16	--	--	--	0.3	--	--	--		-0.17	--	--	--	0.29	--	--	
treatment by male	0.11	0.61	0.8	-0.02	-0.26	0.61	0.8	0.05		0.27	0.61	0.78	-0.05	0.49	0.61	0.78	-0.08
treatment by age	0.03	1	0	0.03	-0.09	1	0	-0.09		0.02	0.9	0	0.02	-0.02	0.9	0	-0.02
treatment by Diabetes	-0.05	0.23	0.24	0	-0.07	0.23	0.24	0		-0.1	0.23	0.28	0.01	-0.22	0.23	0.28	0.01
treatment by loop diuretics	-0.07	0.9	0	-0.06	-0.03	0.9	0	-0.03		--	--	--	--	--	--	--	--
treatment by sodium	-0.14	-0.38	0	0.05	0.01	-0.38	0	0		--	--	--	--	--	--	--	--

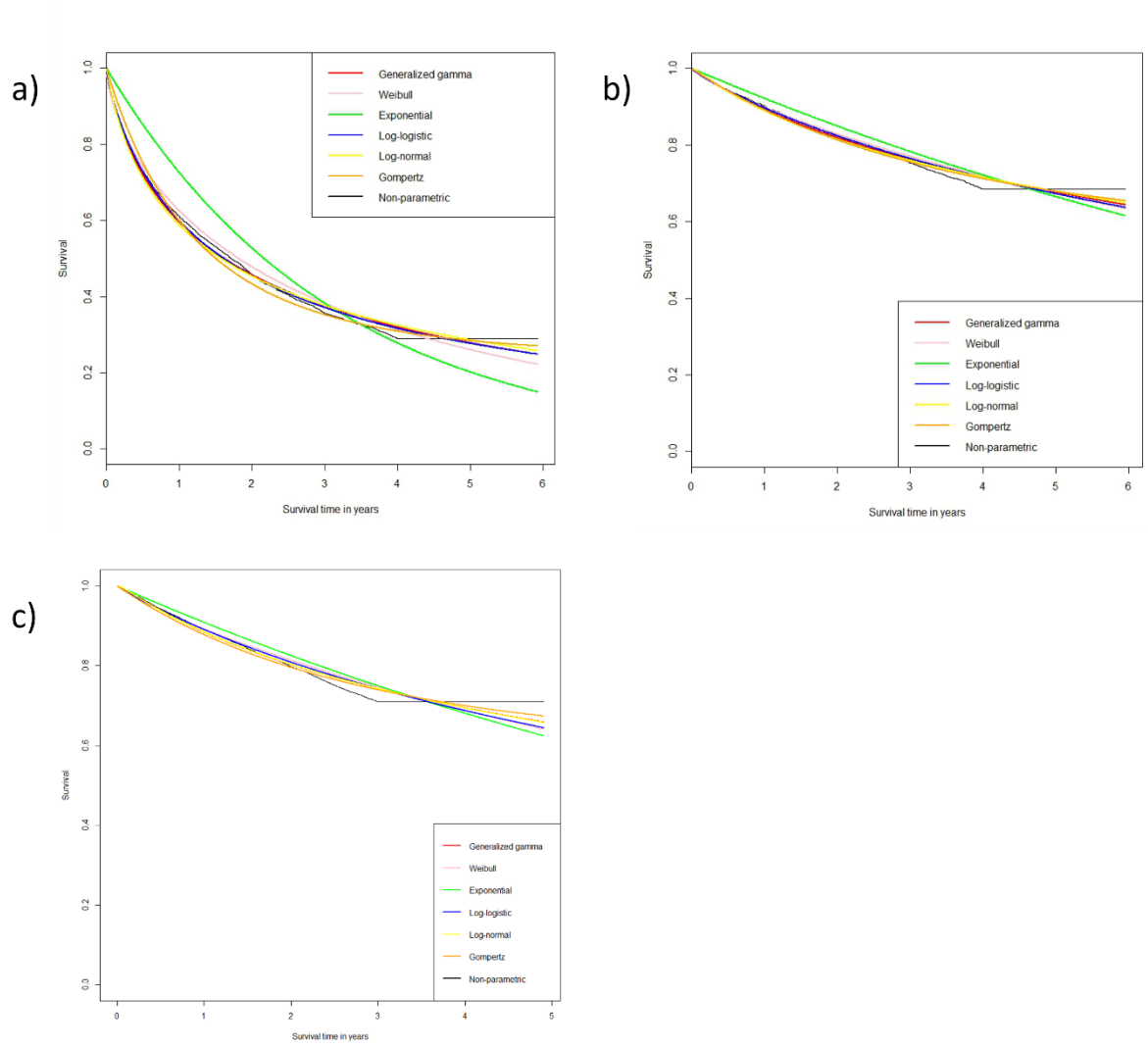
treatment by eGFR	0.23	-0.39	0	-0.09	0.05	-0.39	0	-0.02		-0.04	-0.2	0	0.01	-0.12	-0.2	0	0.02
treatment by SBP	0.12	-0.33	0	-0.04	0.07	-0.33	0	-0.02	treat by low SBP	-0.05	0.43	0	-0.02	0.08	0.43	0	0.03
									treat by high SBP	-0.09	-0.12	0	0.01	-0.12	-0.12	0	0.01
treatment by heart	0.06	-0.59	0	-0.04	0.12	-0.59	0	-0.07		0.07	-0.44	0	-0.03	0.12	-0.44	0	-0.05

