
Article:
Parker, V.L. orcid.org/0000-0002-8748-4583, Winter, M.C. orcid.org/0000-0001-6192-9874, Tidy, J.A. et al. (8 more authors) (2024) PREDICT-GTN 2: Two-factor streamlined models match FIGO performance in gestational trophoblastic neoplasia. Gynecologic Oncology, 180. pp. 152-159. ISSN: 0090-8258

https://doi.org/10.1016/j.ygyno.2023.11.017

Gynecologic Oncology 180 (2024) 152–159

Contents lists available at ScienceDirect

Gynecologic Oncology

journal homepage: www.elsevier.com/locate/ygyno

PREDICT-GTN 2: Two-factor streamlined models match FIGO performance in gestational trophoblastic neoplasia

Victoria L. Parker [a,*], Matthew C. Winter [a,b], John A. Tidy [b,1], Julia E. Palmer [b], Naveed Sarwar [c], Kamaljit Singh [b], Xianne Aguiar [c], Barry W. Hancock [a], Allan A. Pacey [d], Michael J. Seckl [c,1], Robert F. Harrison [e,1]

[a] Division of Clinical Medicine, School of Medicine and Population Health, The University of Sheffield, Level 4 The Jessop Wing, Tree Root Walk, Sheffield S10 2SF, UK
[b] Sheffield Centre for Trophoblastic Disease, Weston Park Cancer Centre, Sheffield Teaching Hospitals NHS Foundation Trust, Whitham Road, Sheffield S10 2SJ, UK
[c] Gestational Trophoblastic Disease Centre, Department of Medical Oncology, Charing Cross Hospital, Fulham Palace Road, London W6 8RF, UK
[d] Faculty of Biology, Medicine and Health, Core Technology Facility, 46 Grafton Street, University of Manchester, Manchester, M13 9NT, UK
[e] Department of Automatic Control and Systems Engineering, The University of Sheffield, Mappin Street, Sheffield S1 3JD, UK

HIGHLIGHTS

- The FIGO scoring system in GTN patients can be streamlined from eight risk factors to two.
- Three logistic regression models containing two risk factors match FIGO performance across several performance measures.
- Models favoured for ongoing validation are:
- Model 2 (M2): pre-treatment hCG + site of metastases; and
- Model 3 (M3): pre-treatment hCG + number of metastases.

ARTICLE INFO

Article history:
Received 31 August 2023
Received in revised form 7 November 2023
Accepted 15 November 2023
Available online 12 December 2023

Keywords:
Gestational trophoblastic disease
Gestational trophoblastic neoplasia
FIGO
Streamline
Refine
Two-factor model

ABSTRACT

Objective. The International Federation of Gynecology and Obstetrics (FIGO) scoring system uses the sum of eight risk-factors to predict single-agent chemotherapy resistance in Gestational Trophoblastic Neoplasia (GTN). To improve ease of use, this study aimed to generate: (i) streamlined models that match FIGO performance and; (ii) visual-decision aids (nomograms) for guiding management.

Methods. Using training ($n = 4191$) and validation datasets ($n = 144$) of GTN patients from two UK specialist centres, logistic regression analysis generated two-factor models for cross-validation and exploration. Performance was assessed using true and false positive rate, positive and negative predictive values, Bland-Altman calibration plots, receiver operating characteristic (ROC) curves, decision-curve analysis (DCA) and contingency tables. Nomograms were developed from estimated model parameters and performance cross-checked upon the training and validation dataset.

Results. Three streamlined, two-factor models were selected for analysis: (i) M1, pre-treatment hCG + history of failed chemotherapy; (ii) M2, pre-treatment hCG + site of metastases and; (iii) M3, pre-treatment hCG + number of metastases. Using both training and validation datasets, these models showed no evidence of significant discordance from FIGO (McNemar's test $p > 0.78$) or across a range of performance parameters. This behaviour was maintained when applying algorithms simulating the logic of the nomograms.

Conclusions. Our streamlined models could be used to assess GTN patients and replace FIGO, statistically matching performance. Given the importance of imaging parameters in guiding treatment, M2 and M3 are favoured for ongoing validation. In resource-poor countries, where access to specialist centres is problematic, M1 could be pragmatically implemented. Further prospective validation on a larger cohort is recommended.

