Appleyard Thomas W (Orcid ID: 0000-0002-0404-4997)
Thomas Martin J (Orcid ID: 0000-0002-4951-9925)
Peat George (Orcid ID: 0000-0002-9008-0184)

PREDICTION MODELS TO ESTIMATE THE FUTURE RISK OF OSTEOARTHRITIS IN THE GENERAL POPULATION: A SYSTEMATIC REVIEW

Running Title: Future Osteoarthritis Risk Prediction Models: Systematic Review

Tom Appleyard MBChB MPhil[1] * https://orcid.org/0000-0002-0404-4997

Martin J Thomas PhD MCSP[1, 2] https://orcid.org/0000-0002-4951-9925

Deborah Antcliff PhD[1, 3, 4] https://orcid.org/0000-0002-9771-8232

George Peat PhD[1, 5] https://orcid.org/0000-0002-9008-0184

1 Primary Care Centre Versus Arthritis, School of Medicine, Keele University, Staffordshire, ST5 5BG, UK

2 Haywood Academic Rheumatology Centre, Midlands Partnership NHS Foundation Trust, Haywood Hospital, Burslem, Staffordshire, ST6 7AG, UK

3 Bury Integrated Pain Service, Northern Care Alliance NHS Foundation Trust, Bury Care Organisation, Manchester, M26 2SP, UK

4 School of Healthcare, University of Leeds, Leeds, LS2 9JT, UK

5 Centre for Applied Health & Social Care Research, Sheffield Hallam University, Sheffield, S10 2BP, UK

* Corresponding author: Tom Appleyard, t.appleyard@outlook.com

FUNDING

MJT and DA were supported by a National Institute for Health and Care Research (NIHR) Development and Skills Enhancement Award (NIHR300818 and NIHR301005 respectively). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Conflicts of interest: GP held an Academic Consultant Contract with Office for Health Improvement & Disparities. TA, DA, MJT: No conflicts of interest to declare.

Registration Number (Prospero): CRD42020220446

ABSTRACT

Objective: To evaluate the performance and applicability of multivariable prediction models for osteoarthritis (OA).

*Methods***:** Systematic review and narrative synthesis using three databases (EMBASE, PubMed, Web of Science; inception to December 2021). We included general population longitudinal studies reporting derivation, comparison, or validation of multivariable models to predict individual risk of OA incidence, defined by recognised clinical or imaging criteria. We excluded studies reporting prevalent OA and joint arthroplasty outcome. Paired reviewers independently performed article selection, data extraction, and risk of bias assessment. Model performance, calibration and retained predictors were summarised.

Results: 26 studies were included reporting 31 final multivariable prediction models for incident knee (23), hip (4), hand (3) and any-site OA (1), with a median of outcome events of 121.5 (range: 27-12,803), median prediction horizon of 8 years (2-41), and a median of 6 predictors (3-24). Age, body mass index, previous injury, and occupational exposures were among the most commonly included predictors. Model discrimination after validation was generally acceptable to excellent (Area Under the Curve = 0.70 to 0.85). Either internal or external validation processes were used in most models although risk of bias was often judged to be high with limited applicability to mass application in diverse populations.

Conclusion: Despite growing interest in multivariable prediction models for incident OA, there remains a predominant focus on the knee, reliance on data from a small pool of appropriate cohort datasets, and concerns over general population applicability.

Systematic review registration: PROSPERO (CRD42020220446).

Keywords: osteoarthritis; systematic review; risk prediction; validation; general population predictive prevention

Significance and Innovations

- Prediction models support earlier intervention in several diseases but, to date, not in osteoarthritis (OA).
- Our systematic review provides a comprehensive and critical synthesis of 31 published multivariable models derived by international research teams to predict the future individual risk of developing OA.
- We found generally good performance and evidence of increasing use of internal and external validation. However, a focus on knee OA, a reliance on a restricted number of cohort datasets, mainly from higher-income countries, and use of data sources that may be challenging to scale up in routine practice, may limit the applicability of many existing prediction models in general populations.
- The emergence of prediction modelling using routine healthcare data and improvements in analytic methods may help address some, but not all, of these limitations.

Introduction

In recent years, there has been an increasing popularity and utilisation of risk prediction models to estimate the likelihood of the incidence of health-related outcomes. These models assist clinicians, complementing clinical decision making and aiding the provision of information to patients, as well as contribute to public health, identifying future healthcare needs for the wider at-risk population [1]. With older people constituting a growing proportion of the global population, disease burden is increasingly associated with non-communicable diseases, for example cardiovascular, cancer, diabetes and musculoskeletal disorders [2]. Models predicting an individual's future risk of developing these conditions, permitting the modification of risk factors whilst patients remain free of disease, may contribute to their prevention. Preventing non-communicable disease is a global priority – in 2015 the UN Sustainable Development Goals programme outlined this as a 15-year aim [3]. Whilst risk prediction models have been derived, validated and implemented in clinical medicine and public health screening programmes to predict the incidence of cardiovascular disease [4]–[6], their use in musculoskeletal disorders, such as osteoarthritis (OA), remains uncommon. The current systematic review was motivated by a desire to understand how close such a prospect may be in OA and what remaining limitations may need to be addressed.

OA is a chronic, painful condition that poses significant challenges to public health. In recent years, the disability-adjusted life-years associated with OA has risen markedly, estimated by the Global Burden of Disease project to have increased 34% between 1990 and 2015 [2]. OA also has significant impacts on healthcare utilisation, including surgical intervention [7], [8], and total costs are estimated to represent up to 0.5% of high-income nations' gross

domestic product [9]. With healthcare-associated costs of OA predicted to rise [10], [11], there is a need for validated risk prediction models to identify high-risk patients, permitting the communication of risk, stratification of care and attempts at risk-informed prevention. Furthermore, models to predict disease incidence may provide insight into the classification and diagnosis of "early OA", of which there is increasing interest due to its chronic, progressive nature, alongside a growing focus on the prognosis, rather than solely the diagnosis, of the condition [12], [13].

Several studies have developed risk prediction models for OA outcomes, but to date, there has been no systematic synthesis of this evidence. Our systematic review identifies and critically synthesises published studies deriving and validating multivariable risk prediction models for predicting individualised risk of OA incidence within general populations. The motivating questions for our review were to summarise currently published models, evaluating their applicability to large-scale use in clinical practice.

Materials and Methods

Literature Search

We conducted preliminary literature searches before finalising our search strategy and specifying our research protocol following PRISMA-P guidelines[14] (PROSPERO registration number: 4220446; approved November 2020). We searched PubMed (Medline), EMBASE and Web of Science from inception to December 2021. Our searches used the modified Ingui filter, a generic filter for clinical prediction modified by Geersing et al. for greater

sensitivity, together with condition-specific terms relating to OA [15]. To increase specificity for risk, rather than prognosis/progression, the terms [onset OR inciden*] were also applied (see Supplementary Data for full search strategy). We screened the reference lists of included articles and published abstracts of recent international annual conferences of Osteoarthritis Research Society International (conference years 2017-2019), the European Alliance of Associations for Rheumatology (2018-2020) and The American College of Rheumatology (2018-2020) to identify models in development that may have subsequently been published.

Eligibility Criteria

We included any original study of a longitudinal design (including randomised controlled trials, cohort and (nested) case-control studies) conducted in a general population sample that developed, compared, or validated a multivariable prediction model to predict an individual's risk of future OA incidence, irrespective of time span for prediction. Articles presenting a clinical prediction score based upon a model, as well as those evaluating prediction model 'impact', were also eligible. Eligible definitions of OA included symptomatic, radiographic, and symptomatic-radiographic. We excluded studies of hospital inpatients and other selective settings, prognostic models of patients with existing symptomatic or radiographic disease, and those utilising arthroplasty as the sole outcome. Cross-sectional studies, case reports/series and conference abstracts were excluded. Titles and abstracts of studies were required in English; no language restriction was placed on articles eligible for full text review.

Screening

Search results of the three searched databases were exported to the reference software Rayyan [16]. TA undertook de-duplication. Title screening was undertaken by a single reviewer (DA, GP, MJT or TA), with a sample of decisions checked by a second reviewer. Authors then worked in pairs (DA & GP; TA & MJT, blinded to the other's decision) to screen abstracts, with conflicts resolved by a third reviewer (GP or MJT) not involved in the original decision. Upon full-text review, paired authors again worked independently with a third reviewer to resolve conflict; reasons for exclusion were documented.

Data Extraction

Data extraction of eligible studies was performed by paired reviewers (DA & GP, MJT & TA), using a shared Microsoft Excel [17] worksheet incorporating items for extraction as outlined by Cochrane [18] and the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMs) checklist [19]. Data extracted included general study information, followed by model-specific data, for instance relating to study design, sample size, outcome definition and included predictors. Lastly, we extracted performance metrics (overall fit, discrimination, calibration). The data extraction template is available in Supplementary Data. Where studies presented final models for multiple eligible outcomes, information was extracted for each model. In those with more than one model per outcome, a "final model" was identified, based upon the authors' own designation or inferred from their description of the model-building process and intended application.