--: Not available; SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate; Diabetes and Male are categorical variables, others are all numerical variables and scaled by dividing by the standard deviation (SD) after subtracting the mean (both SD and mean referring to the HF registry); the means in this table are the standardised mean; * multiply

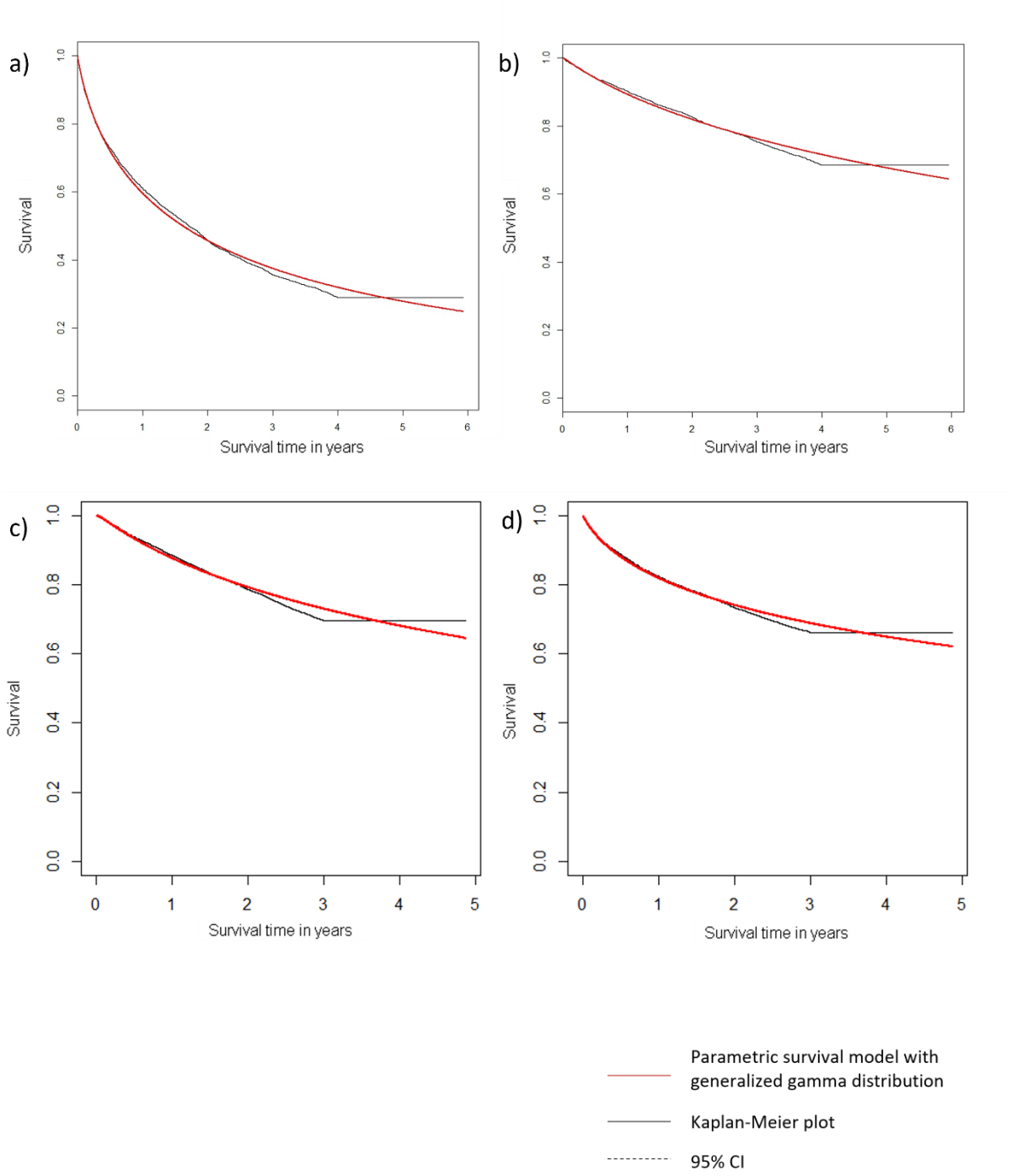
Plots for model fit in trials.

