'Real-world' observational studies in arrhythmia research

data sources, methodology, and interpretation. A position document from European Heart Rhythm Association (EHRA), endorsed by Heart Rhythm Society (HRS), Asia-Pacific HRS (APHRS), and Latin America HRS (LAHRS)

Torp-Pedersen, Christian; Goette, Andreas; Nielsen, Peter Bronnum; Potpara, Tatjana; Fauchier, Laurent; John Camm, Alan; Arbelo, Elena; Boriani, Giuseppe; Skjoeth, Flemming; Rumsfeld, John; Masoudi, Frederick; Guo, Yutao; Joung, Boyoung; Refaat, Marwan M; Kim, Young-Hoon; Albert, Christine M; Piccini, Jonathan; Avezum, Alvaro; Lip, Gregory Y H; ESC Scientific Document Group; Lane, Deirdre

Published in:
Europace

DOI (link to publication from Publisher):
10.1093/europace/euz210

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

Real world' observational studies in arrhythmia research: data sources, methodology and interpretation. A position document from **EHRA, endorsed by HRS, APHRS, LAHRS**

Torp-Pedersen (Chair) .. Jonathan P. Piccini, MD, MHS; others (alphabetical) … Lip (co-Chair)

Christian Tobias Torp-Pedersen: Department of clinical investigation and cardiology, Nordsjaellands Hospital, Hillerød and Department of Cardiology, Aalborg University, Hospital, Aalborg, Denmark

Andreas Goette: St. Vincenz-Krankenhaus Gmbh, Paderborn, Germany

Peter Bronnum Nielsen: Aalborg University, Health Science And Technology, Aalborg, Denmark

Tatjana Potpara: School of Medicine, Belgrade University; Cardiology Clinic, Clinical Center of Serbia

Laurent Fauchier: Service de Cardiologie, Centre Hospitalier Universitaire Trousseau et Université de Tours, Faculté de Médecine. Tours, France.
Consultant or speaker for Bayer, BMS/Pfizer, Boehringer Ingelheim, Medtronic, Novartis.

A. John Camm: St. George's, University of London, Molecular And Clinical Sciences Research Inst, St George's, University of London, London, United Kingdom

Elena Arbelo:
aArrhythmia Section, Cardiology Department, Hospital Clínic, Universitat de Barcelona. Barcelona (Spain).
bIDIBAPS, Institut d'Investigació August Pi i Sunyer (IDIBAPS). Barcelona (Spain).
cCentro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), Madrid (Spain)

Giuseppe Boriani: Cardiology Division, Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Policlinico di Modena, Modena, Italy.

Flemming Skjoeth: Aalborg University, Health Science And Technology, Aalborg, Denmark

John Rumsfeld: University of Colorado School of Medicine, Aurora, CO, USA

Frederick Masoudi: University of Colorado Anschutz Medical Campus, Department of Medicine, Division of Cardiology, Aurora, CO USA

Yutao Guo : Chinese PLA General Hospital, Cardiology, Beijing China People's Republic of

Boyoung Joung : Yonsei University College of Medicine, Cardiology Department, Seoul, Korea Republic of

Marwan M. Refaat: American University of Beirut Medical Center, Department of Internal Medicine, Beirut, Lebanon

Young-Hoon Kim (APHRS representative): Korea University Medical Center, Cardiology Department, Seoul, Korea Republic of

Christine M. Albert (HRS representative): Brigham And Women's Hospital, Boston, MA United States of America

Jonathan Piccini (HRS representative): Duke Center for Atrial Fibrillation, Duke University Medical Center, Duke Clinical Research Institute, Durham, USA

Alvaro Avezum (LAHRS representative): Dante Pazzanese Institute Of Cardiology, Sao Paulo, Brazil

Gregory YH Lip: Liverpool Centre for Cardiovascular Science, University of Liverpool and Liverpool Heart & Chest Hospital, Liverpool, United Kingdom; and Aalborg Thrombosis Research Unit, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

ESC Scientific Document Group:
Dr. Nikolaos Dagres, Dr. Serge Boveda, Dr. Kevin Vernooy, Prof. Zbigniew Kalarus, Prof. Gulmira Kudaiberdieva, Dr. Georges H Mairesse, Prof. Valentina Kutyifa, Prof. Thomas Deneke, Pof. Jesper Hastrup Svendsen, Dr. Vassil B Traykov, Prof. Arthur Wilde, Prof. Frank R. Heinzel

External reviewers
Bulent Gorenek (Review coordinator): Eskisehir Osmangazi University, Eskisehir, Turkey

Nikolaos Dagres: Heart Center Leipzig, Department of Electrophysiology, Leipzig, Germany

Francesco Violi: Sapienza Università di Roma, Dipartimento di Medicina Interna e Specialità Mediche, Viale del Policlinico, 155 - 00162 Roma

Xin Du: Anzhen Hospital, Department of Cardiology, An Ding men wai, Anzhenli, 100029 Beijing, China People's Republic of

Masaharu Akao: Department of Cardiology, National Hospital Organization Kyoto Medical Center, 1-1, Mukaihata-cho, Fukakusa, Fushimi-ku, Kyoto 612-8555, Japan

Eue-Keun Choi: Seoul National University College of Medicine, Cardiovascular Center, Seoul National University Hospital, 101 Daehak-ro, Chongno-gu Seoul 03080, Republic of Korea

Renato D. Lopes: Duke University Medical Center, Duke Clinical Research Institute, 200 Morris St., Durham, NC 27701, P.O. Box 17969, Durham, NC 27715, DUMC#: 3850

Evren Emin Ozcan: Dokuz Eylul University Department of Cardiology, 35340 Balçova / İzmir / Turkey

Deirdre Lane: Liverpool Centre for Cardiovascular Science, University of Liverpool and Liverpool Heart & Chest Hospital, Liverpool, United Kingdom; and Aalborg Thrombosis Research Unit, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

Francisco Marin: Hospital Universitario Virgen de la Arrixaca, IMIB-Arrixaca, CIBERCV, University of Murcia, Murcia, Spain

Christopher Peter Gale:University Of Leeds, LICAMM, Medical Research Counc Bioinformatics Centre, LICAMM, Worsley building, LS2 9JT Leeds, West Yorkshire United Kingdom
Kevin Vernooy: Maastricht University Medical Center, department of Cardiology, PO BOX 5800, 6202 AZ, Maastricht, Netherlands

Gulmira Kudaiberdieva : Adana, Turkey Center for Postgraduate Education and Research, Bishkek, Kyrgyzstan Scientific Research Institute of Heart Surgery and Organ Transplantation, Bishkek, Kyrgyzstan, Editor in Chief of Heart Vessels and Transplantation

Valentina Kutyifa: University of Rochester Medical Center, Rochester, New York & Adjunct Position at Semmelweis University Heart Center, Budapest, Hungary

Vassil B Traykov: Department of Invasive electrophysiology, Acibadem City Clinic Tokuda Hospital

Carolina Guevara: Instituto Nacional Cardiovascular Incor, Jirón Coronel Zegarra 417, Jesús María 15072, Peru

Tze-Fan Chao: Division of Cardiology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

Sana M. Al-Khatib: Duke University Medical Center, NC 27715 Durham, North Carolina United States of America

Key words

Abbreviations

Introduction

The criterion standard for demonstrating the efficacy of a clinical intervention is the randomized clinical trial (RCT). Randomization supports equal distribution of known as well as unknown confounders, and therefore the relationship between the intervention and the outcome may be considered causal. Nevertheless, RCTs have limitations such as cost and cohort selection, and data from such trials are not available to provide evidence for the majority of clinical decisions. Most of recommendations in international cardiology guidelines are not based on randomised trials and there appears no improvement over the last 10 years[1].

