Predicting the outcomes of assisted reproductive technology treatments: A systematic review and quality assessment of prediction models

Ian Henderson, MSc, Michael P. Rimmer, MSc, Stephen D. Keay, FRCOG, Paul Sutcliffe, PhD, Khalid S. Khan, FRCOG, Ephia Yasmin, PhD, Bassel H. Al Wattar, PhD

PII: S2666-5719(20)30010-4

DOI: https://doi.org/10.1016/j.xfnr.2020.11.002

Reference: XFNR 8

To appear in: *F&S Reviews*

Received Date: 8 October 2020

Revised Date: 13 November 2020

Accepted Date: 23 November 2020

Predicting the outcomes of assisted reproductive technology treatments: A systematic review and quality assessment of prediction models

Ian Henderson MSc[1,2], Michael P Rimmer MSc[3], Stephen D Keay FRCOG[2], Paul Sutcliffe PhD[1], Khalid S Khan FRCOG[4], Ephia Yasmin PhD[5], Bassel H.Al Wattar PhD[1,2,5]

[1]Warwick Medical School, Warwick University, Coventry, UK.

[2]Centre for Reproductive medicine, University Hospital Coventry and Warwickshire, Clifford Bridge Road, Coventry, UK.

[3]MRC Centre for Reproductive Health, Queens Medical Research Institute, Edinburgh BioQuarter, University of Edinburgh, UK.

[4]Department of Preventive Medicine and Public Health, University of Granada, 18071 Granada, Spain.

[5]Reproductive medicine unit, University College London Hospitals, London, UK.

Corresponding author: Bassel H.Al Wattar - Warwick Medical School, Warwick University, Coventry, UK. Email: dr.basselwa@gmail.com.

Short title: Predicting assisted conception outcomes

Capsule (30):

We reviewed and evaluated 120 prediction models published over the last 24 years. We identified twelve externally validated models that could be used to advise couples undergoing fertility treatments.

Abstract (250):

Objective: Predicting the outcomes of assisted reproductive technology (ART) treatments is desirable, but adopting prediction models into clinical practice remains limited. We aimed to review available prediction models for ART treatments by conducting a systematic review of the literature to identify the best performing models for their accuracy, generalisability and applicability.

Evidence review: We searched electronic databases (MEDLINE, EMBASE, and CENTRAL) until June 2020. We included studies reporting on the development or evaluation of models predicting the reproductive outcomes before (pre-ART) or after starting (Intra-ART) treatment in couples undergoing any ART treatment. We evaluated the models' discrimination, calibration, type of validation, and any implementation tools for clinical practice.

Results: We included 69 cohort studies reporting on 120 unique prediction models. Half the studies reported on pre-ART (48%) and half on intra-ART (56%) prediction models. The commonest predictors used were maternal age (90%), tubal factor subfertility (50%), and embryo quality (60%).

Only fourteen models were externally-validated (14/120, 12%) including eight pre-ART models (Templeton, Nelson, LaMarca, McLernon, Arvis, and the Stolwijk A/I,C,II models), and five intra-ART models (Cai, Hunault, van Loendersloot, Meijerink, Stolwijk B, and the McLernon post-treatment model) with a reported c-statistics ranging from 0.50 to 0.78. Ten of these models provided implementation tools for clinical practice with only two reported online calculators.

Conclusion: We identified externally validated prediction models that could be used to advise couples undergoing ART treatments on their reproductive outcomes. The quality of available models remains limited and more research is needed to improve their generalizability and applicability into clinical practice.

Keywords: infertility, prediction, assisted reproduction, systematic review.

Highlights:

- Over the last 24 years a high number of studies attempted to produce useful prediction models and decision aids for clinicians and patients undergoing ART.
- In this review we evaluated 69 studies reporting on 120 unique prediction models, but only a minority of these models were externally validated or useful in clinical practice.
- Most of these models suffered from a high risk of bias driven by poor model development, data sampling and analysis methodology.
- More research is needed to leverage available data, refine published models, and increase their applicability in clinical practice using novel technology such as artificial intelligence and dynamic intra-treatment prediction modelling.

Introduction

Assisted reproductive technology (ART) has evolved over the last 40 years offering hope to a record number of infertile couples worldwide (1–3). Currently ART is the first port of call for many couples inclusive of those experiencing unexplained and reversible causes of subfertility such as mild male factor and unilateral tubal pathology. The birth rate with assisted conception increased steadily over the last few decades from an average of 9% in 1991 to 23% in 2018 (4). This mass adoption of ART, however, sparked the debate on the ethical use of some ART treatments (5), their cost-effectiveness, and the risk of profiteering to certain patient groups (6). Accurate prediction of clinical outcomes and any mitigating risk factors could help to rationalize the use of ART treatments and improve their clinical effectiveness (7). While many prediction models have been produced to aid clinicians and couples in planning their fertility treatments, implementing those models remains limited in practice (8).

To be used effectively, prediction models should undergo rigorous development, validation, and impact assessment (9,10). Unsurprisingly, few published models complete this process which limits their clinical value and increase research wastage (7,8,11). Advances in data gathering and statistical methodology using machine learning and artificial intelligence could help to streamline the development and validation process of prediction models, but such practice remains limited in reproductive medicine (12).

Our aim was to systematically review and evaluate the performance, generalisability and applicability of published prediction models for ART treatments to identify the best performing models that could be used in clinical practice.

Methods

We conducted this systematic review using a prospectively registered protocol (CRD42019156606) and reported the findings following standard guidelines (13).

Search strategy and study selection

We searched electronic databases (MEDLINE, EMBASE, and Cochrane CENTRAL) from inception until June 2020 for all studies reporting on the development or evaluation of any prediction model for the outcome of any ART treatments (in vitro fertilization (IVF) and/or intracytoplasmic sperm injection (ICSI)). We did not apply any search filters or language restrictions. Articles in non-English were translated if deemed relevant. We conducted supplementary searches in Google Scholar and Scopus for any additional articles of interest in the grey literature. We also searched the bibliographies of relevant articles to identify any missing citations.

We included longitudinal studies that reported on the development or evaluation of any model for predicting clinical pregnancy (confirmed on ultrasound) or live birth following any ART treatments. We excluded studies reporting on the crude association between a single independent variable and the outcomes of interest, those reporting on non-predictive models, and those not reporting on the model performance measures. Models predicting non-reproductive outcomes or solely predicting biochemical pregnancy were also excluded. Similarly, we excluded models that used solely embryological or seminal parameters to predict the outcomes of interest. Finally, we also excluded case series, conference abstracts and review articles.

Assessment of study quality

We assessed the risk of bias and applicability of the included studies in duplicate using the PROBAST tool (14). Studies were assessed in four domains: population, predictors, outcome, and analysis. Studies were deemed low risk of bias if they were cohort studies, defined and measured predictors consistently and independently of the pre-specified outcome, included sufficient events per variable with appropriate parameterisation of predictors, included all participants in the analysis, treated missing data appropriately, did not include predictors based on univariable analyses, assessed the model's discrimination and calibration appropriately, and accounted for model overfitting and optimism based on the use of an appropriate validation procedure and shrinkage of estimates in the presence of optimism which were evaluated in the context of events per variable, appropriate parameterisation and modelling strategy (14). We produced an overall assessment of both the risk of bias and model applicability per study.

Models performance, generalizability and applicability

We evaluated models' performance by their reported discrimination (the model's ability to separate those with and without the outcome of interest) and calibration (the concordance between predicted and observed outcome frequency) measures (15). Discrimination is commonly described using the rank order statistic 'area under the receiver operating characteristic curve' (AUROC), which is equivalent to the concordance-statistic (c-statistic). We considered a c-statistic value of 0.5 to represent no discriminative ability, a value of 1 to represent perfect discriminative ability (15). Calibration is often assessed using the Hosmer-Lemeshow statistic (16). A model is considered well-calibrated when the average predicted probability per sub-group matches the observed proportion. Calibration is more informatively assessed graphically by the calibration plot, where the predicted probability per ordered sub-group is plotted against the observed proportion, demonstrating the nature and magnitude of any miscalibration. An intercept of 0 and a slope of 1 therefore represents perfect calibration (17).