Risk of Bias Assessment

For all included articles, we assessed risk of bias across four pre-determined domains - participants, outcome, predictors and analysis - using the Prediction study Risk of Bias Assessment Tool (PROBAST) [20]. Risk of bias and applicability were scored as low, high, or of unclear risk, with applicability appraised with respect to large-scale use in diverse general populations and in contexts where imaging may not be routinely available or recommended. Risk of bias assessments were conducted by one reviewer, checked by a second reviewer.

Narrative Synthesis

Given the heterogeneous nature of study designs and model content, we conducted a narrative synthesis of results. Final models were grouped by the outcome joint of interest (index joint: knee; hip; hand; other; any) to reflect prior evidence of joint-specific risk factors and synthesised then by specific outcome definition (e.g. radiographic, symptomatic, symptomatic-radiographic). Performance measures were summarised; calibration relating to the agreement between predicted versus observed risk, discrimination assessing whether patients with the outcome (at a given threshold) have higher risk prediction scores [21]. The area under the receiver operating characteristic curve (AUC), as the most commonly reported standard metric for discrimination, was displayed (with 95% confidence intervals (95% CI), where reported) in a forest plot for each model across derivation (i.e. 'apparent performance'), internal validation (assessment of performance typically within a subset of the original dataset, for instance by bootstrapping or cross-validation) and external validation (different sample to derivation) phases [22]. An AUC of 0.5 suggests the model demonstrates no discrimination, ≥0.7 and ≥0.8 were deemed 'good' and 'excellent',

respectively, accepting that such thresholds are quite arbitrary. Predictors included in final models were tabulated and colour-coded by mode of assessment. To reduce the volume of information presented, we grouped different or multiple measurements that had been used to capture the same construct. However, a spreadsheet of this tabulation with minimal grouping of predictors was retained as supplementary data. Throughout the synthesis, studies were generally presented in order of year of publication to help discern trends over time.

Patient and public involvement

Patients and members of the public were not involved in this systematic review.

Results

The search yielded 10,129 articles. After de-duplication, followed by title and then abstract screening, 62 articles were taken forward for full-text screening, of which 20 were eligible for inclusion. A further study was added during data extraction through reference searches, and five were added upon re-running of searches in December 2021. As a result, 26 studies were included in the final analysis (Figure 1).

General Characteristics of Included Studies

We included 26 eligible studies reporting 31 final multivariable prediction models for incident OA, published between 2010 and 2022, using study populations from 15 unique

data sources in the USA (9 studies), Netherlands (8), UK (4), Sweden (2), Canada, China and Norway (Table 1). Median prediction horizon was 8 years (range 2 to 41 years), median number of participants/joints with the outcome of interest was 121.5 (range 27 to 12,803), and the median number of predictors included in final models was 6 (range 3 to 24). Regression analysis was used in 24 models (commonly logistic regression, generalised estimating equations), while 7 involved machine learning approaches (e.g. (deep) neural networks). Internal validation was undertaken for 19 models, 7 were externally validated, and 2 were both internally and externally validated.

Knee OA

Incident radiographic knee OA

Of 23 models predicting incident knee OA, 13 defined the outcome radiographically by plain radiography, typically Kellgren-Lawrence (KL) grade 2 or more, although two models selected the more severe threshold of KL≥3 [23], [24]. The median number of participants/joints with the outcome of interest for these 13 models was 95 (range: 27 to 474). Median AUC following internal and external validation was 0.77 (range: 0.69 to 0.82, 6 models) and 0.76 (range: 0.60 to 0.86, 4 models, 6 populations), respectively (Figure 3). All 13 models included predictors obtainable from clinical assessment. Most common predictors, featuring in more than four final models, were, age, sex, body mass index (BMI), previous knee injury as well as self-reported pain, stiffness, and function scores from the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). Eight models solely used predictors available from clinical assessment [25]–[30], a further six models included predictors sourced from plain radiographs at baseline [23], [24], [31]–[34], three

used magnetic resonance imaging (MRI) [34]–[36] and five incorporated serum or urinary biomarkers (included predictors for all models presented in Figure 2) [24], [35], [37]–[39].

Incident symptomatic OA (frequent knee pain)

A symptomatic knee OA outcome was used in five models, most commonly defined as the onset of frequent knee pain. Median number of participants/joints with the outcome of interest for these models ranged from 51 to 2,103. Median AUC following internal validation was 0.71 (range: 0.70 to 0.78, 5 models). None of these models utilised predictors beyond those obtainable from clinical assessment or plain radiography. Fernandes et al. [27] and Landsmeer et al. [28] externally validated their symptomatic models but with contrasting results. External validation in OAI of the model derived from Nottingham data resulted in poor discrimination and calibration (AUC=0.54 (95%CI 0.50, 0.58)) [27]. Landsmeer et al. found relatively little reduction in their model discrimination upon external validation in the Rotterdam Study-III cohort (AUC=0.71 (0.62, 0.79)) [28] (Figure 3).

Symptomatic-radiographic knee OA

A definition of knee OA combining both radiographic and symptomatic criteria was used in four models. Models by Zhang et al. [26] and Landsmeer et al. [28] were derived solely using clinical assessment predictors; Chan et al. [40] included radiographic predictors, with Lazzarini et al. [35] incorporating radiographic, MRI and biomarker predictors. All models were internally or externally validated. Lazzarini et al. [35] and Chan et al. [40] internally validated by cross-validation and bootstrapping, respectively. Again, the model derived

within the Nottingham cohort did not perform well in OAI (AUC: 0.60 (95% CI: 0.58, 0.63)), although discrimination remained high in the Genetics of OA and Lifestyle case-control study (AUC: 0.79 (0.77, 0.81)) [26]. Landsmeer et al. [28] also validated their combined-criteria model; AUC in Rotterdam Study-III was 0.81 (0.71, 0.90) (Figure 3).

Recorded diagnosis of knee OA

Magnusson et al.'s 2019 study was the only identified study predicting knee OA using Electronic Health Record (EHR) data [29]. They identified male patients undergoing military conscription within Sweden in 1969 (n=40,118), and, using the National Patient Register, who subsequently developed knee OA (n=2,052). Knee OA was defined as the entry of a relevant ICD-9 or ICD-10 code within the patient's EHR between 1987 and 2010. AUC of the model following internal validation was 0.60 (0.59, 0.61) (Figure 3).

Hip OA

Of four models derived to predict hip OA [41]–[44], all originated in the Netherlands and used the same definition: composite outcome of KL ≥2 or total hip replacement (THR). Median number of participants/joints was 994.5. The two earliest models were derived in the Rotterdam Study-I cohort [41], [42], the latter two used the Cohort Hip and Cohort Knee study (CHECK) [43], [44]. All models featured age, sex and BMI, together with radiographic parameters. Baseline KL (0/1) was used in three models [41]–[43], the remainder used the presence of joint space narrowing and osteophytes (upon which KL is calculated) [44]. The most recently published models incorporated trabecular bone texture [43] and patented

automated hip shape via machine learning algorithm [44]. Discrimination of the latter model in particular was high (AUC from internal validation = 0.86; 95%CI: 0.83, 0.90)) with near-perfect calibration [44]. External validation was undertaken only by Saberi Hosnijeh et al., finding a reduction in performance in both Rotterdam Study-II (AUC 0.75; 0.72, 0.79), and more notably in the CHECK cohort dataset (AUC 0.71; 0.66, 0.75) [42] (Figure 4).

Hand OA

We identified two Scandinavian studies (deriving three models) predicting the incidence of hand OA [45], [46]. Magnusson et al. [45] used the aforementioned military conscription cohort and linked National Patient Register to identify participants who developed hand OA (using ICD-10 codes) between 1998-2010. A small proportion (212 (0.5%)) of their study population (n=40,118) developed the outcome. Their final model included education level, BMI and sleep problems; AUC following internal validation was 0.62 (0.58, 0.64) [45]. Building upon this, Johnsen et al. [46] used the Norwegian Nord-Trøndelag Health Study (HUNT2) cohort to externally validate Magnusson et al.'s original model, as well as a simplified algorithm [46]. Model performance was comparable, with AUCs of 0.60 (0.56, 0.64) and 0.62 (0.57, 0.65), respectively (Figure 4). Johnsen et al. [46] then developed their own prediction models for hand OA in male subjects and female subjects, separately, using diagnostic ICD-9 and ICD-10 codes within the Norwegian National Patient Register. Of note, no improvement in performance was observed with addition of a genetic risk score in models for either patient sex, or with reproductive and hormonal factors in female participants [46]. Whilst both studies underwent internal validation, neither was externally validated.