For many clinical scenarios, observational data may be the highest level of evidence available[2]. Observational data can also be of particular use in evaluating care delivery, and effectiveness and safety of care in clinical practice. However, observational studies also carry significant limitations, especially when applied to therapeutic interventions (i.e. trying to determine effectiveness). Observational data is subject to underlying biases such as selection bias and are prone to unmeasured confounding. In an overview, 25% of observational studies were contradicted when the findings were tested in a randomized design [3]. Over the last decade there has been an exponential growth of observational data (e.g. from electronic health records, clinical registries, and other sources). This has been coupled with advances in the conduct and interpretation of observational studies to minimize these issues and guidelines/checklists have been developed for the conduct of observational studies (https://www.strobe-statement.org). In parallel, there is tremendous interest in utilizing observational, or 'real world' data to inform clinical care.

In recognizing these issues, European Heart Rhythm Association (EHRA), with additional contributions from Heart Rhythm Society (HRS), Asia Pacific Heart Rhythm Society (APHRS) and the Latin America Heart Rhythm Society (HRS) proposed a position document describing contemporary techniques for optimal conduct and presentation of observational studies. An additional aim was to provide recommendations to encourage implementation of new designs.
This review first describes the usual data sources for observational studies, reviews common and important techniques, overviews the proper interpretation of results, and finally makes appropriate recommendations regarding the design, conduct, and interpretation of observational data. The intended reader is the clinical cardiologist that wishes to get an overview of current methodology. It is hoped that it will aid the discussion between clinicians and cardiologists. It has been attempted to cover briefly the most used current methods with focus on more recent methodology. It is a very large area that is covered and therefore many details are not touched in this overview.

Evidence Review

This document was prepared by the Task Force with representation from EHRA, with additional contributions from HRS, APHRS, LAHRS and CASSA, and has been peer-reviewed by official external reviewers representing all these bodies. A detailed literature review was conducted, weighing

the strength of evidence for or against a specific treatment or procedure, and where data exist including estimates of expected health outcomes.

We have used a simple and user-friendly system of grading recommendations using 'coloured hearts' (Table 1). This EHRA grading of consensus statements does not use separate definitions of the level of evidence. This categorization, used for consensus statements, must not be considered as directly similar to that used for official society guideline recommendations, which apply a classification (Class I-III) and level of evidence (A, B and C) to recommendations used in official guidelines.

The routine use of hearts is changed for this publication which addresses statistical methods rather than interventions. Thus, a green heart indicates recommended strategies, a yellow heart something that can be considered and a red heart something to be avoided.

Table 1

Definitions where related to a treatment or procedure	Consensus statement instruction	Symbol
Scientific evidence that a treatment or procedure is beneficial and effective. Requires at least one randomized trial, or is supported by strong observational evidence and authors' consensus (as indicated by an asterisk).	'Should do this'	
General agreement and/or scientific evidence favour the usefulness / efficacy of a treatment or procedure. May be supported by randomized trials based on a small number of patients or which is not widely applicable.	'May do this'	
Scientific evidence or general agreement not to use or recommend a treatment or procedure.	'Do not do this'	

*This categorisation for our consensus document should not be considered as being directly similar to that used for official society guideline recommendations which apply a classification (I-III) and level of evidence (A, B and C) to recommendations.

Data sources

A selection of common and important data sources follow and table 2 highlights their main strengths and weaknesses. It should be noted that the categories are not completely independent with considerable overlap in some regions.

Registries for regulatory sponsored studies

Registries play an important role in the evaluation of safety and effectiveness of medical devices and pharmaceutical agents. In the case of pharmacotheapeutics, these registries are also referred to as phase IV observational studies, which gather information on drug safety and effectiveness after regulatory approval. Regulatory agencies such as the United States Food and Drug Administration (FDA) may request a registry as a condition of approval for a device approved under a premarket approval (PMA) order. Post-approval registries help assess several aspects of therapeutic interventions, including safety, effectiveness, reliability in clinical practice or "real world" settings, and long-term outcomes. The European Medicines Agency (EMA) launched an initiative for patient registries in 2015 to support more systematic approach to their conduct and use in estimating benefit-risk assessment for pharmaceutical agents in the European Economic Area. Similarly, the EMA also established a European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) and an associated registry database to synergize registry efforts. The ENCePP has also published a Guide on Methodological Standards in Pharmacoepidemiology. (http://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml)

There has also been particular interest in the use of registry data to help monitor post-market performance of medical devices.[4] The FDA has established the unique identifier (UDI) system to incorporate UDI into electronic health information in order to help track individual devices and facilitate tracking outcomes so as to improve nationwide surveillance of device performance. However, the approach to integrating the UDI into data sources has not been established. The FDA is also promoting the development of national and international device registries in several therapeutic areas and interventions. A relevant program is the National Cardiovascular Data Registry for Implantable Cardioverter Defibrillators (NCDR ICD, www.ncdr.com). This registry was developed in conjunction with Centers for Medicare and Medicaid Services (CMS) to serve a coverage with evidence development decision for primary prevention defibrillators in CMS beneficiaries. This program has also been employed by FDA and industry for post-market analysis. The NCDR Left atrial appendix occlusion (LAAO) Registry (www.ncdr.com) was also developed in conjunction with FDA and CMS both to fulfil post-marketing requirements (of FDA) and coverage with evidence development (for CMS).

Registries sponsored by learned societies

The EURObservational Research Programme on Atrial Fibrillation (EORP-AF) was an independent initiative promoted by ESC in order to systematically collect data regarding the management and treatment of AF in ESC member countries. The first registry (EORP-AF Pilot Survey) enrolled 3119 patients in 67 centers from February 2012 to March 2013 and showed that the uptake of oral anticoagulation (mostly vitamin K antagonist therapy) had improved since the Euro

Heart Survey performed 10 years before, although antiplatelet therapy (especially aspirin) was still used in one-third of the patients and elderly patients were commonly undertreated with oral anticoagulation.[5-7] Follow up data showed that 1-year mortality and morbidity remained high in AF patients, particularly in patients with heart failure or chronic kidney disease.[7, 8] Additionally, asymptomatic AF was particularly common (around 40% of patients) and associated with elderly age, more comorbidities, an high thromboembolic risks and a higher 1-year mortality as compared with symptomatic patients.[9] As a consequence of the characteristics of the registry some centres did not participate to long-term follow up, so only 2119 (68%) patients were included into the 3-year follow up analysis.[10]

The second EORP registry was the EORP-AF Long-Term General Registry, a prospective, observational, large-scale multicentre registry of ESC, that enrolled more than 11 000 AF patients in 250 centres from 27 participating ESC countries from October 2013 to September 2016 [11]. This registry showed that around 85% of AF patients are currently treated with oral anticoagulants, with an increase as compared to the past mostly due to the progressive uptake of NOACs.[11, 12] Overall, the registries promoted by ESC over a decade allowed to document significant changes in AF epidemiology in Europe, with an increased complexity of AF patients due to comorbidities, with an impact on both morbidity and mortality.[12]

The American College of Cardiology's PINNACLE Registry is an outpatient, longitudinal clinical quality program that captures data from ambulatory electronic health records among cardiovascular practice across the United States, and some practices from other countries (e.g. Brazil, India). One of the primary patient cohorts is atrial fibrillation. There have been a number of publications on AF patients from PINNACLE. Recent examples include: Sex Differences in the Use of Oral Anticoagulants, showing that women were less likely to receive anticoagulant therapy at all levels of CHA_2DS_2-VASc score;[13] Predictors of oral anticoagulant non-prescription in patients with atrial fibrillation and elevated stroke risk, highlighting the prevalence of anti-platelet use;[14] and Influence of Direct Oral Anticoagulants on Rates of Oral Anticoagulation for Atrial Fibrillation, demonstrating that the growing use of DOACs is associated with higher overall oral anticoagulation rates in the U.S., although significant practice variation still exists.[15] There have also been nascent efforts to collaborate among global professional society AF registries, with initial participants from the United States, Europe, China, Brazil, South Korea, Taiwan, Singapore, Japan, and the Balkan countries, in order to advance global research insights on AF care and outcomes.[16]
The First Brazilian Cardiovascular Registry of Atrial Fibrillation (the RECALL study) will assess demographic characteristics and evidence-based practice of a representative sample of patients with AF in Brazil. Results are expect in 2020[17].

