DANISH MEDICAL JOURNAL

Invited State-Of-The-Art Review

Dan Med J 2023;70(4):A12220796

Health registries as research tools: a review of methodological key issues

Sissel Toft Sørensen[1, 2], Frederik Pagh Kristensen[1, 2], Frederikke Schønfeldt Troelsen[1, 2], Morten Schmidt[1, 2] & Henrik Toft Sørensen[1, 2]

1) Department of Clinical Epidemiology, Aarhus University Hospital, 2) Department of Clinical Epidemiology, Aarhus University, Denmark

Dan Med J 2023;70(4):A12220796

ABSTRACT

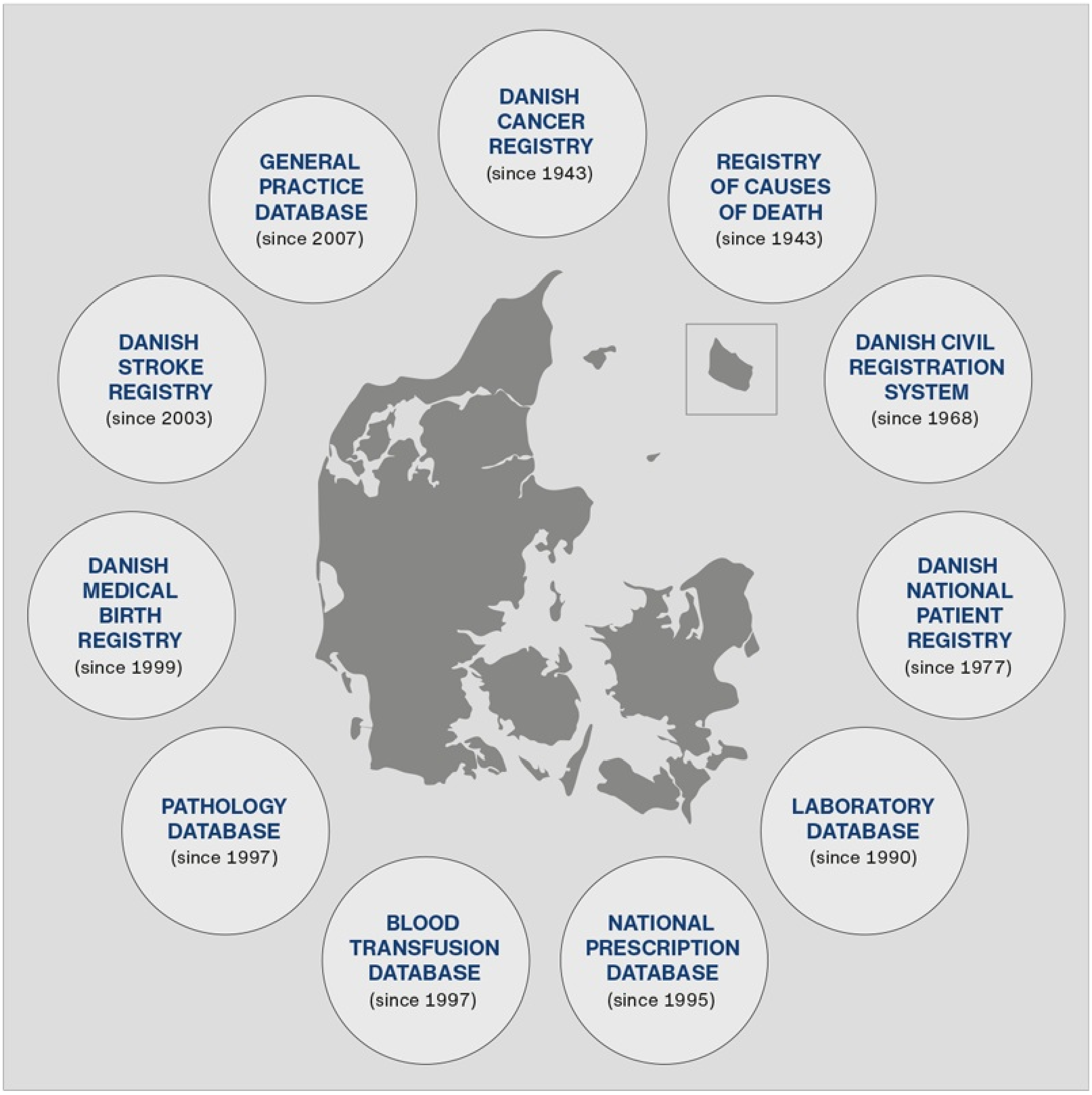
Health registries provide opportunities for conducting large-scale, population-based studies, but attention must be devoted to their specific limitations. Herein, we describe potential limitations that may affect the validity of registry-based research. Our review includes descriptions of 1) populations, 2) variables, 3) medical coding systems for medical information and 4) selected key methodological challenges. Knowledge of such factors and epidemiological study designs in general is likely to increase the quality of registry-based research and reduce potential biases.

