Stratified care in hip arthroscopy: can we predict successful and unsuccessful outcomes? Development and external temporal validation of multivariable prediction models

Lasse Ishøi [1], Kristian Thorborg [1], Thomas Kallemose,[2] Joanne L Kemp [3], Michael P Reiman [4], Mathias Fabricius Nielsen [1], Per Hölmich [1]

[1]Sports Orthopaedic Research Center–Copenhagen (SORC-C), Arthroscopic Center, Department of Orthopedic Surgery, Hvidovre Hospital, Copenhagen University Hospital, Hvidovre, Denmark
[2]Department of Clinical Research, Hvidovre Hospital, Copenhagen University Hospital, Hvidovre, Denmark
[3]Latrobe Sports Exercise Medicine Research Centre, School of Allied Health, Human Services and Sport, La Trobe University, Bundoora, Victoria, Australia
[4]Department of Orthopedic Surgery, Duke University, Duke University Medical Center, Durham, North Carolina, USA

Correspondence to
Mr Lasse Ishøi, Sports Orthopaedic Research Center–Copenhagen (SORC-C), Arthroscopic Center, Department of Orthopedic Surgery, Copenhagen University Hospital, Hvidovre Hospital, Hvidovre, Denmark; lasse.ishoei@regionh.dk

Check for updates

ABSTRACT

Objective Although hip arthroscopy is a widely adopted treatment option for hip-related pain, it is unknown whether preoperative clinical information can be used to assist surgical decision-making to avoid offering surgery to patients with limited potential for a successful outcome. We aimed to develop and validate clinical prediction models to identify patients more likely to have an unsuccessful or successful outcome 1 year post hip arthroscopy based on the patient acceptable symptom state.

Methods Patient records were extracted from the Danish Hip Arthroscopy Registry (DHAR). A priori, 26 common clinical variables from DHAR were selected as prognostic factors, including demographics, radiographic parameters of hip morphology and self-reported measures. We used 1082 hip arthroscopy patients (surgery performed 25 April 2012 to 4 October 2017) to develop the clinical prediction models based on logistic regression analyses. The development models were internally validated using bootstrapping and shrinkage before temporal external validation was performed using 464 hip arthroscopy patients (surgery performed 5 October 2017 to 13 May 2019).

Results The prediction model for unsuccessful outcomes showed best and acceptable predictive performance on the external validation dataset for all multiple imputations (Nagelkerke R^2 range: 0.25–0.26) and calibration (intercept range: −0.10 to −0.11; slope range: 1.06–1.09), and acceptable discrimination (area under the curve range: 0.76–0.77). The prediction model for successful outcomes did not calibrate well, while also showing poor discrimination.

Conclusion Common clinical variables including demographics, radiographic parameters of hip morphology and self-reported measures were able to predict the probability of having an unsuccessful outcome 1 year after hip arthroscopy, while the model for successful outcome showed unacceptable accuracy. The externally validated prediction model can be used to support clinical evaluation and shared decision making by informing the orthopaedic surgeon and patient about the risk of an unsuccessful outcome, and thus when surgery may not be appropriate.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Level 1 evidence exists for the effectiveness of hip arthroscopy for femoroacetabular impingement syndrome.
⇒ Many patients have residual symptoms after hip arthroscopy.
⇒ It is unknown if presurgical clinical variables can be used to predict the outcome after hip arthroscopy

WHAT THIS STUDY ADDS

⇒ Common clinical variables obtained prior to hip arthroscopy can predict the risk of having an unsuccessful outcome (not achieving the patients acceptable symptom state) 1 year after hip arthroscopy.
⇒ Patients with a successful outcome (achieving the patients acceptable symptom state) were less accurately predicted.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The prediction model can be used to support the shared decision-making process by informing the orthopaedic surgeon and patient of when surgery may not be appropriate.
⇒ This may improve the overall outcomes after hip arthroscopy and improve the cost-effectiveness of the procedure.

INTRODUCTION

Hip-related pain causes disability and low quality of life in young to middle-aged individuals.[1–3] Since the conceptualisation of femoroacetabular impingement in the early 2000s by Ganz et al,[4] large advances have been made in relation to definitions, diagnosis, classifications and treatment of hip-related pain.[1 2 5] This has led to an exponential rise in the number of hip arthroscopies performed globally.[3 6] Several studies have investigated outcomes after hip arthroscopy, showing favourable short-term to long-term results.[7–9] However, residual symptoms and activity limitations are common,[10] and up to 50% of patients seem to have unacceptable symptoms[11–13] or are unable to return to preinjury sports activities after hip arthroscopy.[14–16] These results suggest that, although considered effective at a group level, not all patients are suited for hip arthroscopy. Consequently, there has been recent focus on identifying prognostic factors (such as age and sex) associated with good and poor

outcomes after hip arthroscopy to aid surgical candidate selection and improve surgical outcomes.[17 18 19] While identification of prognostic factors can be used to guide preliminary decision making at a group level, development and external validation of clinical prediction models are needed for individual outcome prediction.[20 21 22] Several prediction models have been published recently for hip arthroscopy patients,[23–31] yet, only one model, predicting conversion to hip arthroplasty, has been externally validated.[24] However, this study included intra-articular findings identified during hip arthroscopy as predictor variables, limiting the utility of the model prior to surgery.[24] In addition, most existing prediction models attempted to predict achievement of the minimal clinically important difference (MCID),[23 25 26 28–30] although achieving an acceptable symptom state or not matters more to patients than an improvement.[32]

We aimed to advance the field of individual prognosis in patients undergoing hip arthroscopy by developing and validating prediction models. These models were applied preoperatively to determine the probability of achieving an unsuccessful or successful outcome defined by the patient's acceptable symptom state, as a primary aim and improvement or not, based on MCID, as a secondary aim, at 1-year post hip arthroscopy. In additional exploratory analyses, all models were reconstructed using intra-articular findings from the arthroscopic procedure to investigate the potential added benefit of such information.

METHODS
For the current study, we followed the initial three steps of The PROGnosis RESearch Strategy (PROGRESS) framework (figure 1).[33] The PROGRESS[33] is a four-step framework for prognostic research: (step 1) description of outcomes of current care (fundamental prognosis research), (step 2) identification of factors associated with outcomes (prognostic factor research), (step 3) development and validation of prediction models (prognostic model research) and (step 4) utilisation of the information to tailor treatment (stratified care research). The final step in the PROGRESS framework (step 4: stratified care research) is beyond the scope of the present study.

The reporting of the present study adheres to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines,[34 35] supplemented with recommendations from Prediction model Risk Of Bias Assessment Tool.[36 37] We developed and validated four multivariable prediction models to determine 1-year outcomes of patients with hip-related pain undergoing hip arthroscopy in Denmark using only predictor variables that are available at the initial consultation before undergoing hip arthroscopy (demographic data, radiological data, patient-reported outcome measures) to reflect the intended use of the models.[34–36] As supplementary and exploratory analyses, all models were also constructed including perioperative predictor variables (information in hip-joint cartilage and labral injury identified during surgery). These models were considered supplementary since the additional predictor variables were not available at the time the models are intended to be used clinically; that is before undergoing surgery,[36] and thus merely serve as explorative analyses to scientifically understand the potential role of perioperative findings.

Source of data
Data were collected retrospectively from the Danish Hip Arthroscopy Registry (DHAR) for both development and temporal validation (steps 1C, 2B, 2C; figure 1).[37] DHAR is a national database initiated in 2012 with ongoing web-based prospective registration of hip arthroscopies performed at 11 specialised public and private hospitals/clinics, including 21 orthopaedic surgeons, in Denmark (detailed information on DHAR is provided in references[37–39]). Hip arthroscopies included in the present study were performed between 25 April 2012 and 4 October 2017 (development sample) and 5 October 2017 and 13 May 2019 (validation sample).

Participants
All participants for the development and validation models were included from DHAR database.[37 39] Inclusion criteria were as follows: Male/female who had a hip arthroscopy at the age of 15–50 years. Exclusion criteria were as follows: A previous periacetabular osteotomy; revision hip arthroscopy within 1 year (mean time to revision in DHAR: 17 months)[38]; previous hip pathology such as Perthes' disease, slipped capital femoral epiphysis and/or avascular necrosis of the femoral head; any rheumatoid disease in the hip joint such as synovial chondromatosis, incompleteness of data regarding preoperative and postoperative self-reported hip and groin function and pain (see table 1 for key characteristics related to the development and validation sample).