* Corresponding author at: Division of Clinical Medicine, School of Medicine and Population Health, The University of Sheffield, Level 4 The Jessop Wing, Tree Root Walk, Sheffield S10 2SF, UK.
E-mail address: v.parker@sheffield.ac.uk (V.L. Parker).
[1] Joint senior authors.

https://doi.org/10.1016/j.ygyno.2023.11.017

1. Introduction

The International Federation of Gynecology and Obstetrics scoring system (FIGO) is commonly used to determine the risk of primary, single-agent chemotherapy resistance to methotrexate or actinomycin D and select between these versus multi-agent chemotherapy in patients diagnosed with Gestational Trophoblastic Neoplasia (GTN). The scoring system uses the sum of eight prognostic risk-factor scores: (i) maternal age; (ii) type of antecedent pregnancy; (iii) time interval between the end of the index pregnancy and treatment start; (iv) pre-treatment level of human chorionic gonadotrophin (hCG); (v) size of the largest tumour; (vi) site of metastases; (vii) number of metastases; and (viii) a history of previous failed chemotherapy in the treatment of GTN [1]. In the United Kingdom (U.K.), low-risk patients (total score ≤ 6) receive primary intramuscular Methotrexate, whereas high-risk patients (total score ≥ 7) receive first-line combination treatment, usually EMA-CO (intravenous Etoposide, Methotrexate, Actinomycin D/Cyclophosphamide and Vincristine) [2]. The scoring system cannot be used for rare histological subtypes of GTN, including Placental site- (PSTT) and Epithelioid trophoblastic tumours (ETT) owing to the differing behaviours of these tumours [2–5].

Historically, numerous differing scoring systems were used internationally, to assess GTN patients, favouring predominantly histological, anatomical or clinical risk-factors [6]. Based upon expert clinical opinion, the FIGO prognostic scoring system was formulated in 2000; designed to harmonise classification, facilitate data comparison and reduce variability in the management of patients diagnosed with GTN [7]. However, prior to their inclusion, none of the risk-factors were subject to rigorous statistical retrospective or prospective validation, in part owing to the rare nature of the disease and small patient numbers. The revised system of 2000 was only tested retrospectively upon a small patient cohort ($n = 201$) from the Sheffield Trophoblastic Disease Centre, U.K (STDC) [8]. Therefore, perhaps unsurprisingly, FIGO is imperfect, with 25–35% patients overall and 75–80% of those scoring 5 or 6 being resistant to primary, single-agent chemotherapy [2,9–11].

Attempts to improve upon FIGO performance through statistical modelling of the available data are flawed, because FIGO is used a priori to guide primary chemotherapy treatment. Consequently, low- and high-risk patients cannot truly be compared having received different first-line agents. Furthermore, any apparent 'improvements' over FIGO performance would re-categorise low-risk patients that are resistant to single-agent treatment to the high-risk group. However, approximately 50% of those re-categorised would be overtreated and unnecessarily exposed to toxic, multi-agent chemotherapy, which is preferable to avoid given the young patient population [12,13].

Clinicians have therefore scrutinised the eight risk-factors within FIGO scoring to assess their individual prognostic significance and determine whether the system can be refined. In 2017, a five-factor logistic regression prognostic model comprising age, antecedent pregnancy, interval, pre-treatment hCG and number of metastases was proposed to match the performance of the FIGO score, developed using 793 low- and high-risk patients treated at the Charing Cross Trophoblastic Disease Centre, U.K. (CCTDC), however the system is yet to be externally or prospectively validated [14]. In this study, we combine data from the two U.K. specialist treatment centres (STDC and CCTDC) to create what we believe to be the world's largest dataset of GTN patients in order to generate: (i) streamlined models that statistically match FIGO performance and; (ii) visual-decision aids (nomograms) for guiding the management of GTN patients.

