HSR Health Services Research

© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2011.01325.x
METHODS BRIEF

Estimation of Disease Incidence in Claims Data Dependent on the Length of Follow-Up: A Methodological Approach

Sascha Abbas, Peter Ihle, Ingrid Köster, and Ingrid Schubert

Objective. To analyze the impact of the length of disease-free intervals on incidence estimation.

Data Source. Statutory health insurance sample in Germany.

Study Design. Overestimation of the incidence in the first quarter of 2008 for three selected diseases, diabetes mellitus, colorectal cancer, and heart failure, depending on different lengths of preceding disease-free intervals.

Data Collection/Extraction Methods. Continuously insured from 2000 until 2008 $\geq$ 18 years ($N =$ 144,907).

Principal Findings. Compared with an 8-year disease-free period, incidence overestimations for diabetes, colorectal cancer, and heart failure were 40, 23, and 43 percent defining a 1-year, and 5, 9, and 5 percent defining a 5-year disease-free period, respectively.

Conclusions. Depending on the specific disease, caution has to be taken while using short disease-free periods because incidence estimates may be extremely overestimated.

Key Words. Incidence estimation, claims data

Various scientific publications use claims data for administrative prevalence and incidence definition (Whittle et al. 1991; Cooper et al. 2002; Walker et al. 2011). However, this is a sophisticated task, as only diagnoses and not the existing disease status is documented. Furthermore, whether the disease is incident or prevalent is not documented in Germany. Moreover, only treatment prevalence or incidence, often called administrative prevalence/incidence, but not the population prevalence/incidence can be estimated within routine data. Thus, when defining a certain case based on claims data, one important task is the external and internal validation of the claims diagnoses

(Herrett et al. 2010; Schubert, Ihle, and Köster 2010). After validation and specification of criteria for case definition, disease prevalence can be estimated. Following the definition of a prevalent case in a certain year, a common procedure to ascertain incident cases is the exclusion of cases with documentation of the respective diagnoses in the preceding periods. Claims data, however, do not allow for assessing lifetime incidence due to the limited time span of available data. To get close to a population under risk for incidence estimation, the disease-free interval should be as large as possible. However, due to data limitation, numerous studies report diagnosis-free intervals of 1, 2, or 3 years only (Margolis et al. 2002; Sloan et al. 2003; Linsell et al. 2006), and sometimes only one quarter of a year (Ziegler and Doblhammer 2009). Furthermore, the choice of the length of the disease-free intervals for incidence estimation may depend on the respective disease and its specific trajectory.

The aim of this study was to analyze the impact of the length of disease-free intervals on incidence estimation in three selected common diseases, namely diabetes mellitus, colorectal cancer, and heart failure. The focus is not on the estimation of incidence for the respective disease, but rather on presenting a method on how the length of disease-free intervals impacts incidence case definition and, consequently, incidence estimation. As the database, we use a statutory health insurance sample in Germany allowing a follow-up of 9 years for insurants.

METHODS

This study is based on claims data from a regional health insurance fund in the state of Hesse, Germany, the AOK Hesse, with 1.9 million insurants at the start of data collection in 1998 covering approximately one-third of the regional population. Data were obtained from the "Statutory Health Insurance sample (SHI) AOK Hesse/KV Hesse," a 18.75 percent random sample of all insurants from the AOK Hesse (Ihle et al. 2005). The SHI sample currently covers data from 1998 to 2008. For the present study, the following data were used: master data (e.g., age, gender, time insured) and ICD-9 or ICD-10

Address correspondence to Dr. Sascha Abbas, PMV Research Group at the Department of Child and Adolescent Psychiatry and Psychotherapy, University of Cologne, Herderstr. 52, 50931 Cologne, Germany; e-mail: Sascha.Abbas@uk-koeln.de. Peter Ihle, Ingrid Köster, and Ingrid Schubert are with the PMV Research Group at the Department of Child and Adolescent Psychiatry and Psychotherapy, University of Cologne, Cologne, Germany.

(German Institute of Medical Documentation and Information 2008) coded diagnoses (outpatient and inpatient care). Several health service research studies have been performed using this database (Köster et al. 2006; Lehmkuhl, Köster, and Schubert 2009; Schubert and Lehmkuhl 2009; Schubert, Köster, and Lehmkuhl 2010). Data on diagnoses were available for analysis from the year 2000 onward.

