



# Incidence and Prevalence of Knee Osteoarthritis Using Codified and Narrative Data From Electronic Health Records: A Population-Based Study

Ilgin G. Arslan,  Jurgen Damen, Marcel de Wilde, Jacqueline J. van den Driest,  Patrick J. E. Bindels, Johan van der Lei, Dieuwke Schiphof, and Sita M. A. Bierma-Zeinstra

**Objective.** To determine the incidence and prevalence of knee osteoarthritis (OA) using codified and narrative data from general practices throughout The Netherlands.

**Methods.** This retrospective cohort study was conducted using the Integrated Primary Care Information database. Patients with codified knee OA were selected, and an algorithm was developed to identify patients with narratively diagnosed knee OA only. Point prevalence proportions and incidence rates among people age  $\geq 30$  years were assessed from 2008 to 2019. The association of comorbidities with codified knee OA was analyzed using multivariable logistic regression.

**Results.** The positive predicted value of narratively diagnosed knee OA only was 94.0% (95% confidence interval [95% CI] 87.4–100%) and for codified knee OA 96.0% (95% CI 90.6–100%). Including narrative data in addition to codified data resulted in a prevalence 1.83–2.01 times higher (over the study years); prevalence increased from 5.8% to 11.8% between 2008 and 2019. The incidence rate was 1.93–2.28 times higher and increased from 9.98 per 1,000 person-years to 13.8 per 1,000 person-years between 2008 and 2019. Among patients with codified knee OA, 39.4% were previously diagnosed narratively with knee OA, on average  $\sim 3$  years earlier. Comorbidities influenced the likelihood of being recorded with codified knee OA.

**Conclusion.** Our study of a Dutch primary care database showed that current incidence and prevalence estimates based on codified data alone from electronic health records are underestimated. Narrative data can be incorporated in addition to codified data to identify knee OA patients more accurately.

## INTRODUCTION

Osteoarthritis (OA) has been ranked as the tenth leading contributor to global disability, with the knee as the most commonly affected joint (1–3). Between 2007 and 2017, the years lived with disability attributed to knee OA increased by 30.8%, which was a large increase for noncommunicable diseases (4). The prevalence is expected to increase significantly in the coming years due to the increasing age and obesity population.

Population-based incidence and prevalence estimates and predictions concerning the disease burden of knee OA are mostly based on electronic health records (EHRs). EHRs consist of codified data (i.e., specific codes for specific diseases) and narrative data (i.e., free-text notes by general practitioners [GPs] and

correspondence between GPs and other health care providers). Current epidemiologic research on knee OA is largely limited to codified data (5–9). However, diagnoses may not be codified by the GP or updated after disease progression or a change in the final diagnosis. Earlier research (10–12) suggested that patients in general practice may present with multiple health problems, and GPs may not be inclined to code for OA in circumstances where other health problems appear more urgent during the consultation, leading to under-recording of knee OA. In addition, diagnoses may include misclassification of codes due to various reasons, such as lack of time (13,14). These misclassifications and under-recording of codes may have an impact on the accuracy of epidemiologic estimates of knee OA. Earlier research showed significant under-recording of OA in primary care EHRs

Ilgin G. Arslan, MSc, Jurgen Damen, PhD, Marcel de Wilde, MSc, Jacqueline J. van den Driest, MSc, Patrick J. E. Bindels, PhD, Johan van der Lei, PhD, Dieuwke Schiphof, PhD, Sita M. A. Bierma-Zeinstra, PhD: Erasmus MC University Medical Center, Rotterdam, The Netherlands.

Author disclosures are available at <https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Facr.24861&file=acr24861-sup-0001-Disclosureform.pdf>.

Address correspondence to Ilgin G. Arslan, MSc, Erasmus Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands. Email: [i.arslan@erasmusmc.nl](mailto:i.arslan@erasmusmc.nl)

Submitted for publication September 22, 2021; accepted in revised form January 11, 2022.

### SIGNIFICANCE & INNOVATIONS

- This study, using a Dutch primary care database, showed that current incidence and prevalence estimates of knee osteoarthritis (OA) based on codified data from electronic health records are underestimated.
- The prevalence and incidence of knee OA are approximately twice as high when adding narrative data to codified data.
- Narrative data in addition to codified data can be used to obtain more accurate incidence and prevalence estimates of knee OA by developing algorithms containing keywords related to knee OA, as applied in the current study.

in the UK. One-fourth of the patients who underwent a total knee or hip replacement did not have codified joint pain or a codified OA diagnosis in the previous 10 years (12).

Including narrative data in addition to codified data can help to provide more reliable estimations of the burden of knee OA. Reliable estimates are needed for health policy makers in order to respond to the increase in the demand for health care relating to knee OA, but also to enable researchers and health care providers to identify patients with knee OA more accurately.

Therefore, the aim of this study was determine the incidence and prevalence of knee OA using the complete EHR consisting of both codified and narrative data from a large primary care database from The Netherlands in the period 2008–2019. By combining narrative and codified data, this study aims to detect patients with knee OA more accurately than the standard approach of using codified data alone.