The parametric survival model with treatment interactions by using Generalised gamma, Weibull, Exponential, Log-normal, Log-logistic, Gompertz distribution was built respectively in each trial and was displayed in Fig 1. In the trial, the generalised gamma distribution lines closer with the non-parametric compared with other distributions and was then used for the HF register.

Supplementary II Fig 1. Parametric survival model with different distributions in a) COMET all-cause death or hospitalisation; b) COMET all-cause death; c) DIG all-cause death.



Supplementary II Fig 2. Visualisation of model fit (with generalised gamma distribution) for trials for a) COMET all-cause death or hospitalisation; b) COMET all-cause death; c) DIG all-cause death; d) DIG death or hospitalisation due to worsening HF.



Inverse Odds of Sampling Weights (IOSW)

Coefficients from regression model for estimating inclusion odds.

Supplementary II Table S8. Coefficients from regression model for generating inclusion of odds.

	HF registry & COMET (all-cause death)			HF registry & COMET (all-cause death or all-cause hospitalisation)			HF registry & DIG (all-cause death)			HF registry & DIG (death or hospitalisation due to worsening heart failure)		
	overall data	lowest risk decile	highest risk decile	overall data	lowest risk decile	highest risk decile	overall data	lowest risk decile	highest risk decile	overall data	lowest risk decile	highest risk decile
Intercept	-4.85	-13.05	6.59	-4.85	-3.31	-7.86	0.44	-3.95	-12.86	0.44	4.85	-2.43
age	-0.05	0.06	-0.25	-0.05	0.02	-0.13	-0.06	0.07	-0.21	-0.06	-0.01	-0.09
men	1.02	1.07	0.65	1.02	0.9	1.3	0.78	0.91	0.3	0.78	0.69	0.97
SBP	0.01	0	0.03	0.01	0	0.03	0.01	0	0.03	0.01	-0.03	0.08
heart rate	0.02	0.04	-0.01	0.02	0.03	0.02	0.02	0.02	0.01	0.02	0.02	0.01

sodium	0.03	0.04	0.08	0.03	0	0.08	--	--	--	--	--	--
eGFR	0	0.01	0.02	0	0	0.02	0	-0.01	0.02	0	-0.01	0.01
furosemide	-0.01	-0.01	-0.01	-0.01	-0.01	0	--	--	--	--	--	--
diabetes	0.05	0.17	-1.57	0.05	0.28	-1.71	0.14	0.82	-1.73	0.14	1.82	-1.41

--: Not available; SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate.

Supplementary II Table S9. Measure of effects in uncalibrated and calibrated analyses.

	Uncalibrated				Regression-based				IOSW			
	arm1	arm 2	ARR	OR	arm1	arm2	ARR	OR	arm1	arm2	ARR	OR
COMET all-cause death	34%	40%	0.06	0.83 (0.74, 0.93)	48% (41%, 55%)	49% (47%, 50%)	0.01 (-0.06, 0.08)	0.97 (0.72, 1.27)	39% (33%, 44%)	51% (40%, 61%)	0.12 (0, 0.23)	0.62 (0.39, 0.99)
COMET all-cause death or hospitalisation	74%	76%	0.02	0.94 (0.86, 1.02)	87% (84%, 90%)	87% (86%, 88%)	-0.01 (-0.03, 0.03)	1.08 (0.81, 1.39)	79% (76%, 82%)	81% (75%, 86%)	0.02 (-0.04, 0.08)	0.87 (0.59, 1.30)
DIG all-cause death	35%	35%	0	0.99 (0.91, 1.07)	43% (38%, 48%)	42% (40%, 43%)	-0.01 (-0.06, 0.04)	1.05 (0.86, 1.28)	33% (31%, 35%)	32% (30%, 34%)	-0.01 (-0.04, 0.02)	1.06 (0.92, 1.21)
DIG death or hospitalisation due to worsening heart failure	31%	38%	0.07	0.75 (0.69, 0.82)	32% (29%, 34%)	36% (34%, 38%)	0.04 (0.02, 0.05)	0.84 (0.78, 0.91)	29% (27%, 31%)	36% (34%, 38%)	0.07 (0.04, 0.1)	0.73 (0.64, 0.83)