All included patients were treated arthroscopically for various causes of hip-related pain.[1] The DHAR contains data from several surgeons and the specific surgical techniques and indications for surgery may vary; and are not captured in DHAR.[37] Commonly, surgeries were performed under general anaesthesia in supine position using a standard two-portal technique (anterolateral and inferior mid-anterior),[10 40] with surgical procedures (eg, rim trimming, labral repair, chondral debridement and capsular closure) performed as indicated by the surgeon. Information on postoperative management is not contained in DHAR,[37] however, all patients were offered physiotherapist-led rehabilitation, either at the surgical facility or at a local community/private physical therapy centre for a period of 3–5 months.[10 40–42]

Patient and public involvement
We did not include patients or public stakeholder as part of the design and reporting of this study. We aim to include patients in the further validation of the prediction models to get insights of patient perspective in terms of use and potential implementation.

Outcomes predicted by the models
Successful and unsuccessful outcome (PASS or not)
The primary outcomes of interest to be predicted were defined a priori (step 1As and 1B; figure 1) and included patients who, at 1 year after surgery, had: (1) a successful or (2) an unsuccessful outcome. To determine a successful and unsuccessful outcome, we used previously established cut-off scores of the Patient Acceptable Symptom State (online supplemental file 1 [13]) based on the Copenhagen Hip and Groin Outcome Score (HAGOS).[43] Patients were categorised as having a successful outcome if all HAGOS subscale scores at 1 year, extracted from DHAR, surpassed the individual subscale PASS cut-off scores. In contrast, patients were categorised as having an unsuccessful outcome if none of the HAGOS subscale scores surpassed the PASS cut-off scores. This means that patients who only had achieved PASS cut-off scores for some HAGOS subscales were included as comparator group in both prediction models.

The primary endpoint of 1-year post hip arthroscopy was decided and based on previous literature indicating that patient-reported outcomes seem to plateau from 1 to 5 year postoperatively,[44] as well as 1-year outcomes being associated with
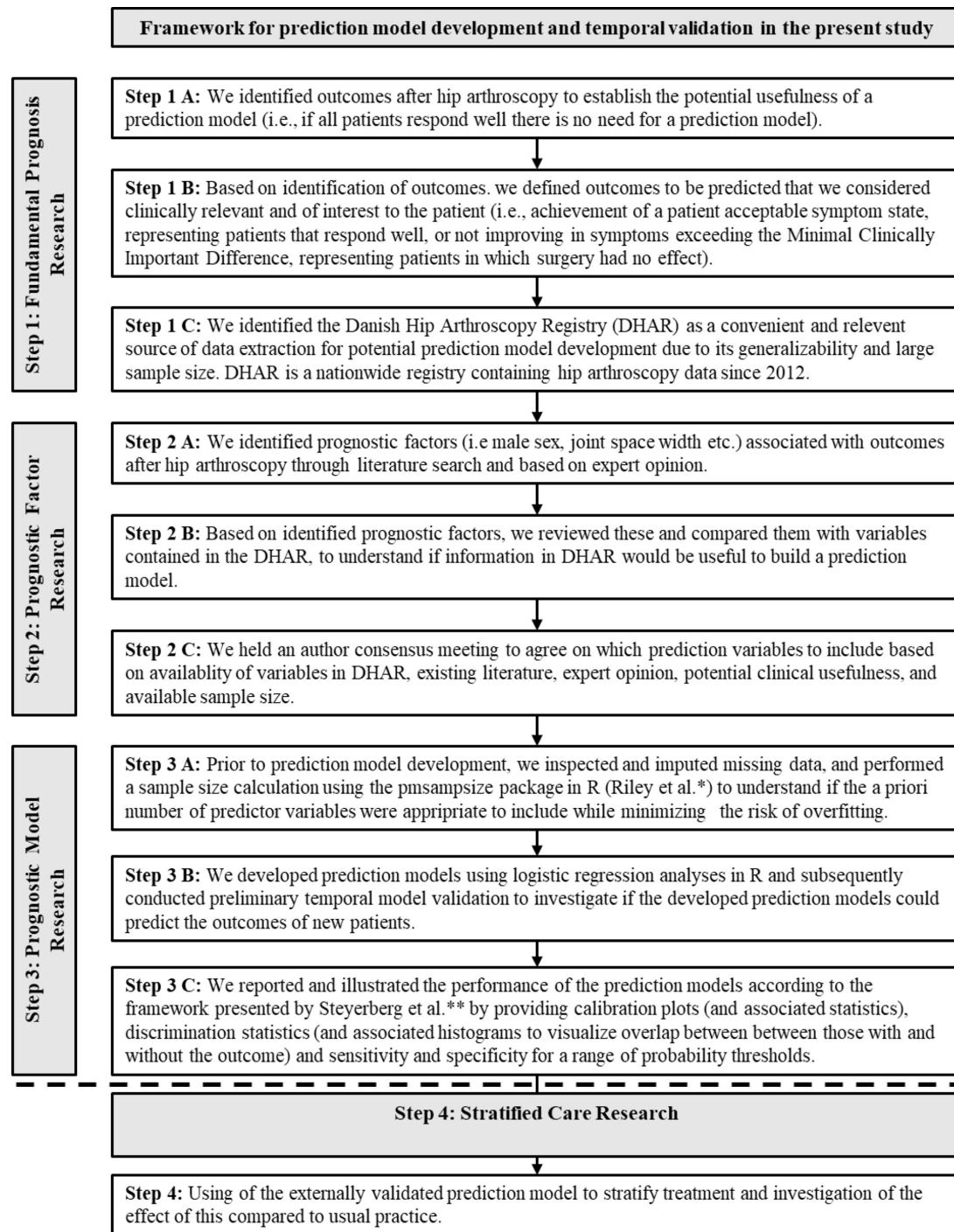
Figure 1 Study process from initial idea to prediction model development inspired by The PROGnosis RESearch Strategy (PROGRESS) Framework.[5] * and ** refers to reference [50] [55] respectively.

both revision surgery[45 46] and 5-year outcomes.[47] Definitions of outcomes were a priori agreed on among all authors of the present study.

Clinical improvement (MCID or not)

The secondary outcomes to be predicted were patients who had (1) an improvement or (2) not an improvement in self-reported hip and groin function and pain from before to 1 year after hip arthroscopy. To determine an improvement or not an improvement, we used the MCID[48] of the HAGOS questionnaire. We calculated MCID for each HAGOS subscale as 0.5 SD of the preoperative HAGOS subscale values (online supplemental file 1).[10] Patients were categorised as having an improvement if the change from preoperation to 1-year postoperation on all HAGOS subscales surpassed the MCID scores, whereas patients were categorised as not having an improvement if no change above the MCID scores in any HAGOS subscale were observed from preoperation to 1-year postoperation.

Predictor variables

All predictor variables were extracted from DHAR a priori.[37] Based on availability of predictor variables, we decided on 26

Table 1 Summary of key study characteristics for the development and temporal validation samples (n=1546)