### MATERIALS AND METHODS

**Design and setting.** A retrospective cohort study was conducted using the Integrated Primary Care Information (IPCI) database. A detailed description of the IPCI database has been given elsewhere (15,16). In summary, the IPCI database is a dynamic database and contains primary care EHRs for ~2.5 million patients in The Netherlands. The EHRs contain detailed clinical information in a medical journal documented using free-text notes by the GP, diagnoses according to the International Classification of Primary Care (ICPC) codes, laboratory findings, drug prescriptions, and referrals and correspondence with other health care providers in primary and secondary care. In The Netherlands, all citizens are obliged to register with a GP. GPs are the first point of contact and act as a gatekeeper to secondary care (17,18). We therefore assume that EHRs from the IPCI database contain all relevant medical information, including medical findings and diagnoses from secondary care. This study was approved by the Board of Directors of the IPCI database.

**Study cohort.** Patients were included during each study year from January 1, 2008, until December 31, 2019, if they were age  $\geq 30$  years. To increase the reliability of the data, the first year that a patient is included in the IPCI database was not included as new medical information (i.e., this information was included as part of medical history). Patients with a codified diagnosis of knee OA were selected. The codified diagnosis of knee OA was based on the ICPC code L90. In addition, an algorithm was developed by the research group, which included GPs, to identify patients with keywords referring to knee OA in narrative data (i.e., the free text in their EHR) without any record of codified knee OA based on the ICPC code L90. The algorithm included patients with an ICPC code L15 (i.e., knee symptoms) plus keywords related to OA or keywords related to knee plus OA without ICPC code L15, for example ‘knee’ plus ‘osteoarthritis,’ ‘gonarthrosis,’ and ‘knee’ plus ‘prosthesis.’ Keywords combined with terms indicating negation (e.g., ‘not’ or ‘no’) were excluded, as were combinations with relatives (e.g., ‘father has,’ ‘mother has’), patient’s anxiety about a possible diagnosis of OA, and expressions of uncertainties regarding the OA diagnosis by the GP or other health care providers in primary care or secondary care (e.g., ‘probably,’ ‘differential diagnoses’). A random sample of 100 patients identified by the algorithm was assessed by one author (IGA) to check for terminology variations and misspellings of keywords. Textual alternations were made after discussion with all authors to improve the algorithm. Full details of the algorithm are provided in Supplementary Table 1, available on the *Arthritis Care & Research* website at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>.

**Validity of the algorithm.** Two authors, IGA (a physical therapist and researcher) and JD (an academic GP), independently assessed the positive predictive value (PPV) of the algorithm by reading the full EHRs of a random selection of 50 narratively diagnosed knee OA patients without any record of codified knee OA. Patients were defined as true-positive when there was supporting evidence that the GP, the health care provider in primary care (e.g., a physical therapist), or a health care provider in secondary care (e.g., an orthopedist or radiologist) reported a knee OA diagnosis in the free text of the EHR; this is a commonly used reference standard to identify the PPV in EHRs (10). Inconsistencies were resolved by consensus and, if necessary, through discussion with a coauthor (DS, a senior researcher who has wide experience with the IPCI database). To compare the validity of the algorithm to that of codified knee OA, one author (IGA) assessed the PPV of a random selection of 50 patients identified with codified knee OA (i.e., ICPC code L90) by reading the full EHRs, with scrutiny by the coauthors (JD or DS) if necessary. Similar to the PPV assessment for narratively diagnosed knee OA, patients with codified knee OA were defined as true-positive when there was supporting evidence that the GP or the health care provider in primary care or in secondary care reported a knee

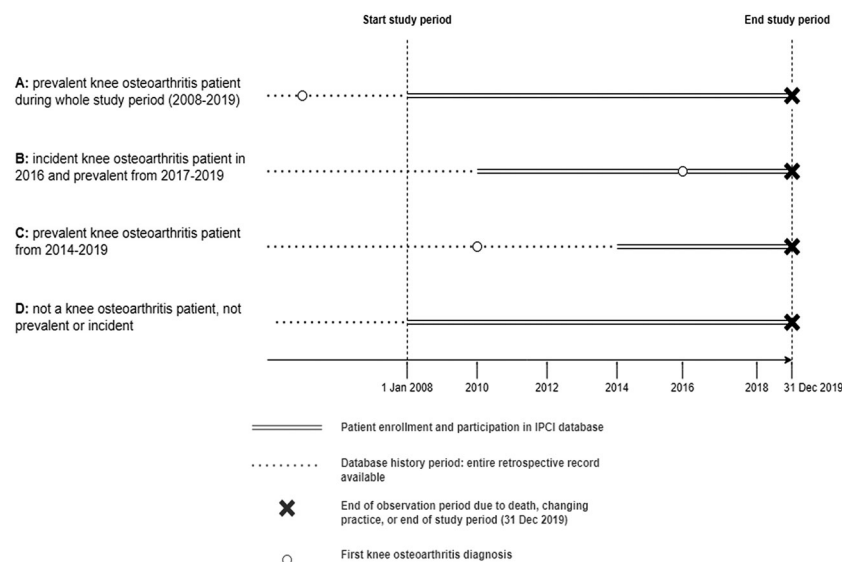
OA diagnosis in the free text of the EHR. PPVs were calculated as the proportion of patients who were confirmed as having knee OA based on the information reported in the EHR.