2. Methods

2.1. Data collection

Patients diagnosed with GTN were identified from the University of Sheffield and National Health Service (NHS) registries of patients maintained by STDC (February 1973–July 2019) and CCTDC (August 1958–July 2019) containing 1294 and 4393 patients respectively. This formed the training dataset. A validation dataset consisted of $n = 100$ patients treated at CCTDC (August 2019–August 2020) and $n = 44$ STDC (May 2019–December 2020) (Fig. 1). Patients were included if they had: (i) a diagnosis of GTN; (ii) received treatment (chemotherapy or additional surgery) for GTN beyond the initial uterine evacuation(s); (iii) a full complement of scored and raw data (where possible) for the eight-prognostic risk-factors constituting the FIGO score; (iv) details of the primary chemotherapy received; and (v) the response to primary chemotherapy (treatment resistance (TR) versus complete response (CR)). TR to primary chemotherapy was defined as a rise in two or more serial serum hCG levels over four weeks, or three or more consecutive hCG readings that did not fall as expected (by approximately 25%) over the same time period [11]. Excluded data included patients with: (i) duplicate data entries; (ii) histology inconsistent with a diagnosis of Gestational Trophoblastic Disease following review by expert pathologists; (iii) rare histological subtypes of GTN including PSTT, ETT or placental-site nodule (PSN); and (iv) a risk category that changed following data cleaning and checking (Fig. 1).

Each dataset was extensively and independently cleaned and checked by two individuals (VP, RF) to ensure complete coverage. This involved identifying and correcting where possible, non-sense (e.g., words/inappropriate numbers written in the scored or raw data columns) or human-error data entries (e.g., incorrect score calculated based upon the raw data) and populating missing data. To achieve this, the datasets were cross-referenced against additional history and treatment information held by the centres, and where necessary, NHS records were consulted. This was done for all included patients to ensure data integrity. Where discrepancies occurred, the total FIGO score was re-calculated using the 'checked' data and used in subsequent analyses. Patients whose FIGO risk-category (low- or high-risk) changed as a result were excluded from the analysis, because their treatment decisions had been made using the 'original' data. Scored data was available for all eight risk-factors and raw data for three parameters: (i) maternal age; (ii) time interval (in months) between the end of the index pregnancy and date of treatment start (defining a month as 28 days); and (iii) pre-treatment hCG level.

2.2. Diagnosis, treatment and follow-up

The methodology for the diagnosis, treatment and follow-up of GTN patients was as previously described [15]. Over time, imaging modalities used to stage GTN patients has changed. In the 1960s, pelvic disease was assessed with arteriography and replaced by ultrasonography in the 1970s. Similarly, chest disease was assessed by Chest X-Ray alone until the 1970s, after which chest computed tomography (CT) imaging has been used to clarify equivocal disease on chest X-Ray. Only lesions >1 cm in diameter on either chest X-Ray or CT are counted within the FIGO score. In the 1970s, contrast-enhanced CT head imaging was introduced for patients with lung metastases, being replaced by magnetic resonance imaging (MRI) brain in 1970s.

2.3. Statistical analysis

Multivariate logistic regression (LR) analysis of the combined datasets from the two centres was used to assess the relative importance of the eight FIGO risk-factors in the prediction of primary chemotherapy resistance. We adopted a bottom-up strategy, exhaustively searching all one- then two-factor models, terminating when a match was found for FIGO. Five-fold cross validated (5FCV) performance was used to avoid over-specialization during model selection [16]. Specifically, the dataset was randomly divided into five non-overlapping subgroups of approximately equal sample size and stratified for prevalence of TR; using four subgroups as a training set and one for validation, repeating the process five times. Each stage of 5FCV resulted in a new