Characteristics	Development sample (n=1082)	Temporal validation sample (n=464)
Study setting		
Data collection period	25 April 2012 to 4 October 2017	5 October 2017 to 13 May 2019
Study design	Retrospective	
Setting	Secondary care (public and private hospitals in Denmark)	
Inclusion criteria	Male/female undergoing a hip arthroscopy at the age of 15–50 years	
Demographic data		
Sex, female, no (%)	640 (59.1)	283 (61.0)
Age at surgery	34.8 (10.0)	34.6 (10.2)
Hip Sports Activity Scale†	2.5 (IQR: 1–4)	2.6 (IQR: 1–4)
Radiographic data		
Alpha Angle	68.0 (13.4)	66.2 (14.2)
Lateral Centre Edge Angle	31.6 (5.1)	29.8 (5.3)
Joint Space Width >4.0 mm, no (%)	713 (65.9)	326 (70.3)
Acetabular Index Angle	5.2 (3.8)	4.7 (4.0)
Perioperative data		
Beck classification, grade 0–1, no (%)	811 (84.3)	296 (72.4)
ICRS xlassification, grade 0–1, no (%)	150 (15.6)	67 (16.3)
Labral injury, no. (%)	994 (91.9)	428 (92.2)
Diagnostic entity based on bony morphology*		
Normal, no (%)	134 (13.7)	86 (22.4)
Isolated cam, no (%)	717 (73.4)	233 (60.7)
Cam and pincer, no (%)	64 (6.6)	9 (2.3)
Cam and dysplasia, no (%)	39 (4.0)	39 (10.2)
Isolated pincer, no (%)	16 (1.6)	7 (1.8)
Isolated dysplasia, no (%)	7 (0.7)	10 (2.6)
Preoperative self-reported hip and groin function		
HAGOS Pain	52.2 (IQR: 37.5–65.0)	50.3 (IQR: 35.0–65.0)
HAGOS Symptoms	49.8 (IQR: 35.7–64.3)	47.1 (IQR: 35.7–60.7)
HAGOS Activities of Daily Living	55.0 (IQR: 40.0–75.0)	52.6 (IQR: 35.0–70.0)
HAGOS Sport and Recreational activities	37.1 (IQR: 18.8–53.1)	35.3 (IQR: 15.6–53.1)
HAGOS Physical Activities	20.3 (IQR: 0.00–37.5)	21.1 (IQR: 0.00–37.5)
HAGOS Quality of Life	30.3 (IQR: 20.0–40.0)	29.0 (IQR: 20.0–40.0)
Overall rating of hip function‡, mean (SD)	42.40 (18.53)	39.87 (18.63)
Anxiety or depression (EQ-5D-3L Health questionnaire)		
Not depressed (%)	83.3	81.9
Moderately depressed (%)	16.3	17.0
Extremely depressed (%)	0.5	1.1

*Only based on patients with full data on alpha angle and lateral centre edge angle.
†Hip Sports Activity scale is measured on a 1–9 scale.
‡Measured on a 0–100 Visual Analogue Scale.
EQ-5D-3L, EuroQoL five-dimensional instrument; HAGOS, Copenhagen Hip And Groin Outcome Score; ICRS, International Cartilage Repair Society.

predictor variables for the primary prediction models (and 5 additional variables for the supplementary models). Selection of variables were based on previous studies regarding prognostic factors for outcomes after hip arthroscopy[17] combined with consensus among the authors (hip arthroscopy surgeon (n=1), physiotherapist (n=5), steps 2A–C, figure 1). This was done by listing all the potential predictor variables contained in DHAR including items from HAGOS, and subsequently relating them to existing literature on risk factors for a poor or good outcome (table 2) combined with clinical experience of the authors (LI, 5 years; KT, 23 years; JLK, 28 years; MPR, 30 years; MFN, 3 years, PH, 40 years). A full list of predictor variables and reasons for selection is presented in table 2. Preoperative radiographies were assessed by the operating surgeon and included Lateral Centre Edge Angle, Ischial Spine Sign, Alpha Angle, Joint Space Width and Acetabular Index Angle as these represent common radiological measures to

determine femoral head-neck and acetabular morphology.[2 49] For a description of each measure, we refer to online supplemental file 2. Preoperative self-reported variables related to hip function, pain severity, psychosocial state was obtained using patient-reported outcome measures (table 2). We prioritised to include specific items as predictors rather than composite scores, as single items can represent specific constructs and be easily implemented in the history-taking process (table 2). Finally, perioperative findings of cartilage and labral injury were assessed during hip arthroscopy (online supplemental file 2), but these variables were only included in the supplementary prediction models.

Sample size considerations

An a priori sample size calculation was not performed as the sample size was determined by eligible patients in DHAR (n=1546). However, to minimise the risk of overfitting and

Table 2 Overview of a priori defined predictor variables included in the prediction models

Predictor variables	Scale	Reasons for selection of predictor variable
Demography		
Age	Continuous (years)	Younger age is associated with improved self-reported outcome and lower revision rates.[17]
Sex	Dichotomous	Male sex is associated with improved self-reported outcome.[17]
Hip Sports Activity Scale	Ordinal (9-point scale)	Sports participation reflects overall hip function, which is associated with self-reported outcomes.[17]
Context		
Hospital setting	Dichotomous (private vs public)	Patients in a private setting seems to have better preoperative symptoms, which may reflect a specific subgroup of patients.[65]
Preoperative radiography		
Lateral Centre Edge Angle	Continuous (angle)	Higher angle is associated with lower failure rates.[17]
Ischial Spine Sign	Dichotomous	Acetabular version is associated with outcome.[17]
Alpha Angle	Continuous (angle)	Larger cam morphology is associated with revision surgery[17] and severe acetabular cartilage injuries.[66]
Joint Space Width	Ordinal (5-point scale)	Narrow joint space width is associated with severe cartilage injury,[66] worse self-reported outcomes and conversion to total hip replacement,[17]
Acetabular Index Angle	Continuous (angle)	Less than 3 degrees is associated with revision surgery.[17]
Preoperative self-reported hip function		
Overall rating of hip function	Continuous (0–100 VAS)	Better overall hip function preoperatively is generally associated with better postoperative outcome.[17]
Problems during running*	Ordinal (5-point scale)	Activities that reflects overall hip function and load bearing capacity,[67] and thus may be associated with self-reported outcome.[17] The ability to walk, run, participate in sport and get in or out of a car represent everyday activities which is often part of the history-taking process.
Problems during walking*	Ordinal (5-point scale)	
Problems get in/out of car*	Ordinal (5-point scale)	
Sports participation*	Ordinal (5-point scale)	
Preoperative self-reported pain severity		
Pain frequency*	Ordinal (5-point scale)	Having less pain preoperatively, indicative of better preoperative status may be associated with better postoperative outcome.[17] Pain characteristics, such as stabbing and stiffness, as well as pain intensity during specific activities is considered important for the diagnosis of hip pain[5] and is often part of the history-taking process.
Pain in other areas*	Ordinal (5-point scale)	
Stabbing sensation*	Ordinal (5-point scale)	
Morning stiffness*	Ordinal (5-point scale)	
Stiffness after sitting*	Ordinal (5-point scale)	
Night pain*	Ordinal (5-point scale)	
Pain during rest	Continuous (0–100 NRS)	
Pain during walking	Continuous (0–100 NRS)	
Preoperative self-reported psychosocial factors		
Anxiety or depression†	Ordinal (3-point scale)	Preoperative mental status and depression state are associated with worse self-reported outcomes.[17]
Awareness of hip*	Ordinal (5-point scale)	
Lifestyle changes*	Ordinal (5-point scale)	
Mood changes*	Ordinal (5-point scale)	
Perioperative findings‡		
ICRS Femoral head	Ordinal (5-point grading)	Degeneration of intra-articular structures, such as severe cartilage and/or labral injury, is associated with worse self-reported outcomes and conversion to total hip replacement.[17]
Femoral size lesion	Ordinal (4-point grading)	
Beck acetabulum	Ordinal (5-point grading)	
Acetabulum size lesion	Ordinal (4-point grading)	
Labral injury	Dichotomous	

*Represent single Items from the Copenhagen Hip And Groin Outcome Score. Items are scored on a 5-point Likert scale ranging from extreme problems/pain to no problems/pain.[43]

†Represent the anxiety and depression Item from the EQ-5D-3L Health questionnaire, which is scored on a 3-point Likert scale ranging from no anxiety/depression to extreme anxiety/depression.[68]

‡Predictor variables representing intra-articular findings identified during hip arthroscopy. These variables are only included in the supplementary prediction models.