KEY POINTS

- Danish health registries offer an opportunity for performing large-scale, population-based studies with complete follow-up.
- The research value of a registry depends on the quality of target population registration and the registry variables.
- Left truncation, right and left censoring and changes in the diagnostic criteria over time may affect the research value of registries.

Denmark and the remaining Nordic countries have a long history of registry-based research. A step forward for registry-based research was the introduction of the personal registration number (CPR number) in 1968. Since then, the CPR number has allowed recording of information on an individual-level basis in various registries [1]. In comparison with many other countries, this is a unique situation that has improved registry-based research in Denmark. Denmark has a large network of population-based registries and databases routinely collecting data as a byproduct of healthcare administration (**Figure 1**) [2, 3]. The Danish health registries and databases may broadly be divided into disease registries (e.g., the Danish Cancer Registry), administrative registries (e.g., the Danish National Patient Registry (DNPR)) and clinical quality databases (e.g., the Stroke Database) [2]. In the following, we use the term registry for both health registries and databases.

FIGURE 1 Examples of Danish health registries, serving as valuables research tools.


DANISH CANCER REGISTRY (since 1943)
REGISTRY OF CAUSES OF DEATH (since 1943)
DANISH CIVIL REGISTRATION SYSTEM (since 1968)
DANISH NATIONAL PATIENT REGISTRY (since 1977)
LABORATORY DATABASE (since 1990)
NATIONAL PRESCRIPTION DATABASE (since 1995)
BLOOD TRANSFUSION DATABASE (since 1997)
PATHOLOGY DATABASE (since 1997)
DANISH MEDICAL BIRTH REGISTRY (since 1999)
DANISH STROKE REGISTRY (since 2003)
GENERAL PRACTICE DATABASE (since 2007)

Health registries offer an opportunity for large-scale, population-based studies with several advantages [4-8]: 1) Their large size improves the precision of estimates and enables the study of rare exposures and outcomes with long-term latency, 2) Inclusion of nearly all individuals in the target population ensures that the data reflect routine clinical care and all clinical segments of the source population, 3) Data are collected independently of each research study, thus minimising certain types of bias, e.g., non-response, and the influence from attention to the research question on the diagnostic process.

However, the limitations of using health registries are often ignored. Important problems encountered in research based on registry data include the degree of sensitivity and specificity for capturing patients with certain diseases and the validity of the information contained in the data. Poor data quality may permanently impede research and pose ethical problems. The timeliness and relevance of the data are important for many research questions. If a substantial lag time exists, this may possibly delay the recognition of important new trends or findings. Data must be collected and reported in a manner that is relevant to the needs of clinicians, public health officers and policy makers.

Below, we review the following health registry characteristics affecting the value of registry-based research: 1) the population, 2) the variables, 3) the coding systems of medical information and 4) selected key methodological challenges.

POPULATION

In Denmark, the entire population is recorded in the Civil Registration System through the CPR number [1]. The Civil Registration System contains information on migration and vital status, thus enabling studies on the entire population with nearly complete follow-up [1]. However, the ability to identify specific patient populations varies across registries. For example, examining diseases diagnosed and treated primarily by general practitioners, such as migraine and hypertension, poses a challenge in Denmark because the diagnoses made by general practitioners are unavailable in Danish administrative registries [1]. The concept of coverage is often used to describe the proportion of eligible cases included in the database.