All continuously insured men and women over the age of 18 years from 2000 until 2008 were defined as the population for disease prevalence and incidence estimation (N = 144,907). The analyses were performed separately for the following three diseases: diabetes mellitus: ICD-10: E10-E14, and ICD-9: 250; colorectal cancer: ICD-10: C18-C20 and ICD-9: 153, 154.0, 154.1; and heart failure: ICD-10: I11.0, I13.0, I13.2, I50, and ICD-9: 428).

In Germany, diagnoses from outpatient care are coded since 2005 with a diagnostic modifier for diagnostic certainty including the modifiers "suspected," "assured," "status post," and "excluded." In inpatient care, documented diagnoses include an admission and a discharge diagnosis and various secondary diagnoses without diagnostic modifiers.

All diagnoses from in- and outpatient care were included, except for those with the diagnostic modifiers "suspected" or "excluded," or "status post" used in outpatient diagnosis documentation. For remuneration purposes, diagnoses in outpatient care are coded quarterly in Germany. Thus, the time unit used in the present analysis is a quarter of the year.

As a diagnosis entry is not equivalent to a disease status of a person, all quarterly documented diagnoses were internally confirmed as follows: a person was defined as prevalent for the disease in a quarter of the year if there was (i) documentation of a diabetes/colorectal cancer/heart failure diagnosis from either in- or outpatient care in the respective quarter, and (ii) at least one further out- or inpatient documentation of the respective diagnosis in the following three quarters.

In a sensitivity analysis, we applied a more specific case definition exemplary for diabetes including antidiabetic medications: A person was defined as prevalent for the disease in a quarter if there was a documented prescription of an antidiabetic drug (ATC: A10) or a documented diagnosis (in- or outpatient care) in the respective quarter and one of the following conditions in the quarter and the following three quarters: (1) at least two antidiabetic prescriptions on different days or (2) a single prescription and at least one further documentation of a diagnosis or (3) a documentation of a hospital discharge diagnosis, or (4) at least three quarters with a documented diagnosis (Köster et al. 2011).

We used the first quarter in 2008 (2008/I) for prevalent case definition. Insurants who fulfilled the case definition in 2008/I were defined as prevalent. To ascertain potential incident cases among the prevalent cases, we defined a disease-free interval before 2008/I by excluding those cases with a documented confirmed diagnosis in the respective interval. To assess the impact of the disease-free interval on the incidence estimation, we widened the interval previous to the quarter of prevalence (2008/I) by consecutively adding a quarter of the year, that is, one quarter disease-free interval, two quarters, three quarters, 1 year, 1 year and one quarter, etc.

The best incidence estimation ("internal gold standard") was defined as that using the largest disease-free interval available in our data, that is, eight disease-free years (32 quarters). We present the overestimation of the incidence in relation to the "internal gold standard" incidence when shortening the disease-free interval.

Results

Mean age (standard deviation) of the study population in 2000, the first year in the follow-up, was 53.3 (16.0) years. Fifty-four percent of the population was women. In the first quarter of 2008, 24,097, 937, and 10,449 cases were defined as prevalent for diabetes mellitus, colorectal cancer, and heart failure, respectively. We then calculated the number of potential incident cases among the prevalent cases in 2008/I depending on the length of the preceding disease-free interval.

Figure 1 presents the overestimation of the incidence in the first quarter of 2008 when using different lengths of disease-free intervals when compared with the gold standard (an 8-year disease-free interval). Defining a disease-free period of only one quarter results in an overestimation of the incidence of 159, 108, and 118 percent for diabetes, colorectal cancer, and heart failure, respectively. The vertical continuous lines indicate the overestimation when using 1, 2, and 5 years of disease-free intervals. Defining a 1-year disease-free period for incidence definition would result in a 40, 23, and 43 percent overestimation of the incidence for diabetes, colorectal cancer, and heart failure, respectively (continuous lines crossing the x-axis at four quarters, Figure 1). The resulting overestimations were 19, 17, and 24 percent defining a 2-year period and 5, 9, and 5 percent defining a 5-year disease-free period for diabetes, colorectal cancer, and heart failure, respectively (continuous lines crossing the x-axis at -8 and -20 quarters, respectively, Figure 1).