ICRS, International Cartilage Repair Society; VAS, Visual Analogue Scale.

ensure precise estimations, we performed the four-step sample size calculation approach suggested by Riley et al[50] using the 'pmsampsize' (V.1.1.0) package in R. This helped us to identify if the number of a priori defined predictor variables were reasonable to include in the development of the models before overfitting becomes a concern.[50] With an outcome proportion of 0.3 for the primary outcome measures, 26 predictor variables, an expected shrinkage factor of ≤10 % and a C-statistics of 0.78 (estimated R^2: 0.20) based on previous models,[25 26 29 30] 1043 patients were deemed adequate for model development, corresponding to 313 events and 12.03 events per predictor; we included 1082 patients in the development sample.[50] The remaining 464 patients were used in the temporal external validation sample, which secured at least 100 events as recommended for the primary outcome measures.[51] However, larger sample sizes may be needed for precise estimates of calibration.[52]

Missing data

Missing data for predictor variables were imputed by multiple imputation with chain equations on both development and validation sample with 20 imputations. Two radiological variables (alpha angle and acetabular index angle) had ∼10% missing data (online supplemental file 1). Imputations models were based on all available data from the 26 predictor variables
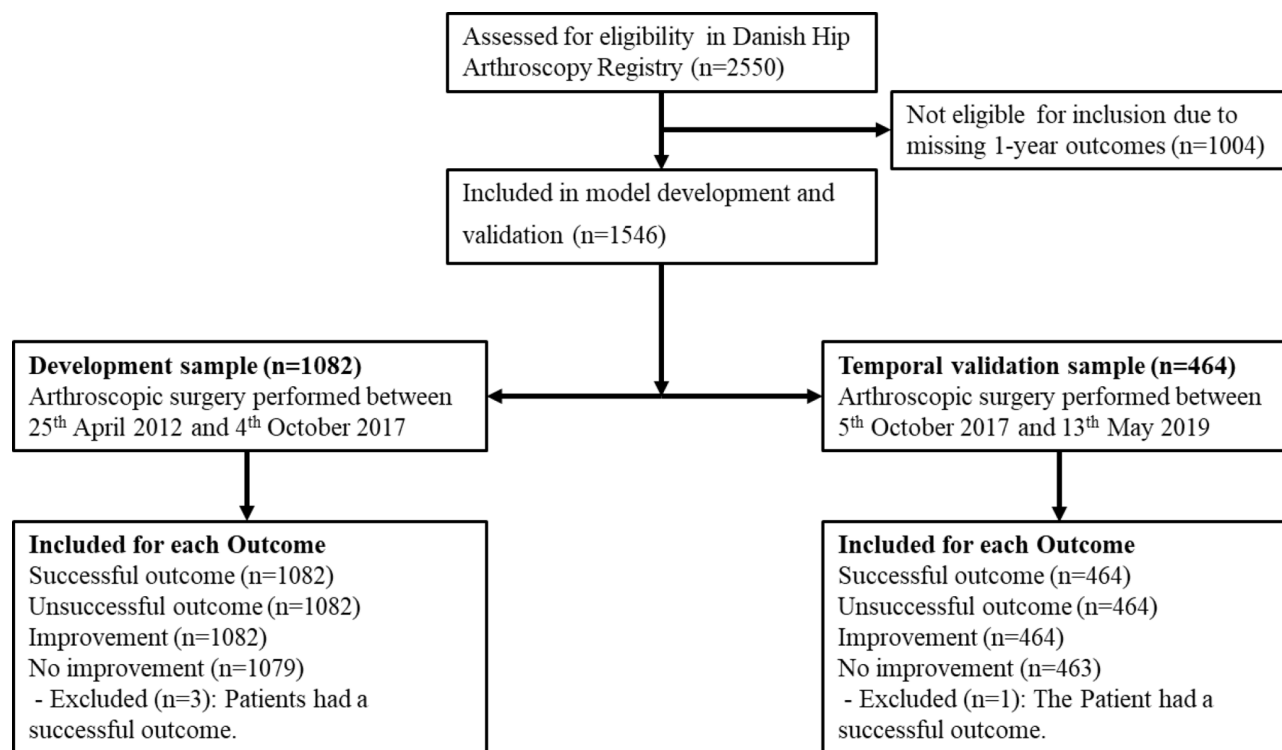
Figure 2 Flow of patients.

and outcome variables. Continuous variables were imputed by predictive mean matching and categorical variables by polytomous logistic regression. Prediction models were fitted by both imputed data as well as complete case to evaluate impact of the missing values.[34]

Statistical methods

Development and temporal validation of prediction models were analysed using logistic regression models including all 26 prediction variables as single term with no interactions to minimise risk of overfitting. Linearity of continuous variables was evaluated by comparing models with single term against restricted cubic spline terms, using 1–10 knots, for each variable. Models were compared using Akaike information criterion (AIC) and likelihood-ratio test. For most variables, spline models were not significantly different from single term models and in these cases the reduction in AIC was less than 0.5%. Because of this all variables are included as single term to reduce complicity and risk of overfitting, as the impact of possible violation of linearity is likely very small.[34] The supplementary prediction models included five additional predictors related to perioperative findings (table 2). We chose a logistic regression model approach over machine learning, although machine learning is popular in hip arthroscopy research,[23 25 28–30] a recent systematic review found similar predictive performance between the two approaches for clinical prediction models.[53] In addition, logistic regression requires far less events per variable compared with machine learning strategies.[54] All continuous variables were kept continuous[34] and ordinal scales were treated as continuous, except JSW and HSA (table 2). Uniform shrinkage by bootstrapping, with 1000 replications, were applied to regression coefficients.[34] All analyses were performed in R (R Foundation for Statistical Computing, Vienna, Austria, V.3.6.3).

External temporal validation

To evaluate the performance of the prediction models on the temporal external validation set, we obtained the predicted probability for each patient in the validation data set using the intercept and regression coefficients derived from the development data. Model performance was investigated in line with the TRIPOD recommendations[34] using the framework presented by Steyerberg et al.[55] We report the explained variance (Nagelkerke R^2), calibration plots (and associated statistics) and[56] discrimination statistics (area under the receiver operating characteristics curve, AUC).[57] In addition, we report histograms to visualise the distribution of predicted probability between patients with and without the outcome[55] and sensitivity and specificity for a range of probability thresholds.[57]

Calibration refers to the agreement between observed outcomes and outcome predictions, and thus is a measure of the model's ability to provide unbiased estimates.[34] We assessed calibration as defined by Van Calster et al[56] as: (1) mean calibration (calibration-in-the-large) reflecting if the observed outcome rate equals the average predicted risk, (2) weak calibration reflecting if the model, on average, overestimates or underestimates the risk assessed by calibration intercept and slope, with a target value of 0 and 1, respectively, and (3) moderate calibration, reflecting if the estimated risks corresponds to the observed proportions, assessed graphically using a calibration plot, with the target being a smoothed calibration curve lying closely around the 45° line.[56] Calibration plots and associated parameters were produced using 'val.prob.ci.2' package in R.[58] Discrimination was assessed using AUC (c-statistics), which quantifies the model's discriminative ability, that is the probability that the model estimates higher risks for patients with the outcome than patients without the outcome.[57] AUC ranges from 0.5 to 1, representing no and perfect discriminative ability, respectively.[57] All validation plots and values were generated for each imputed dataset and presented as ranges of values.
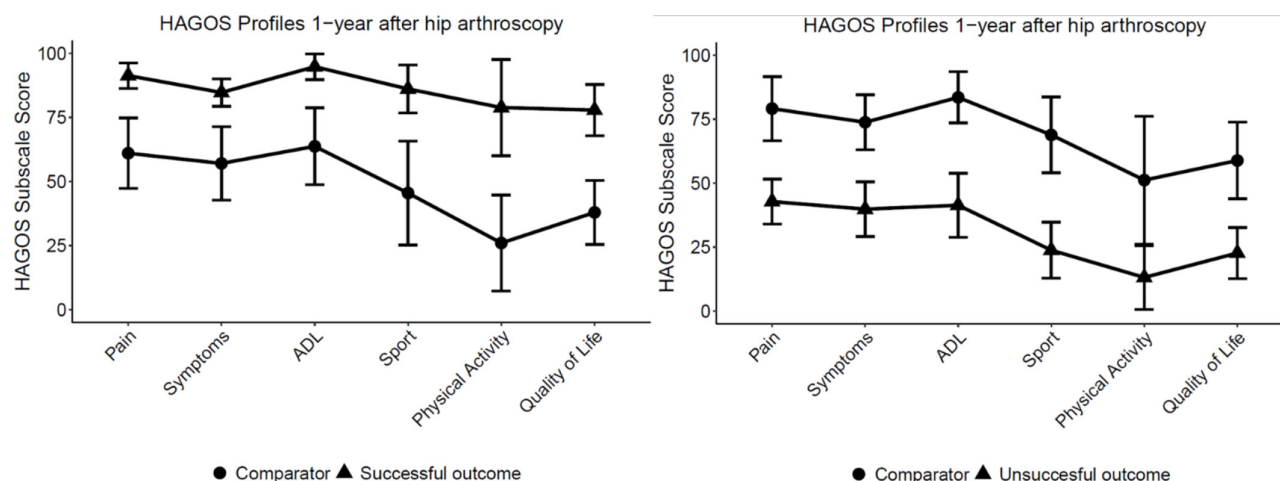
Figure 3 Self-reported hip and groin pain and function measured using the Copenhagen Hip and Groin Outcome Score (HAGOS) in patients with a successful outcome defined as having a Patients Acceptable Symptom State (PASS) in all HAGOS subscales versus in some/no subscales (left figure), and patients with an unsuccessful outcome defined as having PASS in no HAGOS subscales versus in some/all subscales (right figure). Error bars show IQR. ADL, activities of daily living.