OA (Any Joint)

Black et al., using the Canadian Primary Sentinel Surveillance Network, was the only study that sought to predict the incidence of OA irrespective of joint [47]. They identified 383,117 eligible patients, of whom 12,803 received a billing or problem-list code for OA within five years of cohort entry. Their model consisted of five predictors routinely collected within EHR data – age, sex, BMI, prior leg injury, and osteoporosis diagnosis. Both discrimination (AUC=0.84 (95% CI 0.83, 0.85) and calibration were good following 10-fold cross-validation (Figure 4).

Calibration Summary

Of 12 models internally validating their model derived for knee OA outcomes, five included calibration assessment. Earlier models [27], [31] appraised this using Hosmer-Lemeshow statistic, and four studies presented and appraised calibration plots [27], [29], [30], [38]. Five of the six models for knee OA undergoing external validation assessed calibration using Hosmer-Lemeshow, only Fernandes et al. [27] presented findings visually.

All models for hip (except the earliest [41]), hand and any-joint OA were internally validated. All but one model presented calibration plots visually, reporting reasonable or good agreement between expected and observed outcomes. Models by Gielis et al. [44] and Black et al. [47] demonstrated excellent calibration.

Risk of Bias Summary

The most common sources of potential bias included the extensive use of univariable analysis to select predictors for inclusion in final models, and suboptimal handling of missing data, competing risks, and cohort attrition (domain 4). The inclusion of THR (a measure of both incidence and progression) in composite outcome definitions of incident hip OA was also flagged (Figure 5).

We judged model applicability in terms of their ability to be implemented at scale in diverse general (adult) populations. Within those terms, applicability was typically judged to be poor. This was most often due to the need for predictors obtained by imaging or biological samples which may not be routinely available, recommended, or affordable for such application (domain 2). Ethnicity and other social stratifiers were seldom considered or included in final models and it was often unclear whether datasets used to derive and validate models had drawn from a sufficiently diverse population to be applied at scale in general populations (domain 1).

Discussion

Summary of main findings

We sought to systematically identify and critically evaluate existing multivariable risk prediction models for OA incidence and to consider their potential application at scale in diverse populations to advance individual risk-informed preventive action. Our review identified 26 studies deriving 31 multivariable risk prediction models. 16 models published since 2018 suggests a growing field, attracting machine learning approaches and novel

biomarkers, but one that remains centred around a relatively small number of mature cohort datasets of knee (and to a lesser extent hip) OA incidence in high-income countries.

Importantly, our review identifies a general lack of inclusion of social stratifiers beyond age, sex and occupation-associated risk. Of note, until Chan et al.'s study in 2021 (and subsequently Guan et al. in 2022) [40], [48] there was a complete absence of stratifiers relating to ethnicity and markers of deprivation, factors that are associated with disparities in both incidence and prevalence of OA [49], [50]. The lack of such predictors, as well as income, education and geographic location, may also contribute to a lack of applicability, and usability in, wider populations.

With some exceptions, notably when predicting a future recorded diagnosis of OA across very long prediction horizons, model discrimination after validation ranged from AUC=0.70 to 0.85. This range of performance relates to heterogeneous models with prediction horizons from 2 to 12 years and predictors whose collection and processing varies in cost and complexity. In several models undergoing internal or external validation, calibration was either not reported or relied on the Hosmer-Lemeshow test statistic which is recognized as problematic and no longer recommended [51]. Better approaches, including the visual display of calibration plots [29], [38], [42]–[47], and reported intercept and slope [38], [43], [47] were however used in several more recent studies. Poor calibration of some models was attributed to inherent unpredictability of incident OA over very long prediction horizons [45], and also to challenges in identifying suitably comparable cohort datasets for external validation [27], [42]. We identified no recent examples of externally validated models supported by moderate or strong evidence of good calibration. Acknowledging that the development of a single prediction model for OA may be challenging, particularly across

different target populations, within different healthcare systems, or across long prediction horizons, from adolescence to disease onset, is an important implication. Predicting individual risk and preventive intervention remain achievable, but may require several models in different settings and contexts.

Amid our critical appraisals of risk of bias and areas for methodological improvement were several positives: the use of internal or external validation was common – greatly facilitated by data sharing; the foresight to design overlapping data points across different cohort studies; the shift towards more careful evaluation of model calibration; a common practice of including certain core predictors – age, sex, BMI, previous injury. We would also encourage others to emulate, where possible, Johnsen et al.'s initial approach of testing and adapting a previously published model rather than assuming the need to derive another new model [46].

We found little evidence of patient and public involvement and engagement (PPIE) in included studies. This may contribute to a lack of clarity on potential applications and routes to implementation, and studies may be strengthened with clear rationale statements alongside integration of PPIE, for instance by following Guidance for Reporting Involvement of Patients and the Public (GRIPP2) [52]. Our own review can be criticised on this point, limited by being an unfunded project, with no means for required remuneration. It is an area for future strengthening.

Strengths and weaknesses of our review

Our prospectively registered review used a replicable search strategy without language restriction in the search phase across three electronic databases which was re-run prior to submission and supplemented by searches of reference lists and conference abstracts. Pairs of reviewers working independently and using recommended checklists and risk of bias tools performed study selection, data extraction, and risk of bias assessments. Our review has several limitations. Firstly, risk of bias tools specifically for prediction modelling studies using machine learning techniques were under development at the time of our review [53]. Secondly, several studies derived multiple models for the same outcome. The designation of a 'final model' relied on two reviewers' independent judgement based first on the authors' description but was not necessarily the best performing or most applicable model reported. Thirdly, while it would be of interest to know all of the candidate predictors considered in model development, this information was often lacking or partially reported. We opted not to try to synthesise this information. We also refrained from attempting to calculate events per variable for each model because of this and because it is no longer recommended as a guide to sample size [54], [55]. Fourth, we did not undertake meta-analysis due to study heterogeneity, nor meta-regression due to insufficient number of models. Finally, models were developed by their authors for many reasons. Models judged by us to have low applicability to large-scale use in diverse general populations may be highly applicable for other purposes such as enriched recruitment to clinical trials or within selected clinical settings.

Our review in relation to previous research

We are unaware of any other previous published review of multivariable risk prediction models for OA incidence with which to compare our findings, although we note a published protocol of a review in development relating to prognostic models for knee OA [56]. Reviews in other fields have found similar concerns over methodological quality and applicability of multivariable prediction models for disease incidence [57]–[60]. An excess of model development and a lack of rigorous external validation by independent research teams is a recurrent theme. In the more established field of cardiovascular risk prediction, use of registry and EHR datasets appears more common and efforts are underway to adapt models for application in low and middle income countries [4]. These may signal directions for future development of individual risk prediction in OA. Challenges on the validity and completeness of coding and the availability of information on important predictors within routine EHR data are well-recognised, yet the approach of Black et al. [47] suggests that it may be possible to mitigate some of these and produce prediction models with good performance and a prospect of implementation within existing national health systems. However, it is unclear whether routine EHR data can support accurate risk prediction models, specifically for hip OA and hand OA. Aspects of hip morphology appear to add important predictive value but will have limited availability in routine records for general populations. Furthermore, the consistent use of composite outcomes including THR may limit both applicability and accuracy in predicting incident disease; an implication for hip OA model development that may be highlighted following more extensive external validation. Substantial under- and mis-diagnosis of hand OA poses a different challenge.

Johnsen et al. [46], in their separate prediction models of incident hand OA in both males and females, did not find a significant improvement in model performance with the addition of a genetic risk score. Genetic association within OA is a growing field, with ongoing

identification of associated variants [61]. Whilst the predictive value of these novel variants remains unknown, we believe a model feasible for widespread implementation in clinical practice should utilise routinely available predictors. Furthermore, previous literature suggests that associations of the strength rarely observed in studies are typically required for the accurate prediction of outcomes [62], especially as this is in combination with other predictors, and may explain the apparent "null result" in Johnsen et al.'s models [46].

Our review excluded several studies that we feel deserve specific mention. We excluded studies that relied solely on joint replacement as the outcome because of the risk of conflating predictors of incidence and progression. However, the separation of 'incidence' from 'progression' in OA can be contested. Approaches to modelling changes in symptom and disease severity, classifying cohort enrolment or censorship as a spectrum rather than binary events, such as by Halilaj et al. [63] and Widera et al. [64], may still contain relevant information. In addition, the linked studies of Losina et al. [65] and Michl et al. [66] provide evidence that is highly relevant to introducing individual risk models for OA in one scalable format: patient self-evaluation using an online OA risk calculator. Of note, their knee OA risk calculator used relatively simple-to-report predictors based on the earlier Nottingham risk prediction model, derived by Zhang et al. [26]. Beyond this, there is a lack of evaluation of the impact of risk prediction models for OA used in clinical practice.