IOSW: Inverse Odds of Sampling Weights; ARR: Absolute Risk Reduction; OR: Odds Ratio; Some are not OR, eg, Hazard Ratio for uncalibrated COMET, Risk Ratio for uncalibrated DIG; arm1 vs arm2: carvedilol vs metoprolol in COMET, digoxin vs placebo in DIG.

Exploratory analyses

Baseline profiles of lowest and highest risk deciles individuals in the HF Registry and trials.

The coefficients (Table S5) from the registry natural history model were also used to determine the lowest and highest risk subgroups for the trial. These risk subgroups in the trial were used to calculate the uncalibrated treatment effects using the Cox proportional-hazards model.

The baseline profiles of the lowest and highest risk subgroup individuals in the HF register and trials were showed as Table S10. Generally, either in the HF register or the trials, age, frusemide dose, history of diabetes in the lowest risk group were much lower than those in the highest risk group, while the eGFR levels were much higher (e.g. in the HF register for COMET all-cause death outcome, the age was 49.10 in the lowest risk group vs 86.83 in the highest risk group), which is in line with the main effects showed in Figure 10.

Supplementary II Table S10. Baseline profiles of lowest and highest risk deciles individuals in the HF Registry and trials.

	COMET (all-cause death)				COMET (all-cause death or hospitalisation)				DIG (all-cause death)				DIG (composite outcome for death or hospitalisation due to worsening heart failure)			
	Registry		Trial		Registry		Trial		Registry		Trial		Registry		Trial	
	lowest risk decile	highest risk decile	lowest risk decile	highest risk decile	lowest risk decile	highest risk decile	lowest risk decile	highest risk decile	lowest risk decile	highest risk decile	lowest risk decile	highest risk decile	lowest risk decile	highest risk decile	lowest risk decile	highest risk decile
Age (years)	49.10 (9.91)	86.83 (6.33)	41.34 (7.89)	74.31 (6.58)	55.36 (13.86)	81.53 (8.13)	46.37 (11.00)	70.09 (7.82)	48.82 (9.64)	87.73 (5.58)	43.50 (7.38)	77.70 (5.51)	66.87 (14.74)	77.03 (9.89)	43.50 (7.38)	67.69 (9.49)
men sex (%)	65.54 %	60.80 %	83.55 %	78.15 %	73.66 %	51.06 %	89.14 %	73.18 %	64.67 %	61.30 %	75.29 %	72.79 %		55.56 %	75.29 %	65.15 %
Loop_diuretics (mg)	52.16 (23.95)	83.18 (48.30)	7.72 (21.11)	71.53 (93.41)	47.62 (20.87)	88.86 (48.86)	5.79 (17.97)	79.74 (94.22)	--	--	--	--	--	--	--	--
History of diabetes (%)	7.74%	40.32 %	7.24%	49.34 %	3.62%	56.80 %	4.28%	61.92 %	7.74%	41.95 %	7.21%	53.09 %	4.62%	55.31 %	7.21%	73.24 %

Systolic blood pressure (mm Hg)	120.61 (24.08)	109.25 (18.18)	128.57 (20.38)	117.12 (17.92)	132.73 (26.40)	105.27 (16.34)	137.31 (21.41)	113.51 (16.41)	120.37 (23.07)	109.20 (20.24)	125.40 (19.34)	119.13 (18.84)	154.63 (20.24)	92.49 (9.98)	125.40 (19.34)	103.22 (11.82)
Heart rate (beats per minute)	74.23 (13.10)	74.76 (13.51)	82.21 (14.06)	83.57 (13.27)	73.69 (12.89)	74.30 (13.46)	80.51 (13.70)	83.59 (13.47)	74.43 (13.10)	74.27 (13.48)	79.42 (12.99)	81.46 (12.88)	72.78 (13.09)	74.69 (14.40)	79.42 (12.99)	83.00 (12.99)
Sodium (mmol/l)	138.64 93.15)	136.26 (4.47)	140.07 (3.08)	137.40 (4.94)	139.84 (3.01)	134.88 (4.97)	141.22 (3.06)	136.49 (4.92)	--	--	--	--	--	--	--	--
eGFR (mL/min/ 1.73m2)	83.38 (23.59)	41.27 (15.60)	90.61 (22.00)	47.69 (14.76)	84.97 (22.51)	39.55 (15.40)	93.14 (23.05)	48.14 (15.52)	78.07 (14.46)	39.03 (13.91)	76.42 (12.79)	43.50 (13.32)	67.16 (19.86)	47.00 (17.83)	76.42 (12.79)	51.28 (16.59)
Categorical variables are shown as counts (%s) and continuous variables as means (standard deviations); --: Not available; eGFR: estimated glomerular filtration rate.																