RESULTS

Of 2550 eligible patients, we included 1546 patients with complete outcome data at 1-year follow-up (figure 2). In general, very small differences were observed between included and patients with missing outcome data for demographics, radiology, operative findings and preoperative symptoms (online supplemental file 1).

Participants

In total, 1082 patients were used for developing the models, whereas 464 patients were used for validation, with samples being comparable in terms of demographics, radiology, operative findings, preoperative symptoms and outcomes (table 1; see online supplemental file 1 for a summary of the distribution of predictor variables in the development and validation sample). Since missing data in predictor variables were imputed, all patients with complete HAGOS at baseline and 1-year follow-up were included. The proportion of events were similar between the development and validation samples; successful outcome (development: 339 events (31.3%), validation: 137 events (29.5%)), unsuccessful outcome (development: 294 events (27.2%), validation: 117 events (25.2%)), improvement (development: 333 events (30.8%), validation: 161 events (34.7%)) and no improvement (development: 140 events (13.0%), validation: 51 events (11.0%)). Clear differences were found between groups in postoperative HAGOS scores and change in HAGOS score from presurgery to postsurgery (figures 3 and 4; similar findings were observed in the development sample; online supplemental file 4).

Model specification and performance

For the development models, calibration plots and associated statistics are presented in online supplemental file 3. The validation models, the best model performance was found for the primary outcome measure, an unsuccessful outcome (Nagelkerke R^2 range: 0.25–0.26), which also showed adequate calibration (predicted mean probability vs actual mean probability: 27.0% vs 25.2%; intercept range: −0.10 to −0.11; slope range:
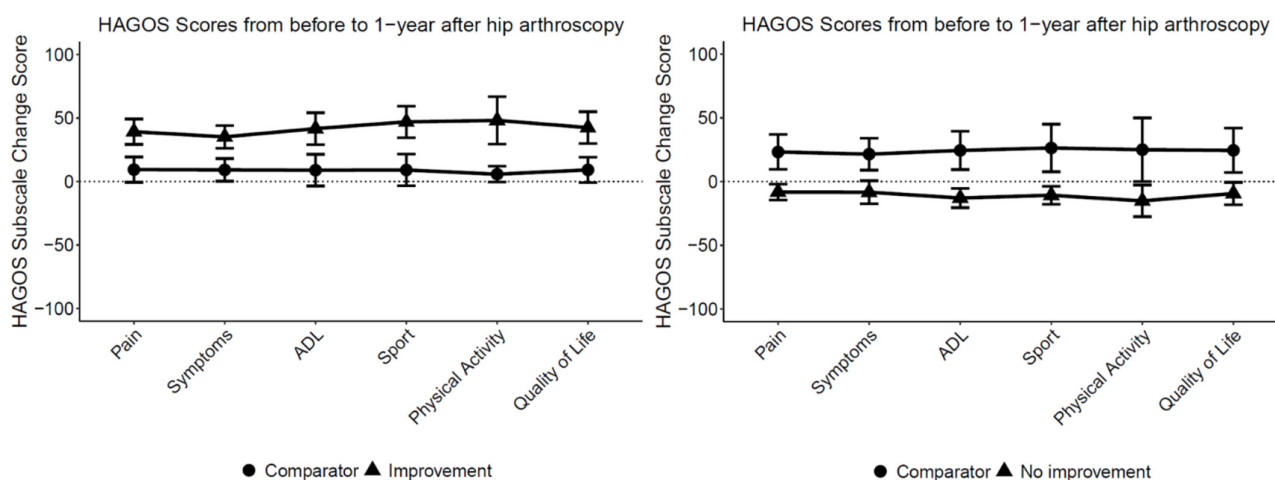
Figure 4 Changes in self-reported hip and groin pain and function measured using the Copenhagen Hip and Groin Outcome Score (HAGOS) in patients who have achieved an improvement defined as exceeding the Minimal Clinically Important Difference (MCID) in all HAGOS subscales versus in some/no subscales (left figure), and patients who have not achieved an improvement defined as not exceeding MCID in any HAGOS subscale versus in some/all subscales (right figure). Error bars show IQR. ADL, activities of daily living.
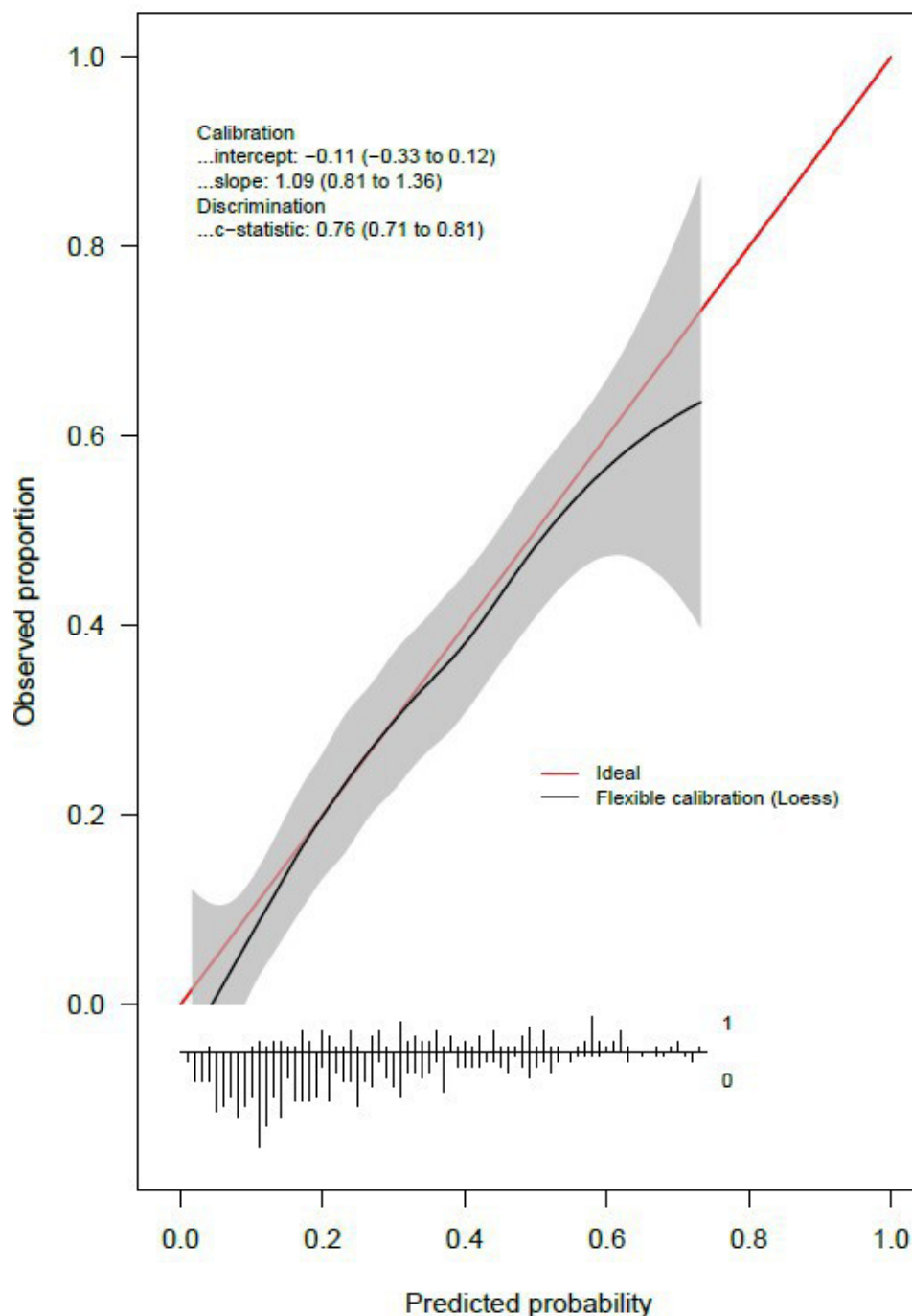
Figure 5 Calibration plot for predicting patients who have achieved an unsuccessful outcome (PASS in no HAGOS subscale)Shaded area in calibration plots depicts 95% CIs. HAGOS; Copenhagen Hip and Groin Outcome Score.