Eligibility for inclusion in some registries may also change over time, thereby increasing the risk of loss to follow-up. Even in the best registries, some patients will not be included due to data error, failure to record the relevant diagnosis or procedure or administrative error [8]. The term consistency is often related to the use of diagnostic codes over time and whether they have been used in the same way by different hospitals and clinicians over time. These problems also relate to Danish registries although they are a larger concern in many other countries such as the US where, e.g., inclusion in the US Medicaid programme is continuously determined on the basis of health status and socioeconomic factors [9].

VARIABLES

Health registries may be used to obtain data on exposures, outcomes and potential confounders [2]. Registry records contain information on several variables, including diagnoses, surgical procedures, selected hospital-based treatments and prescribed drugs [2]. The validity of a variable is the extent to which it measures what it is intended to measure. Lack of validity is referred to as bias. Routinely conducted systematic validation of variables is not performed in many registries. The two key validity measures used in registry research are sensitivity and positive predictive value. The sensitivity is the ability of the registry to capture individuals with, e.g., a disease. The positive predictive value is the probability that an individual recorded with a characteristic has the given characteristic. Specificity is often very difficult to examine. Review of medical records is often used to examine the positive predictive value of diagnoses in hospital discharge summaries, birth and cancer registry data [10]. Reassuringly, many studies have indicated a high or adequate predictive value for many variables in the main Danish health registries [11, 12]. Nonetheless, predictive value is only one of several validity parameters.

Exposure example

A major strength of the Danish prescription data is that drug use is designated according to actual dispensing rather than prescribing. However, using prescription registry data to assess drug exposure still requires consideration of potential exposure misclassification. Non-steroidal anti-inflammatory drugs (NSAIDs) may provide an example of the strengths and weaknesses that must be considered in assessing drug exposure data from the Danish National Prescription Registry [13], specifically misclassification of true NSAID use as non-use or vice versa [14, 15].

Misclassification of true NSAID use as non-use in registry-based studies results from over-the-counter (OTC) drug sales or in-hospital use. OTC NSAIDs include aspirin in all preparations, diclofenac (during 2007-2008) and low-dose ibuprofen (200 mg tablets) since 27 March 1989 [15]. Regular users of NSAIDs have an economic incentive to obtain the drugs by prescription so that they qualify for reimbursement through the programme of the National Health Service. Moreover, the Danish Health Authorities have restricted dispensing of OTC ibuprofen to adults and 20 tablets per dispensing [15]. Thus, the proportions of total NSAID sales dispensed by prescription, and consequently captured in prescription registries, are high at approximately 90% for low-dose aspirin, 75% for ibuprofen and 100% for all other non-aspirin NSAIDs (as of 2019) [14, 15]. In practice, misclassification due to OTC NSAID use therefore rarely has any impact on effect estimates in registry-based studies [14, 15]. Misclassification of non-use as NSAID use results primarily from patient non-adherence or stockpiling of prescribed NSAIDs or aspirins [15].

Outcomes

Administrative health registries, such as the DNPR, provide a rich data source for identifying outcomes such as diseases and procedures [10]. Thus, the DNPR may be used to study outcomes and prognostic factors in well-defined patient groups (e.g., diagnostic examinations, recurrence and complications) [10]. These patient groups may be identified from the DNPR, other registries or from surveys [10]. The DNPR is also used to gather long-term follow-up data for randomised trials using clinically driven outcome detection [16]. The automated event-detection feature of the DNPR allows large, low-cost randomised trials that reflect daily clinical practice, cover a broad range of patients and endpoints and include lifelong follow-up [17, 18]. As with cohort studies, case-control studies [19] and ecological studies [20], DNPR data may also be used to identify exposures and cases/outcomes.

Confounding example

Confounding may be a challenge in registry-based research due to a lack of recording of key covariates. For example, obesity is a potential confounder in many studies. However, exact BMI measurements cannot be entered into the DNPR, which only allows the diagnosis codes for overweight and obesity. In clinical settings, these diagnoses are rarely used, thus leading to potential unmeasured confounding [21].