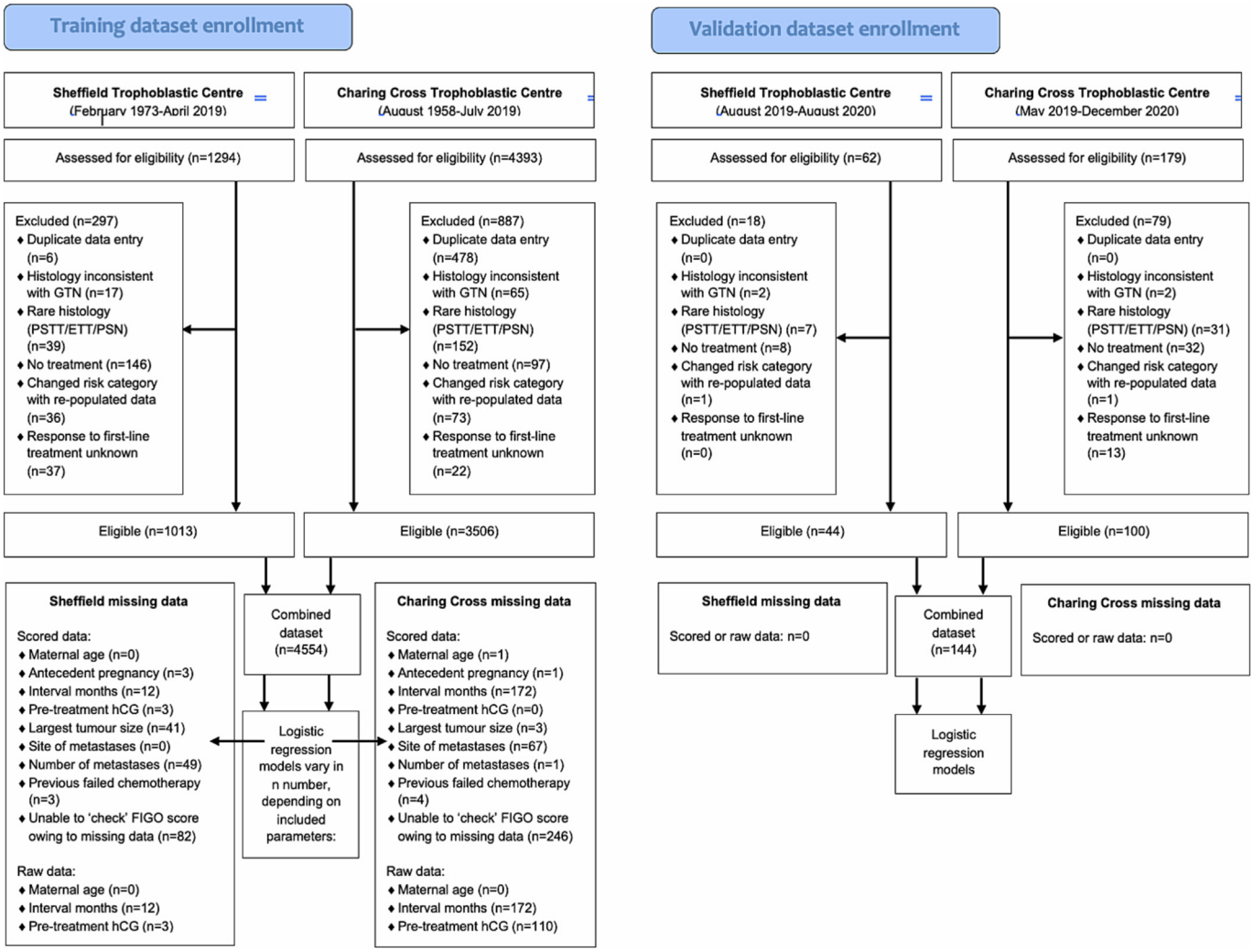

Fig. 1. CONSORT diagram. GTN, Gestational Trophoblastic Neoplasia; PSTT, Placental site trophoblastic tumour; ETT, Epithelioid trophoblastic tumour; PSN, Placental site nodule; hCG, human chorionic gonadotrophin; FIGO, International Federation of Gynecology and Obstetrics.

model estimate. For LR models, the underlying linear relationship between the log odds of TR and its predictors permitted the averaging of the (five sets of) model parameters, yielding a single, averaged model. The variability of the parameter estimates about this average was quantified in their coefficient of variation (CoV), defined as the standard deviation divided by the mean, for each parameter. The lower the CoV, the greater the confidence that the averaged model is representative of the dataset.