Conclusion

In summary, we identify 31 multivariable prediction models for OA from 26 published studies. Whilst there is growing interest amongst researchers, there often remains a lack of applicability to clinical practice in diverse general populations, as well as a paucity of

evaluation of impact of implementation. We suggest that models may benefit from clearly stating their rationale, and by integrating PPIE and predictors such as ethnicity. Furthermore, the progression towards viable risk prediction models for OA, that are applicable across a number of settings, would be aided with a focus on routinely available predictors and with wider external validation of models in varied populations. Lastly, growing interest in machine-learning techniques, as well as the classification of OA disease as progression rather than dichotomous incidence, warrants updated research guidelines to better appraise such innovative approaches.

CONTRIBUTOR STATEMENTS

GMP formulated the research question. GMP and TA developed search strategy. All authors were involved in screening and data extraction; DA, GMP and TA appraised risk of bias. All authors contributed to the manuscript and provided their approval prior to submission.

References

[1] L. Chen, "Overview of clinical prediction models," *Ann. Transl. Med.*, vol. 8, no. 4, pp. 71–71, 2020, doi: 10.21037/atm.2019.11.121.

[2] T. Vos *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015," *Lancet*, vol. 388, no. 10053, pp. 1545–1602, 2016, doi: 10.1016/S0140-6736(16)31678-6.

[3] United Nations, "Transforming our world: the 2030 Agenda for Sustainable Development," 2015. doi: 10.1163/157180910X12665776638740.

[4] The WHO CVD Risk Chart Working Group, "World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions," *Lancet Glob. Heal.*, vol. 7, pp. e1332–e1345, 2019, doi: 10.1016/S2214-109X(19)30318-3.

[5] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle, "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study," *Br. Med. J.*, vol. 335, no. 7611, pp. 136–141, 2007, doi: 10.1136/bmj.39261.471806.55.

[6] B. W. Sahle, A. J. Owen, K. L. Chin, and C. M. Reid, "Risk Prediction Models for Incident Heart Failure: A Systematic Review of Methodology and Model Performance," *J. Card. Fail.*, vol. 23, no. 9, pp. 680–687, 2017, doi: 10.1016/j.cardfail.2017.03.005.

[7] C. Pabinger, H. Lothaller, and A. Geissler, "Utilization rates of knee-arthroplasty in OECD countries," *Osteoarthr. Cartil.*, vol. 23, no. 10, pp. 1664–1673, 2015, doi: 10.1016/j.joca.2015.05.008.

[8] F. Xie, B. Kovic, X. Jin, X. He, M. Wang, and C. Silvestre, "Economic and Humanistic Burden of Osteoarthritis: A Systematic Review of Large Sample Studies," *Pharmacoeconomics*, vol. 34, no. 11, pp. 1087–1100, 2016, doi: 10.1007/s40273-016-0424-x.

[9] J. Puig-Junoy and A. Ruiz Zamora, "Socio-economic costs of osteoarthritis: A systematic

review of cost-of-illness studies," *Semin. Arthritis Rheum.*, vol. 44, no. 5, pp. 531–541, 2015, doi: 10.1016/j.semarthrit.2014.10.012.

[10] D. Culliford, J. Maskell, A. Judge, C. Cooper, D. Prieto-Alhambra, and N. K. Arden, "Future projections of total hip and knee arthroplasty in the UK: Results from the UK Clinical Practice Research Datalink," *Osteoarthr. Cartil.*, vol. 23, no. 4, pp. 594–600, 2015, doi: 10.1016/j.joca.2014.12.022.

[11] I. N. Ackerman *et al.*, "The projected burden of primary total knee and hip replacement for osteoarthritis in Australia to the year 2030," *BMC Musculoskelet. Disord.*, vol. 20, no. 1, pp. 1–10, 2019, doi: 10.1186/s12891-019-2411-9.

[12] C. A. Emery *et al.*, "Establishing outcome measures in early knee osteoarthritis," *Nat. Rev. Rheumatol.*, vol. 15, no. 7, pp. 438–448, 2019, doi: 10.1038/s41584-019-0237-3.

[13] P. Croft *et al.*, "The science of clinical practice: Disease diagnosis or patient prognosis? Evidence about 'what is likely to happen' should shape clinical practice," *BMC Med.*, vol. 13, no. 1, pp. 1–8, 2015, doi: 10.1186/s12916-014-0265-4.

[14] L. Shamseer *et al.*, "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation OPEN ACCESS," *BMJ*, vol. 349, p. g7647, 2015, doi: 10.1136/bmj.g7647.

[15] G. J. Geersing, W. Bouwmeester, P. Zuithoff, R. Spijker, M. Leeflang, and K. Moons, "Search filters for finding prognostic and diagnostic prediction studies in medline to enhance systematic reviews," *PLoS One*, vol. 7, no. 2, p. e32844, 2012, doi: 10.1371/journal.pone.0032844.

[16] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, "Rayyan - a web and mobile app for systematic reviews," *Syst. Rev.*, vol. 5, no. 1, p. 210, 2016, doi: 10.1186/s13643-016-0384-4.

[17] Microsoft Corporation, "Microsoft Excel." 2021.

[18] Cochrane, "Prognosis Tools," 2021. [Online]. Available:

https://methods.cochrane.org/prognosis/tools.

[19] K. G. M. Moons *et al.*, "Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist," *PLoS Med.*, vol. 11, no. 10, 2014, doi: 10.1371/journal.pmed.1001744.

[20] R. F. Wolff *et al.*, "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, 2019, doi: 10.7326/M18-1376.

[21] E. W. Steyerberg *et al.*, "Assessing the performance of prediction models: A framework for traditional and novel measures," *Epidemiology*, vol. 21, no. 1, pp. 128–138, 2010, doi: 10.1097/EDE.0b013e3181c30fb2.

[22] D. Hosmer and S. Lemeshow, *Assessing the Fit of the Model*. 2000.

[23] D. L. Riddle, P. . Stratford, and R. . Perera, "The incident tibiofemoral osteoarthritis with rapid progression phenotype: development and validation of a prognostic prediction rule," *Osteoarthr. Cartil.*, vol. 24, no. 12, pp. 2100–2107, 2016, [Online]. Available: http://www.elsevier.com/inca/publications/store/6/2/3/0/5/5/index.htt.

[24] J. Driban *et al.*, "Risk factors can classify individuals who develop accelerated knee osteoarthritis: Data from the osteoarthritis initiative.," *J. Orthop. Res.*, vol. 36, no. 3, pp. 876–880, 2018, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28776751/.

[25] G. Peat, E. Thomas, R. Duncan, and L. Wood, "Is a 'false-positive' clinical diagnosis of knee osteoarthritis just the early diagnosis of pre-radiographic disease?," *Arthritis care & research*, vol. 62, no. 10. United States, pp. 1502–1506, 2010, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20506491/.

[26] W. Zhang *et al.*, "Nottingham knee osteoarthritis risk prediction models," *Ann. Rheum. Dis.*, vol. 70, no. 9, pp. 1599–1604, 2011, doi: 10.1136/ard.2011.149807.

[27] G. S. Fernandes, A. Bhattacharya, D. F. McWilliams, S. L. Ingham, M. Doherty, and W. Zhang, "Risk prediction model for knee pain in the Nottingham community: A Bayesian modelling approach," *Arthritis Res. Ther.*, vol. 19, no. 59, pp. 1–8, 2017, doi: 10.1186/s13075-017-1272-

6.

[28] M. Landsmeer *et al.*, "Predicting knee pain and knee osteoarthritis among overweight women," *J. Am. Board Fam. Med.*, vol. 32, no. 4, pp. 575–584, 2019, [Online]. Available: https://www.jabfm.org/content/jabfp/32/4/575.full.pdf.

[29] K. Magnusson, A. Turkiewicz, S. Timpka, and M. Englund, "A Prediction Model for the 40-Year Risk of Knee Osteoarthritis in Adolescent Men," *Arthritis Care Res. (Hoboken).*, vol. 71, no. 4, pp. 558–562, 2019.

[30] L. Wang, H. Lu, H. Chen, S. Jin, M. Wang, and S. Shang, "Development of a model for predicting the 4-year risk of symptomatic knee osteoarthritis in China: a longitudinal cohort study," *Arthritis Res. Ther.*, vol. 23, no. 1, pp. 1–13, 2021, doi: 10.1186/s13075-021-02447-5.

[31] M. Kinds *et al.*, "Evaluation of separate quantitative radiographic features adds to the prediction of incident radiographic osteoarthritis in individuals with recent onset of knee pain: 5-year follow-up in the CHECK cohort," *Osteoarthr. Cartil.*, vol. 20, no. 6, pp. 548–556, 2012.