Results for exploratory analyses.

The odds ratio (OR), absolute risk reduction (ARR), and risk in each arm in the lowest, highest risk deciles and overall group when each trial was calibrated to the HF registry were displayed as the Table S11 and Fig 3. The CIs of the efficacy estimates (ORs) and absolute risk reduction (ARR) estimates where the trials were calibrated to the high-risk and low-risk deciles are wider than those for the whole register calibration, while within each decile, the results are still similar. And generally, they were not far away from the overall calibration.

Supplementary II Table S11. Calibrated results for overall, lowest and highest risk subgroup individuals.

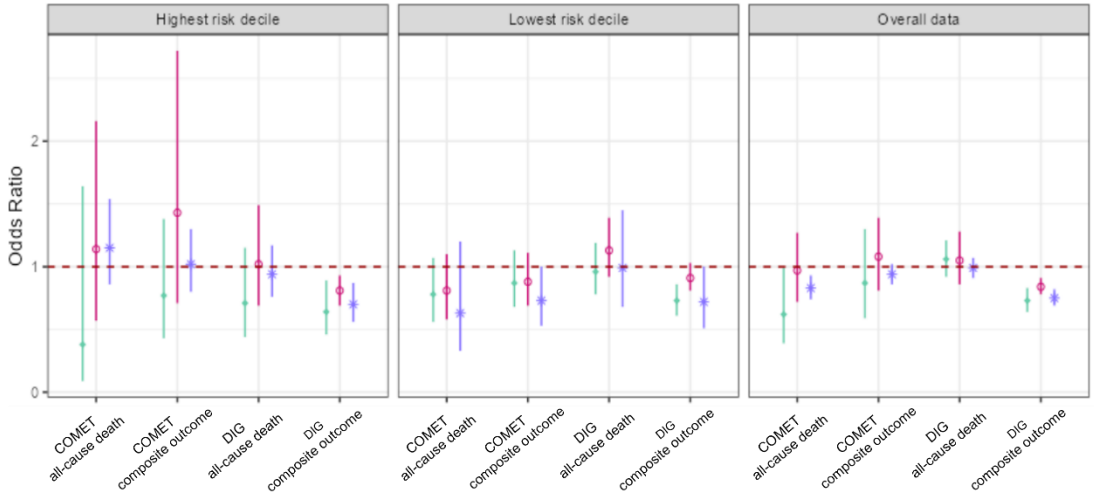
	Uncalibrated				Regression-based				IOSW				IOSW (trimming the largest 1% of weights)			
	arm1	arm2	ARR	OR	arm1	arm2	ARR	OR	arm1	arm2	ARR	OR	arm1	arm2	ARR	OR
Overall																
COMET all-cause death	34%	40%	0.06	0.83 (0.74, 0.93)	48% (41%, 55%)	49% (47%, 50%)	0.01 (-0.06, 0.08)	0.97 (0.72, 1.27)	39% (33%, 44%)	51% (40%, 61%)	0.12 (0, 0.23)	0.62 (0.39, 0.99)	37% (33%, 40%)	38% (35%, 42%)	0.02 (-0.03, 0.07)	0.93 (0.75, 1.15)
COMET all-cause death or hospitalisation	74%	76%	0.02	0.94 (0.86, 1.02)	87% (84%, 90%)	87% (86%, 88%)	-0.01 (-0.03, 0.03)	1.08 (0.81, 1.39)	79% (76%, 82%)	81% (75%, 86%)	0.02 (-0.04, 0.08)	0.87 (0.59, 1.30)	76% (74%, 79%)	78% (75%, 81%)	0.02 (-0.02, 0.06)	0.91 (0.73, 1.13)
DIG all-cause death	35%	35%	0	0.99 (0.91, 1.07)	43% (38%, 48%)	42% (40%, 43%)	-0.01 (-	1.05 (0.86, 1.28)	33% (31%, 35%)	32% (30%, 34%)	-0.01 (-	1.06 (0.92, 1.21)	32% (30%, 34%)	31% (29%, 33%)	-0.01 (-	1.04 (0.91, 1.18)