**Outcomes.** Point prevalence proportions and incidence rates were presented for 2 independent groups: 1) patients with a codified diagnosis of knee OA, and 2) patients with narratively diagnosed knee OA without any record of codified knee OA in their EHR according to the algorithm. The point prevalence proportion was calculated for each year between 2008 and 2019 as the total number of people ever diagnosed with knee OA as of July 1 each calendar year, divided by the total number of patients in the population as of July 1 of that calendar year, and multiplied by 100. The entire retrospective record available for patients was used to estimate the prevalence proportion. The annual incidence rate was calculated for each year between 2008 and 2019 by the number of new cases between January 1 and December 31 in each calendar year, divided by the number of person-years at risk between January 1 and December 31 each calendar year. The at-risk period is the period that a patient was participating in the IPCI database (i.e., from the moment of enrollment in the IPCI database) and not recorded with a knee OA diagnosis until the time of a knee OA diagnosis, death, changing practice, or the end of participation in IPCI database. When estimating the incidence rates, the entire retrospective record available for patients

was used to exclude prevalent knee OA. Thus, patients with a diagnosis in their medical history (i.e., before enrollment in the IPCI database) were defined as having prevalent knee OA. Patients with a diagnosis before January 1, 2008, were also defined as having prevalent knee OA. See Supplementary Table 2, available on the *Arthritis Care & Research* website at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>, for more information. A codified knee OA diagnosis was defined as at least 1 diagnostic code for knee OA (ICPC code L90). A narrative knee OA diagnosis was defined as at least 1 narrative diagnosis according to our algorithm. Incidence and prevalence estimates were calculated stratified by sex. Detailed information regarding the study design is illustrated in Figure 1.

To determine the effect of including narrative data in addition to codified data, annual rate ratios between the point prevalence proportions and incidence rates of codified knee OA and codified plus narratively diagnosed knee OA were calculated. Furthermore, some of the patients identified with codified knee OA may have been identified with knee OA at an earlier date based on narrative data. We explored the proportion of patients with a narrative knee OA diagnosis prior to a codified knee OA diagnosis. The number of days between the first narrative knee OA diagnosis and the first codified knee OA diagnosis was calculated.

Differences in descriptive characteristics between patients with a codified knee OA diagnosis and patients with narratively



**Figure 1.** Details of the study design. The figure shows 4 examples of patients in the study cohort (patients A–D). The study period started January 1, 2008, and ended December 31, 2019. The Integrated Primary Care Information (IPCI) database is an open cohort, meaning that patients can also enter the database after the start of the study period and stop before the end of study period due to death or changing practice. Patients were followed from the start of the study period (patients A and D) or from the moment they entered the IPCI database if this moment was after January 1, 2008 (patients B and C). Patients were followed until the end of the study period (patients A, B, C, and D) or until the moment of death or changing practice when this moment was before December 31, 2019. The entire retrospective record available for patients was used to exclude prior knee osteoarthritis (OA) when estimating the incidence rates (patients A and C). A first knee OA diagnosis was defined as incident when the first diagnosis was given within the study period and participation in the IPCI database (patient B). The entire retrospective record available for patients was used to estimate the prevalence proportions. Patients with a knee OA diagnosis before January 1, 2008, were defined as having prevalence knee OA (patient A), as were patients with a first knee OA diagnosis before participation in the IPCI database (patient C).

diagnosed knee OA were determined. Furthermore, as described earlier, comorbidities in patients with OA may be a reason why codified knee OA is under recorded. Among patients with prevalent knee OA (either codified or narratively diagnosed) during the observation period (i.e., January 1, 2008 to December 31, 2019), we analyzed the association of concurrent comorbidities (i.e., occurring before the first knee OA diagnosis) with a codified knee OA diagnosis. Frequently occurring comorbidities in patients with OA were selected based on an earlier systematic review (19): 1) hypertension, hyperlipidemia, being overweight, diabetes mellitus (i.e., disorders related to metabolic syndrome); 2) heart/vascular diseases and events (i.e., stroke/transient ischemic attack, peripheral arterial disease, and myocardial infarction/angina pectoris); 3) asthma; 4) chronic obstructive pulmonary disease; 5) a small selection of OA related to joints other than the knee (i.e., spinal OA and hip OA) (see Supplementary Table 3, available on the *Arthritis Care & Research* website at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>, for the full list of ICD codes). Analysis of the association of concurrent comorbidities with codified knee OA was adjusted for age and sex.

