Impact of variable look-back periods on the incidence rates of chronic diseases using real world data

Mats Rosenlund1,2 | Nils Ekström1 | Michael Törnblom1 | Viktor Wintzell3 | James H. Stark4 | Lina Titievsky4

1IQVIA, Real World Evidence Solutions, Nordics, Solna, Sweden
2Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Solna, Sweden
3Clinical Epidemiology Division, Department of Medicine, Solna, Karolinska Institutet, Solna, Sweden
4Pfizer Inc, New York, New York

Correspondence
Mats Rosenlund, Dept of Learning, Informatics, Management, and Ethics, Karolinska Institutet, Solna, Sweden.
Email: mats.rosenlund@ki.se

Funding information
Pfizer

Abstract

Purpose: Estimating disease incidence based on secondary data requires a look-back period to exclude patients with pre-existing disease from the incidence risk set. However, the optimal length of the look-back period and its impact on incidence rates are often unknown. We studied the impact of the length of the look-back period on incidence rates of 24 different chronic diseases.

Methods: Everyone residing in Sweden between January 1, 2005 and December 31, 2013 were identified from national registries and followed up to 2 years (through December 31, 2015). Outcome events were identified from inpatient and outpatient hospital contacts and incidence rates were calculated per 100 000 person-years. The length of the look-back period was varied with 6-month increments, starting at 6 months. The maximum look-back period of 9 years was used as reference period.

Results: There were 7 943 807 individuals with a look-back period of at least 9 years (mean age 46.5 years) and a mean follow-up time of 1.97 years. Incidence rates were higher across all diseases when restricting the look-back to 1 year compared to 9 years, with a magnitude of overestimation of the incidence rates between 13% (temporal arteritis) and 174% (type 1 diabetes). However, for most diseases the effect of extending the look-back period beyond 3-5 years appeared comparably small.

Conclusions: This study illustrates how short look-back periods cause overestimation of the incidence rates of chronic diseases, suggesting that sensitivity analyses with respect to look-back period are considered, particularly using data sources with limited information on past medical history.

KEYWORDS
chronic disease, epidemiology, incidence, methodology, pharmacoepidemiology
1 | INTRODUCTION

Estimating disease incidence based on secondary data typically requires a look-back period to exclude patients with pre-existing disease from the incidence risk set. This is particularly important for chronic diseases where the identification of new cases requires a look-back period sufficiently long to exclude patients with past diagnosis of the same condition. Consequently, in studies using real world data such as patient registries, researchers need to consider the length of the look-back period when collecting retrospective data for the disease under study. However, the look-back period is often chosen based on several factors including data availability, sample size considerations, and duration of follow up. Studies using data with limited historical information, may have little flexibility to vary the look-back period.

Others have reported varying incidence rates depending on the length of the look-back period. However, there are few other studies that have investigated the degree of overestimation of the incidence rates due to the choice of look-back period across a wide range of diseases in a common data source. Previous studies have involved single conditions or used claims data, with less completeness and historical clinical information on individual patients. Despite the extensive use of the Nordic healthcare registers in real world evidence and pharmacoepidemiology owing to its completeness of data, ability to link across various registers, long follow up times and small attrition rates, there are few assessments of the impact of the look-back period on incidence rates using such data. Therefore, there is a lack of understanding and guidance regarding the choice of optimal look-back period in relation to the outcome under study. While the varying lengths of look-back periods used by researchers may contribute to conflicting estimates of incidence rates, the impact in terms of accuracy of disease incidence estimations has not been thoroughly examined across multiple diseases in the same data source. In addition, there are no standards or guidelines for the optimal length of a look-back period by disease or data source.

The aim of this study was to characterize the impact of varying lengths of the look-back period on the observed incidence rates in a systematic way for multiple chronic diseases using Swedish national healthcare registers.

2 | METHODS

2.1 | Study population

This is a nationwide register-based cohort study including all persons who were alive and lived in Sweden between January 1, 2005 and December 31, 2013. The study cohort was identified in the Total Population Register and followed for a maximum of 2 years (between January 1, 2014 [index date] and December 31, 2015). For the purposes of this analysis, we restricted the study population to persons who had 9 years of available look-back prior to index date. Thus, persons who were younger than 9 years on January 1, 2014 were not included in the analysis.