Nationwide cohorts

Large population-based studies can inform on the incidence, prevalence, natural history, treatment, correlates, outcomes, and patterns of health care utilization. A special type of large population study encompasses the population of an entire nation. Advantages include very large sample size and lack of selection and participation bias. These advantages are enhanced further when the databases are rich in clinical, personal, and risk factor information and when different pieces of information are linked to permit joint analysis. Once the process for data access is established, vast amounts of information can be obtained at minimal cost, especially when additional collection and update of information is carried out routinely for purposes inherent in medical care and/or insurance coverage

and reimbursement. Nationwide cohorts differ from "Claims data" described below by covering all citizens in an entire region as opposed to e.g. an insurance provider where the sample to be examined is defined very differently than a region.

Large nationwide registries are further valuable for examining temporal changes over prolonged time[18, 19]. A recent example is analysis of recurrence of AF following ablation in the Danish register[20]. For example, Denmark, Taiwan, Sweden and Korea have well-established and validated nationwide health insurance (NHI) databases, other national dataset resources, and the capacity for cross-linking some of these databases and/or resources for aetiologic information, outcomes, and other data. Supplement table 1 shows some main features of the national databases of these countries [21-30] Currently, the Nationwide Research Database includes data files containing information on personal characteristics (sex, date of birth, place of residence, details of insurance, employment); family relationships; details of clinical information, including date, expenditures, and diagnosis related to both inpatient and outpatient procedures; prescription details; examinations; and operations. While these registries differ in length of retrospective period and specific health data information, their primary strengths include lack of use of selection criteria for enrollment and minimal loss to follow-up. Their weakness is generally lack of obviously important factors such as smoking habits, body weight, etc. except for Korea. Korea database contains lifestyle and habits (body weight, height, smoking, alcohol, and exercise), and basic laboratory data including creatinine, and lipid parameters, etc[31].

By law, all residents of these countries have a unique personal identification number that is used also for tax returns, bank accounts, and all transactions. Thus, NHI Research Database data are linkable to multiple national databases maintained by other departments, including drug prescriptions, registries of births, deaths, households, immunizations, cancer, reportable infectious diseases, and environmental exposures. In addition, the data in the biobank will be linked with Nationwide Research Database data.

While these sources are highly useful it is also important to point out that access is restricted. Each country has legal restraints to who may access the data. While understandable that the world cannot freely access health information on individuals from a whole population it is important to recognise that anyone wishing to challenge a result from these sources can only do so in collaboration with researchers with proper access authorisation.

Claims data

Healthcare systems with access to administrative dataset based on claims data provide an opportunity for observational studies. Examples include insurance data in the US, such as CMS, which is the payer for services for older persons and the disabled. Claims analyses are limited by appropriateness of coding (usually based on ICD-9 or ICD-10 codes) and whether particular individuals maintain enrollment with the same insurer. Studies that merge multiple claims datasets may identify patients that have been included in >1 insurance datasets. Another important limitation is that patients may not be available for follow-up if they change insurance provider. As for nationwide registers the level of detail is limited to the information collected, and important and granular clinical data are often missing.

The data have been the basis of recent large comparative effectiveness studies on various NOACs versus warfarin, or against each other using claims data from the USA. Examples include papers that have investigated NOACs vs warfarin, and for NOAC vs NOACs from independent academic

groups[32]. Claims data have also been used by industry-sponsored studies, for example, those by Lip et al [33]

Registries from Industry sponsored cohort studies

Industry sponsorship has led to drug based registries (eg. XANTUS, XALIA) and disease-based registries (GARFIELD-AF, GLORIA-AF, PREFER in AF, ORBIT-AF, etc). There are also several examples of government funded observational multicenter prospective cohort studies (PROSE-ICD, PREDETERMINE, Long QT registry, etc). As these are sponsored efforts, the investigator is often reimbursed for including patients into a particular registry or study, so some element of channelling bias is possible. Nonetheless by design, there would be including selected patients in (also selected) enrolling centres, but has the positive aspect of careful protocol-based follow-up. In addition to these centre-patient based studies, there are a variety of population-based studies that have been utilized to study arrhythmic endpoints (FHS, ARIC, CHS, MESA, WHS, NHS, REGARDS).

Hospital cohorts (vs community)

Hospital cohorts are referred to prospective, or retrospective, observational cohort studies of patients with or at risk for arrhythmia or cardiac conditions and usually receiving a specific treatment or intervention (anticoagulants, ablation, devices, surgery, etc.). They may be local cohorts or wider scale regional or national cohorts covering a global healthcare system. Nationwide hospital cohorts can provide real-world evidence of clinical practice, patient outcomes, safety, comparative effectiveness and cost-effectiveness of interventions. A systematic robust research design, with accurate measurement of appropriate outcomes and control variables is needed for protecting the quality of data.

Both hospital and community-based cohorts can be used to evaluate the outcomes of patients exposed to a particular program or management strategy and are useful for understanding the real-world safety and effectiveness of specific treatments and may provide the analysis of the relative effectiveness of a given treatment among alternative patients' subgroups. Compared to hospital cohorts, the communities' cohorts can provide the advantage of longitudinal data collection on considerable number of unselected patients. The key end points, such as mortality information, could be attained from the hospital cohort, which are variably missing in administrative claims databases. By contrast, nationwide administrative databases may identify outcomes recorded on different healthcare facilities on a larger scale and may reduce channelling bias (see below).

Hospital cohorts have important limitations. Hospital uptake may be highly selective resulting in patients for study being of higher or lower risk than the average patient. Such weaknesses may also vary over time as treatments change from in-hospital to out-patient treatment.

Table 2.
Strengths and weaknesses of common data sources

Strengths	Weaknesses
Regulatory Sponsored studies	
Arrives early after marketing	Patient selection may not to be representative
Targeted data collection	
Learned Society academic studies	
Targeted data collection	Patient selection need not to be representative
Usually wide geographical representation	Quality of outcome registration can vary

Nationwide or Regional registries	
Large scale Less bias in patient selection Low cost	Data quality may be limited given use of clinical documentation International generalisability uncertain
Claims data	
Complete selection of data within an administrative unit Low cost	Many clinically important data (both independent and outcome variables) may not be available. Quality of data may be limited
Investigator Initiated and Industry sponsored studies	
Multiple centres Careful monitoring of data collected Targeted data collection	Reimbursement for participation can influence patients who consent to intervention. Centre selection can result in unrepresentative patients. Questions may be designed to ensure a higher probability of a favourable outcome.
Hospital cohorts	
Uniform patient selection Similar expertise to all patients	Patient selection not representative Data quality may not be high Expertise of selected centres may not be generalised

Bias and Confounding

Bias

All studies including randomized studies are potentially subject to processes that may cause a study to report results that may not be generalized or may even be incorrect. These processes are referred to as bias and nearly all bias is related to the selection of the study population (selection bias) or recording of data from a study (information bias). Sacket lists 35 types of bias[34] and the list is far from complete. Table 3 is a selected list of either common or commonly overlooked sources of bias.

In addition to bias that can at least be listed as limitations there are other sources. Data dredging bias is when multiple analyses are performed on a dataset and only the apparently interesting ones are reported. It is related to publication bias, where journals are more likely to accept potentially interesting positive findings, but once an interesting finding has been published the absence of the same finding may become interesting enough for publication. Cognitive dissonance bias is when strong beliefs prevail in spite of evidence.

So, what can be done about bias? The always important limitations of observational studies is that unknown or unaccounted bias can never be completely excluded. There is no mathematical technique

to adjust for bias that is potentially present but not known. On occasion subgroup analyses and other sensitivity analyses may cast light on the problems in a study.

In many cases bias is complex. One example is comparison of treatments and allowing both prevalent and new users in an analysis. This introduces several sources of bias. There is a selection bias towards patients that tolerate a certain therapy and information bias that therapy can change the covariates. A new user design is prefarable for examination of the importance of any treatment[35].