							0.06, 0.04)				0.04, 0.02)				0.04, 0.02)	
DIG death or hospitalisation due to worsening heart failure	31%	38%	0.07	0.75 (0.69, 0.82)	32% (29%, 34%)	36% (34%, 38%)	0.04 (0.02, 0.05)	0.84 (0.78, 0.91)	29% (27%, 31%)	36% (34%, 38%)	0.07 (0.04, 0.1)	0.73 (0.64, 0.83)	28% (26%, 30%)	35% (33%, 37%)	0.07 (0.04, 0.10)	0.72 (0.63, 0.82)
Low risk decile																
COMET all-cause death	10%	16%	0.06	0.63 (0.33, 1.20)	16% (12%, 20%)	19% (17%, 20%)	0.03 (-0.02, 0.07)	0.81 (0.58, 1.10)	27% (23%, 31%)	32% (27%, 38%)	0.05 (-0.02, 0.12)	0.78 (0.56, 1.07)	28% (26%, 31%)	29% (26%, 33%)	0.02 (-0.03, 0.06)	0.92 (0.73, 1.16)
COMET all-cause death or hospitalisation	46%	58%	0.12	0.73 (0.53, 1.00)	71% (66%, 76%)	74% (72%, 76%)	0.03 (-0.02, 0.08)	0.88 (0.69, 1.11)	71% (67%, 74%)	73% (69%, 77%)	0.03 (-0.02, 0.08)	0.87 (0.68, 1.13)	69% (66%, 72%)	72% (69%, 75%)	0.04 (-0.01, 0.08)	0.84 (0.69, 1.03)
DIG all-cause death	16%	16%	0	0.99 (0.68, 1.45)	17% (14%, 21%)	16% (14%, 17%)	-0.02 (-0.05, 0.01)	1.13 (0.92, 1.39)	21% (19%, 24%)	22% (19%, 24%)	0.01 (-0.03, 0.04)	0.96 (0.78, 1.19)	23% (21%, 25%)	23% (21%, 25%)	0 (-0.03, 0.03)	1.00 (0.84, 1.18)
DIG death or hospitalisation due to worsening heart failure	18%	23%	0.05	0.72 (0.51, 1.00)	28% (24%, 31%)	30% (27%, 32%)	0.02 (-0.01, 0.04)	0.91 (0.81, 1.03)	22% (20%, 24%)	28% (26%, 30%)	0.06 (0.03, 0.09)	0.73 (0.61, 0.86)	22% (20%, 24%)	28% (26%, 30%)	0.06 (0.03, 0.09)	0.73 (0.63, 0.85)
High risk decile																
COMET all-cause death	62%	58%	-0.04	1.15 (0.86, 1.54)	74% (60%, 85%)	73% (71%, 75%)	-0.01 (-0.12, 0.12)	1.14 (0.57, 2.16)	64% (46%, 78%)	82% (56%, 94%)	0.18 (-0.07, 0.43)	0.38 (0.09, 1.64)	60% (52%, 67%)	56% (48%, 64%)	-0.04 (-0.15, 0.08)	1.16 (0.73, 1.83)
COMET all-cause death or hospitalisation	88%	86%	-0.02	1.02 (0.80, 1.30)	95% (91%, 98%)	94% (93%, 94%)	-0.01 (-0.04, 0.02)	1.43 (0.71, 2.72)	85% (79%, 89%)	88% (83%, 92%)	0.03 (-0.04, 0.1)	0.77 (0.43, 1.38)	81% (75%, 85%)	86% (82%, 89%)	0.06 (-0.01, 0.12)	0.67 (0.43, 1.03)