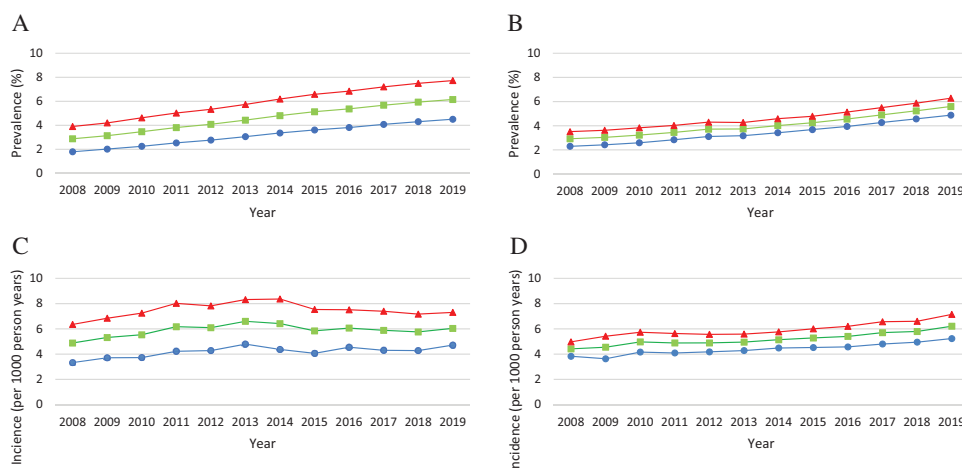
**Statistical analysis.** Binomial 95% confidence intervals (95% CIs) were calculated for the PPV of the algorithm. Prevalence and incidence estimates were standardized for the changing annual age and sex structure of the Dutch population as given by the StatLine database of Statistics Netherlands from 2008 to 2019 (20). Poisson distribution was used to provide 95% CIs for prevalence and incidence estimates. Descriptive characteristics were reported as means  $\pm$  SDs, medians and interquartile ranges (IQRs), and counts and percentages, as appropriate. Multivariable logistic regression was performed to determine the association of comorbidities with the codified diagnosis in patients with knee OA, adjusted for age and sex; the

results were expressed as odds ratios (ORs) including 95% CIs. Prior to the multivariable regression analysis, a variance inflation factor (VIF) was leveraged to detect the collinearity of comorbidities in the multivariable logistic regression analysis. A VIF of  $>5$  was considered indicative of multicollinearity. Nonlinearity between age and the logit of the outcome was observed using the Box-Tidwell test and restricted cubic spline plot. A model with linear splines with 4 knots at the 5th, 35th, 65th, and 95th percentiles based on the recommendations of Harrell (21) showed the best model fit based on Akaike's information criterion and was used as the final multivariable logistic regression model. The significance level throughout was set at 2-tailed  $P$  values less than 0.05. Statistical analyses were performed using RStudio software, version 4.0.2. The aggregated data are available on request from the corresponding author.

## RESULTS

**Validity of the algorithm.** The PPV of the algorithm based on narrative data without a record of codified knee OA was estimated to be 94.0% (95% CI 87.4–100%). Reasons for the 3 false positives were physician typing errors ( $n = 1$ ), patient's anxiety about a possible diagnosis of OA ( $n = 1$ ), and expression of uncertainty about the OA diagnosis ( $n = 1$ ), which could not be excluded by the algorithm (see Supplementary Table 1, available at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>, for more details). The PPV of codified knee OA (i.e., ICD code L90) was estimated to be 96.0% (95% CI 90.6–100%). Reasons for the 2 false-positives were expressions of uncertainty about the OA diagnosis.

**Trends in prevalence and incidence estimates.** Of the 180,986 patients with knee OA included in the cohort, 94,969 were diagnosed with codified knee OA, and 86,017 with



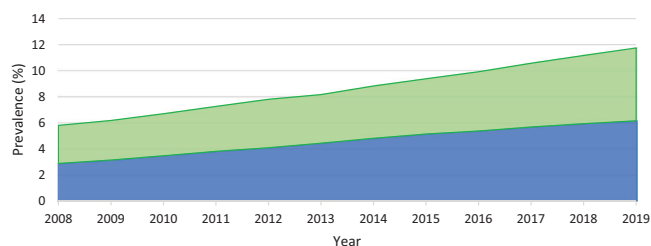
**Figure 2.** Prevalence of knee osteoarthritis (OA) based on codified data (A) and narrative data alone (B), and incidence of knee OA based on codified data (C) and narrative data alone (D) for men (blue), women (red), and both (green).

narratively diagnosed knee OA only without any record of codified knee OA.

**Prevalence proportions.** The standardized prevalence of codified knee OA increased from 2.88% (95% CI 2.87–2.89) in 2008 to 6.15% (95% CI 6.14–6.17) in 2019 (Figure 2A). The standardized prevalence of narratively diagnosed knee OA only without any record of codified knee OA increased from 2.92% (95% CI 2.91–2.93) in 2008 to 5.60% (95% CI 5.58–5.61) in 2019 (Figure 2B). The annual crude and standardized prevalence are presented in Supplementary Table 4, available on the *Arthritis Care & Research* website at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>.

**Incidence rates.** The standardized incidence rate of codified knee OA increased from 4.88 per 1,000 person-years (95% CI 4.84–4.93) in 2008 to 6.04 per 1,000 person-years (95% CI 6.00–6.09) in 2019 and peaked around the year 2013 with 6.60 per 1,000 person-years (95% CI 6.55–6.65) (Figure 2C). The standardized incidence of narratively diagnosed knee OA only without any record of codified knee OA increased consistently over the years from 4.42 per 1,000 person-years (95% CI 4.38–4.46) in 2008 to 6.21 per 1,000 person-years (95% CI 6.16–6.26) in 2019 (Figure 2D). The annual crude and standardized incidence rates are presented in Supplementary Table 4, available on the *Arthritis Care & Research* website at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>. Both the prevalence and incidence rates were higher for women than for men at any given time point (Figure 2 and Supplementary Table 5, available on the *Arthritis Care & Research* website at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>).