To evaluate generalizability, we reported on the validation process for each model including the validation type, procedures, and characteristics of the validation population. We divided validation efforts into ‘internal’, ‘temporal’, or ‘external’ depending the type of validation population.

To evaluate the models’ applicability and translation into clinical practice, we reported on efforts to increase the model’s accessibility to both health professionals and lay consumers, and the availability of any decision support tools including predicted probabilities based on patient profile, score-based decision aids, score-based nomograms, to end-user web-based predictive calculators.

Data extraction

Two independent reviewers (IH and MPR) extracted data onto a custom designed collection database guided by the CHARMS checklist (18) to identify relevant data points for extraction and reporting. We extracted data on the study design, outcome, sample size, population characteristics, model development methods, performance and validation statistics, and clinical application. We divided models into (pre-ART) where outcome prediction was possible prior to commencing ovarian stimulation, and (intra-ART) where outcome prediction was possible after commencing ovarian stimulation. We categorized the included studies as per the TRIPOD guidelines into: type 1a studies developing a model and evaluating its predictive performance using the same data (apparent performance), type 1b studies developing a prediction model using the entire dataset with resampling (e.g. bootstrapping or cross-validation) techniques to evaluate the performance and optimise the developed model, type 2a studies with data randomly split to develop the model and then to evaluate its predictive performance, type 2b studies with data non-randomly split (e.g. by location or time) to develop the prediction model and then to evaluate its predictive performance, type 3 studies developing a prediction model using one dataset and an evaluation of its performance on separate data (e.g. from a different population), and type 4 studies which are only evaluating the predictive performance of an existing prediction model in a separate dataset (19).

Statistical analysis

We summarised data using descriptive statistics and reported on continuous data using means or medians with standard deviations where relevant. For dichotomous data we reported using frequencies and natural percentages. All analyses and figures were produced using RStudio version 1.2.1335 (RStudio, Boston, MA) (20).

Results

Study characteristics

Our search revealed 8052 potentially relevant unique citations; of these, we reviewed 483 in full and included 69 studies in our review reporting on the development of 120 ART prediction models (Figure 1). All included studies were cohort studies, 55 of which were retrospective (55/69, 79.7%) and 14 prospective (14/69, 20.3%). As per TRIPOD classification, 18 (18/69, 26.1%) of these studies were type 1a studies, 20 (20/69, 29.0%) were type 1b, 6 (6/69, 8.7%) were type 2a, 10 (10/69, 14.5%) were type 2b, 5 (5/69, 7.2%) were type 3, and 10 (10/69, 14.5%) type 4 (Figure 2). The majority were from Europe (49/69, 71.0%) with only eleven from Asia (11/69, 15.9%), and three from North America (3/69, 4.3%).

There were variations in the population characteristics across included studies. Nine studies (13.4%) included unselected couples (for age, cycle cancellation, maternal comorbidity, aetiology, and sperm source), seven included unselected couples but excluded women using donor gametes (10.4%), and twelve studies (17.9%) included couples with selected baseline characteristics (Supplementary Table 1). About half of the included studies explicitly excluded donor oocyte cycles (29/69, 42.0%), and a third explicitly excluded cancelled cycles (21/69, 30.4%), and a quarter explicitly excluded women outside a specific age range (18/69, 26.1%).

Most of the included studies reported on the development (with or without validation) of novel models (62/69, 89.9%), with the remainder uniquely reporting on the validation of pre-existing models (7/69, 10.1%). Half of these studies (30/62, 48.3%) reported on pre-ART predictive models (21–47), and 56% (35/62, 56.5%) reported on intra-ART (48–78). Only three studies (3/62, 4.8%) reported on both pre and intra-ART predictive models (79–81). Three quarters of these developmental studies (47/62, 75.8%) involved IVF/ICSI treatments, twelve IVF treatment only (12/62, 19.4%), and two ICSI treatment only (2/62, 3.2%), with 1 unspecified by the authors. Two-thirds included only cycles using a fresh embryo transfer (41/62, 66.1%), while both fresh and frozen embryo cycles were included in 21 studies (21/62, 33.9%).

Predictors and outcomes

For studies that developed pre-ART models, the commonest included predictor was maternal age (27/30, 90.0%) followed by tubal factor subfertility (15/30, 50.0%), gravidity (13/30, 43.3%), and the duration of subfertility (12/30, 40.0%) (Figure 3a). A similar trend was seen for intra-ART models as the commonest included predictor was also maternal age (33/35, 94.3%), followed by embryo quality (21/35, 60.0%), previous ART success (16/35, 45.7%), duration of subfertility (12/35, 34.3%), and tubal factor subfertility (10/35, 28.6%) (Figure 3b).

Live birth was the outcome of interest across all studies, for those that developed both pre-ART (20/30, 66.7%) and intra-ART (18/35, 51.4%) models. A quarter of studies that developed intra-ART models focused on clinical pregnancy (10/35, 28.6%) and ongoing pregnancy (8/35, 22.9%) which were less frequently reported in pre-ART models (clinical pregnancy (5/30, 16.7%), ongoing pregnancy (5/30, 16.7%)).

Sample size and modelling method

The median sample size for developing pre-ART models was 757 for participants (range 85-113,873) and 1,061 for ART cycles (range 113-443,202). For intra-ART models, the median participant sample size was 1,419 (range 90-113,873) and median ART cycles was 1,676 (range 110-184,269). Most studies (48/69, 69.6%) had ≥10 events per candidate variable (degrees of freedom). The majority of studies developed models using logistic regression (pre-ART (24/30, 80.0%), intra-ART (30/35, 85.7%)). Only a minority used other methods, including generalized estimating equations, Bayesian networks, Cox regression, machine learning techniques and deep learning techniques (Supplementary Table 2).

Performance, generalizability and applicability

Discrimination was reported for most of the included studies (109/120, 90.8%) while calibration was reported for over half (72/120, 60.0%). Both discrimination and calibration were reported in only 61 studies (61/120, 50.8%). The commonest methods to assess calibration were the Hosmer-Lemeshow statistic (27/72, 37.5%), calibration plot (24/72, 33.3%), slope test (14/72, 19.4%), and calibration-in-the-large (11/72, 15.3%).

We captured 31 unvalidated models from type 1a studies without subsequent validation (31/120, 25.8%), as well as six models that were locally refit from validation studies (6/120, 5.0%). Fifty-five models were internally-validated from 1b/2a studies without subsequent validation (55/120, 45.8%), 15 were temporally-validated models from 2b studies without subsequent validation (15/120, 12.5%). There were seven external validation studies (7/120, 5.8%). Four were type 4 studies by a team that overlapped with the model development team (4/120, 3.3%)(35,80,82),(22,23,79), and three studies were performed by independent validation teams (30,37,57)

We captured eight externally validated pre-ART models: the Templeton model (n=6 validations), Nelson model (n=3), LaMarca model (n=1), McLernon pre-treatment model (n=1), Arvis model (n=1), and The Stolwijk models A/I, C, and II (n=7). All models showed similar performance with c-statistics ranging from 0.53 to 0.78. The Stolwijk models A/I and II were declared invalid (Table 1).

Among the intra-ART models, only five were externally validated: the Cai model (n=1), Hunault model (n=1), van Loendersloot model (n=1), Meijerink model (n=1), and the McLernon post-treatment model (n=1). All models showed similar performance with c-statistics ranging from 0.63 to 0.78. However, only the McLernon model was validated in a good quality external validation study with low risk of bias showing a c-statistic of 0.71 (95%CI 0.69-0.74) and reportedly good calibration (Table 1).