Table 3 Selected sources of bias

Bias	Description
Selection bias	subjects chosen for the study are not representative of the population of interest
Prevalence-incidence (Neyman) bias	A late look at those with a disease or condition will miss early problems and those that have died
Admission rate (Berkson) bias	A hospital based study of the relation between a disease and some exposure will be biased if patients with the disease are more or less admitted to hospital depending on the exposure of interest.
Immortal lifetime bias	When future events are included as baseline data those that have the future event will be immortal until the time when the future data were recorded.
Unmasking (detection signal) bias	An innocent exposure may become associated with disease if it triggers search for a disease.
Volunteer bias	Individuals volunteering for studies or seeking early help for symptoms may be more healthy than non-volunteers or late-comers
Response bias	People who agree to take part in a study have different characteristics from those that do not, and this distorts the results when making conclusions about the whole population
Withdrawal bias	If patients that discontinue a study differ importantly from those that remain in a study the final result may be severely distorted, in particular when only measurements at the end of the study such as rhythm control can enter the analyses
Channeling bias	the propensity of "sicker" or selected patients to be prescribed disproportionately the newer and perceived to be more potent medications differentially.
Confounding by indication, nearly identical to channeling bias	When studying an intervention such as a pharmaceutical drug it may be impossible to distinguish between the risk of the intervention and the risk of the condition that triggered the intervention.
Protopathic bias (reverse causation)	The exposure changes as a result of early disease manifestions. If patients change lifestyle because of early disease signs a wrong direction between lifestyle and disease may be observed.
Information bias	
Recall bias	Information that relies on patient memory may be influenced by their condition. If a relation between a disease and a symptom is

	available to the patient that may help the patient remember a condition.
Insensitive measure bias	If the measurement used in a study does not detect what it is supposed to detect and underestimation of that measurement will be the result.
Regression dilution bias	If a measurement is inaccurate the relation between the measurement and outcome is weakened. For comparison of continuous variables the slope will be reduced.
Follow-up bias	If follow-up depends on the presence of a condition this can create a false relation between a condition and a disease, the direction depending on whether the condition improves or worsens follow-up.
Assessment bias	The assesment and thus collected data on a subject is influence by other factors
Interviewer bias	if an interviewer is aware of the subject's health status, this may influence the questions asked, or how they are asked, which consequently affects the response

Confounding

A confounder is classically defined as a factor which influences both the exposure and the outcome. If for example a study of implantable defibrillators for heart failure is randomized, then we would expect all characteristics of the patients to be equally distributed in the two groups. Factors such as age and sex would be expected to be (nearly) identical in the two groups. And also factors of importance that we do not know (unknown confounders/residual confounders) would be expected to be similar in the two groups. If, on the other hand, the study was observational, then we would expect age and sex to be differently distributed between the two groups. Age and sex would also be expected to be important for survival. In this case age and sex are examples of the classical definition of a confounder: they are unevenly distributed between the treatment groups and they have importance for the outcome.

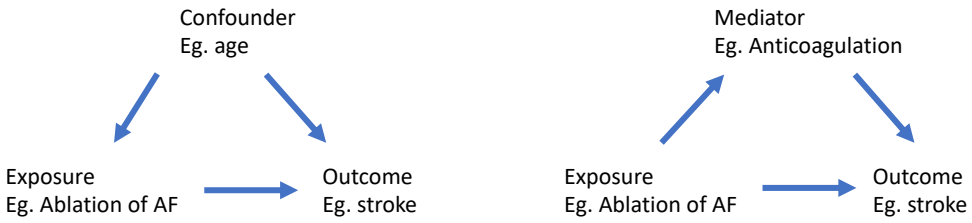
Classical confounders such as age and sex are accounted for by including them as covariates in a multivariable model. The distinction between confounders and model covariates can easily become blurred. Usually we have to select a reasonable number of known factors as potential confounders and use them as covariates in analysis. Directed acyclic graphs (see online supplement) is often a helpful instrument. For example, socioeconomic status of patients could also influence survival and in an observational study socioeconomic status could also influence whether a patient received a defibrillator. If we do not have a recording of socioeconomic status it would be a classical example of an unknown confounder. Ultimately, all observational analyses are potentially subject to bias from unknown confounders.

If we further have a recording of myocardial infarction after implantation, such a variable should not be used in analysis of the importance of the defibrillator. First, the infarction comes after study start. A patient obviously cannot die before the infarction and therefore, an immortal lifetime bias is introduced in a simple analysis. Further, the infarction lies on the pathological pathway between having a defibrillator and the outcome of mortality. It is an intermediate and intermediates should not be used as confounder. Because of its position on the pathway between defibrillator and death it might distort the result if by some mechanism there was an association between getting a defibrillator

and the risk of a myocardial infarction. For a more technical approach to confounding we refer to previous literature.[36, 37]

Mediation

A mediator or intermediate variable is a variable/factor which lies on the pathological path between the exposure of interest and the outcome. Figure 1 shows the major difference between a mediator and a confounder. Appropriate analysis of mediators is complex and there is further explanation in the online Appendix. Mediators should not be treated as confounders.



Legend: Figure 1 – Directed acyclic graphs of a confounder and a mediator

Causal inference

Causal inference is a framework to derive average treatment effects from observational studies with the ultimate aim (or hope) of demonstrating a causal interpretation. If the above study of defibrillators to patients with heart failure was randomized and we after a year found that the mortality with a defibrillator was 4% and 7% without a defibrillator. We could then calculate the **average treatment effect** at one year of 3%. Assuming that the trial was also statistically significant that average treatment effect would be a very important message and easily used to calculate the number of patients to treat to save a life (over one year).
If our study on the other hand was observational we might also have a difference in mortality of 3% after one year. But we would have age, sex and other factors being different in the two group, so we could not expect the 3% to hold for the average patient even if we have no unknown confounders. We could present a multivariable model with hazard ratios or odds ratios, but the average treatment effect from the randomized trial and the number needed to treat would not be available.

Causal inference is a framework to derive the average treatment effect of an observational study providing that we have perfect adjustment for **all** confounders. From a clinical perspective two methods from causal inference are useful and used: **Propensity adjustment** and the **g-formula**. The reader interested in further detail including formal assumptions is referred to an excellent book on the subject: "Causal inference".[38]

In the case of **propensity score matching**, using regression analysis, we would calculate the "propensity" for getting a defibrillator for the entire cohort, including those with and without a defibrillator. This is simply the probability of getting a defibrillator given the covariates. We would then match patients with and without a defibrillator as having very similar probability of getting one.

We would discard patients from the analysis when they cannot be reasonably matched. When the technique is successful, we have a moderately smaller sample than we started with and a demographic table that shows similar covariate distribution in both groups. We can then use the same instruments as we used in the randomized study to obtain **average treatment effect** (actually average treatment effect of the treated) and number needed to treat. The **pitfalls** of this method arrive when the covariates actually do not predict treatment and the demographic table after matching does not show a good balance.

Causal inference provide average treatment effects as do randomised studies, but observational studies are not randomised and therefore the presence of unknown or unmeasured confounders may drive differences. Only large randomised studies assure control of unmeasured confounders..

A technique related to propensity score matching is **inverse probability weighting.** With this technique cases are given a weight corresponding their probability of receiving the treatment of interest. This technique can also provide average treatment effect. It has the advantage that all patients are included in the analysis[39].

While propensity matching is commonly used it has the important disadvantage that not all patients can be matched and commonly not all covariates are evenly distributed after matching. Another technique that has become available is to simulate a randomized trial where first **all** the patients in the study receive a defibrillator and afterwards all patients do not get a defibrillator. This technique is called the **G-formula** and it relies on using statistical models to predict the outcome of every patient first with a defibrillator and then without a defibrillator. Using this simulated study we can calculate **average treatment effect** and number needed to treat using suitable techniques.[38] In propensity score matching of the defibrillator study it was a requirement that the covariates predict whether a patients gets a defibrillator. The G-model does not have this requirement, but the requirement that the covariates predict the outcome accurately and that there are no unknown confounders.

The G-formula and propensity based techniques are not competing techniques, but each has advantages and disadvantages – and both allow calculation of average treatment effects and numbers needed to treat.