**Effect of adding narrative data to codified data.** Adding narrative data to codified data resulted into a prevalence that was 1.83–2.01 times higher over the study period (Table 1). The standardized prevalence was 5.80% (95% CI 5.79–5.82) in 2008, and it increased to 11.75 (95% CI 11.73–11.77) in 2019 (Figure 3). The standardized incidence was 1.93 to 2.28 higher over the study period when adding narrative data to codified data



**Figure 3.** Point prevalence of knee osteoarthritis (OA) based on narrative data alone (green) in addition to codified data (blue). Among patients identified with codified knee OA, 39.4% were diagnosed narratively with knee OA at an earlier stage, which was ~3 years prior to the first codified knee OA diagnosis. These patients are not counted in the prevalence proportions of the narrative data alone. Color figure can be viewed in the online issue, which is available at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>.

(Table 1) and increased from 9.98 per 1,000 person-years (95% CI 9.92–10.04) in 2008 to 13.78 per 1,000 person-years (95% CI 13.71–13.84) in 2019 (Figure 4). Both the prevalence and incidence rates were higher for women than for men at any given time point (see Supplementary Table 6, available on the *Arthritis Care & Research* website at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>).

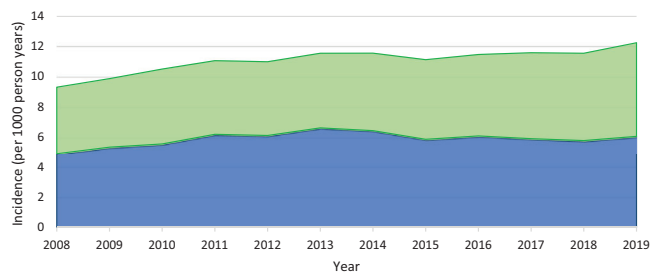
**Narrative diagnosis prior to codified diagnosis.** Among patients identified with codified knee OA ( $n = 94,969$ ), 39.4% ( $n = 37,375$ ) were diagnosed narratively with knee OA at an earlier stage. This was ~3 years on average prior to the first codified knee OA diagnosis (median number of days 1,111 [IQR 143–2,836]).

**Characteristics associated with codified knee OA diagnosis.** The VIF of all independent variables was  $<1.20$ , indicating that there is no collinearity between variables. Multivariable analysis adjusted for age and sex showed that the presence of hypertension, hyperlipidemia, diabetes mellitus, and especially being overweight (OR 1.37 [95% CI 1.32–1.42]) prior to knee OA

**Table 1.** Prevalence and incidence of knee osteoarthritis based on codified data versus a combination of codified and narrative data\*

Year	Standardized point prevalence (95% CI)			Standardized incidence (95% CI)		
	Codified data	Codified + narrative data	RR	Codified data	Codified + narrative data	RR
2008	2.88 (2.87–2.89)	5.80 (5.79–5.82)	2.01	4.88 (4.84–4.93)	9.98 (9.92–10.04)	2.04
2009	3.14 (3.13–3.15)	6.18 (6.17–6.20)	1.97	5.32 (5.27–5.36)	10.61 (10.55–10.67)	1.99
2010	3.47 (3.46–3.48)	6.70 (6.68–6.71)	1.93	5.53 (5.49–5.58)	11.31 (11.24–11.37)	2.04
2011	3.81 (3.80–3.82)	7.26 (7.25–7.28)	1.90	6.17 (6.13–6.22)	11.95 (11.88–12.01)	1.93
2012	4.09 (4.07–4.10)	7.80 (7.79–7.82)	1.91	6.10 (6.05–6.14)	11.92 (11.86–11.99)	1.96
2013	4.43 (4.42–4.45)	8.17 (8.15–8.19)	1.84	6.60 (6.55–6.65)	12.57 (12.50–12.64)	1.90
2014	4.81 (4.80–4.83)	8.83 (8.81–8.85)	1.83	6.42 (6.37–6.47)	12.63 (12.57–12.70)	1.97
2015	5.14 (5.12–5.15)	9.38 (9.36–9.40)	1.83	5.84 (5.80–5.89)	12.24 (12.17–12.30)	2.09
2016	5.37 (5.36–5.38)	9.92 (9.90–9.94)	1.85	6.07 (6.02–6.11)	12.67 (12.60–12.74)	2.09
2017	5.68 (5.66–5.69)	10.58 (10.56–10.59)	1.86	5.89 (5.84–5.93)	12.89 (12.82–12.96)	2.19
2018	5.94 (5.92–5.95)	11.17 (11.15–11.19)	1.88	5.76 (5.72–5.80)	12.93 (12.86–12.99)	2.24
2019	6.15 (6.14–6.17)	11.75 (11.73–11.77)	1.91	6.04 (6.00–6.09)	13.78 (13.71–13.84)	2.28

\* Standardized point prevalence proportions and incidence rates are standardized for age and sex distribution of the total population from The Netherlands. 95% CI = 95% confidence interval; RR = rate ratio.