Only a quarter of all published models (33/120, 25.4%) were presented in full either offering the regression formula, coefficients with intercept, or baseline hazard. Seven models presented nomograms or score charts (7/120, 5.8%), and seven were adapted into online risk prediction calculators (7/120, 5.8%). Of these, only three calculators were functional at the time of writing this review(83–85). Overall, half of the included studies (35/62, 56.5%), reporting on 47 models (47/120, 39.2%), enabled the reader to generate a personalised prediction in a useful format. All the externally validated models offered an implementation tool except the Cai model and the invalid Stolwijk models. But only two presented an online calculator for use by health professionals and patients (the Nelson and the McLernon calculators) (Table 1).

Quality and risk of bias

Overall, a majority of the included studies were at high risk of bias (56/69, 81.2%) and only ten studies at low risk (10/69, 14.5%) (Figure 4, Supplementary Table 3). Within the ‘participant’ domain, three-quarters of the included studies were at low risk (50/69, 72.5%) and nine at high risk (9/69, 13.0%). Similarly, within the ‘outcome’ domain, the majority were at low risk (66/69, 95.7%). In contrast, within the ‘predictor’ domain only half were at low risk (32/69, 46.4%), with 36 studies of unclear risk due to providing inadequate definitions, namely for candidate predictors (36/69, 52.2%). For the ‘analysis’ domain, less than a fifth were of low risk of bias (12/69, 17.4%). Half (35/69, 50.7%) assessed model performance appropriately, by discrimination and an informative measure of calibration. Only a quarter reported and handled missing data appropriately (16/69, 23.2%); only 19 studies (19/69, 27.5%) addressed overfitting and optimism; only 48 had sufficient events per candidate predictor (≥20 events (14)) (48/69, 69.6%), and only 38 parameterized predictors appropriately (38/69, 55.1%).

Discussion

Summary of main findings

Our findings depict an overall high investment in producing working prediction models and decision aids for clinicians and patients undergoing ART treatments with 120 models produced over the last 24 years, an average of 5 models produced per year. However, while huge resources and patient data were committed to producing these models, only a minority of these studies offered externally validated models that could be used in everyday practice.

The majority of the included studies had a high risk of bias, largely driven by poor model development methodology specifically in data sampling and analysis (Figure 4). Only a minority of models were developed within large sizes cohorts (only 9 studies included >10,000 women/cycles) and most were selected ART populations, thus reducing model's applicability in practice. In contrast, with much prediction data available several clinical and biochemical markers are now well established as reliable predictors of reproductive outcomes (Figure 3a, 3b). Leveraging this large body of evidence could facilitate the process of developing and validating future models to minimize duplication of efforts. Logistic regression modelling remains the commonest method for model development, though alternative methodology is becoming popular such as artificial intelligence aided techniques (29,34,38,46,48,49,54,65,69,75,86).

Strengths and limitations

The strengths of our review are several. In contrast to previously published reviews (7,8,11), we used a prospectively registered protocol, applied a comprehensive search strategy, extracted data in duplicate, assessed quality according to PROBAST criteria, and included all types of studies as per TRIPOD (both model development and validation studies) to evaluate models' applicability into clinical practice. Consequently, our findings offer a robust assessment of the current state-of-the-art in ART prediction modelling and the remaining knowledge gap. To aid their adoption in practice, we identified top performing models referencing their quantitative assessment markers, relevant population of interest and how they can be accessed online (Table 1).

Our research was inclusive with almost double the number of studies included in the most recent review (11) offering a more comprehensive and systematic assessment of the literature. A previous review by Ratna et al adopted an arbitrary quality threshold of 80% adherence to TRIPOD (19) in their inclusion criteria which could have limited the generalizability of their findings. We refrained from imposing any reporting thresholds and assessed the methodological quality of all published models to offer a comprehensive and objective assessment of the literature.

Our findings still have some limitations. Several of the studies reported vaguely on the measures of calibration using terms like “good calibration” which limited our ability to provide an objective assessment of these models. Furthermore, given the lack of a universally adopted definition of what constitutes good calibration for ART models, it is difficult to preferentially select top performing models. Clearly, most subfertile couples have some probability of conceiving independent of any treatment, similarly the chance of conception in healthy couples is never 100% in every cycle. As the methodological standards for model development improved over time, our contemporary PROBAST assessment of risk of bias might differ from older reviews and the findings are therefore not completely reproducible.

Implications for clinical practice

Introducing prediction modelling into clinical practice was aimed to tailor treatments to each patient's individual needs, thus maximising effectiveness and reducing personal harm (9). Models can aid decision making on starting treatment (87) or to adjust a treatment to the patient characteristics (88). Whilst most treatments are static (e.g., medication or surgery), the process of undergoing IVF or ICSI treatments is heterogeneous and dynamic, continuously changing through a series of interconnected complex decisions made to optimise successful conception. Coupled with the rapid progress in ART, it is likely that most models will be over-simplistic and become outdated. This applies especially to pre-ART models which are dependent on a limited range of predictors that cannot adjust for initial treatment response (e.g., ovulation stimulation and embryo fertilisation). Consequently, the clinical value of available models is currently limited to counselling patients on the value of starting ART treatment rather than tailoring those treatments to maximize chances of conception. A solution could lie in the development, validation and continuous update of dynamic models that could adjust for the within-treatment changes and offer a refined estimate of successful conception throughout the ART treatment process (89).

The process of IVF/ICSI is emotionally and psychologically demanding with patients often having to make difficult decisions such as the use of frozen embryos or consider add-on therapies (90). Predicting the chances of conception in itself can be stressful (91) which could limit the adoption of these models in practice. As such, developing any prediction models should be guided by expressed patients' needs (92), a practice we did not observe in the models included in this review. Future model development should take into account the various decision-making processes involved in the ART treatment process and the associated predictors that could add cumulative information to aid patients and their caring clinicians in the decision-making process. Lastly, successful model implementation into clinical practice could be facilitated by improved interpretability (93) and user-friendly interfaces that enable end users to input and access data effortlessly in jargon-free outputs such as online risk calculators or decision aid tools hosted on mobile apps (83–85).

Future research need

Our findings illustrate an abundance of data dedicated to predict ART outcomes, yet translation into practice remains limited. As our ability to collect and analysis large datasets improves over time, perhaps future steps should focus more on harmonizing data collection across institutions, regulators and countries to facilitate streamlined model development, validation, and update while reducing associated costs. Crucially, there is a need to focus available resources on combining data from published models (e.g., using individual patient data meta-analysis methodology) and externally validating ensuing ones rather than on developing newer models.

We captured a recent trend towards using artificial intelligence (AI) technology in model development (29,34,38,46,48,49,54,65,69,75,86). While promising, most of these models did not achieve improved prediction performance nor followed sound methodology compared to older ones (94). Specifically, the work on many of these models seem to be driven by an experimental approach evaluating the different AI technologies rather than a multi-disciplinary approach aiming to address real patients' needs. Still, leveraging the power of AI technology and big data research methods to simulate the complex decision making process involved in ART treatments could be a game changer to provide accurate individualized fertility assessment to couples in need (95). Large multi-national multi-disciplinary teams are best equipped to address this complex and important health problem.

Conclusions

We identified externally validated prediction models that could be used to advise couples undergoing ART treatments on their reproductive outcomes. The quality of available models remains limited and more research is needed to improve their generalisability and applicability in clinical practice.

Acknowledgement: None

Funding: None

Conflict of interest: All authors have nothing to disclose. BHA holds a personal Lectureship from the UK National Health Institute of Research. KSK is Distinguished Investigator at the University of Granada with a grant awarded by the Beatriz Galindo Program (senior modality) of the Spanish Ministry of Science, Innovation and Universities.