CODING SYSTEMS

Most registries use coding systems, and researchers must be familiarised with these. Here, we discuss several of the most commonly used coding systems in Danish and Nordic registry-based research.

International Classification of Diseases

For more than a century, the International Classification of Diseases (ICD) has been the basis for comparable statistics on causes of mortality and morbidity across locations and over time. The most recent version of the ICD, ICD-11, was used by the 72nd World Health Assembly in 2019. In Denmark, ICD-10 is still used. Researchers must be aware that changes in codes and diagnostic criteria may impede comparison of data over long time periods. Other challenges that may limit the utility of information coded according to the ICD are: 1) coding variations among institutions and coders, 2) coding errors, 3) lack of coding of co-morbidities and lifestyle factors, 4) limitations regarding the specificity of available codes and 5) errors and variations in the clinical diagnoses on which the coding is based [4, 22].

Coding of surgical procedures

From 1977 to 1995, surgical procedures were registered in the DNPR according to three consecutive editions of the Danish Classification of Surgical Procedures and Therapies [23]. These codes were divided into groups according to specific organ systems. Since 1996, surgical procedures have been coded on the basis of a Danish Version of the Nordic Medico-Statistical Committee Classification of Surgical Procedures (NOMESCO) [24]. Every NOMESCO code consists of three alphabetic characters that reflect the 1) general and 2) specific anatomic region and the 3) general method of the procedure. These characters are combined with two numerical characters (at positions 4 and 5) identifying the exact procedure. For example, the code for colonoscopy with biopsy includes "U" for transluminal endoscopy, "J" for gastrointestinal tract, "F" for colonoscopy and "35" for the specific procedure. Hence, the combined code is "UJF35".

Systemized Nomenclature of Medicine coding

In the Danish Pathology Data Bank and the National Pathology Registry, specimens are classified according to the Systematized Nomenclature of Medicine (SNOMED) [25]. The SNOMED classification is based on six axes of codes. Each axis is identified with an alphabetic character followed by a five-digit number. The first three axes (T, M, and Æ) reflect the topography, morphology and aetiology of the specimen, respectively. The fourth axis (F) reflects all normal and abnormal functions of the specimen (e.g., expression of mismatch repair proteins for colorectal cancer) and the fifth axis (S) reflects all diseases and syndromes associated with the specimen (e.g., Crohn's disease). The sixth axis (P) reflects all procedures associated with the specimen. More information on the use of SNOMED codes, including a description the of five-digit number system for each axis, can be found at [26].

Anatomical Therapeutic Chemical code system

The Anatomical Therapeutic Chemical (ATC) code system classifies medicinal products by their active substances and pharmacological and therapeutic subgroups [27]. The ATC classification system is a hierarchical classification based on the active substances, with five sublevels [27]. The first level is the main anatomical/pharmacological group, which is labelled with a letter (A, B, C, etc.). The second level describes the therapeutic subgroups (e.g., drugs used in diabetes) and the third and fourth levels are associated with a chemical, pharmacological or therapeutic subgroup. The fifth level is the chemical substance (e.g., metformin) [27]. A given ATC code is also assigned a unit of measurement, route of administration and a defined daily dose [13, 27].

International System of Nomenclature, Properties and Units

Individual biomarker data from point-of-care testing and biological samples (e.g., blood, urine, joint fluid or cerebrospinal fluid) obtained by general practitioners and hospitals in Denmark are routinely recorded in the Register of Laboratory Results for Research and coded according to the International System of Nomenclature, Properties, and Units (the NPU system) [28-31]. The NPU system requires that laboratory results are coded with information including: 1) the part of the human body undergoing examination (e.g., urine, plasma, secret, etc.), 2) component measured in the sample (e.g., calcium, ethanol, glucose, etc.), 3) relevant kind-of-property (e.g., substance concentration, mass fraction, arbitrary content, etc.) and 4) unit of measurement [30].