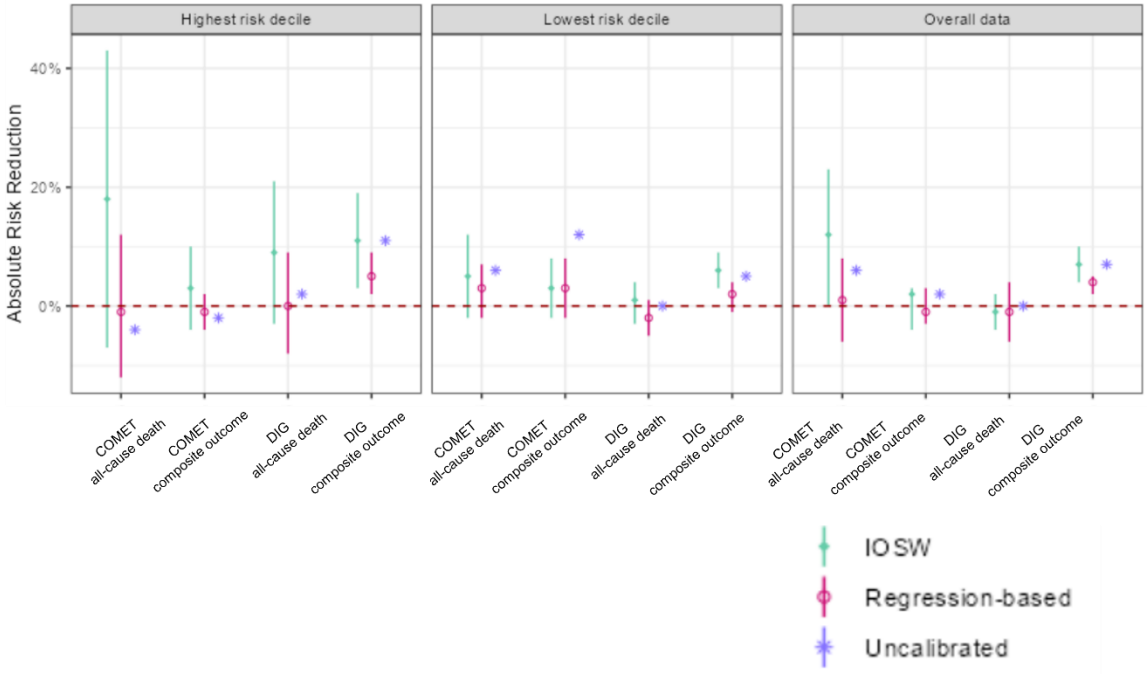
DIG all-cause death	47%	49%	0.02	0.94 (0.76, 1.17)	67% (58%, 75%)	67% (65%, 69%)	0 (-0.08, 0.09)	1.02 (0.69, 1.49)	47% (41%, 54%)	56% (45%, 65%)	0.09 (-0.03, 0.21)	0.71 (0.44, 1.15)	44% (40%, 49%)	44% (40%, 49%)	1 (-0.07, 0.06)	1.01 (0.77, 1.31)
DIG death or hospitalisation due to worsening heart failure	42%	53%	0.11	0.70 (0.56, 0.87)	36% (32%, 40%)	42% (40%, 43%)	0.05 (0.02, 0.09)	0.81 (0.69, 0.93)	42% (36%, 47%)	53% (46%, 59%)	0.11 (0.03, 0.19)	0.64 (0.46, 0.89)	35% (32%, 39%)	45% (42%, 49%)	0.10 (0.05, 0.16)	0.65 (0.52, 0.82)
IOSW: Inverse Odds of Sampling Weights; ARR: Absolute Risk Reduction; OR: Odds Ratio; Some are not OR, eg, Hazard Ratio for uncalibrated COMET, Risk Ratio for uncalibrated DIG; arm1 vs arm2: carvedilol vs metoprolol in COMET, digoxin vs placebo in DIG.																