Contribution to Authorship: BHA conceived the idea. BHA and IH wrote the final protocol and manuscript. IH conducted the search. IH and MR conducted the data extraction and 1st draft of the manuscript. BHA and IH conducted the statistical analysis and data interpretation. SK and KSK contributed to data interpretation and final editing of the manuscript.

References:

1. wwwhfeagovuk. HFEA Fertility treatment 2017: trends and figures [Internet]. 2017. Available from: www.hfea.gov.uk
2. Society for Assisted Reproductive Technology. National Summary Report [Internet]. 2017;Available from: https://www.sartcorsonline.com/rptCSR_PublicMultYear.aspx
3. Zegers-Hochschild F, Schwarze JE, Crosby J, Musri C, Urbina MT. Assisted reproductive techniques in Latin America: the Latin American Registry, 2015. Reprod Biomed Online 2018;37:685–92.
4. Human Fertilisation & Embryology Authority. Fertility treatment 2018: trends and figures. 2020.
5. te Velde E, Habbema D, Nieschlag E, Sobotka T, Burdorf A. Ever growing demand for in vitro fertilization despite stable biological fertility—A European paradox. Eur. J. Obstet. Gynecol. Reprod. Biol. 2017;214:204–8.
6. Heng BC. Can the difference in medical fees for self and donor freeze-thaw embryo transfer cycle, be in fact a cover-up for the sale of donated human embryos? Philos. Ethics, Humanit. Med. 2007;
7. Leushuis E, van der Steeg JW, Steures P, Bossuyt PMM, Eijkemans MJC, van der Veen F, et al. Prediction models in reproductive medicine: a critical appraisal. Hum Reprod Update 15:537–52.
8. van Loendersloot L, Repping S, Bossuyt PMM, van der Veen F, van Wely M. Prediction models in in vitro fertilization; where are we? A mini review. J. Adv. Res. 2014;
9. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology. 2010;21:128–38.

10. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J. Clin. Epidemiol. 2016;

11. Ratna MB, Bhattacharya S, Abdulrahim B, McLernon DJ. A systematic review of the quality of clinical prediction models in in vitro fertilisation. Hum Reprod 2020;35:100–16.

12. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:1–9.

13. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. BMJ. 2009;339:332–6.

14. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51–8.

15. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–35.

16. HOSMER DW, HOSMER T, CESSIE S LE, LEMESHOW S. A COMPARISON OF GOODNESS□OF□FIT TESTS FOR THE LOGISTIC REGRESSION MODEL. Stat Med 1997;16:965–80.

17. Miller ME, Langefeld CD, Tierney WM, Hui SL, Mcdonald CJ. Validation of Probabilistic Predictions. Med Decis Mak 1993;

18. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med 2014;11:e1001744.

19. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, 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.

20. RStudio Team. RStudio: Integrated Development for R. 2018;

21. Alebić MŠ, Stojanović N, Zuvić-Butorac M. The IVF Outcome Counseling Based on the Model Combining DHEAS and Age in Patients with Low AMH Prior to the First Cycle of GnRH Antagonist Protocol of Ovarian Stimulation. Int J Endocrinol 2013;2013:637919.

22. Arvis P, Lehert P, Guivarc'h-Levêque A. Simple adaptations to the Templeton model for IVF outcome prediction make it current and clinically useful. Hum Reprod 2012;27:2971–8.

23. La Marca A, Nelson SM, Sighinolfi G, Manno M, Baraldi E, Roli L, et al. Anti-Müllerian hormone-based prediction model for a live birth in assisted reproduction. Reprod Biomed Online 2011;22:341–9.

24. Li HWR, Lee VCY, Lau EYL, Yeung WSB, Ho PC, Ng EHY. Role of Baseline Antral Follicle Count and Anti-Mullerian Hormone in Prediction of Cumulative Live Birth in the First In Vitro Fertilisation Cycle: A Retrospective Cohort Analysis. PLoS One 2013;8:e61095.

25. Lintsen AME, Eijkemans MJC, Hunault CC, Bouwmans CAM, Hakkaart L, Habbema JDF, et al. Predicting ongoing pregnancy chances after IVF and ICSI: a national prospective study. Hum Reprod 2007;22:2455–62.

26. Luke B, Brown MB, Wantman E, Stern JE, Baker VL, Widra E, et al. A prediction model for live birth and multiple births within the first three cycles of assisted reproductive technology. Fertil Steril 2014;102:744–52.

27. McLernon DJ, Lee AJ, Maheshwari A, van Eekelen R, van Geloven N, Putter H, et al. Predicting the chances of having a baby with or without treatment at different time points in couples with unexplained subfertility. Hum Reprod 2019;34:1126–38.

28. Metello JL, Tomás C, Ferreira P. Can we predict the IVF/ICSI live birth rate? J Bras Reprod Assist 2019;23:402–7.

29. Nelson SM, Fleming R, Gaudoin M, Choi B, Santo-Domingo K, Yao M. Antimüllerian hormone levels and antral follicle count as prognostic indicators in a personalized prediction model of live birth. Fertil Steril 2015;104:325–32.

30. Nelson SM, Lawlor DA. Predicting live birth, preterm delivery, and low birth weight in infants born from in vitro fertilisation: A prospective study of 144,018 treatment cycles. PLoS Med 2011;8.

31. Pettersson G, Nyboe Andersen A, Broberg P, Arce JC. Pre-stimulation parameters predicting live birth after IVF in the long GnRH agonist protocol. Reprod Biomed Online 2010;20:572–81.

32. Porcu G, Lehert P, Colella C, Giorgetti C. Predicting live birth chances for women with multiple consecutive failing IVF cycles: a simple and accurate prediction for routine medical practice. Reprod Biol Endocrinol 2013;11:1.

33. Ballester M, Oppenheimer A, d'Argent EM, Touboul C, Antoine J-M, Coutant C, et al. Nomogram to predict pregnancy rate after ICSI-IVF cycle in patients with endometriosis. Hum Reprod 2012;27:451–6.

34. Qiu J, Li P, Dong M, Xin X, Tan J. Personalized prediction of live birth prior to the first in vitro fertilization treatment: A machine learning method. J Transl Med 2019;17.

35. Rongieres C, Colella C, Lehert P. To what extent does Anti-Mullerian Hormone contribute to a better prediction of live birth after IVF? J Assist Reprod Genet 2015;32:37–43.

36. Stolwijk AM, Straatman H, Zielhuis GA, Jansen CA, Braat DD, van Dop PA, et al. External validation of prognostic models for ongoing pregnancy after in-vitro fertilization. Hum Reprod 1998;13:3542–9.

37. Templeton A, Morris JK, Parslow W. Factors that affect outcome of in-vitro fertilisation treatment. Lancet 1996;348:1402–6.

38. Wald M, Sparks AET, Sandlow J, Van-Voorhis B, Syrop CH, Niederberger CS. Computational models for prediction of IVF/ICSI outcomes with surgically retrieved spermatozoa. Reprod Biomed Online 2005;11:325–31.

39. van Weert J-M, Repping S, van der Steeg JW, Steures P, van der Veen F, Mol BW. A prediction model for ongoing pregnancy after in vitro fertilization in couples with male subfertility. J Reprod Med 2008;53:250–6.

40. Tarín JJ, Pascual E, García-Pérez MA, Gómez R, Hidalgo-Mora JJ, Cano A. A predictive model for women's assisted fecundity before starting the first IVF/ICSI treatment cycle. J Assist Reprod Genet 2020;

41. Bancsi LFJMM, Huijs AM, Den Ouden CT, Broekmans FJM, Looman CWN, Blankenstein MA, et al. Basal follicle-stimulating hormone levels are of limited value in predicting ongoing pregnancy rates after in vitro fertilization. Fertil Steril 2000;73:552–7.