1.06 to 1.09 and discrimination (AUC range: 0.76 to 0.77) (see figure 5 for a representative calibration plot and online supplemental file 8 for all calibration plots derived from the multiple imputations).

The model for successful outcomes showed poor calibration and discrimination. A complete summary of model performance for all four models are available in online supplemental file 1); while sensitivity and specificity for probability thresholds (from 0.1 to 0.9) are presented in online supplemental file 5.

For usage of the prediction models, the full models with estimates are resented in online supplemental file 6 while an excel calculator is provided online https://bit.ly/3avOcjJ. The complete-case analyses showed similar model performance

for all outcomes. For the supplementary models, the addition of perioperative findings (information on cartilage and labrum injuries) did not improve model performance.

DISCUSSION

The present study is the first to develop and externally temporal validate clinical prediction models to identify those hip arthroscopy patients who at 1 year after surgery can be considered having a successful (having achieved PASS) or unsuccessful (not having achieved PASS) outcome. Our findings indicate that by using 26 common clinical variables, including demographics, radiographic parameters of hip morphology and self-reported

measures, the probability of patients with an unsuccessful outcome (1-year mean HAGOS Subscales scores ranging 13–43 points; figure 2) can be predicted with acceptable discrimination and adequate calibration. The calibration, however, becomes imprecise towards higher predicted probability due to few events (figure 3). The prediction model for successful outcomes showed less accuracy, and thus is unlikely to impact clinical practice.

The present study extends existing knowledge regarding prediction modelling for hip arthroscopy. Although several models have been published, these are associated with important methodological shortcomings, which may result in too optimistic and/or unstable predictive performance.[24–26 28–31] First, only one of eight existing prediction models has been attempted externally validated,[24] however, this was only based on 13 patients with the outcome of interest (a minimum of 100 events are recommended for external validation).[51 52] Since prediction models show best performance on the development sample, external validation is needed to adjust initial optimism and improve application to future patients.[34] In the present study, this is illustrated by C-statistics for all models being lower in the validation sample than the development sample. Compared with the present study, no sample size consideration has been made in any previous study, resulting in events per predictor ranging between 3 and 8.[25 26 28–30] While this may not seem very different from the present study (events per predictor for the primary outcome: 11–13), the majority of published prediction models has been developed using machine learning strategies,[25 28–30] which require >200 events per predictor before low optimism and stable performance measures are reached.[54] Thus, the existing previous prediction models for hip arthroscopy patients are associated with a high risk of overfitting, and thus potentially unreliable predictions when applied on future patients.[50]

Clinical usefulness of the prediction models
The present study suggests that the probability of having an unsuccessful outcome, defined as not having PASS in any of the HAGOS subscales, can be predicted. While hip arthroscopy is considered an effective procedure for treatment of hip joint-related pain,[8] up to 50% of patients do not achieve an acceptable symptom state at 1–2 years follow-up[13] highlighting the clinical relevance of identifying patients for whom surgery may not be helpful. The proportion of patients with residual symptoms may thereby decrease and the overall outcome of hip arthroscopy improve. Thus, the prediction model is an initial step towards stratified care for patients with hip-related pain.[59] However, before clinical adoption and stratified care is implemented (step 4 of the PROGRESS Framework),[59] the model should be externally validated in a true external data set, while the effectiveness of the model is tested in a randomised controlled trial with stakeholder involvement.

How should the prediction models be used
The prediction model can support clinical evaluation and shared decision making by informing the orthopaedic surgeon and the individual patient about the risk of an unsuccessful outcome. In practice, the probability is derived using the prediction formula (presented in online supplemental file 6), and available as an Excel calculator (https://bit.ly/3avOcjJ) for illustrative purposes, which combines the ORs for all 26 predictors into a single probability from 0% to 100% (An example can also be found in table 3). It is important to state that single predictors, although statistically significant, should not be used in isolation, as the performance of the prediction model relies on all predictors regardless of p values for individual predictors. Since the prediction model is developed

Table 3 Overview of predictive variables for two patients with either a low or high probability of having an unsuccessful outcome according to the prediction model

Characteristics	Low probability (1.4%) for an unsuccessful outcome	High probability (56.5%) for an unsuccessful outcome
Demographic data		
Sex	Male	Male
Age at surgery	21	37
Hip Sports Activity Scale*	Level 8 (Elite level)	Level one sport (Recreational)
Radiographic data		
Alpha Angle	71	78
Lateral Centre Edge Angle	31	29
Ischial Spine Sign	Yes	Yes
Joint Space Width	>4.0 mm	3.1–4.0 mm
Acetabular Index Angle	0	2
Preoperative self-reported hip function		
Overall rating of hip function	31	39
Problems during running‡	Extreme	Extreme
Problems during walking‡	Moderate	Moderate
Problems get in/out of car‡	Severe	Extreme
Sports participation‡	Sometimes	Never
Preoperative self-reported pain severity		
Pain frequency‡	Weekly	Always
Pain in other areas‡	Never	Always
Stabbing sensation‡	Never	All the time
Morning stiffness‡	None	Severe
Stiffness after sitting‡	None	Severe
Night pain‡	None	Extreme
Pain during rest	44	66
Pain during walking	23	84
Preoperative self-reported psychosocial factors		
Anxiety or depression§	Not depressed	Not depressed
Awareness of hip‡	Weekly	Always
Lifestyle changes‡	Severely	Totally
Mood changes‡	Sometimes	Sometimes
Perioperative findings¶		
ICRS femoral head	Grade 0	Grade 0
Femoral size lesion	No lesion	No leasion
Beck acetabulum	Grade 2	Grade 3
Acetabulum size lesion	1–2 cm²	<1 cm²
Labral injury	Yes	No
1-year HAGOS Scores		
Pain	90	40
Symptoms	78.5	32
Activities of daily living	100	50
Sport and recreational activities	100	25
Physical activities	75	25
Quality of life	85	25

*Hip Sports Activity scale is measured on a 1–9 scale.
†Measured on a 0–100 Visual Analogue Scale.
‡Represent single Items from the Copenhagen HAGOS. Items are scored on a 5-point Likert scale ranging from extreme problems/pain to no problems/pain.[43]
§Represent the anxiety and depression Item from the EQ-5D-3L Health questionnaire, which is scored on a 3-point Likert scale ranging from no anxiety/depression to extreme anxiety/depression.[68]
¶Predictor variables representing intra-articular findings identified during hip arthroscopy. These variables are not included in the calculation of probability.
HAGOS, Hip And Groin Outcome Score; ICRS, International Cartilage Repair Society.

and validated on patients who underwent surgery, the prediction model is best used once the orthopaedic surgeon has decided for surgery. In such instances, the model could be used as a data-driven 'second opinion' to estimate the risk of an unsuccessful outcome, and understand when surgery may not provide enough benefit to proceed. In clinical practice, this means that the prediction model is suited to be used in the final stages of a stepped-care approach[60] starting with targeted exercise-based treatment followed by potential surgery if symptoms have not resolved.[2 61] If used for dichotomous decisions in clinical practice (surgery vs no surgery), we advise that the predicted probability is combined with the sensitivity and specificity measures presented in online supplemental file 5, to understand the false positive and negative rates of the specific probability threshold, that is, misclassification of patients.