Statistical Modelling

Addressing again an observational study of defibrillators to patients with heart failure we would expect to find that age, sex and other variables would differ among patients with and without a defibrillator. The most basic technique for handling this is stratification – to study independently young versus old and men versus women etc. This is useful if there are few variables with few values which is rarely the case. Another technique is to match patients with and without defibrillators and having the same age, sex etc. This is a very efficient technique but usually fails because it is not possible to find a match for many patients. Instead of matching on each variable we could turn to propensity score matching above which may or may not solve our matching problem.

The alternative to matching and stratification is a statistical model and table 4 lists commonly used models. Such models output parameter estimates which after transformation provide odds ratios, hazard ratios or rate ratios. If these measures are statistically significant there is an association between a factor of interest and the outcome of interest. This may be entirely useful for a study of whether a factor has some importance for an outcome, but it is important to realize that this importance cannot be interpreted as prediction. It is therefore important to determine whether the object of a study is to explain or to predict[40]. Some uncertainty arises from the fact that "risk"

and "prediction" do not have universally defined mathematical equivalents. For the current account **prediction** is defined as the absolute risk at a defined time horizon. There is a recent example from the hypertension field[41]. This study used hazard ratios to argue for a value of ambulatory blood pressure, but the aim was to examine whether ambulatory blood pressure improved prediction of cardiovascular outcomes. When encouraged to actually calculate a change in prediction the actual improvement in predictive value was very small.[41, 42] For a study of this nature it would be natural to focus on predictive value rather than on hazard ratios.[43] There is plenty of literature to show that even very high or low hazard ratios may have little relation to prediction. [44-48] In general, whenever the importance of a new treatment or a new biomarker is involved it should be considered whether prediction is the more important estimate to calculate.

C-index / Area under a receiver operator curve.

Let us assume that we want to examine whether late potentials add to prediction of cardiovascular mortality in patients with heart failure. A simple approach would be to present the hazard ratio of some cutoff of late potentials. If this was significant, we could assume late potentials to have some importance. But as described above in the section on hazard ratio and below with competing risk we would not have assurance that we can predict cardiovascular mortality at e.g. 5 years. The right method to show the benefit of a "new" biomarker such as the suggested late potentials demonstrate that a properly selected C-index or area under a receiver operator curve is significantly changed by a new biomarker.[44, 46] This is a field in development with several pitfalls. Thus the commonly used methods of integrated discrimination improvement (IDI) and net reclassification index(NRI)[49] are not valid. Addition of random data to datasets can improve the parameters. The C-index from a Cox model should also not be used to indicate discriminative improvement at specific times.[50]

The bottom line for selection of statistical models is to ensure such a discussion between statisticians and clinicians that the statistical methods used match the clinical question. If the aim e.g. is to estimate the survival benefit of a defibrillator in heart failure after 5 years then a model that address prediction should be used. If it is sufficient to know that the defibrillator does "something", then models that provide hazard ratio, rate ratio or odds ratio may suffice.

Table 4 – Common epidemiological modelling methods

Model	Description	Critical assumptions
Cox proportional hazard	Models risk as hazard ratio, there is a single non-parametric time scale	Proportional hazard assumption – the ratio between hazards needs to be constant
Poisson regression	Time is split into interval as dependent of up to many time scales and timing of covariates	The rate of events needs to be constant in intervals
Logistic Regression	Examines only the outcome as usually a bivariate outcome	Can be used in outcome studies when there is no censoring
G-modelling	Causal inference - One of the above models is used to predict outcome at a time point for the WHOLE study population	Simulates a randomized experiment where the whole study population is subjected to all treatments – assumes no residual confounding

Matching on covariates prior to modelling	Reduces modeling assumptions by perfect adjustment for the matching covariates. The sample size may be reduced	Requires that the selected covariates define necessary confounding and lack of important unknown confounders.
Propensity stratified models	Uses covariates to calculate the probability of receiving one of two treatments and then compares outcome in strata of that probability	Assumes that the difference in treatment is perfectly explained by the probability of receiving treatment
Propensity matched models	The propensity is calculated as above and then cases with same or very similar probability in two groups are matched	Same as above, depending on the matching the sample size may be reduced

Competing risk

Let us assume in the study of defibrillators for heart failure that we were not so much interested in all cause mortality but rather in cardiovascular mortality. This would not be unreasonable since defibrillators can only influence cardiovascular mortality. This has important consequences for the analysis. The competing risk of death from other causes than cardiovascular mortality cannot be ignored and the cumulative cardiovascular mortality presentation needs to take into account the competing risk with proper technique.[51]

Competing risk has for technical reasons no influence on the calculation of hazard ratios, but the interpretation of hazard ratio becomes complex. In fact, there is no certainty that a significant hazard ratio influences long term prediction such as 5 year cardiovascular mortality and dedicated analysis of prediction is necessary if this is the goal.

Instrumental variable analysis

A good instrument is a variable that affects an outcome and is not affected by confounders. The only common example in clinical medicine is “mendelian randomization”. With this technique genes that influence a factor of interest is used instead of directly addressing the factor. Since genes have been there prior to establishing the influence of important confounders that could be age and smoking the confounding by these can be avoided. More detail is provided in the online Appendix. It is important to appreciate the limitations and a good reference is Federspiel et al[52].

Missing data

Missing data are common in observational studies and most statistical procedures exclude individuals with missing data. If in the study of defibrillators for heart failure and important variable such as age is missing for some patients it could bias the interpretation of the study if these patients are simply removed from the analysis. There are a number of useful techniques to include as much information as possible from cases with missing data and these are described further in the online Appendix.

Common problems

Causality versus association

Observational studies will by their nature always include a risk of bias from unknown or unobserved confounders. Causal language is common and a very common task for reviewers is to request the removal of causal language from observational manuscripts. It can be argued that in stating the objective of a study a causal language should be used.[53]

Conditioning on the future

Conditioning on the future is when information is obtained some time in the future compared to baseline is included as baseline information. Patients that pick up a prescription cannot die before that day while patients dying prior to reaching the pharmacy never pick up a prescription. Using the prescription information at baseline will bias survival towards those that pick up a prescription – the immortal lifetime bias.[54] It is a very similar problem if patients are excluded from a study because of events after baseline – This will in a very similar manner bias survival towards those that do not have the factor that caused exclusion. Friberg et al.[55, 56] studied stroke in atrial fibrillation not treated with anticoagulation. By excluding patients who received anticoagulation during the study a bias was introduced. This particular bias was examined in a different study[57] that demonstrated a bias towards lower stroke rate with low CHA_2DS_2-VASc by excluding after baseline.

Metaanalysis of observational studies

Meta-analyses of RCTs assume that each individual study provides an unbiased estimate of the effect and any variability between study results is attributed to random variation[58, 59]. The overall effect will provide an unbiased estimate, as long as the studies are representative and wisely combined[58, 59]. While RCTs, if properly designed, are expected to have a high internal validity, they traditionally have the limitations of smaller sample sizes, very selected populations, shorter follow-up time, ethical constrains and high cost[60, 61]. Incorporating non-randomised trials into meta-analyses can overcome some of these limitations by improving generalisability (more diverse populations), allowing larger sample sizes, allowing exploring aetiological hypothesis (unethical to deliberately expose patients to harmful risk factors in an RCT), and evaluating less common adverse effects[60-62].

Observational studies, however, have a higher risk of bias and confounding and, as a consequence, the association estimates may differ from the truth beyond the effect of chance[63, 64]. The individual studies may measure and control for known confounding factors during the analysis. However, even if this is case, bias and residual confounding (i.e. when the confounding factor cannot be measured with sufficient precision[65, 66]) remain a relevant threat to validity in observational research[67]. As a consequence, using non-randomised studies in meta-analysis could (more often than not) perpetuate the biases that are unknown, unmeasured or uncontrolled in these observational studies, and threaten the validity of the entire meta-analysis[64, 67, 68]. Furthermore, reporting in observational studies is frequently not sufficiently detailed to judge their limitations[67, 69-71], they show significant heterogeneity[72-74] and deficiencies in methodology[68, 75, 76]. Network meta-analyses (i.e. meta-analyses that compare simultaneously multiple treatment options) incorporating non-randomised trials, face similar challenges[77].