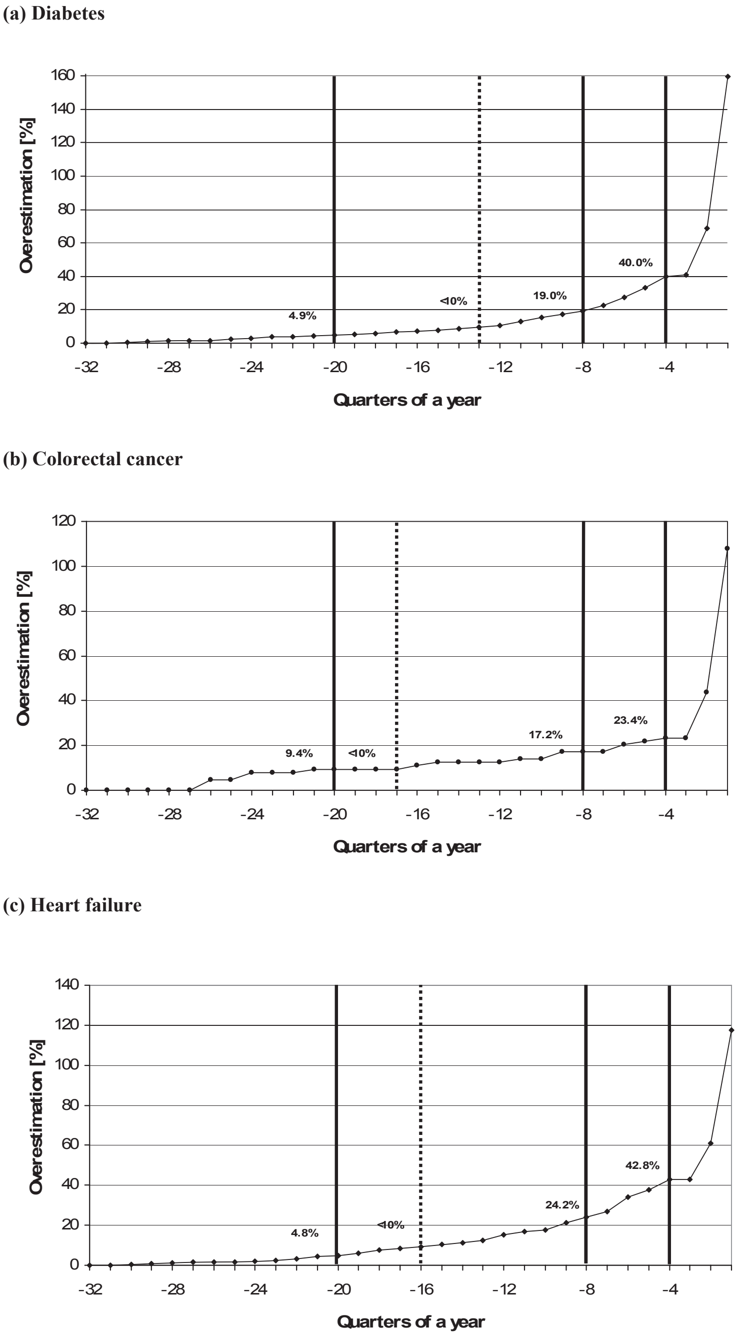
Figure 1: Overestimation of the Incidence in the First Quarter of 2008 (2008/I; Quarter 0) by Different Lengths of Disease-Free Intervals. (a) Diabetes. (b) Colorectal Cancer. (c) Heart Failure. Overestimation of the incidence in the first quarter of 2008 (2008/I; quarter 0) when using different lengths of disease-free intervals when compared with the gold standard (an 8-year disease-free interval; quarter −32). The dotted line shows the quarter X where overestimation is below 10 percent, meaning 10 percent more incident cases in 2008/I would have been defined using a disease-free interval up to the quarter X when compared with the incidence using the longest disease-free interval up to the first quarter in 2000. The vertical continuous lines indicate the overestimation when using 1, 2, and 5 years (i.e., 4, 8, and 20 quarters) of disease-free intervals, respectively.

The dotted vertical line indicates the minimum length of a disease-free interval, where overestimation is <10 percent. To allow a maximum of 10 percent overestimation of the incidence in our examples, a minimum of 13, 17, and 16 quarters as a disease-free period is necessary for diabetes, colorectal cancer, and heart failure, respectively (dotted lines, Figure 1).

Discussion

This study assessed the impact of the length of a disease-free interval for incidence estimation by a given case definition for three selected diseases. The main result is that short disease-free intervals applied for incidence estimation may lead to strong overestimations of the incidence. Interestingly, incidence overestimation was higher for diabetes and heart failure compared with colorectal cancer when using a short disease-free period (e.g., only 1 year), but vice versa when using a long disease-free period of 5 years. Thus, the patterns of incidence overestimation strongly depend on the course of the respective disease.