[32] H. J. M. Kerkhof *et al.*, "Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors," *Ann. Rheum. Dis.*, vol. 73, no. 12, pp. 2116–2121, 2014, doi: 10.1136/annrheumdis-2013-203620.

[33] T. Janvier, R. Jennane, H. Toumi, and E. Lespessailles, "Subchondral tibial bone texture predicts the incidence of radiographic knee osteoarthritis: data from the Osteoarthritis Initiative.," *Osteoarthr. Cartil.*, vol. 25, no. 12, pp. 2047–2054, 2017, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28935435/.

[34] L. Sharma *et al.*, "Knee tissue lesions and prediction of incident knee osteoarthritis over 7 years in a cohort of persons at higher risk," *Osteoarthr. Cartil.*, vol. 25, no. 7, pp. 1068–1075, 2017, [Online]. Available: http://www.elsevier.com/inca/publications/store/6/2/3/0/5/5/index.htt.

[35] N. Lazzarini *et al.*, "A machine learning approach for the identification of new biomarkers for

knee osteoarthritis development in overweight and obese women," *Osteoarthr. Cartil.*, vol. 25, no. 12, pp. 2014–2021, 2017, [Online]. Available: http://www.elsevier.com/inca/publications/store/6/2/3/0/5/5/index.htt.

[36] G. B. Joseph, C. E. McCulloch, M. C. Nevitt, T. M. Link, and J. H. Sohn, "Machine learning to predict incident radiographic knee osteoarthritis over 8 Years using combined MR imaging features, demographics, and clinical factors: data from the Osteoarthritis Initiative," *Osteoarthr. Cartil.*, vol. 30, no. 2, pp. 270–279, 2022, doi: 10.1016/j.joca.2021.11.007.

[37] M. Camacho-Encina *et al.*, "Discovery of an autoantibody signature for the early diagnosis of knee osteoarthritis: Data from the Osteoarthritis Initiative," *Ann. Rheum. Dis.*, vol. 78, no. 12, pp. 1699–1705, 2019, [Online]. Available: http://ard.bmj.com/.

[38] C. Garriga *et al.*, "Predicting Incident Radiographic Knee Osteoarthritis in Middle-Aged Women Within Four Years: The Importance of Knee-Level Prognostic Factors," *Arthritis Care Res.*, vol. 72, no. 1, pp. 88–97, 2020, [Online]. Available: http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2151-4658.

[39] L. Lourido *et al.*, "A clinical model including protein biomarkers predicts radiographic knee osteoarthritis: a prospective study using data from the Osteoarthritis Initiative," *Osteoarthr. Cartil.*, vol. 29, no. 8, pp. 1147–1154, 2021, doi: 10.1016/j.joca.2021.04.011.

[40] L. C. Chan, H. H. T. Li, P. K. Chan, and C. Wen, "A machine learning-based approach to decipher multi-etiology of knee osteoarthritis onset and deterioration," *Osteoarthr. Cartil. Open*, vol. 3, no. 1, p. 100135, 2021, doi: 10.1016/j.ocarto.2020.100135.

[41] M. Castano-Betancourt *et al.*, "The contribution of hip geometry to the prediction of hip osteoarthritis," *Osteoarthr. Cartil.*, vol. 21, no. 10, pp. 1530–1536, 2013.

[42] F. Saberi Hosnijeh *et al.*, "Development of a prediction model for future risk of radiographic hip osteoarthritis," *Osteoarthr. Cartil.*, vol. 26, no. 4, pp. 540–546, 2018, doi: 10.1016/j.joca.2018.01.015.

[43] J. Hirvasniemi *et al.*, "Bone texture analysis for prediction of incident radiographic hip

osteoarthritis using machine learning: data from the Cohort Hip and Cohort Knee (CHECK) study.," *Osteoarthr. Cartil.*, vol. 27, no. 6, pp. 906–914, 2019, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30825609/.

[44] W. Gielis *et al.*, "An automated workflow based on hip shape improves personalized risk prediction for hip osteoarthritis in the CHECK study.," *Osteoarthr. Cartil.*, vol. 28, no. 1, pp. 62–70, 2020, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31604136/.

[45] K. Magnusson, A. Turkiewicz, S. Timpka, and M. Englund, "Prediction of midlife hand osteoarthritis in young men," *Osteoarthr. Cartil.*, vol. 26, no. 8, pp. 1027–1032, Aug. 2018, doi: 10.1016/j.joca.2018.05.010.

[46] M. B. Johnsen *et al.*, "Development and validation of a prediction model for incident hand osteoarthritis in the HUNT study," *Osteoarthr. Cartil.*, vol. 28, no. 7, pp. 932–940, Jul. 2020, doi: 10.1016/j.joca.2020.04.005.

[47] J. Black, A. Terry, and D. Lizotte, "Development and evaluation of an osteoarthritis risk model for integration into primary care health information technology," *Int. J. Med. Inform.*, vol. 141, p. 104160, 2020, [Online]. Available: http://www.elsevier.com/inca/publications/store/5/0/6/0/4/0/.

[48] B. Guan *et al.*, "Deep learning approach to predict pain progression in knee osteoarthritis," *Skeletal Radiol.*, vol. 51, no. 2, pp. 363–373, 2022, doi: 10.1007/s00256-021-03773-0.

[49] L. F. Callahan, R. J. Cleveland, K. D. Allen, and Y. Golightly, "Racial/Ethnic, Socioeconomic, and Geographic Disparities in the Epidemiology of Knee and Hip Osteoarthritis," *Rheum. Dis. Clin. North Am.*, vol. 47, no. 1, pp. 1–20, 2021, doi: 10.1016/j.rdc.2020.09.001.

[50] C. Reyes *et al.*, "Socio-economic status and the risk of developing hand, hip or knee osteoarthritis: A region-wide ecological study," *Osteoarthr. Cartil.*, vol. 23, no. 8, pp. 1323–1329, 2015, doi: 10.1016/j.joca.2015.03.020.

[51] B. Van Calster *et al.*, "Calibration: The Achilles heel of predictive analytics," *BMC Med.*, vol. 17, no. 1, pp. 1–7, 2019, doi: 10.1186/s12916-019-1466-7.

[52] S. Staniszewska *et al.*, "GRIPP2 reporting checklists: Tools to improve reporting of patient and public involvement in research," *Res. Involv. Engagem.*, vol. 3, no. 1, pp. 1–11, 2017, doi: 10.1186/s40900-017-0062-2.

[53] G. S. Collins *et al.*, "Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence," *BMJ Open*, vol. 11, no. 7, pp. 1–7, 2021, doi: 10.1136/bmjopen-2020-048008.

[54] M. Van Smeden *et al.*, "No rationale for 1 variable per 10 events criterion for binary logistic regression analysis," *BMC Med. Res. Methodol.*, vol. 16, no. 1, pp. 1–12, 2016, doi: 10.1186/s12874-016-0267-3.

[55] R. D. Riley *et al.*, "Calculating the sample size required for developing a clinical prediction model," *BMJ*, vol. 368, no. March, pp. 1–12, 2020, doi: 10.1136/bmj.m441.

[56] J. Zhong *et al.*, "Prognostic models for knee osteoarthritis: a protocol for systematic review, critical appraisal, and meta-analysis," *Syst. Rev.*, vol. 10, no. 1, pp. 1–9, 2021, doi: 10.1186/s13643-021-01683-9.

[57] I. Kaiser, A. B. Pfahlberg, W. Uter, M. V. Heppt, M. B. Veierød, and O. Gefeller, "Risk prediction models for melanoma: A systematic review on the heterogeneity in model development and validation," *Int. J. Environ. Res. Public Health*, vol. 17, no. 21, pp. 1–25, 2020, doi: 10.3390/ijerph17217919.

[58] J. C. L. Himmelreich *et al.*, "Prediction models for atrial fibrillation applicable in the community: A systematic review and meta-analysis," *Europace*, vol. 22, no. 5, pp. 684–694, 2020, doi: 10.1093/europace/euaa005.

[59] J. A. A. G. Damen *et al.*, "Prediction models for cardiovascular disease risk in the general population: Systematic review," *BMJ*, vol. 353, 2016, doi: 10.1136/bmj.i2416.

[60] W. Xu, J. Huang, Q. Yu, H. Yu, Y. Pu, and Q. Shi, "A systematic review of the status and methodological considerations for estimating risk of first ever stroke in the general

population," *Neurol. Sci.*, vol. 42, no. 6, pp. 2235–2247, 2021, doi: 10.1007/s10072-021-05219-w.

[61] C. G. Boer *et al.*, "Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations," *Cell*, vol. 184, pp. 4784–4818, 2021, doi: 10.1016/j.cell.2021.07.038.

