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

Please cite this article as: Henderson I, Rimmer MP, Keay SD, Sutcliffe P, Khan KS, Yasmin E, Al Wattar BH, Predicting the outcomes of assisted reproductive technology treatments: A systematic review and quality assessment of prediction models, F&S Reviews (2021), doi: https://doi.org/10.1016/j.xfnr.2020.11.002.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 American Society for Reproductive Medicine. Published by Elsevier Inc. All rights reserved.
Predicting the outcomes of assisted reproductive technology treatments: A systematic review and quality assessment of prediction models

Ian Henderson MSc\textsuperscript{1,2}, Michael P Rimmer MSc\textsuperscript{3}, Stephen D Keay FRCOG\textsuperscript{2}, Paul Sutcliffe PhD\textsuperscript{1}, Khalid S Khan FRCOG\textsuperscript{4}, Ephia Yasmin PhD\textsuperscript{5}, Bassel H.A. Wattar PhD\textsuperscript{1,2,5}

\textsuperscript{1}Warwick Medical School, Warwick University, Coventry, UK.
\textsuperscript{2}Centre for Reproductive medicine, University Hospital Coventry and Warwickshire, Clifford Bridge Road, Coventry, UK.
\textsuperscript{3}MRC Centre for Reproductive Health, Queens Medical Research Institute, Edinburgh BioQuarter, University of Edinburgh, UK.
\textsuperscript{4}Department of Preventive Medicine and Public Health, University of Granada, 18071 Granada, Spain.
\textsuperscript{5}Reproductive medicine unit, University College London Hospitals, London, UK.

Corresponding author: Bassel H.A. 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
(171) 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:


10. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-

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.


PROBAST: A tool to assess the risk of bias and applicability of prediction model

15. Cook NR. Use and misuse of the receiver operating characteristic curve in risk

16. HOSMER DW, HOSMER T, CESSIE S LE, LEMESHOW S. A COMPARISON OF
GOODNESS-OF-FIT TESTS FOR THE LOGISTIC REGRESSION MODEL. Stat

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

al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction

al. Transparent reporting of a multivariable prediction model for individual prognosis


vitro fertilization (IVF): a new outcomes counselling tool for IVF providers and

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

for predicting ongoing pregnancy after in vitro fertilization and embryo transfer.

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

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

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

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


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


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;
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