The graphical approach presented here may serve for sensitivity analysis when estimating incidences for certain diseases, that is, showing how incidence estimates may change when defining different lengths of disease-free periods. We here present a retrospective design by defining a prevalent case and excluding all patients with confirmed diagnoses in the previous years. However, this approach also applies to the classical prospective design, for example, when defining disease-free subjects at cohort entry.



(a) Diabetes

(b) Colorectal cancer

(c) Heart failure

There are, however, several issues one has to keep in mind when estimating incidence in claims data. First of all, a resilient case definition has to be applied, which may not only include documented diagnoses but also specific medication or any benefits or procedures from in- or outpatient care, depending on the objective of the research (Schubert, Ihle, and Köster 2010). As our aim was a methodological approach rather than a sophisticated incidence estimation for the presented diseases, we used a straightforward case definition: a diagnosis in a respective quarter of the year had to be confirmed with at least one further diagnosis documentation in the following three quarters. In a sensitivity analysis, we therefore applied a more specific case definition including antidiabetic medications. In that case, overestimations remained similar with only slight differences. Nevertheless, different validation criteria, for example, including diagnoses with the modifier "suspected," may result in different incidence estimates and thus different overestimations of the incidence.

Secondly, case definition for incidence and prevalence estimation strongly depend on the documentation behavior of a physician. If a physician documents a patient's chronic disease each quarter of a year, the length of the disease-free interval for incidence estimation can be reduced to a minimum. However, besides other influences such as documentation guidelines or electronic medical records software, the documentation behavior strongly depends on the frequency of the patients' visits. No data and thus no diagnoses will be finally transferred to the insurance company if the patient does not seek medical advice. Thus, the different lengths of disease-free periods presented here for incidence estimation in different diseases may in part be explained by different patient behavior, by the nature and stage of the disease and, last but not least, by physician documentation.

We applied the present methodological procedure on three selected diseases, namely diabetes, colorectal cancer, and heart failure. These diseases, however, were selected as examples to present the methodological approach without using sophisticated disease-specific case definitions. An elaborated case definition depending on the individual disease analyzed is a prerequisite for prevalence and incidence definition. The 10 percent cut-off for overestimation is arbitrary and was used to allow comparability of incidence overestimation between diseases. Depending on the specific research question, one has to consider how much overestimation can be accepted.

In our examples, we observed that continually insured patients with a confirmed diagnosis had long periods without any documentation of the disease under observation. On initial consideration, this might be unexpected when focusing on chronic diseases, but possible explanations, depending on the disease, can be hypothesized: our case definition required at least two diagnosis documentations in a four-quarter period. Thus, we might have missed regular check-ups for patients with colorectal cancer where the diagnosis is documented in only one quarter of the year. Furthermore, it is experienced that patients diagnosed with a chronic disease (such as diabetes) sometimes do not comply with regular visits as they do not accept their diagnosis, especially if the disease does not affect them in their daily activities. Moreover, phases of a disease often alternate between phases where medical attendance is high or low. In summary, careful interpretation of the data is necessary, taking into account the clinical course of the disease and thus the frequency of the health care utilization.

A limitation in the present study is the incomplete information of the diagnostic modifier for diagnostic certainty before the year 2005. This may have led to an overestimation of confirmed cases in the respective quarters, which actually were only cases with the modifier "suspected." A strength of the present methodological study is the long follow-up period that allows for assessing incidence estimates using up to 8 years of disease-free periods preceding the diagnosis. The best incidence estimate ("gold standard") was defined by applying the longest disease-free period available, that is, 8 years. Thus, incidence overestimation presented here was carried out in relation to an internal gold standard. We, however, had no clinical data confirming the respective diagnoses.

The approach presented here for incidence overestimation is equally applicable to other health care databases outside Germany. However, the validity of the coded diagnoses including the physician's coding behavior may differ between countries, and thus may lead to shorter or longer intervals for precise incidence estimation depending on the database.

In conclusion, the length of time defining disease-free periods for incidence estimation in claims data is a critical point. Depending on the specific diseases analyzed, caution has to be taken when using disease-free periods of only 1, 2, or 3 years—or even less—as incidence estimates may be extremely overestimated. We therefore recommend visualizing the incidence overestimation depending on the length of the disease-free period as described here. Thereafter, one can decide which length of disease-free period should be applied for incidence estimation for the respective disease under research.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The authors thank the Local Health Care Fund AOK Hesse and the Association of Statutory Health Insurance Physicians (Hesse) for data provision. There was no exertion of influence on manuscript content or manuscript publication by data providers.