[62] M. S. Pepe, H. Janes, G. Longton, W. Leisenring, and P. Newcomb, "Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker," *Am. J. Epidemiol.*, vol. 159, no. 9, pp. 882–890, 2004, doi: 10.1093/aje/kwh101.

[63] E. Halilaj, Y. Le, J. L. Hicks, T. J. Hastie, and S. L. Delp, "Modeling and predicting osteoarthritis progression: data from the osteoarthritis initiative," *Osteoarthr. Cartil.*, vol. 26, no. 12, pp. 1643–1650, 2018, doi: 10.1016/j.joca.2018.08.003.

[64] P. Widera *et al.*, "Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, 2020, doi: 10.1038/s41598-020-64643-8.

[65] E. Losina, K. Klara, G. . Michl, J. E. Collins, and J. . Katz, "Development and feasibility of a personalized, interactive risk calculator for knee osteoarthritis," *BMC Musculoskelet. Disord.*, vol. 16, no. 1, p. 312, 2015, [Online]. Available: http://www.biomedcentral.com/bmcmusculoskeletdisord/.

[66] G. L. Michl, J. N. Katz, and E. Losina, "Risk and risk perception of knee osteoarthritis in the US: A population-based study," *Osteoarthr. Cartil.*, vol. 24, no. 4, pp. 593–596, 2016, doi: 10.1016/j.joca.2015.11.001.

Tables

Table 1. Descriptive Characteristics of Included Studies

General Information				Information Relating to Final Model(s)						
First Author, Year	Cohort Datasets (Nations)^	Modelling Technique	Relevant Models Reported	Outcome Definition	Source of Predictors	Max predict horizon (years)*	No. Predictors	No. Participants (with Outcome) §	Int. Val.	Ext. Val.
Knee Osteoarthritis										
Peat, 2010[23]	**CAS-K** (UK)	Regression	2	ROA	SR, PE	3	10	122 (42)	N	N
Zhang, 2011[24]	**Nott.**, OAI, GOAL (UK, USA)	Regression	2	ROA	SR	12	6	179 (99)	N	Y
				SROA	SR	12	6	179 (99)	N	Y
Kinds, 2012[29]	**CHECK** (NL)	Regression	12	ROA	SR, PE, XR	5	4	985j (189j)	Y	N
Kerkhof, 2014[30]	**RS-I**, RS-II, Ching. (NL, UK)	Regression	6	ROA	SR, PE, XR	9.4	5	2,628 (474)	N	Y
Riddle, 2016[21]	**MOST,** OAI (USA)	Regression	2	ROA	SR, PE, XR	5	4	2,824j (262j)	N	Y
Fernandes, 2017[25]	**Nott.**, OAI (UK, USA)	Regression	1	FKP	SR	12	7	1,203 (?)	Y	Y
Janvier, 2017[31]	**OAI** (USA)	Regression	18	ROA	SR, PE, XR	2	4	344j (79j)	Y	N
Lazzarini, 2017[33]	**PROOF** (NL)	M-L	3	SROA	SR, PE, XR, MRI, SB	2.5	8	354 (39)	Y	N
				FKP	SR, PE, XR	2.5	7	351 (51)	Y	N
				ROA	SR, PE, MRI, SB	2.5	5	321 (27)	Y	N
Sharma, 2017[32]	**OAI** (USA)	Regression	25	ROA	SR, PE, XR, MRI	8	12	841j (53j)	N	N
Driban, 2018[22]	**OAI** (USA)	Regression	1	ROA	SR, PE, XR, SB	4	4	162 (54)	Y	N
Camacho-Encina, 2019[35]	**OAI,** OAI (USA)	Regression	2	ROA	SR, PE, SB	8	7	327 (146)	N	Y
Landsmeer, 2019[26]	**PROOF**, RS-III (NL)	Regression	2	FKP	SR, PE	6.5	7	237 (75)	N	Y
				SROA	SR,PE	6.5	7	235 (70)	N	Y
Magnusson, 2019[27]	**SWE Con.** (SWE)	Regression	2	RDOA	SR, PE, EHR	41	3	40,118 (2,052)	Y	N
Garriga, 2020[36]	**Ching.** (UK)	Regression	2	ROA	SR, PE, UB	4	3	649 (95)	Y	N
Chan, 2021[38]	**OAI** (USA)	M-L	5	SROA	SR, PE, XR	2	24	2,640 (1,254)	Y	N
Joseph, 2021[34]	**OAI** (USA)	M-L	3	ROA	SR, PE, MRI	8	10	710 (124)	Y	N
Lourido, 2021[37]	**OAI** (USA)	Regression	5	ROA	SR, PE, SB	8	6	200 (86)	Y	N
Wang, 2021[28]	**CHARLS** (China)	Regression	1	FKP	SR, PE	4	10	8,193 (815)	Y	N
Guan, 2022[46]	**OAI** (USA)	M-L	4	FKP	SR, PE, XR	2	6	4200j (2103j)	Y	N
Hip Osteoarthritis										
Castan-Betancourt, 2013[39]	**RS-I** (NL)	Regression	14	ROA/THR	SR, PE, XR	11	9	688 (119)	N	N
S Hosnijeh, 2018[40]	**RS-I,** RS-II, CHECK (NL)	Regression	5	ROA/THR	SR, PE, XR, SB	10	13	4,575j (258j)	Y	Y
Hirvasniemi, 2019[41]	**CHECK** (NL)	M-L	10	ROA/THR	SR, PE, XR	10	5	987j (43j)	Y	N
Gielis, 2020[42]	**CHECK** (NL)	Regression	5	ROA/THR	SR, PE, XR	8	4	1,002 (185)	Y	N

Hand Osteoarthritis										
Magnusson, 2018[43]	**SWE Con.** (SWE)	Regression	1	RDOA	SR, PE	41	6	40,118 (212)	Y	N
Johnsen, 2020[44]	**HUNT2** (NOR)	Regression	5	RDOA	SR, PE	23	5	♂: 17,153 (206)	Y	N
				RDOA	SR, PE	23	5	♀: 18,682 (732)	Y	N
Osteoarthritis (Any Joint)										
Black, 2020[45]	**CPCSSN** (CAN)	Regression	1	RDOA	EHR	5	5	383,117 (12,803)	Y	N

CAN Canada; **CAS-K** Clinical Assessment Study – Knee; **C-C** Case-Control; **CHARLS** China Health and Retirement Longitudinal Study; **CHECK** Cohort Hip and Cohort Knee study; **Ching** Chingford Study; **CPCSSN** Canadian Primary Care Sentinel Surveillance Network; **EHR** Electronic Health Record; **FKP** Frequent knee pain; **GOAL** Genetics of Osteoarthritis and Lifestyle study; **HUNT** Helseundersøkelsen i Nord-Trøndelag study; **KOA** Knee osteoarthritis; **M-L** Machine-Learning; **MOST** Multicenter osteoarthritis study; **NL** Netherlands; **NOR** Norway; **Nott** Nottingham Cohort; **OAI** Osteoarthritis Initiative; **PE** Physical examination; **PROOF** PRevention of knee Osteoarthritis in Overweight Females Study; **RCT** Randomised Controlled Trial; **RDOA** Recorded diagnosis of osteoarthritis; **ROA** Radiographic osteoarthritis; **RS** Rotterdam Study; **SB** Serum biomarker; **SR** Self-report; **SROA** Symptomatic-radiographic osteoarthritis; **SWE** Sweden; **SWE Con.** Swedish Conscription Cohort; **THR** Total Hip Replacement; **UK** United Kingdom; **USA** United States of America; **XR** X-ray (Plain radiography); **UB** Urine Biomarker

ˆ Bold denotes derivation dataset, non-bold samples in which external validation was undertaken; * Certain studies had follow-up at interim intervals. § j = joint level of interest i.e. 985j (189j) for Kinds et al. represents 985 knees included, of which 189 knees developed outcome

Figure Legends

Figure 1. Flowchart of Article Screening and Inclusion

Figure 2. Predictor variables and domains represented in final models, by mode of assessment

Figure 3. Forest Plot of Area under the Curve (AUC) with 95% confidence intervals for multivariable prediction models for incident knee OA, stratified by outcome, ordered by year of publication.

First column lists author, year of publication, dataset, prediction horizon.

White marker = model development; grey marker = internal validation; black marker = external validation.

OA Osteoarthritis

Figure 4. Forest Plot of Area under the Curve (AUC) with 95% confidence intervals for multivariable prediction models for incident hip, hand and any joint OA, ordered by year of publication.

First column lists author, year of publication, dataset, prediction horizon.