Limitations

The present study is associated with some limitations. First, we appreciate that dichotomisation of a continuous outcome is generally not recommended due to loss of information[57]; but since HAGOS contains six subscales that cannot be aggregated to a single score,[36] we chose the current approach to avoid dealing with six different predictions models (one for each subscale), that would complicate the clinical utility. We believe that patients who have exceeded the cut-off scores of all HAGOS subscales at 1-year follow-up are likely to represent a subgroup of patients that feel very well after surgery (a successful outcome) and viceversa for patients who do not surpass a single subscale score (an unsuccessful outcome).[32] Second, since the prediction models were developed based on data from the DHAR, predictor variables were limited to those contained in the registry.[37] However, these were included based on their potential association with hip arthroscopy outcomes[17] and represent common, currently used, and easily collectable clinical variables, although we cannot exclude the potential added value of additional variables. Third, although we included at least 100 events in the external validation models for the primary outcome based on rule-of-thumb,[51] we appreciate that this rule may be imprecise.[52] Based on the reviewer response, we were made aware that simulations of CI for C-statistic and slope estimate for calibration curves can be used to estimate the number of events in the validation sample, with R-code for simulations available at https://github.com/gscollins1973/External-validation-sample-size. This was done post hoc with 500 repetitions, an SD of the linear predictor from the model of 0.8 (based on the coefficients from the prediction model fitted by the development data) and probability of event set to 30% (successful outcome, unsuccessful outcome and improvement) and 12% (no improvement). Curves for number of events ranging from 50 to 1000 can be found in online supplemental file 7. Additionally, estimations for sample size calculation were based on the number of variables used in the model not the number of parameters (number of levels minus 1). Because the Hip Sports Activity Scale and Joint Space Width were used as 4 and 3 level variables, respectively, this contributes 3 and 2 levels rather than 1 level each. All other variables were either 2 levels or continues variables each contributing 1 level. Sample size should, therefore, have been based on 29 parameters instead of 26 variables. Fourth, we appreciate that model development and validation was performed using all hospitals combined and that site-specific differences may exist that could impact on the predictive performance when applied in a specific setting. Therefore, further external validation is needed to confirm the present findings at each site. Fifth, like many other registries DHAR contains missing data on postoperative outcome, and these patients were excluded from the present study; thus, we cannot exclude that the study sample represent a selected cohort of patients. Finally, while we have no specific information on the postoperative rehabilitation received, we acknowledge that this is considered an integral part of the hip arthroscopy procedure[61] with potential to affect postoperative outcomes,[62–64] and thus the predictive performance.

CONCLUSION

Common clinical variables including demographics, radiographic parameters of hip morphology and self-reported measures were able to predict the probability of having an unsuccessful outcome 1 year after hip arthroscopy. This temporal externally validated prediction model can be used to support clinical evaluation and shared decision making by informing the orthopaedic surgeon and patient about the risk of an unsuccessful outcome, and thus when surgery may not be appropriate. This may reduce unsuccessful outcomes and could therefore potentially improve the overall outcome of hip arthroscopy in the future. Patients with a successful outcome (achieving the patients acceptable symptom state) was less accurately predicted.

Twitter Lasse Ishøi @LasseIshoei, Kristian Thorborg @KThorborg, Joanne L Kemp @JoanneLKemp, Michael P Reiman @MikeReiman, Mathias Fabricius Nielsen @Physiomathias and Per Hölmich @PerHolmich

ORCID iDs
Lasse Ishøi http://orcid.org/0000-0002-2716-6567
Kristian Thorborg http://orcid.org/0000-0001-9102-4515
Joanne L Kemp http://orcid.org/0000-0002-9234-1923
Michael P Reiman http://orcid.org/0000-0003-4557-3446
Mathias Fabricius Nielsen http://orcid.org/0000-0003-4211-3516
Per Hölmich http://orcid.org/0000-0003-2098-0272

REFERENCES

1 Reiman MP, Agricola R, Kemp JL, et al. Consensus recommendations on the classification, definition and diagnostic criteria of hip-related pain in young and middle-aged active adults from the international hip-related pain research network, zurich 2018. *Br J Sports Med* 2020;54:631–41.

2 Griffin DR, Dickenson EJ, O'Donnell J, et al. The warwick agreement on femoroacetabular impingement syndrome (FAI syndrome): an international consensus statement. *Br J Sports Med* 2016;50:1169–76.

3 Hale RF, Melugin HP, Zhou J, et al. Incidence of femoroacetabular impingement and surgical management trends over time. *Am J Sports Med* 2021;49:35–41.

4   Ganz R, Parvizi J, Beck M, *et al*. Femoroacetabular impingement: a cause for osteoarthritis of the hip. *Clin Orthop Relat Res* 2003;112–20.

5   Reiman MP, Thorborg K, Covington K, *et al*. Important clinical descriptors to include in the examination and assessment of patients with femoroacetabular impingement syndrome: an international and multi-disciplinary delphi survey. *Knee Surg Sports Traumatol Arthrosc* 2017;25:1975–86.

6   Reiman MP, Thorborg K, Hölmich P. Femoroacetabular impingement surgery is on the rise-but what is the next step? *J Orthop Sports Phys Ther* 2016;46:406–8.

7   Kierkegaard S, Langeskov-Christensen M, Lund B, *et al*. Pain, activities of daily living and sport function at different time points after hip arthroscopy in patients with femoroacetabular impingement: a systematic review with meta-analysis. *Br J Sports Med* 2017;51:572–9.

8   Casartelli NC, Valenzuela PL, Maffiuletti NA, *et al*. Effectiveness of hip arthroscopy on treatment of femoroacetabular impingement syndrome: a meta-analysis of randomized controlled trials. *Arthritis Care Res (Hoboken)* 2021;73:1140–5.

9   Kyin C, Maldonado DR, Go CC, *et al*. Mid- to long-term outcomes of hip arthroscopy: A systematic review. *Arthroscopy* 2021;37:S0749-8063(20)30828-8:1011–25.:.

10 Thorborg K, Kraemer O, Madsen A-D, *et al*. Patient-reported outcomes within the first year after hip arthroscopy and rehabilitation for femoroacetabular impingement and/or labral injury: the difference between getting better and getting back to normal. *Am J Sports Med* 2018;46:2607–14.

11 Palmer AJR, Ayyar Gupta V, Fernquest S, *et al*. Arthroscopic hip surgery compared with physiotherapy and activity modification for the treatment of symptomatic femoroacetabular impingement: multicentre randomised controlled trial. *BMJ* 2019;364:l185.

12 Jones DM, Crossley KM, Ackerman IN, *et al*. Physical activity following hip arthroscopy in young and middle-aged adults: A systematic review. *Sports Med Open* 2020;6:7.

13 Ishøi L, Thorborg K, Ørum MG, *et al*. How many patients achieve an acceptable symptom state after hip arthroscopy for femoroacetabular impingement syndrome? A cross-sectional study including PASS cutoff values for the HAGOS and ihot-33. *Orthop J Sports Med* 2021;9:2325967121995267.

14 Ishøi L, Thorborg K, Kraemer O, *et al*. Return to sport and performance after hip arthroscopy for femoroacetabular impingement in 18- to 30-year-old athletes: A cross-sectional cohort study of 189 athletes. *Am J Sports Med* 2018;46:2578–87.

15 Wörner T, Thorborg K, Stålman A, *et al*. High or low return to sport rates following hip arthroscopy is a matter of definition? *Br J Sports Med* 2018;52:1475–6.

16 Reiman MP, Peters S, Sylvain J, *et al*. Femoroacetabular impingement surgery allows 74% of athletes to return to the same competitive level of sports participation but their level of performance remains unreported: a systematic review with meta-analysis. *Br J Sports Med* 2018;52:972–81.

17 Sogbein OA, Shah A, Kay J, *et al*. Predictors of outcomes after hip arthroscopic surgery for femoroacetabular impingement: A systematic review. *Orthop J Sports Med* 2019;7:2325967119848982.

18 McQuivey KS, Secretov E, Domb BG, *et al*. A multicenter study of radiographic measures predicting failure of arthroscopy in borderline hip dysplasia: beware of the Tönnis angle. *Am J Sports Med* 2020;48:1608–15.

19 Kuroda Y, Saito M, Çınar EN, *et al*. Patient-related risk factors associated with less favourable outcomes following hip arthroscopy. *Bone Joint J* 2020;102-B:822–31.

20 Steyerberg EW, Moons KGM, van der Windt DA, *et al*. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381.

21 Domb BG, Chen SL, Go CC, *et al*. Predictors of clinical outcomes after hip arthroscopy: 5-year follow-up analysis of 1038 patients. *Am J Sports Med* 2021;49:112–20.

22 Riley RD, Hayden JA, Steyerberg EW, *et al*. Prognosis research strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10:e1001380.

23 Ramkumar PN, Karnuta JM, Haeberle HS, *et al*. Radiographic indices are not predictive of clinical outcomes among 1735 patients indicated for hip arthroscopic surgery: A machine learning analysis. *Am J Sports Med* 2020;48:2910–8.