For these reasons, some authors recommend abandoning meta-analyses of observational data[64, 78, 79]. Yet, when evaluating effect sizes derived from meta-analyses of RCTs and non-randomised studies, discrepancies have shown to be small in high quality observational studies with little heterogeneity[60, 80-83]. Still, discrepancies beyond chance do happen and it is therefore essential to assess the differences between studies[61, 64]. In our –and other authors'- view, gross statistical combination of data alone should be avoided; rather, a thorough analysis of heterogeneity sources and possible bias should be done[61, 73, 84, 85]; this will probably provide better understanding than an overall effect measure, which can potentially be misleading[73].

In 1999, the *Quality of Reporting of Meta-analyses* (QUOROM) statement was issued "to address standards for improving the quality of reporting of meta-analyses of RCTs"[86]. A similar checklist was published in 2000 for reporting *Meta-analyses Of Observational Studies in Epidemiology* (MOOSE)[73]. However, in the face of persistent poor reporting[69, 70, 87-94], these statements were later on updated in the form of the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) statements[95-101]. Many peer-reviewed journals now require that these guidelines are followed when submitting a systematic review or meta-analyses, as the endorsement of these statements improves both reporting and methodological quality[102, 103]; however, there is still room for improvement[104-107]. For editors, reviewers and readers, a measurement tool to assess the methodological quality of systematic reviews (AMSTAR) has also been published and validated[108-110].

Consensus Statements on Observational Studies

	.	Refs
When reporting the results of an observational study/metaanalysis, the STROBE/PRISMA statement checklists should be used: STROBE – www.strobe-statement.org PRISMA – www.prisma-statement.org		
Prior to analysis an analysis plan should be agreed upon and formally recorded		www.strobe-statement.org
The process of data collection should be clearly presented so that the strengths and limitations are clear to the reader.		
If legally possible data should be available for scrutiny by other researchers.		
Studies should have clear objective and use statistical methods that match the objectives		[38]
The reporting of findings should be complete and the strengths and limitations clearly described		
Sources of bias should be identified and presented to the reader		

Conclusion

Observational studies should in general use transparent and valid methodology and use concise reporting. There are available guidelines for epidemiological studies and the most recent is from the International Society of Pharmacoepidemiology.[111] The guideline from the International Society of Pharmacoepidemiology also cites a number of other guidelines. None of the recommendations are in discordance with the current consensus statement. There does not appear to be widely accepted international guidelines for "good epidemiological practice".[112] Finally, an important intermediate step is to ensure that biostatisticians and clinical practitioners both have sufficient insight into the language and methods of each other to ensure that valid studies are conducted and the many pitfalls avoided.

References

1. Faranoff, A.C., et al., *Levels of Evidence Supporting American College of Cardiology/American Heart Association and European Society of Cardiology Guidelines, 2008-2018.* JAMA, 2019. **321**(11): p. 1069-1080.
2. Frieden, T.R., *Evidence for Health Decision Making - Beyond Randomized, Controlled Trials.* N Engl J Med, 2017. **377**(5): p. 465-475.
3. Shikata, S., et al., *Comparison of effects in randomized controlled trials with observational studies in digestive surgery.* Ann Surg, 2006. **244**(5): p. 668-76.
4. Blake, K., *Postmarket surveillance of medical devices: current capabilities and future opportunities.* J Interv Card Electrophysiol, 2013. **36**(2): p. 119-27.
5. Lip, G.Y., et al., *A prospective survey in European Society of Cardiology member countries of atrial fibrillation management: baseline results of EURObservational Research Programme Atrial Fibrillation (EORP-AF) Pilot General Registry.* Europace, 2014. **16**(3): p. 308-19.
6. Lip, G.Y., et al., *'Real-world' antithrombotic treatment in atrial fibrillation: The EORP-AF pilot survey.* Am J Med, 2014. **127**(6): p. 519-29.e1.
7. Lip, G.Y., et al., *Prognosis and treatment of atrial fibrillation patients by European cardiologists: one year follow-up of the EURObservational Research Programme-Atrial Fibrillation General Registry Pilot Phase (EORP-AF Pilot registry).* Eur Heart J, 2014. **35**(47): p. 3365-76.
8. Boriani, G., et al., *Glomerular filtration rate in patients with atrial fibrillation and 1-year outcomes.* Sci Rep, 2016. **6**: p. 30271.
9. Boriani, G., et al., *Asymptomatic atrial fibrillation: clinical correlates, management, and outcomes in the EORP-AF Pilot General Registry.* Am J Med, 2015. **128**(5): p. 509-18.e2.
10. Boriani, G., et al., *Changes to oral anticoagulant therapy and risk of death over a 3-year follow-up of a contemporary cohort of European patients with atrial fibrillation final report of the EURObservational Research Programme on Atrial Fibrillation (EORP-AF) pilot general registry.* Int J Cardiol, 2018. **271**: p. 68-74.
11. Boriani, G., et al., *Contemporary stroke prevention strategies in 11 096 European patients with atrial fibrillation: a report from the EURObservational Research Programme on Atrial Fibrillation (EORP-AF) Long-Term General Registry.* Europace, 2018. **20**(5): p. 747-757.
12. Proietti, M., et al., *Increased burden of comorbidities and risk of cardiovascular death in atrial fibrillation patients in Europe over ten years: A comparison between EORP-AF pilot and EHS-AF registries.* Eur J Intern Med, 2018. **55**: p. 28-34.

13. Thompson, L.E., et al., *Sex Differences in the Use of Oral Anticoagulants for Atrial Fibrillation: A Report From the National Cardiovascular Data Registry (NCDR((R))) PINNACLE Registry.* J Am Heart Assoc, 2017. **6**(7).
14. Lubitz, S.A., et al., *Predictors of oral anticoagulant non-prescription in patients with atrial fibrillation and elevated stroke risk.* Am Heart J, 2018. **200**: p. 24-31.
15. Marzec, L.N., et al., *Influence of Direct Oral Anticoagulants on Rates of Oral Anticoagulation for Atrial Fibrillation.* J Am Coll Cardiol, 2017. **69**(20): p. 2475-2484.
16. Hsu, J.C., et al., *International Collaborative Partnership for the Study of Atrial Fibrillation (INTERAF): Rationale, Design, and Initial Descriptives.* J Am Heart Assoc, 2016. **5**(11).
17. Lopes, R.D., et al., *Rationale and design of the First Brazilian Cardiovascular Registry of Atrial Fibrillation: The RECALL study.* Am Heart J, 2016. **176**: p. 10-6.
18. Kim, D., et al., *10-year nationwide trends of the incidence, prevalence, and adverse outcomes of non-valvular atrial fibrillation nationwide health insurance data covering the entire Korean population.* Am Heart J, 2018. **202**: p. 20-26.
19. Kim, D., et al., *Increasing trends in hospital care burden of atrial fibrillation in Korea, 2006 through 2015.* Heart, 2018.
20. Pallisgaard, J.L., et al., *Temporal trends in atrial fibrillation recurrence rates after ablation between 2005 and 2014: a nationwide Danish cohort study.* Eur Heart J, 2018. **39**(6): p. 442-449.
21. Nielsen, P.B., et al., *Outcomes Associated With Resuming Warfarin Treatment After Hemorrhagic Stroke or Traumatic Intracranial Hemorrhage in Patients With Atrial Fibrillation.* JAMA Intern Med, 2017. **177**(4): p. 563-570.
22. Nielsen, P.B., et al., *Effectiveness and safety of reduced dose non-vitamin K antagonist oral anticoagulants and warfarin in patients with atrial fibrillation: propensity weighted nationwide cohort study.* BMJ, 2017. **356**: p. j510.
23. Kim, T.H., et al., *CHA2DS2-VASc Score (Congestive Heart Failure, Hypertension, Age >/=75 [Doubled], Diabetes Mellitus, Prior Stroke or Transient Ischemic Attack [Doubled], Vascular Disease, Age 65-74, Female) for Stroke in Asian Patients With Atrial Fibrillation: A Korean Nationwide Sample Cohort Study.* Stroke, 2017. **48**(6): p. 1524-1530.
24. Kim, T.H., et al., *CHA2DS2-VASc Score for Identifying Truly Low-Risk Atrial Fibrillation for Stroke: A Korean Nationwide Cohort Study.* Stroke, 2017. **48**(11): p. 2984-2990.
25. Savarese, G., et al., *Reasons for and consequences of oral anticoagulant underuse in atrial fibrillation with heart failure.* Heart, 2018. **104**: p. 1093-1100.
26. Karayiannides, S., et al., *High overall cardiovascular risk and mortality in patients with atrial fibrillation and diabetes: A nationwide report.* Diab Vasc Dis Res, 2018. **15**(1): p. 31-38.
27. Chao, T.F., et al., *Major bleeding and intracranial hemorrhage risk prediction in patients with atrial fibrillation: Attention to modifiable bleeding risk factors or use of a bleeding risk stratification score? A nationwide cohort study.* Int J Cardiol, 2018. **254**: p. 157-161.
28. Chao, T.F., et al., *Relationship of Aging and Incident Comorbidities to Stroke Risk in Patients With Atrial Fibrillation.* J Am Coll Cardiol, 2018. **71**(2): p. 122-132.
29. Chao, T.F., et al., *Lifetime Risks, Projected Numbers, and Adverse Outcomes in Asian Patients With Atrial Fibrillation: A Report From the Taiwan Nationwide AF Cohort Study.* Chest, 2018. **153**(2): p. 453-466.
30. Hsing, A.W. and J.P. Ioannidis, *Nationwide Population Science: Lessons From the Taiwan National Health Insurance Research Database.* JAMA Intern Med, 2015. **175**(9): p. 1527-9.