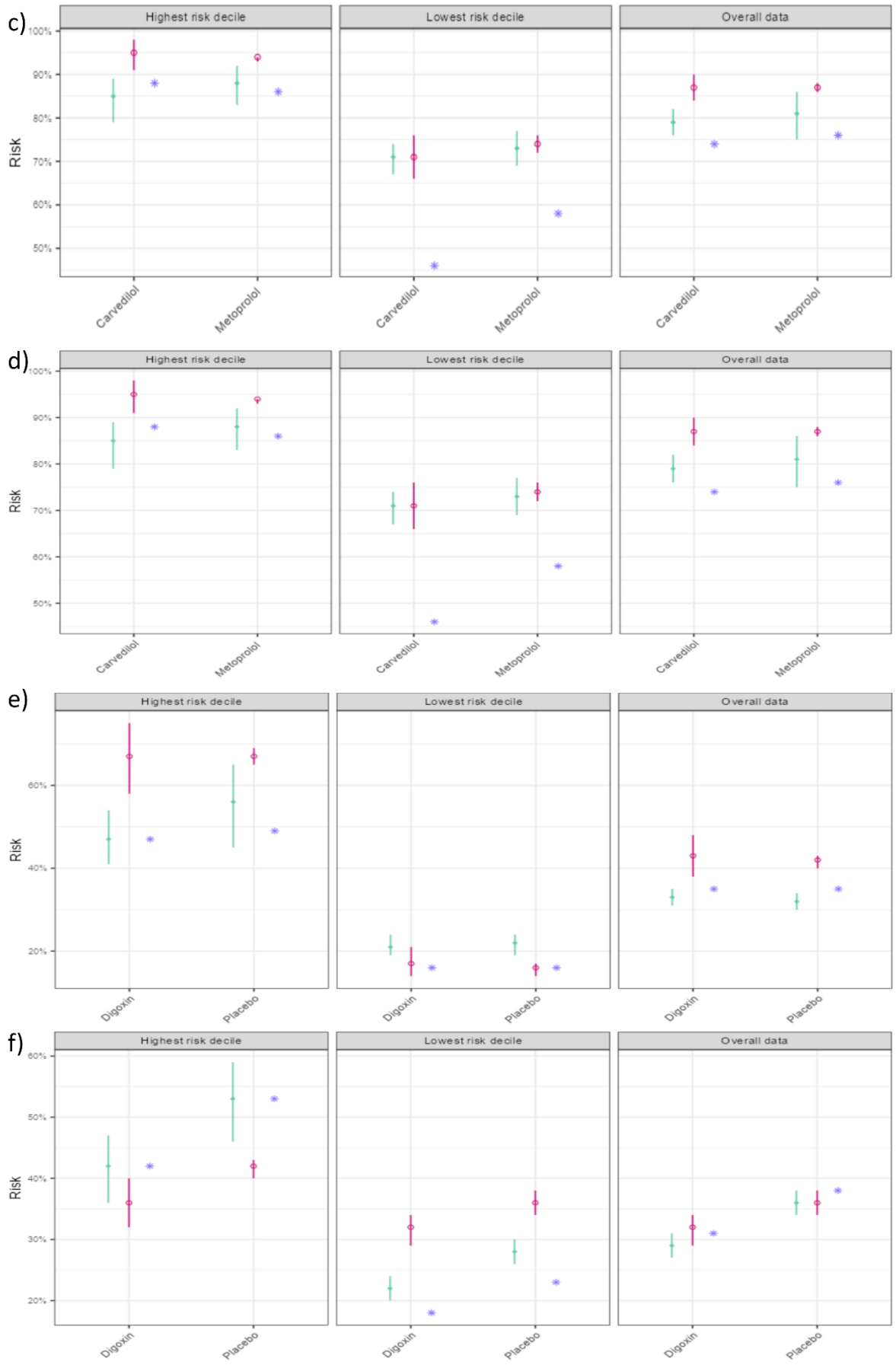
Supplementary II Fig 3. Effect estimates for standard and calibrated analyses in the highest risk, lowest risk deciles and the overall group in 2 trials. a) Odds ratio; b) Absolute risk reduction; c ~ f, risk for c) COMET all-cause death; d) COMET all-cause death or hospitality.

a)



b)




c)
Highest risk decile
Lowest risk decile
Overall data
Risk
100%
90%
80%
70%
60%
50%
Carvedilol
Metoprolol
d)
Highest risk decile
Lowest risk decile
Overall data
Risk
Carvedilol
Metoprolol
e)
Highest risk decile
Lowest risk decile
Overall data
Risk
60%
40%
20%
Digoxin
Placebo
f)
Highest risk decile
Lowest risk decile
Overall data
Risk
60%
50%
40%
30%
20%
Digoxin
Placebo

Target population characteristics.

Supplementary II Table S12. Example of data which could be produced from a registry to reconstructe the joint distribution of patient characteristics.

a) Marginal summary statistics

Categorical variables		Mean for each covariate			Standard deviation for each covariate			n
men	diabetes	age	SBP	…*	age	SBP	…**	
0	0	70.05	120.02	…	10.05	10.03	…	5167
0	1	70.05	119.96	…	9.99	9.94	…	54198
1	0	70.03	120.03	…	10.06	9.99	…	8755
1	1	69.97	120.08	…	10.16	9.94	…	31880
*Mean for the rest numerical variables; **standard deviation for the rest numerical variables; *** correlations between each 2 numerical variables after adjusting for other variables. SBP: systolic blood pressure; eGFR: estimated glomerular filtration rate.								

b) Correlation matrix

	Men	Diabetes	Age	SBP	…
Men	1	0.01	-0.01	0.05	…
Diabetes		1	0.1	0.06	…
Age			1	0.05	…
SBP				1	…
…	…	…	…	…	…