**Figure 4.** Incidence of knee osteoarthritis (OA) based on narrative data alone (green) in addition to codified data (blue). Among patients identified with codified knee OA, 39.4% were diagnosed narratively with knee OA at an earlier stage, which was ~3 years prior to the first codified knee OA diagnosis. These patients are not counted in the annual incidence rates of narrative data alone. Color figure can be viewed in the online issue, which is available at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24861/abstract>.

diagnosis was associated with a greater likelihood of being recorded with a codified knee OA diagnosis (Table 2). Furthermore, knee OA patients with hip OA or spinal OA prior to knee OA diagnosis had a greater likelihood of being recorded with a codified knee OA diagnosis (OR 1.15 [95% CI 1.10–1.19] and OR 1.28 [95% CI 1.23–1.35], respectively).

## DISCUSSION

This study investigated the incidence and prevalence of knee OA using a combination of narrative and codified data in The Netherlands. The point prevalence rate was 1.83–2.01 times higher (over the study years), and the incidence rate was 1.93–2.28 times higher when including narrative data in addition to codified data. Approximately 40% of codified knee OA patients had a previous record of narratively diagnosed knee OA, with the narrative diagnosis being made on average ~3 years earlier. This suggests that a substantial proportion

of patients that we identified with narratively diagnosed knee OA alone without any record of codified knee OA might be diagnosed with codified knee OA in their EHR in the future. Comorbidities influenced the likelihood of a codified knee OA diagnosis being recorded.

The Dutch National Institute for Public Health and the Environment (RIVM) has predicted that the number of people with knee OA in The Netherlands will rise by 41% in the period 2015–2040 (9). The RIVM estimated the prevalence of knee OA based on ICD code L90 in 2019 at 5.1% for women and 3.0% for men. These numbers were based on Nivel Primary Care Registrations, which is an integrated primary care registration. However, the predicted prevalence is seriously underestimated because it is based on codified data alone from EHRs. Our study showed a 2-fold higher prevalence of knee OA when including narrative data in addition to codified data; in 2019, the prevalence of knee OA based on ICD code L90 in the current study was estimated at 4.5% for men and 7.7% for women, but including narrative data to codified data showed a prevalence rate of 9.4% for men and 14.0% for women. To make adequate preparations for the large increase in the prevalence of knee OA that has been predicted, a complete picture of the current and future impact of knee OA is needed. Therefore, health care policy should be more aware that epidemiologic measures of knee OA based on codified data are likely to be underestimated. Incorporating narrative data in addition to codified data can be used to obtain a more adequate picture of the burden of knee OA. We also found that ~40% of codified knee OA patients had a previous record of narratively diagnosed knee OA on average ~3 years earlier. Capturing knee OA patients earlier may help policymakers to plan and prioritize resources more adequately to keep health care affordable.

In the current study, we found incidence and prevalence estimates that were higher than the estimates from the RIVM (i.e., prevalence in 2019, 4.5% for men and 7.7% for women in

**Table 2.** Characteristics associated with codified knee osteoarthritis (OA) diagnosis\*

Characteristic	Narratively diagnosed		
	Codified knee OA (n = 94,969)	knee OA alone (n = 86,017)	Multivariable analysis OR (95% CI)†
Age at knee OA diagnosis (i.e., first hit), mean ± SD years	66.8 ± 11.9	61.3 ± 13.1	–
Men	32,971 (34.7)	35,217 (40.9)	–
Hypertension	33,550 (35.3)	21,945 (25.5)	1.18 (1.15–1.21)
Hyperlipidemia	10,481 (11.0)	7,300 (8.49)	1.04 (1.01–1.08)
Overweight	8,470 (8.92)	5,914 (6.88)	1.37 (1.32–1.42)
Diabetes mellitus	13,182 (13.9)	8,539 (9.93)	1.12 (1.08–1.15)
Myocardial infarction/angina pectoris	9,583 (10.1)	6,013 (6.99)	1.09 (1.05–1.13)
Stroke/TIA	5,372 (5.66)	3,514 (4.09)	0.98 (0.93–1.02)
Peripheral arterial disease	1,535 (1.62)	1,021 (1.19)	1.00 (0.92–1.09)
COPD	5,224 (5.50)	3,512 (4.08)	1.04 (1.00–1.09)
Asthma	8,170 (8.60)	6,832 (7.94)	1.05 (1.01–1.09)
Hip OA	7,312 (7.70)	4,207 (4.89)	1.15 (1.10–1.19)
Spinal OA	5,466 (5.76)	2,976 (3.46)	1.28 (1.23–1.35)

\* Values are the number (%) unless indicated otherwise. 95% CI = 95% confidence interval; COPD = chronic obstructive pulmonary disease; OR = odds ratio; TIA = transient ischemic attack.