31. Lee, S.R., et al., *Edoxaban in Asian Patients With Atrial Fibrillation: Effectiveness and Safety.* J Am Coll Cardiol, 2018. **72**(8): p. 838-853.
32. Noseworthy, P.A., et al., *Comparative effectiveness and safety of non-vitamin K antagonist oral anticoagulants versus warfarin in patients with atrial fibrillation and valvular heart disease.* Int J Cardiol, 2016. **209**: p. 181-3.
33. Lip, G.Y.H., et al., *Effectiveness and Safety of Oral Anticoagulants Among Nonvalvular Atrial Fibrillation Patients.* Stroke, 2018. **49**(12): p. 2933-2944.
34. Sackett, D.L., *Bias in analytic research.* J Chronic Dis, 1979. **32**(1-2): p. 51-63.
35. Ray, W.A., *Evaluating medication effects outside of clinical trials: new-user designs.* Am J Epidemiol, 2003. **158**(9): p. 915-20.
36. Rothman, K.J., S. Greenland, and T.L. Lash, *Modern Epidemiology*. 2012, Philidelphia: Wolters Kluwer.
37. Hernan, M.A., S. Hernandez-Diaz, and J.M. Robins, *A structural approach to selection bias.* Epidemiology, 2004. **15**(5): p. 615-25.
38. Hernán MA, R.J., *Causal Inference*. 2018, Boca Raton: Chapman & Hall/CRC.
39. Austin, P.C. and E.A. Stuart, *Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies.* Stat Med, 2015. **34**(28): p. 3661-79.
40. Shmueli, G., *To Explain or to Predict.* Statistical Science, 2010. **25**(3): p. 289-310.
41. Banegas, J.R., et al., *Relationship between Clinic and Ambulatory Blood-Pressure Measurements and Mortality.* N Engl J Med, 2018. **378**(16): p. 1509-1520.
42. Torp-Pedersen, C., *Ambulatory Blood Pressure and Mortality.* N Engl J Med, 2018. **379**(13): p. 1285-6.
43. Mortensen, R.N., et al., *Office blood pressure or ambulatory blood pressure for the prediction of cardiovascular events.* Eur Heart J, 2017. **38**(44): p. 3296-3304.
44. Hlatky, M.A., et al., *Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association.* Circulation, 2009. **119**(17): p. 2408-16.
45. Kattan, M.W., *Judging new markers by their ability to improve predictive accuracy.* J Natl Cancer Inst, 2003. **95**(9): p. 634-5.
46. Kattan, M.W., *Evaluating a new marker's predictive contribution.* Clin Cancer Res, 2004. **10**(3): p. 822-4.
47. Pepe, M.S., et al., *Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker.* Am J Epidemiol, 2004. **159**(9): p. 882-90.
48. Hernan, M.A., *The hazards of hazard ratios.* Epidemiology, 2010. **21**(1): p. 13-5.
49. Hilden, J. and T.A. Gerds, *A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index.* Stat Med, 2014. **33**(19): p. 3405-14.
50. Blanche, P., M.W. Kattan, and T.A. Gerds, *The c-index is not proper for the evaluation of t-year predicted risks.* Biostatistics, 2018.
51. D.G., K. and K. M., *Competing Risks Suvival Aanalysis*, in *Survival Analysis: A Self-Learning Text*. 2012, Springer: New York. p. 425-95.
52. Federspiel, J.J., et al., *Comparing Inverse Probability of Treatment Weighting and Instrumental Variable Methods for the Evaluation of Adenosine Diphosphate Receptor Inhibitors After Percutaneous Coronary Intervention.* JAMA Cardiol, 2016. **1**(6): p. 655-65.

53. Hernan, M.A., *The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data.* Am J Public Health, 2018. **108**(5): p. 616-619.
54. Lund, J.L., et al., *Conditioning on future exposure to define study cohorts can induce bias: the case of low-dose acetylsalicylic acid and risk of major bleeding.* Clin Epidemiol, 2017. **9**: p. 611-626.
55. Friberg, L., M. Skeppholm, and A. Terent, *Benefit of anticoagulation unlikely in patients with atrial fibrillation and a CHA2DS2-VASc score of 1.* J Am Coll Cardiol, 2015. **65**(3): p. 225-32.
56. Aspberg, S., et al., *Comparison of the ATRIA, CHADS2, and CHA2DS2-VASc stroke risk scores in predicting ischaemic stroke in a large Swedish cohort of patients with atrial fibrillation.* Eur Heart J, 2016. **37**(42): p. 3203-3210.
57. Nielsen, P.B., et al., *Stroke and thromboembolic event rates in atrial fibrillation according to different guideline treatment thresholds: A nationwide cohort study.* Sci Rep, 2016. **6**: p. 27410.
58. Sutton, A.J. and K.R. Abrams, *Bayesian methods in meta-analysis and evidence synthesis.* Statistical Methods in Medical Research, 2001. **10**(4): p. 277-303.
59. Deeks, J.J., J.P.T. Higgins, and D.G. Altman, *Analysing Data and Undertaking Meta-Analyses*, in *Cochrane Handbook for Systematic Reviews of Interventions (Version 5.1.0)*, J.P.T. Higgins and S. Green, Editors. 2011: Available at: http://handbook-5-1.cochrane.org/chapter_9/9_analysing_data_and_undertaking_meta_analyses.htm.
60. Concato, J., N. Shah, and R.I. Horwitz, *Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs.* New England Journal of Medicine, 2000. **342**(25): p. 1887-1892.
61. Ioannidis, J.A., et al., *Comparison of evidence of treatment effects in randomized and nonrandomized studies.* JAMA, 2001. **286**(7): p. 821-830.
62. Lipsett, M. and S. Campleman, *Occupational exposure to diesel exhaust and lung cancer: a meta-analysis.* American Journal of Public Health, 1999. **89**(7): p. 1009-1017.
63. Grimes, D.A. and K.F. Schulz, *Bias and causal associations in observational research.* The Lancet, 2002. **359**(9302): p. 248-252.
64. Deeks, J.J., et al., *Evaluating non-randomised intervention studies.* Health Technology Assessment 2003. **7**(27): p. iii-x, 1-173.
65. Phillips, A.N. and G.D. Smith, *How independent are "independent" effects? relative risk estimation when correlated exposures are measured imprecisely.* Journal of Clinical Epidemiology, 1991. **44**(11): p. 1223-1231.
66. Smith, G.D. and A.N. Phillips, *Confounding in epidemiological studies: why "independent" effects may not be all they seem.* BMJ : British Medical Journal, 1992. **305**(6856): p. 757-759.
67. Groenwold, R.H.H., et al., *Poor Quality of Reporting Confounding Bias in Observational Intervention Studies: A Systematic Review.* Annals of Epidemiology, 2008. **18**(10): p. 746-751.
68. Lijmer, J.G., et al., *Empirical evidence of design-related bias in studies of diagnostic tests.* JAMA, 1999. **282**(11): p. 1061-1066.
69. Hemels, M.E.H., et al., *Quality assessment of meta-analyses of RCTs of pharmacotherapy in major depressive disorder.* Current Medical Research and Opinion, 2004. **20**(4): p. 477-484.
70. Dixon, E., et al., *Evaluating Meta-analyses in the General Surgical Literature: A Critical Appraisal.* Annals of Surgery, 2005. **241**(3): p. 450-459.