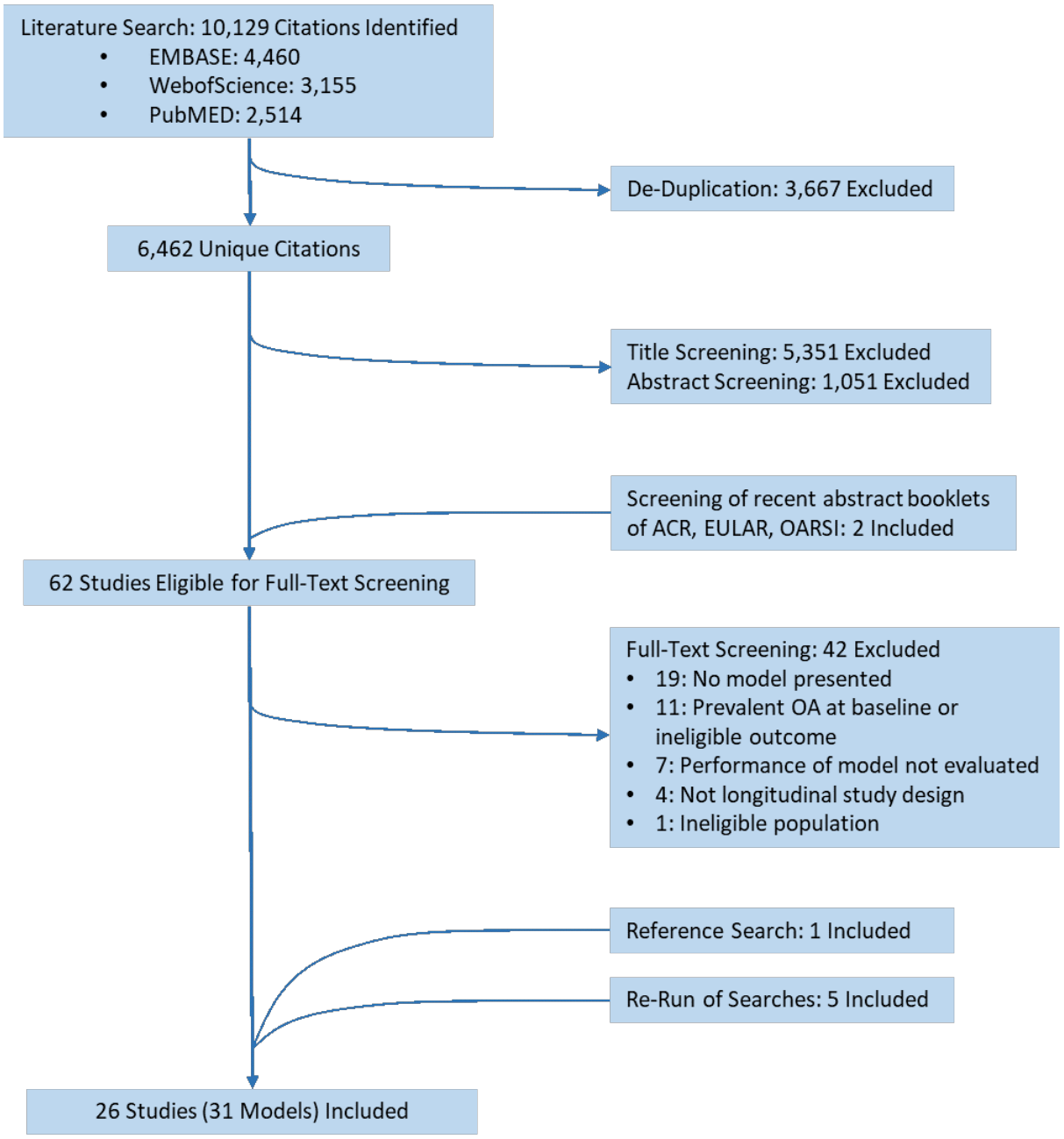
White marker = model development, grey marker = internal validation; black marker = external validation.

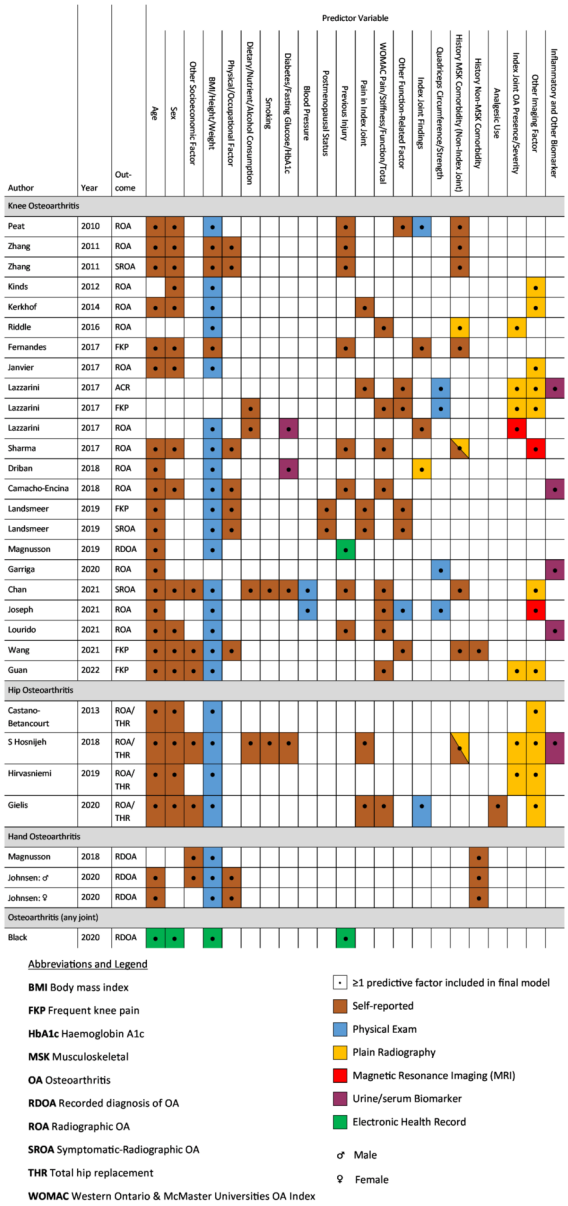
OA Osteoarthritis; **THR** Total hip replacement

Figure 5. Risk of Bias Assessment using PROBAST Tool

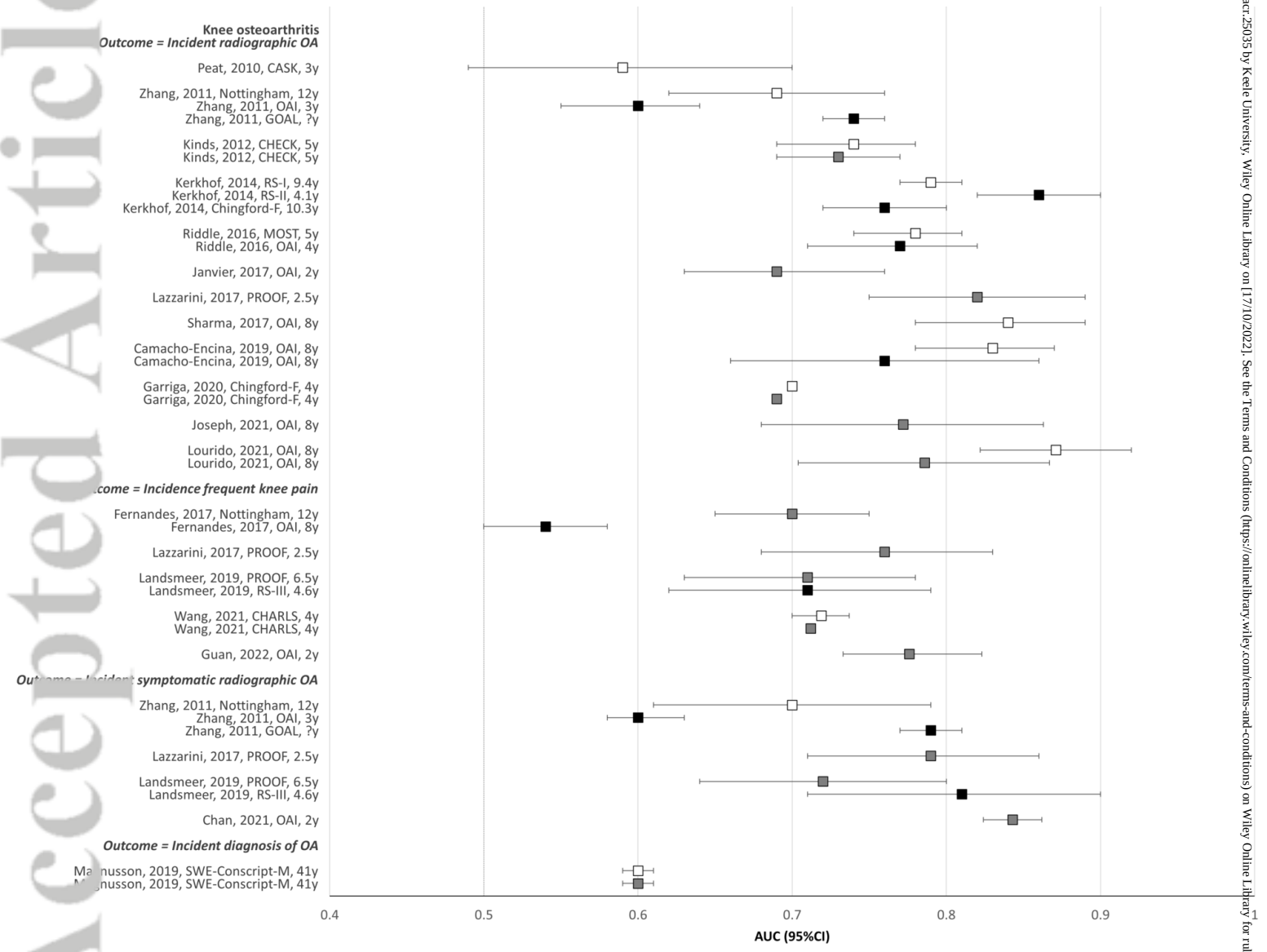
ROB Risk of Bias; **Applic** Applicability