† Codified diagnosis versus narrative diagnosis, with adjustment for age and sex.

the current study versus 5.1% for women and 3.0% for men published by the RIVM). We included patients age  $\geq 30$  years, while estimates from the RIVM were based on all patients regardless of their age. Without restriction on age, our analysis showed similar estimates as those by the RIVM (i.e., crude prevalence in 2019, 3.0% for men and 4.4% for women in the current study). Furthermore, our study showed that the incidence of narratively diagnosed knee OA alone without any record of codified knee OA increased consistently year by year, while this was less pronounced in the incidence of codified knee OA. In contrast, Swain et al (7) found a decline in the incidence of OA using primary care EHR data from the UK. As shown in the current study, the authors acknowledge that their results are open to misclassification bias due to inconsistent recording. To minimize this bias, narrative data in addition to codified data can be used to show the actual trend in the incidence of OA. However, coding systems of diagnoses built into EHRs differ between countries and therefore may require different applications of narrative data. It should also be noted that the use of narrative fields may differ across countries and systems and data protection may limit access to such data fields. There may be other possible alternatives to identify under-recorded knee OA patients, which may be more suitable in countries and systems other than the Dutch GP system, for example, using algorithms that include process, referral, and intervention codes.

The current study showed substantial under-recording of codified knee OA; approximately one-half of the knee OA patients did not have codified knee OA and were identified based on narrative data alone. Yu et al (12) also found under-recording of codified OA in primary care EHRs in the UK. They found that one-fourth of patients with severe OA age 40 years who have had total hip and knee replacements did not have codified joint pain or a codified OA diagnosis in the previous 10 years. However, these results do not apply to the entire spectrum of OA severity, as the average lifetime risk for knee replacement is shown to be  $\sim 30\%$  (22). Moreover, previous research (23) showed that patients with less severe OA are less likely to have a codified OA diagnosis. This suggests that patients with severe knee OA are overrepresented in current epidemiologic research that uses codified data alone. To our knowledge, this is the first study that used both narrative and codified data, and it therefore adds to the current body of knowledge on the incidence and prevalence of knee OA across the entire spectrum of severity.

Similarly to a previous study (23), we found that a record of codified knee OA was associated with being overweight. In addition, our results showed that patients with a concurrent record of hypertension, hyperlipidemia, diabetes mellitus, being overweight, and OA in joints other than the knee were more likely to be diagnosed with codified knee OA. It may be that GPs are more prone to give a codified knee OA diagnosis to patients who fit the risk factor profile of knee OA (metabolic syndrome). In contrast, other studies (10–12) suggested that

multimorbidity may cause GPs to give a lower priority to recording codified knee OA. Our results do not support this hypothesis for hypertension, hyperlipidemia, diabetes mellitus, being overweight, and OA in joints other than the knee.

A strength of the current study is the use of the IPCI database, which contains a representative sample of the Dutch population (15,16). A limitation of this study is that some patients diagnosed with knee OA by physical therapists without a GP referral were not captured within the IPCI database. Since 2006, patients in The Netherlands can access physical therapy care without a GP referral (24). Prevalence and incidence estimates of knee OA might therefore be underestimated in this study. Also, an important aspect to consider when interpreting our results is that changes in the health care system of The Netherlands may have influenced the time trend of the incidence of knee OA. Examples that might have influenced the time trend are GPs' skills for using digital EHRs and changes in permission for data exchange. In addition, to reduce the number of false-positives, we excluded keywords for knee OA combined with expressions of uncertainty (e.g., 'probably' or 'differential diagnoses') from the narrative data algorithm. The restrictiveness of this algorithm may also have led to an underestimation of knee OA. Furthermore, we were not able to request additional information from the GPs to confirm the diagnosis of knee OA in EHRs, which is considered to be the most robust validation method (25). However, this method is subject to selection bias and a low response rate, and it is expensive (25). Instead, we used a manual review of the EHRs, which is more cost effective and a generally accepted method (25). Finally, our findings are limited to primary care EHR data from The Netherlands, and replication of the development of such narrative data algorithms is needed when using EHR data from countries other than The Netherlands.

Under-recording may also be present for OA in joints other than the knee, such as hip OA (i.e., ICPC code L89), and future research into this would be warranted. In The Netherlands, however, OA in other joints does not have specific codes in the EHR data, and GPs use symptomatic codes instead of OA codes, for example, the use of the ICPC code for hand symptoms (i.e., ICPC code L12) in case of hand OA. Developing an algorithm with narrative data in combination with such symptomatic codes can be a solution for identifying patients with OA in joints without an OA code.

In conclusion, the prevalence of knee OA was 1.83–2.01 times higher (over the study years) and the incidence 1.93–2.28 times higher when including narrative data in addition to codified data. Comorbidities influenced the likelihood of being recorded with codified knee OA. Our study of a Dutch primary care database showed that current knowledge and predictions concerning the epidemiology of knee OA based on codified data alone in EHRs from primary care seriously underestimate its prevalence and incidence. Policy makers should be more aware of the underestimated epidemiologic measures of knee OA when

using codified data alone. For a more adequate picture of the current and future impact of knee OA, narrative data in addition to codified data can be used to identify patients with knee OA more accurately.

## ACKNOWLEDGMENTS

The authors thank the reviewers and editors for their valuable comments.

## AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Arslan had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

**Study conception and design.** Arslan, Damen, de Wilde, van den Driest, Bindels, van der Lei, Schiphof, Bierma-Zeinstra.