2.2 | Data sources

Several national healthcare registers were linked to form the study dataset. The Total Population Register, which is maintained by Statistics Sweden and includes data on demographic statistics of the whole Swedish population including birth, death, name, marital status, family relationships and migration since 1967, was used to construct the study population and censor individuals who died or were lost to follow-up (emigration). The National Patient Register, available from 1987 and managed by the National Board of Health and Welfare, includes nationwide data (including discharge diagnoses) from inpatient- and outpatient care. The validity of inpatient diagnoses is high (85-95%) and the long follow-up makes the National Patient Register particularly suitable for large-scale population-based research. The National Patient Register was used to identify outcome events. In addition, we used data from the Prescribed Drug Register available from July 1, 2005 at National Board of Health and Welfare to capture dispensing of prescribed drugs as proxies for certain conditions. The Prescribed Drug Register includes complete data on dispensed prescriptions drugs from pharmacies across Sweden. These registers were linked using a national identification number.

2.3 | Outcome events

Outcome events were identified based on physician-assigned ICD-10 diagnosis codes from inpatient and outpatient contacts, retrieved from the National Patient Register. We selected conditions that represented chronic diseases where the choice of look-back period is expected to influence incidence rates to varying degrees based on routine clinical practice, with some conditions requiring regular healthcare visits and close patient monitoring while others may be managed in primary care and only occasionally lead to specialist visits.

KEY POINTS

- Short look-back periods can cause overestimation of incidence rates of chronic diseases.
- The suitable length of the look-back period to exclude persons with prevalent disease depends on disease and data source.
- The benefit of extending the look-back period beyond 3-5 years to reduce bias in incidence estimations appears comparably small for most diseases.
- Sensitivity analysis with respect to the look-back period should be considered, especially if using data sources with limited information on past medical history.
- Treatment information could be used as proxy for physician-assigned diagnosis in incidence estimations for diseases with limited historical medical data.
Special interest was also put on diseases of concern for vaccine safety, such as autoimmune diseases. Incidence rates were estimated for each outcome event and for multiple look-back periods. Persons with prevalent disease, identified as at least one hospital contact with relevant diagnosis during the look-back period, were excluded from the analysis. The following conditions (ICD-10) were selected for this analysis: acute coronary syndrome (ICD-10: I200, I21, I22), alopecia (ICD-10: L63-L66), alopecia areata (ICD-10: L63), asthma (ICD-10: J45, J46), autoimmune hemolytic anemia (ICD-10: D590, D591), autoimmune hepatitis (ICD-10: K754), autoimmune thyroiditis (ICD-10: E063), celiac disease (ICD-10: K900), chronic inflammatory demyelinating polyradiculoneuropathy (ICD-10: G618, G619), Crohn’s disease (ICD-10: K50), type 1 diabetes (ICD-10: E10), type 2 diabetes (ICD-10: E11), glomerulonephritis (ICD-10: N00, N01, N022-N028, N03, N042-N047, N05, N062-N067), Grave’s disease (ICD-10: E050), idiopathic thrombocytopenic purpura (ICD-10: D693), inflammatory bowel disease (ICD-10: K50, K51, K523, K528), multiple sclerosis (ICD-10: G35), narcolepsy (ICD-10: G474), nephrotic syndrome (ICD-10: N04), psoriasis (ICD-10: L40, M070, M072, M073), Raynaud’s disease (ICD-10: I730), rheumatoid arthritis (ICD-10: M05, M06), systemic lupus erythematosus (ICD-10: M32), and temporal arteritis (ICD-10: M315, M316).