Disclosures: None.

Disclaimers: None.

REFERENCES

Cooper, G. S., Z. Yuan, R. N. Jethva, and A. A. Rimm. 2002. "Use of Medicare Claims Data to Measure County-Level Variation in Breast Carcinoma Incidence and Mammography Rates." *Cancer Detection and Prevention* 26 (3): 197–202.

German Institute of Medical Documentation and Information. 2008. *ICD-10-GM. Systematisches Verzeichnis. Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision—German Modification.* Köln: Deutscher Ärzte-Verlag.

Herrett, E., S. L. Thomas, W. M. Schoonen, L. Smeeth, and A. J. Hall. 2010. "Validation and Validity of Diagnoses in the General Practice Research Database: A Systematic Review." *British Journal of Clinical Pharmacology* 69 (1): 4–14.

Ihle, P., I. Köster, H. Herholz, P. Rambow-Bertram, T. Schardt, and I. Schubert. 2005. "[Sample Survey of Persons Insured in Statutory Health Insurance Institutions in Hessen–Concept and Realisation of Person-Related Data Base]." *Gesundheitswesen* 67 (8–9): 638–45.

Köster, I., L. von Ferber, P. Ihle, I. Schubert, and H. Hauner. 2006. "The Cost Burden of Diabetes Mellitus: The Evidence from Germany—The CoDiM Study." *Diabetologia* 49 (7): 1498–504.

Köster, I., E. Huppertz, H. Hauner, and I. Schubert. 2011. "Direct Costs of Diabetes Mellitus in Germany—CoDiM 2000–2007." *Experimental and Clinical Endocrinology and Diabetes* 119(6): 377–85.

Lehmkuhl, G., I. Köster, and I. Schubert. 2009. "[Outpatient Care for Child and Adolescent Psychiatric Disorders–Data from an Insuree-Related Epidemiological Study]." *Praxis der Kinderpsychologie und Kinderpsychiatrie* 58 (3): 170–85.

Linsell, L., J. Dawson, K. Zondervan, P. Rose, T. Randall, R. Fitzpatrick, and A. Carr. 2006. "Prevalence and Incidence of Adults Consulting for Shoulder Conditions in UK Primary Care; Patterns of Diagnosis and Referral." *Rheumatology (Oxford)* 45 (2): 215–21.

Margolis, D. J., W. Bilker, J. Knauss, M. Baumgarten, and B. L. Strom. 2002. "The Incidence and Prevalence of Pressure Ulcers among Elderly Patients in General Medical Practice." *Annals of Epidemiology* 12 (5): 321–5.

Schubert, I., P. Ihle, and I. Köster. 2010. "[Internal Confirmation of Diagnoses in Routine Statutory Health Insurance Data: Concept with Examples and Case Definitions]." *Gesundheitswesen* 72 (6): 316–22.

Schubert, I., I. Köster, and G. Lehmkuhl. 2010. "The Changing Prevalence of Attention-Deficit/Hyperactivity Disorder and Methylphenidate Prescriptions: A Study of Data from a Random Sample of Insurees of the AOK Health Insurance Company in the German State of Hesse, 2000–2007." *Deutsches Ärzteblatt international* 107 (36): 615–21.

Schubert, I., and G. Lehmkuhl. 2009. "Increased Antipsychotic Prescribing to Youths in Germany." *Psychiatric Services* 60 (2): 269.

Sloan, F. A., D. S. Brown, E. S. Carlisle, J. Ostermann, and P. P. Lee. 2003. "Estimates of Incidence Rates with Longitudinal Claims Data." *Archives of Ophthalmology* 121 (10): 1462–8.

Walker, A. J., T. Card, T. E. Bates, and K. Muir. 2011. "Tricyclic Antidepressants and the Incidence of Certain Cancers: A Study Using the GPRD." *British Journal of Cancer* 104 (1): 193–7.

Whittle, J., E. P. Steinberg, G. F. Anderson, and R. Herbert. 1991. "Accuracy of Medicare Claims Data for Estimation of Cancer Incidence and Resection Rates Among Elderly Americans." *Medical Care* 29 (12): 1226–36.

Ziegler, U., and G. Doblhammer. 2009. "[Prevalence and Incidence of Dementia in Germany—A Study Based on Data from the Public Sick Funds in 2002]." *Gesundheitswesen* 71 (5): 281–90.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.