42. Brodin T, Hadziosmanovic N, Berglund L, Olovsson M, Holte J. Comparing four ovarian reserve markers--associations with ovarian response and live births after assisted reproduction. Acta Obstet Gynecol Scand 2015;94:1056–63.

43. Choi B, Bosch E, Lannon BM, Leveille MC, Wong WH, Leader A, et al. Personalized prediction of first-cycle in vitro fertilization success. Fertil Steril 2013;99(7):1905–11.

44. Dhillon RK, McLernon DJ, Smith PP, Fishel S, Dowell K, Deeks JJ, et al. Predicting the chance of live birth for women undergoing IVF: A novel pretreatment counselling tool. Hum Reprod 2016;31:84–92.

45. Ferlitsch K, Sator MO, Gruber DM, Rücklinger E, Gruber CJ, Huber JC. Body mass index, follicle-stimulating hormone and their predictive value in in vitro fertilization. J

Assist Reprod Genet 2004;21:431–6.

46. Güvenir HA, Misirli G, Dilbaz S, Ozdegirmenci O, Demir B, Dilbaz B. Estimating the chance of success in IVF treatment using a ranking algorithm. Med Biol Eng Comput 2015;53:911–20.

47. Hamdine O, Eijkemans MJC, Lentjes EGW, Torrance HL, Macklon NS, Fauser BCJM, et al. Antimüllerian hormone: prediction of cumulative live birth in gonadotropin-releasing hormone antagonist treatment for in vitro fertilization. Fertil Steril 2015;104:891-898.e2.

48. Banerjee P, Choi B, Shahine LK, Jun SH, O'Leary K, Lathi RB, et al. Deep phenotyping to predict live birth outcomes in in vitro fertilization. Proc Natl Acad Sci U S A 2010;107:13570–5.

49. Blank C, Wildeboer RR, DeCroo I, Tilleman K, Weyers B, de Sutter P, et al. Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. Fertil Steril 2019;

50. Ho V, Pham T, Ho T, Vuong L. Predictive Model for Live Birth at 12 Months After Starting In-Vitro Fertilization Treatment. MedPharmRes 2018;2:5–20.

51. Hunault CC, Eijkemans MJC, Pieters MHEC, Te Velde ER, Habbema JDF, Fauser BCJM, et al. A prediction model for selecting patients undergoing in vitro fertilization for elective single embryo transfer. Fertil Steril 2002;77:725–32.

52. Hunault CC, te Velde ER, Weima SM, Macklon NS, Eijkemans MJC, Klinkert ER, et al. A case study of the applicability of a prediction model for the selection of patients undergoing in vitro fertilization for single embryo transfer in another center. Fertil Steril 2007;87:1314–21.

53. Jones CA, Christensen AL, Salihu H, Carpenter W, Petrozzino J, Abrams E, et al. Prediction of individual probabilities of livebirth and multiple birth events following in

vitro fertilization (IVF): a new outcomes counselling tool for IVF providers and patients using HFEA metrics. J Exp Clin Assist Reprod 2011;8:3.

54. Kaufmann SJ, Eastaugh JL, Snowden S, Smye SW, Sharma V. The application of neural networks in predicting the outcome of in- vitro fertilization. Hum Reprod 1997;12:1454–7.

55. Kim SK, Kim H, Oh S, Lee JR, Jee BC, Kim SH. Development of a novel nomogram for predicting ongoing pregnancy after in vitro fertilization and embryo transfer. Obstet Gynecol Sci 2018;61:669–74.

56. Liao S, Xiong J, Tu H, Hu C, Pan W, Geng Y, et al. Prediction of in vitro fertilization outcome at different antral follicle count thresholds combined with female age, female cause of infertility, and ovarian response in a prospective cohort of 8269 women. Med (United States) 2019;98.

57. van Loendersloot LL, van Wely M, Repping S, Bossuyt PMM, van der Veen F. Individualized decision-making in IVF: calculating the chances of pregnancy. Hum Reprod 2013;28:2972–80.

58. Meijerink AM, Cissen M, Mochtar MH, Fleischer K, Thoonen I, De Melker AA, et al. Prediction model for live birth in ICSI using testicular extracted sperm. Adv Access Publ July 2016;31:1942–51.

59. Ottosen LDM, Kesmodel U, Hindkjær J, Ingerslev HJ. Pregnancy prediction models and eSET criteria for IVF patients - Do we need more information? J Assist Reprod Genet 2007;24:29–36.

60. Cai QF, Wan F, Huang R, Zhang HW. Factors predicting the cumulative outcome of IVF/ICSI treatment: a multivariable analysis of 2450 patients. Hum Reprod 2011;26:2532–40.

61. Roberts SA, Fitzgerald CT, Brison DR. Modelling the impact of single embryo

transfer in a national health service IVF programme. Hum Reprod 2009;24(1):122–31.

62. Roberts SA, Hirst WM, Brison DR, Vail A, towardSET collaboration. Embryo and uterine influences on IVF outcomes: an analysis of a UK multi-centre cohort. Hum Reprod 2010;25:2792–802.

63. Roberts SA, Hann M, Brison DR. Factors affecting embryo viability and uterine receptivity: insights from an analysis of the UK registry data. Reprod Biomed Online 2016;32:197–206.

64. Sunkara SK, Rittenberg V, Raine-Fenning N, Bhattacharya S, Zamora J, Coomarasamy A. Association between the number of eggs and live birth in IVF treatment: an analysis of 400 135 treatment cycles. Hum Reprod 2011;26:1768–74.

65. Uyar A, Bener A, Ciray HN. Predictive Modeling of Implantation Outcome in an in Vitro Fertilization Setting. Med Decis Mak 2015;35:714–25.

66. Verberg MFG, Eijkemans MJC, Macklon NS, Heijnen EMEW, Fauser BCJM, Broekmans FJ. Predictors of ongoing pregnancy after single-embryo transfer following mild ovarian stimulation for IVF. Fertil Steril 2008;

67. Vaegter KK, Lakic TG, Olovsson M, Berglund L, Brodin T, Holte J. Which factors are most predictive for live birth after in vitro fertilization and intracytoplasmic sperm injection (IVF/ICSI) treatments? Analysis of 100 prospectively recorded variables in 8,400 IVF/ICSI single-embryo transfers. Fertil Steril 2017;107:641-648.e2.

68. Vaegter KK, Berglund L, Tilly J, Hadziosmanovic N, Brodin T, Holte J. Construction and validation of a prediction model to minimize twin rates at preserved high live birth rates after IVF. Reprod Biomed Online 2019;38:22–9.

69. Vogiatzi P, Pouliakis A, Siristatidis C. An artificial neural network for the prediction of assisted reproduction outcome. J Assist Reprod Genet 2019;36:1441–8.

70. Wu F, Liu F, Guan Y, Du J, Tan J, Lv H, et al. A nomogram predicting clinical

pregnancy in the first fresh embryo transfer for women undergoing in vitro fertilization and intracytoplasmic sperm injection (IVF/ICSI) treatments. J Biomed Res 2019;33:422.

71. Carrera-Rotllan J, Estrada-García L, Sarquella-Ventura J. Prediction of pregnancy in IVF cycles on the fourth day of ovarian stimulation. J Assist Reprod Genet 2007;24:387–94.

72. Tarín JJ, Pascual E, Gómez R, García-Pérez MA, Cano A. Predictors of live birth in women with a history of biochemical pregnancies after assisted reproduction treatment. Eur J Obstet Gynecol Reprod Biol 2020;

73. Corani G, Magli C, Giusti A, Gianaroli L, Gambardella LM. A Bayesian network model for predicting pregnancy after in vitro fertilization. Comput Biol Med 2013;43:1783–92.