The level of clinical care typically required to diagnose and manage these conditions varies. Most of the outcomes of interest are diagnosed and treated in specialist hospital care (inpatient and outpatient care). However, some outcomes in this study are predominantly diagnosed in primary care (for instance, asthma and type 2 diabetes). While the outcome definition was restricted to primary diagnoses alone, which was intended to maximize specificity of captured diagnoses, the exclusion of prevalent cases was done using both primary and secondary diagnoses, to maximize sensitivity for detection of prevalent disease. In addition, we used the maximum available look-back data from the prescribed drug register (ie, 8.5 years) as a proxy for diagnosis for some of the diseases primarily managed in primary care (type 2 diabetes and asthma) as a sensitivity analysis, since we did not have access to primary care diagnosis in this study. Due to the potential influence of secular trends in prevalence of diagnoses during the look-back period, we also conducted sensitivity analysis estimating incidence rates by calendar year during 2008-2015 using 2008 as reference year and a look-back period of 3 years to exclude most of the prevalent cases.

2.4 Statistical analysis

Incidence rates were calculated as the number of events divided by the total person-time at risk and reported as events per 100,000 person-years with associated 95% confidence intervals (CIs), assuming a Poisson distribution. Hence, in the calculation, the number of events (persons with the event during follow-up) was the numerator and the total person-time in years (the sum of all follow-up for persons with and without event) was the denominator. Given the aim of the analysis, only crude incidence rates were estimated. The study participants were followed from index date until the outcome event, end of study period, emigration, or death, whichever occurred first. The length of the look-back period used to exclude prevalent cases was varied with 6-month increments with a maximum of 9 years, which was used as reference period for the comparison. There were no missing values in any of the registers. The analysis was performed in SAS version 9.4 (SAS Institute Inc.).

3 | RESULTS

Among the 9,610,159 who were alive and resided in Sweden on January 1, 2014, we identified 7,943,807 individuals (82.7%) who had a look-back period of at least 9 years. In the study population, 50.4% were women, mean age was 46.5 years, and mean follow-up time was 1.97 years (Table 1).

3.1 Incidence rates by disease and look-back period

Incidence rates (events per 100,000 person-years) with a 9-year look-back period varied from 1.20 (95% CI, 1.03-1.38) for narcolepsy to 248.75 (246.26-251.25) for acute coronary syndrome. More than half of the conditions had an incidence rate below 10 per 100,000 person-years, including chronic inflammatory demyelinating polyradiculoneuropathy (1.39, 1.21-1.58), autoimmune hemolytic anemia (3.1, 2.4-3.9).

### TABLE 1 Characteristics of the study population including all adults above 18 years in Sweden on the first of January 2014 and followed up through 2015

| Characteristics                  | No. of patients | Percent |
|----------------------------------|----------------|---------|
| Persons alive and registered in Sweden on January 1, 2014 | 9,610,159 |         |
| Excluded due to age <9 years | 1,026,419 | 10.7%   |
| Excluded due to look-back shorter than 9 years | 639,933 | 6.7%    |
| Persons included | 7,943,807 | 82.7%   |
| Gender                          |               |         |
| Male                            | 3,938,742 | 49.6%   |
| Female                          | 4,005,065 | 50.4%   |
| Age, mean (SD), years           | 46.6 (21.4) |         |
| Age groups<sup>a</sup>          |               |         |
| <20                             | 1,051,179 | 13.2%   |
| 20-40                           | 2,136,144 | 26.9%   |
| 41-60                           | 2,348,250 | 29.6%   |
| >60                             | 2,408,234 | 30.3%   |
| Follow-up, mean (SD), years     | 1.97 (0.19) |         |
| Persons with follow-up of 2 years | 7,734,269 | 97.4%   |

<sup>a</sup>Only persons of age 9 years and older were included in the study.
(1.85, 1.64-2.08), alopecia areata (2.53, 2.29-2.79), autoimmune hepatitis (2.54, 2.30-2.80), nephrotic syndrome (2.62, 2.37-2.88), idiopathic thrombocytopenic purpura (4.23, 3.92-4.57), autoimmune thyroiditis (4.76, 4.43-5.12), glomerulonephritis (5.64, 5.27-6.03), alopecia (6.80, 6.40-7.22), and multiple sclerosis (9.30, 8.83-9.79). One third of the diagnoses presented an incidence rate between 10 and 50 per 100,000 person-years, including celiac disease (10.80, 10.29-11.32), Crohn’s disease (17.87, 17.22-18.55), type 1 diabetes (32.65, 31.76-33.57), Grave’s disease (23.17, 22.42-23.94), inflammatory bowel disease (48.04, 46.95-49.14), psoriasis (49.33, 48.23-50.45), rheumatoid arthritis (38.29, 37.32-39.28), and temporal arteritis (10.57, 10.06-11.09). Besides acute coronary syndrome mentioned above, the two other diseases with an incidence rate above 50 events per 100,000 person-years were type 2 diabetes (201.83, 199.58-204.11) and asthma (238, 236.03-240.96).