24 Hevesi M, Leland DP, Rosinsky PJ, *et al*. Risk of conversion to arthroplasty after hip arthroscopy: validation of a published risk score using an independent, prospectively collected database. *Am J Sports Med* 2021;49:1192–8.

25 Nwachukwu BU, Beck EC, Lee EK, *et al*. Application of machine learning for predicting clinically meaningful outcome after arthroscopic femoroacetabular impingement surgery. *Am J Sports Med* 2020;48:415–23.

26 Stephan P, Röling MA, Mathijssen NMC, *et al*. Developing a risk prediction model for the functional outcome after hip arthroscopy. *BMC Musculoskelet Disord* 2018;19:122.

27 Redmond JM, Gupta A, Dunne K, *et al*. What factors predict conversion to THA after arthroscopy? *Clin Orthop Relat Res* 2017;475:2538–45.

28 Kunze KN, Polce EM, Clapp I, *et al*. Machine learning algorithms predict functional improvement after hip arthroscopy for femoroacetabular impingement syndrome in athletes. *Journal of Bone and Joint Surgery* 2021;103:1055–62.

29 Kunze KN, Polce EM, Nwachukwu BU, *et al*. Development and internal validation of supervised machine learning algorithms for predicting clinically significant functional improvement in a mixed population of primary hip arthroscopy. *Arthroscopy* 2021;37:S0749-8063(21)00016-5:1488–97.:.

30 Kunze KN, Polce EM, Rasio J, *et al*. Machine learning algorithms predict clinically significant improvements in satisfaction after hip arthroscopy. *Arthroscopy* 2021;37:S0749-8063(20)30976-2:1143–51.:.

31 Haeberle HS, Ramkumar PN, Karnuta JM, *et al*. Predicting the risk of subsequent hip surgery before primary hip arthroscopy for femoroacetabular impingement syndrome: A machine learning analysis of preoperative risk factors in hip preservation. *Am J Sports Med* 2021;49:2668–76.

32 Tubach F, Dougados M, Falissard B, *et al*. Feeling good rather than feeling better matters more to patients. *Arthritis Rheum* 2006;55:526–30.

33 Hemingway H, Croft P, Perel P, *et al*. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* 2013;346:e5595.

34 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.

35 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.

36 Wolff RF, Moons KGM, Riley RD, *et al*. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170:51–8.

37 Mygind-Klavsen B, Grønbech Nielsen T, Maagaard N, *et al*. Danish hip arthroscopy registry: an epidemiologic and perioperative description of the first 2000 procedures. *J Hip Preserv Surg* 2016;3:138–45.

38 Mygind-Klavsen B, Nielsen TG, Lund B, *et al*. Clinical outcomes after revision hip arthroscopy in patients with femoroacetabular impingement syndrome (FAIS) are inferior compared to primary procedures. results from the danish hip arthroscopy registry (DHAR). *Knee Surg Sports Traumatol Arthrosc* 2021;29:1340–8.

39 Poulsen E, Lund B, Roos EM. The danish hip arthroscopy registry: registration completeness and patient characteristics between responders and non-responders. *Clin Epidemiol* 2020;12:825–33.

40 Kierkegaard S, Mechlenburg I, Lund B, *et al*. Is hip muscle strength normalised in patients with femoroacetabular impingement syndrome one year after surgery? *Journal of Science and Medicine in Sport* 2019;22:413–9.

41 Wahoff M, Ryan M. Rehabilitation after hip femoroacetabular impingement arthroscopy. *Clin Sports Med* 2011;30:463–82.

42 Wörner T, Thorborg K, Moksnes H, *et al*. Similar views on rehabilitation following hip arthroscopy among physiotherapists and surgeons in scandinavia: a specialized care survey. *Knee Surg Sports Traumatol Arthrosc* 2018;26:2519–26.

43 Thorborg K, Hölmich P, Christensen R, *et al*. The copenhagen hip and groin outcome score (HAGOS): development and validation according to the COSMIN checklist. *Br J Sports Med* 2011;45:478–91.

44 Mygind-Klavsen B, Kraemer O, Hölmich P, *et al*. An updated description of more than 5,000 procedures from the danish hip arthroscopy registry. *J Bone Joint Surg Am* 2020;102(Suppl 2):43–50.

45 Shapira J, Bheem R, Kyin C, *et al*. Can patient-reported outcomes predict the need for secondary surgeries after hip arthroscopy? *Am J Sports Med* 2021;49:97–103.

46 Rosinsky PJ, Chen JW, Yelton MJ, *et al*. Does failure to meet threshold scores for mhhs and ihot-12 correlate to secondary operations following hip arthroscopy? *J Hip Preserv Surg* 2020;7:272–80.

47 Akpinar B, Lin LJ, Bloom DA, *et al*. Hip arthroscopy for femoroacetabular impingement: 1-year outcomes predict 5-year outcomes. *Am J Sports Med* 2021;49:104–11.

48 Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean? *Ann Rheum Dis* 2007;66 Suppl 3(Suppl 3):iii40–1.

49 Mascarenhas VV, Castro MO, Rego PA, *et al*. The lisbon agreement on femoroacetabular impingement imaging-part 1: overview. *Eur Radiol* 2020;30:5281–97.

50 Riley RD, Ensor J, Snell KIE, *et al*. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.

51 Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* 2016;35:214–26.

52 Snell KIE, Archer L, Ensor J, *et al*. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *J Clin Epidemiol* 2021;135:S0895-4356(21)00048-2:79–89.:.

53 Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:S0895-4356(18)31081-3:12–22.:.

54 van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014;14:137.

55 Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.

56 Van Calster B, McLernon DJ, van Smeden M, *et al*. Calibration: the achilles heel of predictive analytics. *BMC Med* 2019;17:230.

57 Verbakel JY, Steyerberg EW, Uno H, *et al*. ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *J Clin Epidemiol* 2020;126:S0895-4356(18)31004-7:207–16.:.

58 Van Calster B, Nieboer D, Vergouwe Y, *et al*. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016;74:S0895-4356(15)00581-8:167–76.:.

59 Hingorani AD, Windt DA van der, Riley RD, *et al*. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013;346:e5793.

60 Von Korff M, Moore JC. Stepped care for back pain: activating approaches for primary care. *Ann Intern Med* 2001;134(9 Pt 2):911–7.

61 Kemp JL, Risberg MA, Mosler A, *et al*. Physiotherapist-led treatment for young to middle-aged active adults with hip-related pain: consensus recommendations from the international hip-related pain research network, zurich 2018. *Br J Sports Med* 2020;54:504–11.

62 Ishøi L, Thorborg K, Kemp JL, *et al*. Maximal hip muscle strength and rate of torque development 6-30 months after hip arthroscopy for femoroacetabular impingement syndrome: A cross-sectional study. *J Sci Med Sport* 2021;24:S1440-2440(21)00131-6:1110–5.:.

63 Kemp JL, Mosler AB, Hart H, *et al*. Improving function in people with hip-related pain: a systematic review and meta-analysis of physiotherapist-led interventions for hip-related pain. *Br J Sports Med* 2020;54:1382–94.

64 Kemp JL, Risberg MA, Schache AG, *et al*. Patients with chondrolabral pathology have bilateral functional impairments 12 to 24 months after unilateral hip arthroscopy: A cross-sectional study. *J Orthop Sports Phys Ther* 2016;46:947–56.

65 Diamond L, Murphy N, Eyles J, *et al*. Patient factors predict severity of hip symptoms to a greater extent than abnormal bony hip morphology in femoroacetabular impingement syndrome. *Journal of Science and Medicine in Sport* 2019;22:S62–3.

66 Ishøi L, Thorborg K, Kraemer O, *et al*. Demographic and radiographic factors associated with intra-articular hip cartilage injury: A cross-sectional study of 1511 hip arthroscopy procedures. *Am J Sports Med* 2019;47:2617–25.

67 Ishøi L, Thorborg K, Kraemer O, *et al*. The association between specific sports activities and sport performance following hip arthroscopy for femoroacetabular impingement syndrome: A secondary analysis of A cross-sectional cohort study including 184 athletes. *J Hip Preserv Surg* 2019;6:124–33.

68 EQ-5D-3L – EQ-5D. Available: https://euroqol.org/eq-5d-instruments/eq-5d-3l-about/