SELECTED METHODOLOGICAL CHALLENGES

Truncation and censoring

Health registries include exposures and outcomes during defined time intervals. Exposures and outcomes occurring before the initiation of registration cannot be included in studies. Left truncation occurs if individuals who have already experienced the exposure or event of interest at the beginning of follow-up are not included in a registry [32]. For example, the DNPR has recorded information from all inpatient contacts since 1977 [10]. However, registration of information from outpatient clinics and emergency departments did not become available until 1995. Thus, diagnoses and procedures conducted in an outpatient setting before 1995 are truncated. Should an individual be registered with a diagnosis both before and after initiation of a registry, his/her prevalent disease may possibly be classified as incident disease in the period immediately after initiation of a registry (left censoring), thus resulting in overestimation of disease incidence. To reduce this limitation, incidence studies typically apply a wash-out period. As an example, a Danish study of trends in myocardial infarction occurrence reported incidence from 1984 onwards to have seven years of patient history since the start of the DNPR 1977 to exclude prevalent disease cases [33].

Events may also occur after the end of follow-up, and some diseases may not manifest until years after disease onset [8]. Congenital heart diseases are a classic example, because they often are not diagnosed until many years after onset. A study following children until one year after birth may overlook less severe cases, as a consequence of right censoring.

Changes in diagnostic criteria

Beyond changes in classification systems, other factors may affect coding practice such as changes in diagnostic criteria. Thus, the use of more sensitive diagnostic methods over time (diagnostic drift) may limit the interpretation of secular trends in incidence. For example, a transient increase in the observed rate of myocardial infarction between 2000 and 2004 was probably attributable not to a true increase in occurrence but to new diagnostic criteria introduced with the redefinition of myocardial infarction in 2000, which included troponin as a diagnostic biomarker [33].

CONCLUSION

Health registries provide an opportunity to conduct excellent epidemiological research. However, attention must be devoted to their potential limitations. The study design and whether the data are suitable for answering the given research question must always be critically considered in registry-based research.

Correspondence *Henrik Toft Sørensen*. E-mail: hts@ clin.au.dk

Accepted 27 February 2023

Conflicts of interest Potential conflicts of interest have been declared. Disclosure forms provided by the authors are available with the article at ugeskriftet.dk/dmj

Cite this as Dan Med J 2023;70(4):A12220796

REFERENCES

1. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. Eur J Epidemiol. 2014;29(8):541-9.
2. Schmidt M, Schmidt SAJ, Adelborg K et al. The Danish health care system and epidemiological research: from health care contacts to database records. Clin Epidemiol. 2019;11:563-91.
3. Bonnesen K, Fuglsang CH, Korsgaard S et al. Use of routinely collected registry data for undergraduate and postgraduate medical education in Denmark. J Eur CME. 2021;10(1):1990661.