 Restricting the analysis to a 1-year look-back period resulted in higher incidence rates across all diseases (Figure 1). The overestimation of the incidence rates compared to the analysis using a 9-year look-back varied in magnitude across the diseases. There were four diseases with over 100% overestimation of the incidence rate with the 1-year look-back, including diabetes type 1 and 2, inflammatory bowel disease, and multiple sclerosis. For 12 diagnoses, representing 28% of the diseases included in this study, the incidence rate was overestimated between 50 and 100%, which was the case for psoriasis, asthma, lupus, celiac disease, narcolepsy, and rheumatoid arthritis. The rest, corresponding to 63% of the conditions, had an incidence rate that was overestimated by less than 50%. A few diagnoses, including optic neuritis, encephalitis, Bell’s palsy, and Stevens-Johnson syndrome presented an overestimation of less than 10%. The overestimation of incidence rates decreased when applying a 4-year look-back, where 70% of the diseases had an incidence rate overestimated by less than 10%. Using an 8-year look-back reduced the overestimation of incidence rates to less than 2% for more than 80% of the conditions, indicating small or negligible overestimation of the incidence rates compared to the reference of 9-year look-back.

 The overestimation of incidence rates decreased with longer look-back across all diseases (Figure 2). The decrease of overestimation appeared different for each disease, suggesting a distinct pattern for each condition (Figure 2). For instance, the overestimation of the incidence rate was sharply reduced to 20% when using 2 years of look-back for rheumatoid arthritis while

![FIGURE 1](https://wileyonlinelibrary.com)
FIGURE 2  Percent of overestimated incidence rates compared to 9-year look-back for selected conditions

FIGURE 3  Change in incidence rates by calendar year using 2008 as reference year and a look-back period of 3 years for a subset of diseases with the highest percentage change in incidence rate in 2015 compared to 2008 [Colour figure can be viewed at wileyonlinelibrary.com]
some other diseases, like diabetes and narcolepsy, were still overestimated around 50% and was reduced below 20% after adding at least 3-5 years to the look-back period. Beyond 5 years of look-back, the reduced overestimation appeared less pronounced for all diseases.

3.2 | Sensitivity analysis

Using data on prescribed drugs as proxy for diagnoses expected to be common in primary care resulted in an incidence rate of 508.91 (505.03-512.82) for asthma compared to 238.49 (236.03-240.96) using diagnosis codes with the maximum look-back available. Similarly, defining type 2 diabetes based on prescription data resulted in an incidence of 348.58 (346.59-351.58), compared to 201.83 (200.58-204.11) using diagnosis codes.

Secular trends in disease incidence may impact the comparison of look-back periods across these diseases. Analysis of trends in incidence by year showed that the rates changed less than 20% on average from 2008 to 2015 for the majority (63%) of the diseases. There were six diagnoses that displayed a change in incidence rate of more than 50% for any year compared to 2008 (Figure 3), suggesting that secular trends in disease incidence may be important to consider for these conditions, although most were rare diseases.

4 | DISCUSSION

We estimated incidence rates of various diseases to assess the impact of different length of look-back periods on the incidence rates using national register data. Our results show how longer look-back periods to exclude prevalent cases reduce overestimation of incidence rates of diseases to varying degrees across different diagnoses. There was a near 2-fold overestimation of the incidence rate for some conditions when using short look-backs and the overestimation clearly decreased when applying longer look-backs. These results suggest that the effect of extending the look-back period beyond 3-5 years was comparably small for most diseases. This highlights the importance of considering how far back in time data needs to be collected for studies using real-world data.