2.4. Streamlined model selection

To produce a streamlined model that generated a statistical, but not necessarily a case-by-case match to FIGO, the operational value of the false positive rate (FPR) was fixed to equal that of FIGO (11.9%). Searching the 57 5FCV one- and two-factor models meeting the inclusion criteria, five two-factor candidates were found (Supplementary Table 1, online only). The inclusion criteria were: (i) CoV ≤4% for each included parameter; and (ii) a difference in opinion between the model and FIGO categorisation of <500 cases. The five models were subjected to a further 50 repeats of the 5FCV process to investigate whether the models were reliably independent of different splits in the data. Three models were repeatedly selected and included for ongoing investigation: (i) log pre-treatment hCG (raw data) + previous failed chemotherapy (scored data); (ii) log pre-treatment hCG (raw data) + site of metastases (scored data); and (iii) log pre-treatment hCG (raw data) + number of metastases (scored data) (Supplementary Table 1, online only).

To ensure the LR models matched FIGO, their performance was assessed using a combination of conventional measures such as true (TPR) and false positive rates (FPR), positive (PPV) and negative predictive values (NPV), where a positive was defined as TR to primary chemotherapy. Additional measures included the area under (AUC) the receiver operating characteristic (ROC) curve and decision curve analyses (DCA), the latter to evaluate whether the new model would do more harm than good. DCA compared the net benefit of each model to FIGO and default strategies of treating all patients as high-risk (TAHR) or treating all patients as low-risk (TALR) over a range of clinically applicable probability thresholds [17,18]. Net benefit was calculated as: TPR x Prevalence – (1-FPR) x (1-Prevalence) x (Pt/1-Pt) where Pt was the threshold probability of disease for taking action, in this scenario, administering high-risk treatment [19]. Finally, the correspondence between FIGO and the streamlined models in terms of the categorisation to low- and high-risk groups was assessed, with a detailed analysis of the number of patients that would have changed risk category (low-to-high-risk or vice versa) and been over- or under-treated as a result as determined by primary treatment response. Calibration of the models was assessed using Bland-Altman plots of the observed and predicted deciles of class probability [20].

The three streamlined models were validated using an independent cohort of patients from the STDC and CCTDC as described above. To assist future decision making in clinical practice, nomograms were developed from the estimated model parameters. Their performance was cross-checked upon the training and validation dataset using algorithms simulating the logic of the nomograms.

A Chi-squared test compared the rates of TR versus CR within the two datasets, whilst McNemar's test determined whether the models matched FIGO at a given probability threshold. Statistical analyses were performed in GraphPad Prism (version 9.0.0, San Diego, CA, USA) and MatLab (version R2020a, Natick, MA, USA).

3. Results

3.1. Demographics

The combined STDC and CCTDC dataset included 4554 patients who satisfied the inclusion and exclusion criteria, of which 4191 patients were suitable for ongoing analysis (Fig. 1). A summary of the two datasets is provided in Supplementary Fig. 1 and Supplementary Table 2 (online only). Comparing the two centres, the proportion of TR versus CR patients was not significantly different (Chi-squared test, $p = 0.09$). Similarly, the performance of FIGO was comparable between sites, revealing poor sensitivity and high specificity, with PPV < NPV (Supplementary Table 3, online only).

3.2. Model development

Following step-wise exploration of all one and two factor models satisfying the inclusion criteria and 50 repeats of the 5FCV process, three models were repeatedly selected for ongoing investigation. Supplementary Table 4 summarises the estimated model coefficients, significance and CoV of the variables included within each streamlined model. As expected, each variable proved highly significant, justifying inclusion within the predictive model. The three models matched FIGO performance with respect to TPR, FPR, PPV, NPV and DOR (Table 1).

Model performance was assessed using a variety of measures to ensure a consistent match to FIGO. As a baseline, FIGO has an AUC of 0.62. Calibration analysis revealed the difference between observed and predicted frequencies to lie within the 95% limits of agreement (Fig. 2D). Each streamlined model matched FIGO in terms of the curve shape and AUC (Figs. 2A, 3A, 4A), calibration (Figs. 2B, 3B, 4B) and DCA at a decision probability threshold equivalent to a total FIGO score = 7, the decision point between low- and high-risk groups (Fig. 2C, 3C, 4C).