4. Sørensen HT. Regional administrative health registries as a resource in clinical epidemiology. A study of options, strengths, limitations and data quality provided with examples of use. Int J Risk Saf Med. 1997;10(1):1-22.
5. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol. 1996;25(2):435-42.
6. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005;58(4):323-37.
7. Sørensen HT, Lash TL, Rothman KJ. Beyond randomized controlled trials: a critical comparison of trials with nonrandomized studies. Hepatology. 2006;44(5):1075-82.
8. Baron JA, Weiderpass E. An introduction to epidemiological research with medical databases. Ann Epidemiol. 2000;10(4):200-4.
9. Ray WA, Griffin MR. Use of Medicaid data for pharmacoepidemiology. Am J Epidemiol. 1989;129(4):837-49.
10. Schmidt M, Schmidt SAJ, Sandegaard JL et al. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol. 2015;7:449-90.
11. Adelborg K, Sundbøll J, Munch T et al. Positive predictive value of cardiac examination, procedure and surgery codes in the Danish National Patient Registry: a population-based validation study. BMJ Open. 2016;6(12):e012817.
12. Sundbøll J, Adelborg K, Munch T et al. Positive predictive value of cardiovascular diagnoses in the Danish National Patient Registry: a validation study. BMJ Open. 2016;6(11):e012832.
13. Pottegård A, Schmidt SAJ, Wallach-Kildemoes H et al. Data Resource Profile: The Danish National Prescription Registry. Int J Epidemiol. 2017;46(3):798-798f.
14. Gaster N, Hallas J, Pottegård A et al. The validity of Danish Prescription Data to measure use of aspirin and other non-steroidal anti-inflammatory drugs and quantification of bias due to non-prescription drug use. Clin Epidemiol. 2021;13:569-79.
15. Schmidt M, Hallas J, Friis S. Potential of prescription registries to capture individual-level use of aspirin and other nonsteroidal anti-inflammatory drugs in Denmark: trends in utilization 1999-2012. Clin Epidemiol. 2014;6:155-68.
16. Thuesen L, Jensen LO, Tilsted HH et al. Event detection using population-based health care databases in randomized clinical trials: a novel research tool in interventional cardiology. Clin Epidemiol. 2013;5:357-61.
17. Christiansen EH, Jensen LO, Thayssen P et al. Biolimus-eluting biodegradable polymer-coated stent versus durable polymer-coated sirolimus-eluting stent in unselected patients receiving percutaneous coronary intervention (SORT OUT V): a randomised non-inferiority trial. Lancet. 2013;381(9867):661-9.
18. Sloth AD, Schmidt MR, Munk K et al. Improved long-term clinical outcomes in patients with ST-elevation myocardial infarction undergoing remote ischaemic conditioning as an adjunct to primary percutaneous coronary intervention. Eur Heart J. 2014;35(3):168-75.
19. Schmidt M, Christiansen CF, Horvath-Puho E et al. Non-steroidal anti-inflammatory drug use and risk of venous thromboembolism. J Thromb Haemost. 2011;9(7):1326-33.
20. Hjertholm P, Fenger-Grøn M, Vestergaard M et al. Variation in general practice prostate-specific antigen testing and prostate cancer outcomes: an ecological study. Int J Cancer. 2015;136(2):435-42.
21. Gribsholt SB, Pedersen L, Richelsen B, Thomsen RW. Validity of ICD-10 diagnoses of overweight and obesity in Danish hospitals. Clin Epidemiol. 2019;11:845-54.
22. Steinberg EP, Whittle J, Anderson GF. Impact of claims data research on clinical practice. Int J Technol Assess Health Care. 1990;6(2):282-7.
23. Danish classification of surgical procedures and therapies. 1st, 2nd, 3rd ed. Copenhagen, Denmark: Danish Health and Medicines Authority, 1973, 1980, 1988.
24. Ackerknecht E. Medicine at the Paris Hospital, 1794-1848. Johns Hopkins Press, 1967.
25. Erichsen R, Lash TL, Hamilton-Dutoit SJ et al. Existing data sources for clinical epidemiology: the Danish National Pathology Registry and Data Bank. Clin Epidemiol. 2010;2:51-6.
26. Patobank. www.patobank.dk (2023).
27. WHO Collaborating Centre for Drugs Statistics Methodology. ATC/DDD Index 2023. www.whocc.no/atc_ddd_index/ (2023).
28. Arendt JFH, Hansen AT, Ladefoged SA et al. Existing data sources in clinical epidemiology: Laboratory Information System

Databases in Denmark. Clin Epidemiol. 2020;12:469-75.

29. Sundhedsdatastyrelsen. NPU terminologi for laboratorier. Danish Health Data Authority, 2021. https://sundhedsdatastyrelsen.dk/da/rammer-og-retningslinjer/om-terminologi/npu (2022).

30. Pontet F, Petersen UM, Fuentes-Arderiu X et al. Clinical laboratory sciences data transmission: the NPU coding system. Stud Health Technol Inform. 2009;150:265-9.

31. Nicolaisen SK, Thomsen RW, Lau CJ et al. Development of a 5-year risk prediction model for type 2 diabetes in individuals with incident HbA1c-defined pre-diabetes in Denmark. BMJ Open Diabetes Res Care. 2022;10(5):e002946.

32. Cain KC, Harlow SD, Little RJ et al. Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. Am J Epidemiol. 2011; 173(9):1078-84.

33. Schmidt M, Jacobsen JB, Lash TL et al. 25 year trends in first time hospitalisation for acute myocardial infarction, subsequent short and long term mortality, and the prognostic impact of sex and comorbidity: a Danish nationwide cohort study. BMJ. 2012;344:e356.