The impact of the length of different look-back periods on the incidence rates has been examined by others in studies of various diseases, including diabetes,9 heart disease,10 colorectal cancer,4 epilepsy,8 ocular diseases,2 acute myocardial infarction,2 hypertension,6 osteoporosis,6 and chronic obstructive pulmonary disease.6 However, previous studies either focused on single conditions or used administrative claims data with limitations regarding patient follow-up and completeness over time. In a study using German claims data, restricting the look-back period to 1 year resulted in incidence overestimation for diabetes, colorectal cancer, and heart failure of 40%, 23%, and 43% compared with an 8-year look-back.9 Another study on ocular diseases using US claims data presented a reduction of incidence overestimation of 40% for cataract when expanding the look-back period to 3 years compared to 1 year, which was further reduced below 30% with a 5-year look-back for all of the conditions under study.1 Similar to our study, others have found that 3-5 year may well capture most of the historical events that would otherwise have caused overestimation of the incidence rates.1,3

The magnitude of overestimation due to a short look-back is likely affected by the management of different diseases, where conditions such as temporal arteritis, acute coronary syndrome, and autoimmune hemolytic anemia displayed a modest overestimation of incidence rates even with a short look-back, likely because these patients are monitored and followed up closely in specialist care resulting in frequent events captured in the patient register. However, other diseases that are typically managed by specialists were also overestimated with shorter look-back periods, such as multiple sclerosis, type 1 diabetes, inflammatory bowel disease, and narcolepsy. In contrast, diseases that may be managed outside of specialist care, such as diabetes, psoriasis, and asthma, are not well captured in secondary care and are therefore expected to be more severely overestimated using shorter look-backs, as also indicated by our data. This illustrates the importance to consider the type of data source, disease management, and diagnostic routines for each condition when setting the length of the look-back period.

The incidence rates for the diseases included in this study were largely consistent with other studies.13-18 In another register-based study, the incidence of multiple sclerosis was 10.2 per 100,000 person-years, while the corresponding estimate in our study was 9.30. However, some diseases presented slightly different rates in our study. Any variations in results between studies could be explained by several factors, such as differences in the risk factor profile between the study populations, health care systems, case definitions, time periods of follow-up, and methods to exclude prevalent cases.

The strengths of this study include the size of the population, the nationwide coverage of the study population, and the large number of diseases included in the analysis. The National Patient Register used for this analysis has been shown to have close to complete coverage and high validity (85%-95%) of all somatic care diagnoses.12 Furthermore, the relatively long look-back period used to exclude those with pre-existing disease from the incidence risk set, as well as long duration of follow-up, provided good conditions for estimating precise incidence rates with high reliability. Limitations of this study include the lack of primary care data, mainly relevant for diseases that are managed in primary care. Incidence rates for conditions such as asthma and type 2 diabetes may thus be underestimated if only relying on the diagnosis codes from secondary care and prescription drug data may therefore present a reasonable proxy for such diseases. Using treatments as proxies for primary care diseases from the national prescribed drug register as a sensitivity analysis demonstrated that incidence rates for such conditions are underestimated if ignoring primary care diagnoses. Another limitation is that we did not account for secular trends in disease incidence during the look-back period. While our results indicate that time trends in disease incidence exist, there were few diseases where this would be expected to play an important role, particularly for rare diseases. Even though we
selected diseases that are generally defined as chronic for the present analysis, these diseases can sometimes be reversed or cured. However, this is generally rare and is not expected to have a substantial impact on the results.

These results could be used to guide future research in the choice of how far back in time to collect data on past medical history. Guidelines for look-back periods could be established for various diseases by data source and may help researchers and facilitate comparisons of background incidence rates, for example, for future pharmaco-vigilance assessments. Given the lack of common recommendations, we recommend researchers to conduct sensitivity analyses to determine the optimal look-back period for the outcome under study.

In conclusion, incidence rates were higher across all diseases when applying a 1-year look-back period compared to 9 years, indicating misclassification bias due to short look-back periods. The magnitude of bias varied between diseases and incidence rates were overestimated between 13% (temporal arteritis) and 174% (type 1 diabetes) when applying a 1-year look-back. In general, the effect of extending the look-back period beyond 3-5 years was comparably small. Sensitivity analysis with respect to the look-back period should be considered, especially if using data sources with limited information on past medical history.