71. Vandenbroucke, J.P., et al., *Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration.* Annals of Internal Medicine, 2007. **147**(8): p. W163-94.
72. Maguire, M.J., et al., *Overwhelming heterogeneity in systematic reviews of observational anti-epileptic studies.* Epilepsy Research, 2008. **80**(2): p. 201-212.
73. Stroup, D.F., et al., *Meta-analysis of observational studies in epidemiology: A proposal for reporting.* JAMA, 2000. **283**(15): p. 2008-2012.
74. IntHout, J., et al., *Small studies are more heterogeneous than large ones: a meta-meta-analysis.* Journal of Clinical Epidemiology. **68**(8): p. 860-869.
75. von Elm, E., et al., *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies.* Lancet, 2007. **370**(9596): p. 1453-1457.
76. Simunovic, N., S. Sprague, and M. Bhandari, *Methodological Issues in Systematic Reviews and Meta-Analyses of Observational Studies in Orthopaedic Research.* J Bone Joint Surg Am, 2009. **91**(Supplement_3): p. 87-94.
77. Cameron, C., et al., *Network meta-analysis incorporating randomized controlled trials and non-randomized comparative cohort studies for assessing the safety and effectiveness of medical treatments: challenges and opportunities.* Systematic Reviews, 2015. **4**(1): p. 147.
78. Shapiro, S., *Meta-analysis/Shmeta-analysis.* American Journal of Epidemiology, 1994. **140**(9): p. 771-778.
79. Kunz, R. and A.D. Oxman, *The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials.* British Medical Journal, 1998. **317**(7167)): p. 1185-90.
80. MacLehose, R.R., et al., *A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies.* Health Technol Assess, 2000. **4**(34): p. 1-154.
81. Benson, K. and A.J. Hartz, *A Comparison of Observational Studies and Randomized, Controlled Trials.* New England Journal of Medicine, 2000. **342**(25): p. 1878-1886.
82. Shrier, I., et al., *Should Meta-Analyses of Interventions Include Observational Studies in Addition to Randomized Controlled Trials? A Critical Examination of Underlying Principles.* American Journal of Epidemiology, 2007. **166**(10): p. 1203-1209.
83. Anglemyer, A., H.T. Horvath, and L. Bero, *Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials.* Cochrane Database of Systematic Reviews, 2014(4).
84. Egger, M., M. Schneider, and G.D. Smith, *Spurious precision? Meta-analysis of observational studies.* BMJ, 1998. **316**(7125): p. 140-144.
85. Higgins, J.P.T., et al., *Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice.* Journal of Health Services Research & Policy, 2002. **7**(1): p. 51-61.
86. Moher, D., et al., *Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement.* The Lancet, 1999. **354**(9193): p. 1896-1900.
87. Moher, D., et al., *Epidemiology and Reporting Characteristics of Systematic Reviews.* PLOS Medicine, 2007. **4**(3): p. e78.
88. Moher, D., et al., *Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research.* BMC Medicine, 2008. **6**: p. 13-13.

89. Wen, J., et al., *The reporting quality of meta-analyses improves: a random sampling study.* Journal of Clinical Epidemiology, 2008. **61**(8): p. 770-775.
90. Gianola, S., et al., *Survey of the Reporting Characteristics of Systematic Reviews in Rehabilitation.* Physical Therapy, 2013. **93**(11): p. 1456-1466.
91. Page, M.J., et al., *Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions.* Cochrane Database of Systematic Reviews, 2014(10).
92. Peters, J.P.M., et al., *Reporting Quality of Systematic Reviews and Meta-Analyses of Otorhinolaryngologic Articles Based on the PRISMA Statement.* PLOS ONE, 2015. **10**(8): p. e0136540.
93. Page, M.J., et al., *Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study.* PLOS Medicine, 2016. **13**(5): p. e1002028.
94. Cullis, P.S., K. Gudlaugsdottir, and J. Andrews, *A systematic review of the quality of conduct and reporting of systematic reviews and meta-analyses in paediatric surgery.* PLOS ONE, 2017. **12**(4): p. e0175213.
95. Moher, D., et al., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* British Medical Journal, 2009. **339**(7716): p. 332-336.
96. Moher, D., et al., *Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement.* Systematic Reviews, 2015. **4**(1): p. 1-9.
97. Hutton, B., et al., *The prisma extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: Checklist and explanations.* Annals of Internal Medicine, 2015. **162**(11): p. 777-784.
98. Stewart, L.A., et al., *Preferred reporting items for a systematic review and meta-analysis of individual participant data: The prisma-ipd statement.* JAMA, 2015. **313**(16): p. 1657-1665.
99. Zorzela, L., et al., *PRISMA harms checklist: improving harms reporting in systematic reviews.* British Medical Journal (Online), 2016. **352**: p. i157.
100. Guise, J.-M., et al., *AHRQ series on complex intervention systematic reviews—paper 6: PRISMA-CI extension statement and checklist.* Journal of Clinical Epidemiology, 2017. **90**: p. 43-50.
101. Guise, J.-M., et al., *AHRQ series on complex intervention systematic reviews—paper 7: PRISMA-CI elaboration and explanation.* Journal of Clinical Epidemiology, 2017. **90**: p. 51-58.
102. Panic, N., et al., *Evaluation of the Endorsement of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) Statement on the Quality of Published Systematic Review and Meta-Analyses.* PLOS ONE, 2013. **8**(12): p. e83138.
103. Tunis, A.S., et al., *Association of Study Quality with Completeness of Reporting: Have Completeness of Reporting and Quality of Systematic Reviews and Meta-Analyses in Major Radiology Journals Changed Since Publication of the PRISMA Statement?* Radiology, 2013. **269**(2): p. 413-426.
104. Riado Minguez, D., et al., *Methodological and Reporting Quality of Systematic Reviews Published in the Highest Ranking Journals in the Field of Pain.* Anesthesia & Analgesia, 2017. **125**(4): p. 1348-1354.
105. Pussegoda, K., et al., *Systematic review adherence to methodological or reporting quality.* Systematic Reviews, 2017. **6**(1): p. 131.
106. Page, M.J. and D. Moher, *Evaluations of the uptake and impact of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement and extensions: a scoping review.* Systematic Reviews, 2017. **6**(1): p. 263.

107. Zhang, Z.-w., et al., *Epidemiology, quality and reporting characteristics of meta-analyses of observational studies published in Chinese journals.* BMJ Open, 2015. **5**(12).
108. Shea, B.J., et al., *Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews.* BMC Medical Research Methodology, 2007. **7**(1): p. 1-7.
109. Shea, B.J., et al., *AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews.* Journal of Clinical Epidemiology, 2009. **62**(10): p. 1013-1020.
110. Pieper, D., et al., *Systematic review found AMSTAR, but not R(evised)-AMSTAR, to have good measurement properties.* Journal of Clinical Epidemiology, 2015. **68**(5): p. 574-583.
111. Public Policy Committee, I.S.o.P., *Guidelines for good pharmacoepidemiology practice (GPP).* Pharmacoepidemiol Drug Saf, 2016. **25**(1): p. 2-10.
112. Alba, S. and C. Mergenthaler, *Lies, damned lies and epidemiology: why global health needs good epidemiological practice guidelines.* BMJ Glob Health, 2018. **3**(5): p. e001019.