74. Dessolle L, Fréour T, Ravel C, Jean M, Colombel A, Daraï E, et al. Predictive factors of healthy term birth after single blastocyst transfer. Hum Reprod 2011;

75. Gianaroli L, Magli MC, Gambardella L, Giusti A, Grugnetti C, Corani G. Objective way to support embryo transfer: a probabilistic decision. Hum Reprod 2013;28:1210–20.

76. Goldman RH, Kaser DJ, Missmer SA, Srouji SS, Farland L V, Racowsky C. Building a model to increase live birth rate through patient-specific optimization of embryo transfer day. J Assist Reprod Genet 2016;33:1525–32.

77. Grin L, Mizrachi Y, Cohen O, Lazer T, Liberty G, Meltcer S, et al. Does progesterone to oocyte index have a predictive value for IVF outcome? A retrospective cohort and review of the literature. Gynecol Endocrinol 2018;34:638–43.

78. Hirst WM, Vail A, Brison DR, Roberts SA. Prognostic factors influencing fresh and frozen IVF outcomes: an analysis of the UK national database. Reprod Biomed Online

2011;22:437–48.

79. McLernon DJ, Steyerberg EW, Te Velde ER, Lee AJ, Bhattacharya S. Predicting the chances of a live birth after one or more complete cycles of in vitro fertilisation: Population based study of linked cycle data from 113 873 women. BMJ 2016;355:i5735.

80. Leijdekkers JA, Eijkemans MJC, van Tilborg TC, Oudshoorn SC, McLernon DJ, Bhattacharya S, et al. Predicting the cumulative chance of live birth over multiple complete cycles of in vitro fertilization: an external validation study. Hum Reprod 2018;33:1684–95.

81. Stolwijk AM, Zielhuis GA, Hamilton CJCM, Straatman H, Hollanders JMG, Goverde HJM, et al. Prognostic models for the probability of achieving an ongoing pregnancy after in-vitro fertilization and the importance of testing their predictive value. Hum Reprod 1996;11:2298–303.

82. Khader A, Lloyd SM, McConnachie A, Fleming R, Grisendi V, La Marca A, et al. External validation of anti-Müllerian hormone based prediction of live birth in assisted conception. J Ovarian Res 2013;6:3.

83. Qiu J, Li P, Dong M, Xin X, Tan J. Live birth prediction before the first IVF treatment.

84. University of Aberdeen. Outcome Prediction in Subfertility.

85. Society for Assisted Reproductive Technology. What are my chances with ART? [Internet]. 2020;Available from: https://www.sartcorsonline.com/Predictor/Patient

86. Choi B, Bosch E, Lannon BM, Leveille M-C, Wong WH, Leader A, et al. Personalized prediction of first-cycle in vitro fertilization success. Fertil Steril 2013;99:1905–11.

87. Sperrin M, Martin GP, Pate A, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. Stat Med 2018;

88. Hu YH, Wu F, Lo CL, Tai CT. Predicting warfarin dosage from clinical data: A supervised learning approach. Artif Intell Med 2012;

89. Frank I, Blute ML, Cheville JC, Lohse CM, Weaver AL, Zincke H. An outcome prediction model for patients with clear cell renal cell carcinoma treated with radical nephrectomy based on tumor stage, size, grade and necrosis: The SSIGN score. J Urol 2002;

90. Kaliarnta S, Nihlén-Fahlquist J, Roeser S. Emotions and ethical considerations of women undergoing IVF-treatments. HEC Forum 2011;

91. Mol BW, Verhagen TEM, Hendriks DJ, Collins JA, Coomarasamy A, Opmeer BC, et al. Value of ovarian reserve testing before IVF: A clinical decision analysis. Hum Reprod 2006;

92. Nachtigall RD, Dougall K Mac, Lee M, Harrington J, Becker G. What do patients want? Expectations and perceptions of IVF clinic information and support regarding frozen embryo disposition. Fertil Steril 2010;

93. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;

94. Hassan MR, Al-Insaif S, Hossain MI, Kamruzzaman J. A machine learning approach for prediction of pregnancy outcome following IVF treatment. Neural Comput Appl 2020;

95. Chen JH, Asch SM. Machine learning and prediction in medicine-beyond the peak of inflated expectations. N. Engl. J. Med. 2017;376:2507–9.

Figure legends:

Figure (1): Study selection and inclusion process on prediction models for reproductive outcomes following assisted reproductive technology treatments.

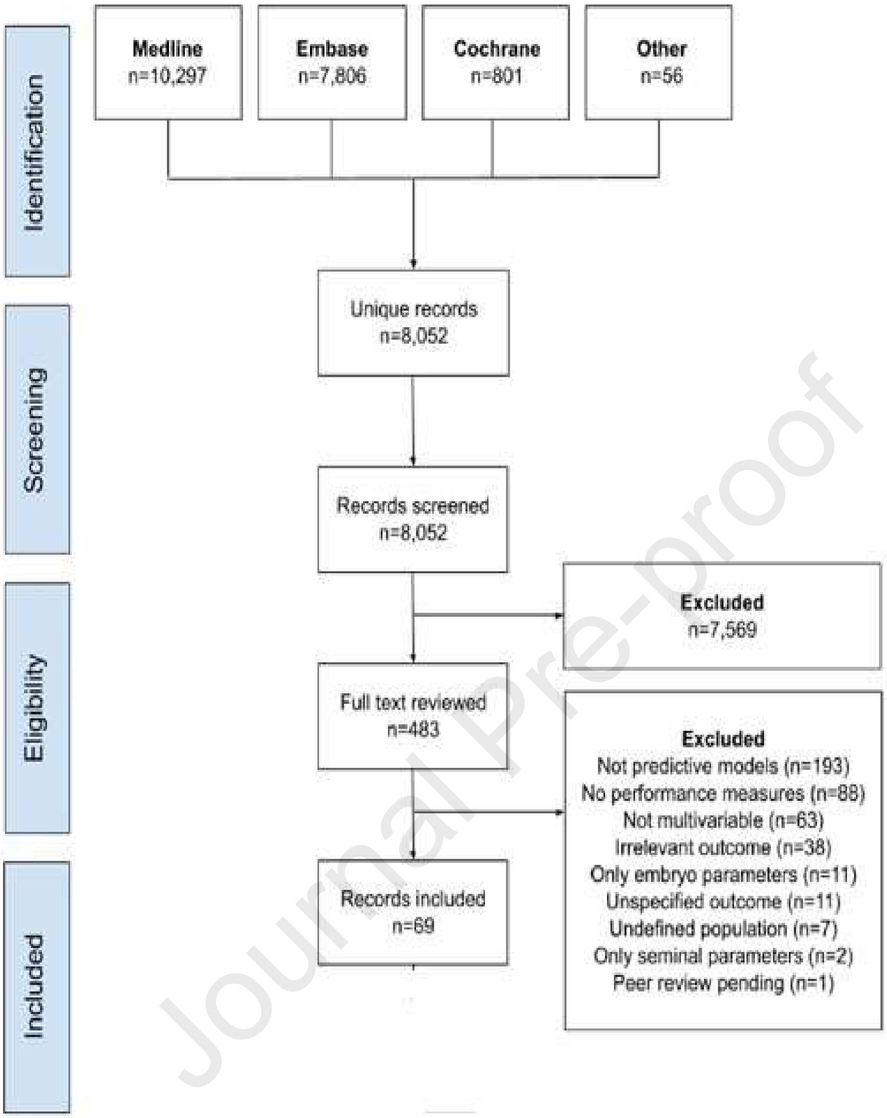
Figure (2): TRIPOD classification of included studies reporting on prediction models for reproductive outcomes following assisted reproductive technology treatments