ETHICS STATEMENT

The study was approved by the regional Ethical Review Board in Stockholm (Dnr 2016/319-31/1).

ACKNOWLEDGEMENTS

Funding was provided by Pfizer, Inc.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

ORCID

Mats Rosenlund  https://orcid.org/0000-0003-3273-9443

REFERENCES

1. Worthington JM, Gattellari M, Goumas C, Jalaludin B. Differentiating incident from recurrent stroke using administrative data: the impact of varying lengths of look-back periods on the risk of misclassification. Neuroepidemiology. 2017;48(3–4):111-118.
2. Sulo G, Igland J, Vollset SE, et al. Effect of the look-back period’s length used to identify incident acute myocardial infarction on the observed trends on incidence rates and survival: cardiovascular disease in Norway project. Circ Cardiovasc Qual Outcomes. 2015;8(4):376-382.
3. Stein JD, Blachley TS, Musch DC. Identification of persons with incident ocular diseases using health care claims databases. Am J Ophthalmol. 2013;156(6):1169-1175. e3.
4. Quantin C, Benzenine E, Hagi M, et al. Estimation of national colorectal-cancer incidence using claims databases. J Cancer Epidemiol. 2012;2012:298369.
5. Kaiboriboon K, Bakaki PM, Lhatoo SD, Koroukian S. Incidence and prevalence of treated epilepsy among poor health and low-income Americans. Neurology. 2013;80(21):1942-1949.
6. Griffiths RL, O’Malley CD, Herbert RJ, Danese MD. Misclassification of incident conditions using claims data: impact of varying the period used to exclude pre-existing disease. BMC Med Res Methodol. 2013;13:32.
7. Camplinan R, Kucharska-Newton A, Cuthbertson CC, Wright JD, Alonso A, Heiss G. Misclassification of incident hospitalized and outpatient heart failure in administrative claims data: the atherosclerosis risk in communities (ARIC) study. Pharmacoepidemiol Drug Saf. 2017;26(4):421-428.
8. Bakaki PM, Koroukian SM, Jackson LW, Albert JM, Kaiboriboon K. Defining incident cases of epilepsy in administrative data. Epilepsy Res. 2013;106(1–2):273-279.
9. Abbas S, Ilhe P, Koster I, Schubert I. Estimation of disease incidence in claims data dependent on the length of follow-up: a methodological approach. Health Serv Res. 2012;47(2):746-755.
10. Kent ST, Safford MM, Zhao H, et al. Optimal use of available claims to identify a medicare population free of coronary heart disease. Am J Epidemiol. 2015;182(9):808-819.
11. Ludvigsson JF, Almqvist C, Bonamy AK, et al. Registers of the Swedish total population and their use in medical research. Eur J Epidemiol. 2016;31(2):125-136.
12. Ludvigsson JF, Andersson E, Ekborn A, et al. External review and validation of the Swedish national inpatient register. BMC Public Health. 2011;11:450.
13. Wettermark B, Hammar N, Fored CM, et al. The new Swedish prescribed drug register-opportunities for pharmacoepidemiological research and experience from the first six months. Pharmacoepidemiol Drug Saf. 2007;16(7):726-735.
14. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2007;16(4):393-401.
15. Allard WJ, Matera J, Miller MC, et al. Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. Clin Cancer Res. 2004;10(20):6897-6904.
16. Trifiro G, Coloma PM, Rijnbeek PR, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J Intern Med. 2014;275(6):551-561.
17. Ahlgren C, Oden A, Lycke J. High nationwide incidence of multiple sclerosis in Sweden. PLoS One. 2014;9(9):e108599.
18. Kivisto JE, Protudjer JL, Karjalainen J, Wickman M, Bergstrom A, Mattila VM. Hospitalizations due to allergic reactions in Finnish and Swedish children during 1999-2011. Allergy. 2016;71(5):677-683.

How to cite this article: Rosenlund M, Ekström N, Tömblom M, Wintzell V, Stark JH, Titievsky L. Impact of variable look-back periods on the incidence rates of chronic diseases using real world data. Pharmacoepidemiol Drug Saf. 2020;29:1086-1092. https://doi.org/10.1002/pds.5066