**Acquisition of data.** Arslan, Damen, de Wilde, Schiphof.

**Analysis and interpretation of data.** Arslan, Damen, Schiphof, Bierma-Zeinstra.

## REFERENCES

- World Health Organization. Musculoskeletal conditions. 2020. URL: <https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions>.
- Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. *Lancet* 2019;393:1745–59.
- Osteoarthritis Research Society International. Osteoarthritis: a serious disease. December 2016. URL: [https://oarsi.org/sites/default/files/docs/2016/oarsi\\_white\\_paper\\_oa\\_serious\\_disease\\_121416\\_1.pdf](https://oarsi.org/sites/default/files/docs/2016/oarsi_white_paper_oa_serious_disease_121416_1.pdf).
- James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018;392:1789–858.
- Cross M, Smith E, Hoy D, Nolte S, Ackerman I, Fransen M, et al. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. *Ann Rheum Dis* 2014;73:1323–30.
- Spitaels D, Mamouris P, Vaes B, Smeets M, Luyten F, Hermens R, et al. Epidemiology of knee osteoarthritis in general practice: a registry-based study. *BMJ Open* 2020;10:e031734.
- Swain S, Sarmanova A, Mallen C, Kuo CF, Coupland C, Doherty M, et al. Trends in incidence and prevalence of osteoarthritis in the United Kingdom: findings from the Clinical Practice Research Datalink (CPRD). *Osteoarthritis Cartilage* 2020;28:792–801.
- Turkiewicz A, Petersson IF, Bjork J, Hawker G, Dahlberg LE, Lohmander LS, et al. Current and future impact of osteoarthritis on health care: a population-based study with projections to year 2032. *Osteoarthritis Cartilage* 2014;22:1826–32.
- National Institute for Public Health and the Environment. Public Health Foresight Study 2018 (VTV-2018): diseases. 2018. URL: <https://www.vtv2018.nl/en/diseases>.
- Shrestha S, Dave AJ, Losina E, Katz JN. Diagnostic accuracy of administrative data algorithms in the diagnosis of osteoarthritis: a systematic review. *BMC Med Inform Decis Mak* 2016;16:82.
- Jencks SF, Williams DK, Kay TL. Assessing hospital-associated deaths from discharge data: the role of length of stay and comorbidities. *JAMA* 1988;260:2240–6.
- Yu D, Jordan KP, Peat G. Underrecording of osteoarthritis in United Kingdom primary care electronic health record data. *Clin Epidemiol* 2018;10:1195–201.
- Johansen MA, Scholl J, Hasvold P, Ellingsen G, Bellika JG. “Garbage in, garbage out”: extracting disease surveillance data from EPR systems in primary care. Proceedings of the 2008 ACM conference on computer-supported cooperative work. San Diego: Association for Computing Machinery; 2008.
- Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. *J Med Internet Res* 2018;20:e185.
- Vlug AE, van der Lei J, Mosseveld BM, van Wijk MA, van der Linden PD, Sturkenboom MC, et al. Postmarketing surveillance based on electronic patient records: the IPCI project. *Methods Inf Med* 1999;38:339–44.
- Van der Lei J, Duisterhout JS, Westerhof HP, van der Does E, Cromme PV, Boon WM, et al. The introduction of computer-based patient records in The Netherlands. *Ann Intern Med* 1993;119:1036–41.
- Kroneman M, Boerma W, van den Berg M, Groenewegen P, de Jong J, van Ginneken E. The Netherlands: health system review. *Health Syst Transit* 2016;18:1–240.
- Kringos D, Boerma W, Bourgueil Y, Cartier T, Dedeu T, Hasvold T, et al. The strength of primary care in Europe: an international comparative study. *Br J Gen Pract* 2013;63:e742–50.
- Swain S, Sarmanova A, Coupland C, Doherty M, Zhang W. Comorbidities in osteoarthritis: a systematic review and meta-analysis of observational studies. *Arthritis Care Res (Hoboken)* 2020;72:991–1000.
- CBS Open Data StatLine. Population dynamics: month and year. URL: [https://opendata.cbs.nl/statline/portal.html?\\_la=nl&catalog=CBS&tableId=37325&theme=91](https://opendata.cbs.nl/statline/portal.html?_la=nl&catalog=CBS&tableId=37325&theme=91).
- Harrell FE. Regression modeling strategies. Vanderbilt University School of Medicine; 2021.
- Burn E, Murray DW, Hawker GA, Pinedo-Villanueva R, Prieto-Alhambra D. Lifetime risk of knee and hip replacement following a GP diagnosis of osteoarthritis: a real-world cohort study. *Osteoarthritis Cartilage* 2019;27:1627–35.
- Jordan KP, Tan V, Edwards JJ, Chen Y, Englund M, Hubertsson J, et al. Influences on the decision to use an osteoarthritis diagnosis in primary care: a cohort study with linked survey and electronic health record data. *Osteoarthritis Cartilage* 2016;24:786–93.
- Groenewegen PP, de Jong JD, Delnoij DM. The Dutch health insurance law; the accumulation of 30 years of reform thought. *Eur J Public Health* 2006;16:34–5.
- Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 2010;69:4–14.