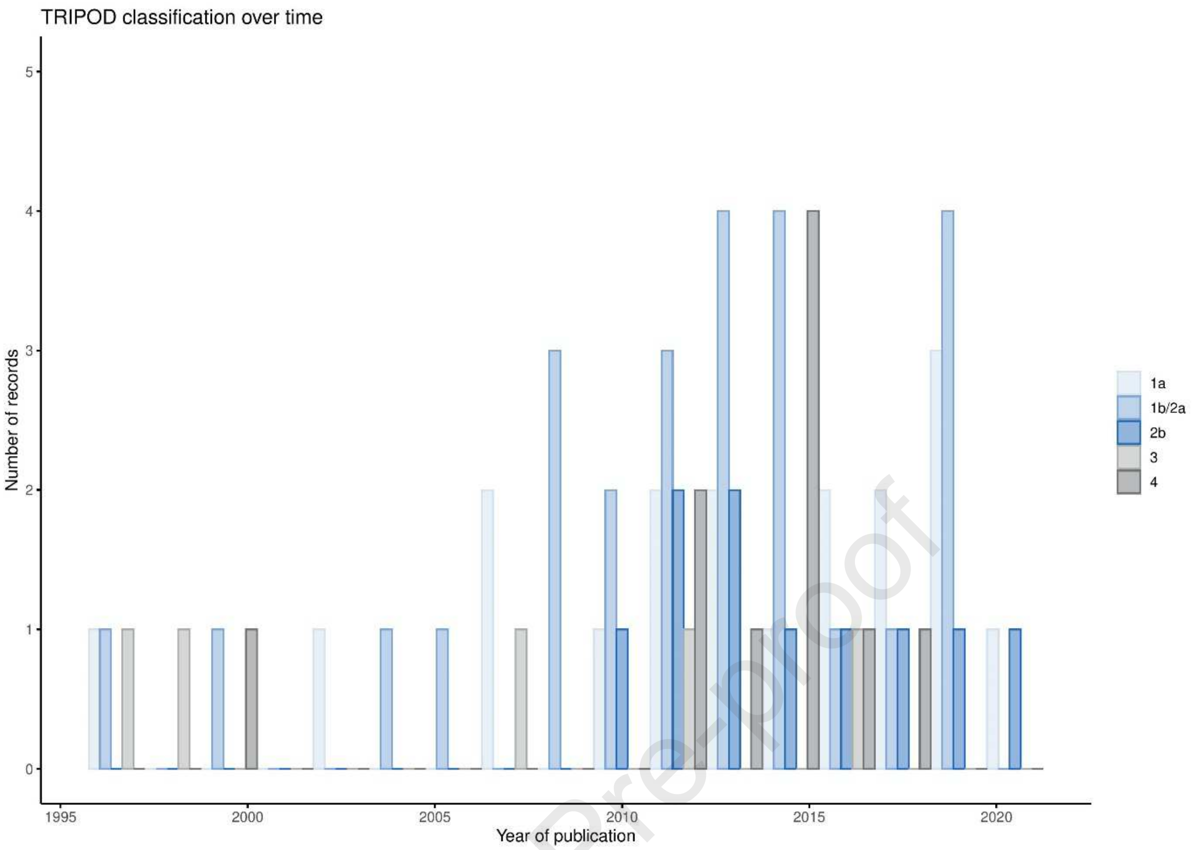
Figure (3): Predictors used in the development of prediction models for reproductive outcomes following assisted reproductive technology treatments.

3a: predictors in pre-ART treatment models

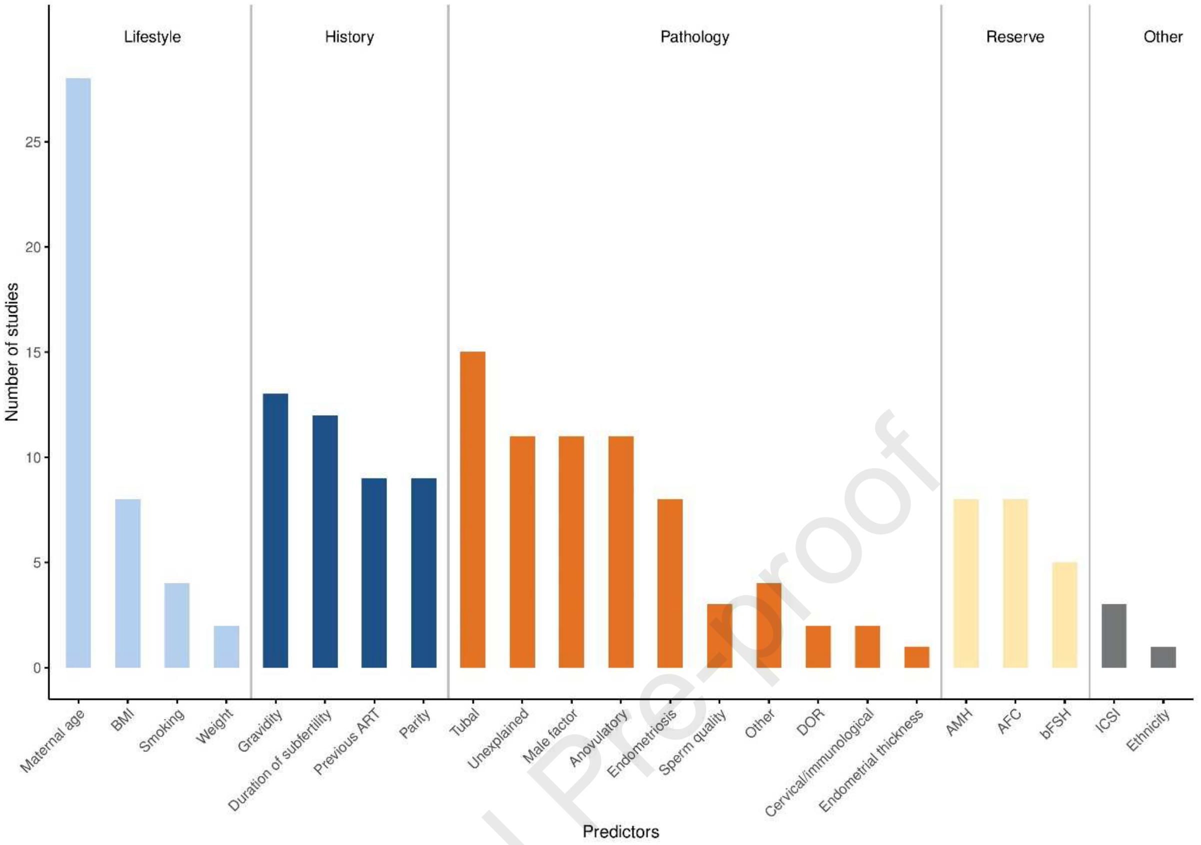
3b: predictors for intra-ART treatment models

Figure (4): Risk of bias assessment in included studies reporting on prediction models for reproductive outcomes following assisted reproductive technology treatments