FKP Frequent knee pain; **ROA** Radiographic osteoarthritis; **SROA** Symptomatic-radiographic osteoarthritis; ♂ Male; ♀ Female

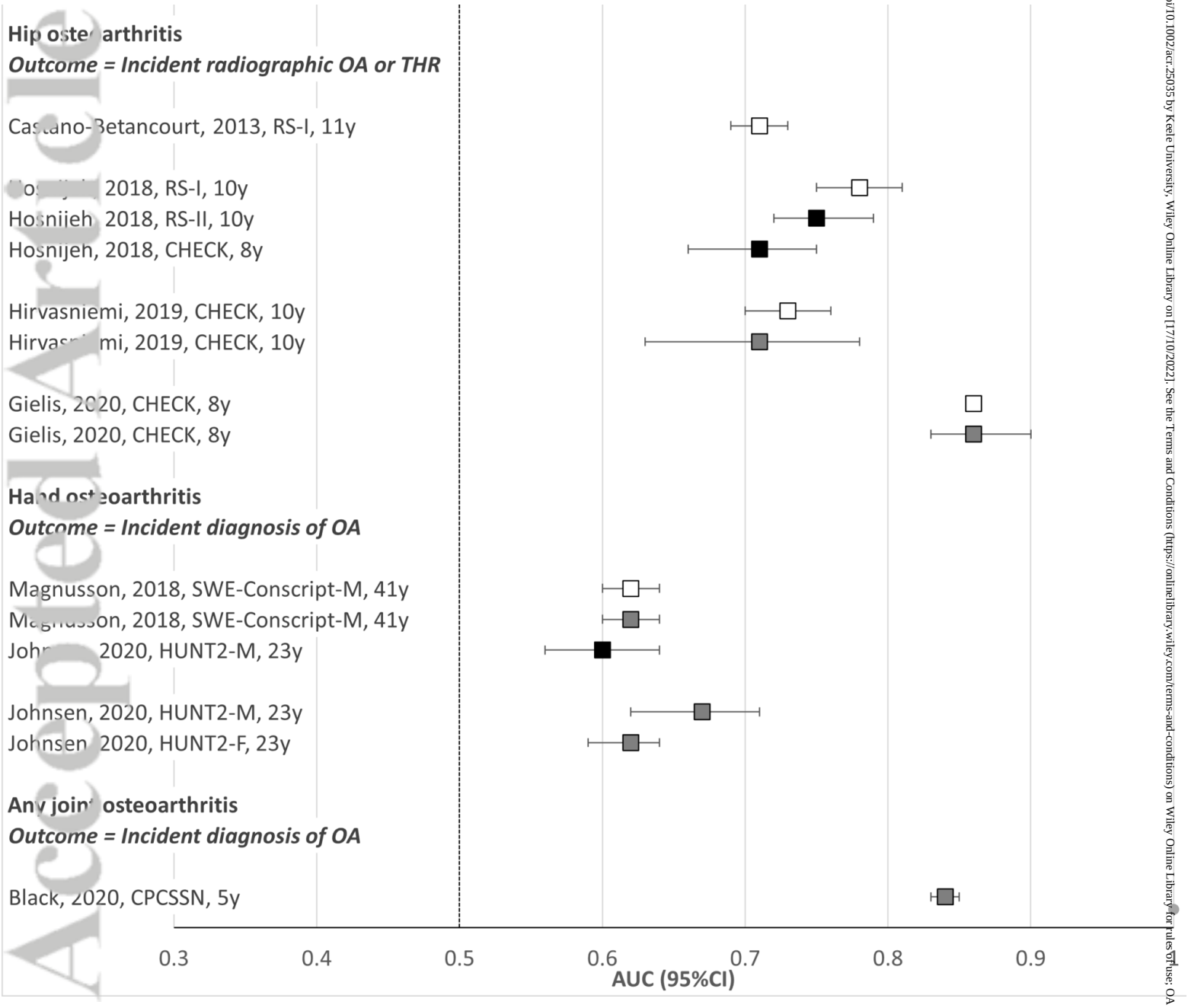




ACR_25035_Figure 2 Predictor Variables.tif



ACR_25035_Figure 3 AUC Forestplot Knee.tif



ACR_25035_Figure 4 AUC HipHandAny.tif

Author (Outcome)	Year	Domain 1 Participants ROB	Domain 1 Participants Applic	Domain 2 Predictors ROB	Domain 2 Predictors Applic	Domain 3 Outcome ROB	Domain 3 Outcome Applic	Domain 4 Analysis ROB	Overall ROB	Overall Applic
Knee Osteoarthritis										
Peat	2010	✓	?	✓	✓	✓	✓	✗	✗	?
Zhang (ROA)	2011	✓	?	✗	?	✓	✓	✗	✗	?
Zhang (SROA)	2011	✓	?	✗	?	✓	✓	✗	✗	?
Kinds	2012	✓	✓	✓	✗	✓	✓	✗	✗	✗
Kerkhof	2014	?	✓	✓	✗	✓	✓	✗	✗	✗
Riddle	2016	✓	✓	✓	✗	✓	?	✗	✗	✗
Fernandes	2017	?	✓	✓	✓	✓	✓	✗	✗	✓
Janvier	2017	✓	✓	✓	✗	✓	✓	✗	✗	✗
Lazzarini (SROA)	2017	?	?	✓	✗	✓	✓	✗	✗	✗
Lazzarini (FKP)	2017	?	?	✓	✗	✓	✓	✗	✗	✗
Lazzarini (ROA)	2017	?	?	✓	✗	✓	✓	✗	✗	✗
Sharma	2017	✓	?	✓	✗	✓	✓	✗	✗	✗
Driban	2018	✓	?	✓	✗	✓	?	✗	✗	✗
Camacho-Encina	2018	?	?	✓	✗	✓	✓	✗	✗	✗
Landsmeer (FKP)	2019	?	?	✓	✓	✓	✓	✗	✗	?
Landsmeer (SROA)	2019	?	?	✓	✓	✓	✓	✗	✗	?
Magnusson	2019	✓	?	✓	✓	✓	✓	✗	✗	?
Garriga	2020	?	?	✓	✗	✓	✓	✗	✗	✗
Chan	2021	✓	?	✓	✗	✓	✓	✗	✗	✗
Joseph	2021	?	✓	✓	✗	✓	✓	✗	✗	✗
Lourido	2021	?	?	✓	✗	✓	✓	✗	✗	✗
Wang	2021	✓	?	✓	✓	✓	✓	✗	✗	?
Guan	2022	?	✗	✓	✗	✓	✓	✗	✗	✗
Hip Osteoarthritis										
Castano-Betancourt	2013	✗	✓	✓	✗	✓	?	✗	✗	✗
S Hosnijeh	2018	✗	✓	✓	✗	✓	?	✗	✗	✗
Hirvasniemi	2019	✗	✓	✓	✗	✓	?	✗	✗	✗
Gielis	2020	✗	✓	✓	✗	✓	?	✗	✗	✗
Hand Osteoarthritis										
Magnusson	2018	✓	?	✓	✗	✓	?	✗	✗	✗
Johnsen: ♂	2020	✓	✓	✓	✓	✓	?	✗	✗	?
Johnsen: ♀	2020	✓	✓	✓	✓	✓	?	✗	✗	?
Osteoarthritis (any joint)										
Black	2020	✓	✓	✓	?	✓	?	✓	✓	?

✓ Low Risk ✗ High Risk ? Of Unclear Risk

ACR_25035_Figure 5 RoB Summary.tif