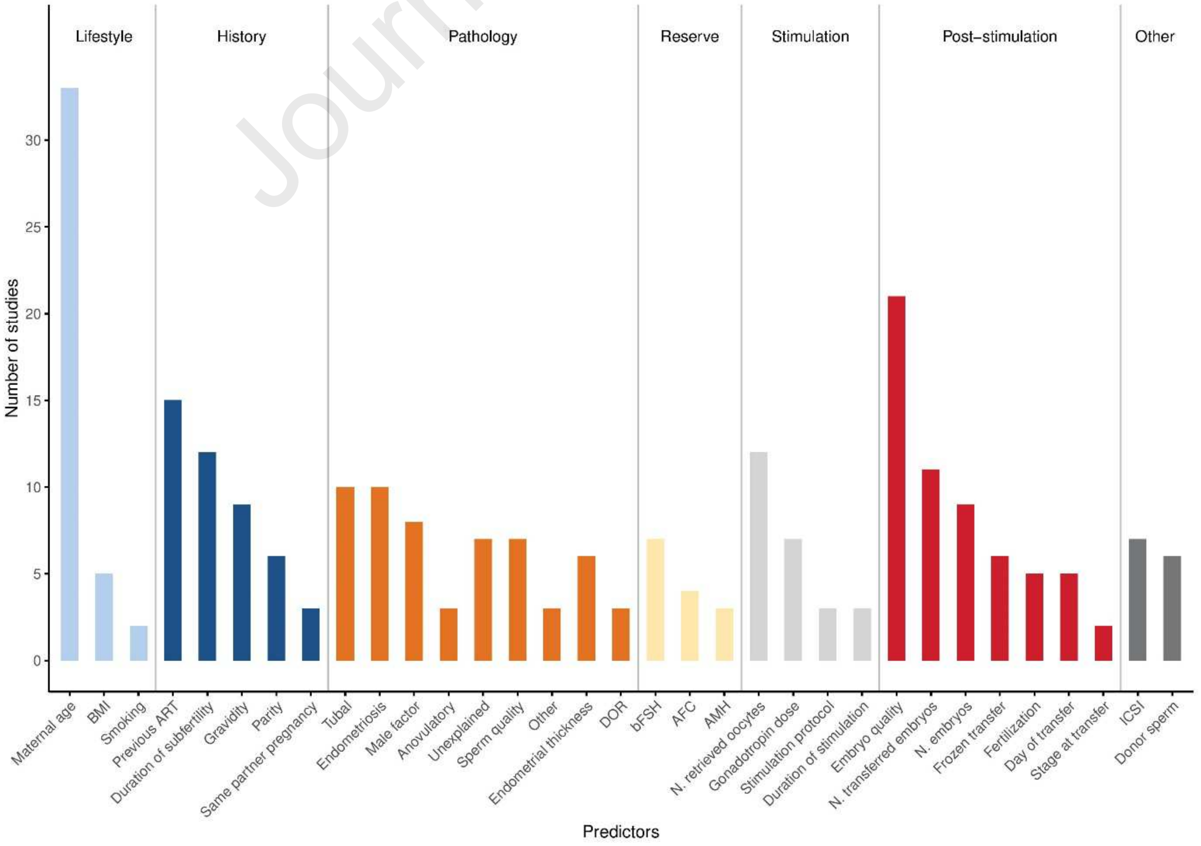

Identification
Screening
Eligibility
Included
Medline
n=10,297
Embase
n=7,806
Cochrane
n=801
Other
n=56
Unique records
n=8,052
Records screened
n=8,052
Excluded
n=7,569
Full text reviewed
n=483
Excluded
Not predictive models (n=193)
No performance measures (n=88)
Not multivariable (n=63)
Irrelevant outcome (n=38)
Only embryo parameters (n=11)
Unspecified outcome (n=11)
Undefined population (n=7)
Only seminal parameters (n=2)
Peer review pending (n=1)
Records included
n=69

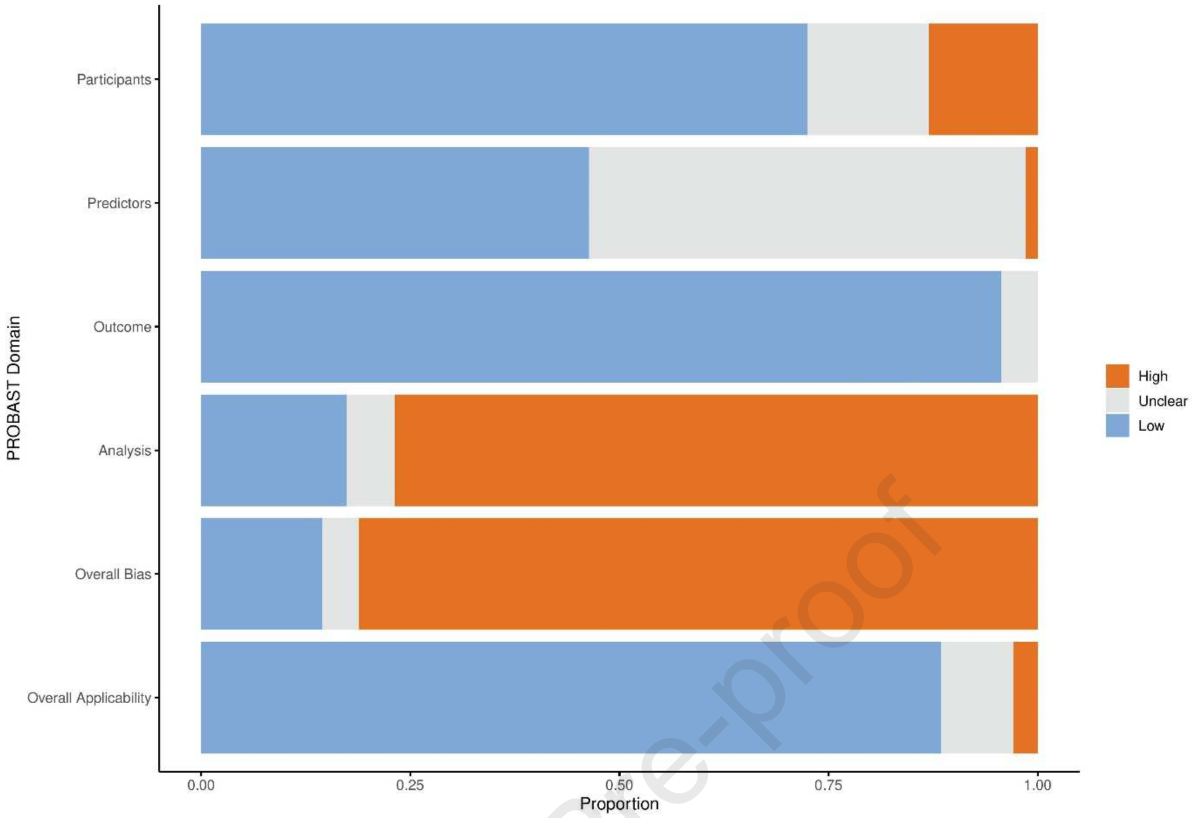

TRIPOD classification over time
Number of records
0
1
2
3
4
5
1995
2000
2005
2010
2015
2020
Year of publication
1a
1b/2a
2b
3
4

3a


Lifestyle
History
Pathology
Reserve
Other
Number of studies
0
5
10
15
20
25
Maternal age
BMI
Smoking
Weight
Gravidity
Duration of subfertility
Previous ART
Parity
Tubal
Unexplained
Male factor
Anovulatory
Endometriosis
Sperm quality
Other
DOR
Cervical/immunological
Endometrial thickness
AMH
AFC
bFSH
ICSI
Ethnicity
Predictors

3b


Lifestyle
History
Pathology
Reserve
Stimulation
Post-stimulation
Other
Number of studies
0
5
10
15
20
25
30
Maternal age
BMI
Smoking
Previous ART
Duration of subfertility
Gravidity
Parity
Same partner pregnancy
Tubal
Endometriosis
Male factor
Anovulatory
Unexplained
Sperm quality
Other
Endometrial thickness
DOR
bFSH
AFC
AMH
N. retrieved oocytes
Gonadotropin dose
Stimulation protocol
Duration of stimulation
Embryo quality
N. transferred embryos
N. embryos
Frozen transfer
Fertilization
Day of transfer
Stage at transfer
ICSI
Donor sperm
Predictors


PROBAST Domain
Participants
Predictors
Outcome
Analysis
Overall Bias
Overall Applicability
0.00
0.25
0.50
0.75
1.00
Proportion
High
Unclear
Low