Contingency table analysis revealed no significant discordance between FIGO and the models (McNemar's test $p > 0.78$), with a disagreement in the classification of patients to low- or high- risk groups affecting ≤10% patients. In summary, the overall disparity between FIGO and M1 involved six patients, whilst no disparity was observed between FIGO versus M2 or M3 (Supplementary Table 5, online only).

3.3. Model validation

Validation using the independent dataset ($n = 144$) revealed that performance was maintained, matching FIGO. Table 1 details the validation dataset performance of using standard measures, whilst Supplementary Fig. 2, 3, 4 (online only) confirm that FIGO performance was matched upon ROC, calibration analysis and DCA. Finally, contingency table analysis showed close correspondence between FIGO and all models, with a disparity of eight cases when applying M1 ($p = 0.04$), five cases using M2 ($p = 0.21$) and three cases with M3 ($p = 0.29$, McNemar's test) (Supplementary Table 6, online only). Supplementary Table 7 (online only) provides a more detailed analysis of the theoretical effect (under- or over-treatment) upon patients that would have changed risk category (low- to high-risk or vice versa) by applying the model. Two patients would have been over-treated by applying M2 and none using M1 and 3. Three patients would have been under-treated using M1 and M2, and two, applying M3. Finally, using both the training and validation dataset, performance was cross-checked with the algorithm simulating the logic of the nomograms (Fig. 5) and found to be equivalent.

4. Discussion

FIGO cannot be improved through statistical modelling [13], hence any attempts to improve it must focus upon streamlining this eight-factor prognostic system to increase ease of use, reliability and efficiency. Using exhaustive search, three streamlined LR models, each containing two variables were selected for exploration. All three models involved raw data for pre-treatment hCG, with M2 and M3 combining this parameter with imaging-based factors, specifically the site and number of metastases. M1 requires no imaging investigations, instead combining pre-treatment hCG with a history of failed chemotherapy (Table 1). The models matched FIGO performance across a range of measures including conventional parameters (TPR, FPR, PPV and NPV) (Table 1), ROC characteristics, DCA, calibration (Fig. 2,3 and 4) and contingency table analysis (Supplementary Table 5, online only). Crucially, model performance was sustained across these parameters when validated upon an independent dataset ($n = 144$) (Table 1, Supplementary Table 6, Fig. 2,3 and 4, online only). Risk change analysis of the validation cohort was equally supportive; with the theoretical over-treatment of two patients using M2 and no patients applying M1 or M3 (Supplementary Table 7, online only). Three visual nomograms were produced to aid and simplify clinical decision making (Fig. 5). Performance was cross checked upon algorithms simulating the logic of the nomograms and found to be equal.

Streamlined models have many benefits. A simplified scoring system is easier and quicker for clinicians to use, with less room for human error in data entry and score calculation. Within the computerised datasets examined here, data entry errors were noteworthy. Inaccurate, missing data entries, non-sense information (e.g., words in a scored/raw data column) and failure to sum the scores correctly changed the total

Table 1
Performance of FIGO and the streamlined logistic regression models in the training ($n = 4191$) and validation datasets ($n = 144$).

Model description	TPR (%)	FPR (%)	PPV (%)	NPV (%)	DOR
FIGO	17.10	11.90	44.40	65.60	1.50
log pre-treatment hCG (raw data) + previous failed chemotherapy (scored data)	17.50	11.90	44.90	65.70	1.60
log pre-treatment hCG (raw data) + site of metastases (scored data)	17.10	11.90	44.40	65.60	1.50
log pre-treatment hCG (raw data) + number of metastases (scored data)	17.10	11.90	44.40	65.60	1.50
VALIDATION DATASET					
FIGO	14.50	12.80	42.10	61.50	1.20
log pre-treatment hCG (raw data) + previous failed chemotherapy (scored data)	21.80	8.10	63.20	64.80	3.10
log pre-treatment hCG (raw data) + site of metastases (scored data)	21.80	11.60	54.50	63.90	2.10
log pre-treatment hCG (raw data) + number of metastases (scored data)	12.70	8.10	50.00	62.20	1.60

FIGO, International Federation of Gynecology and Obstetrics scoring system; log, logarithm; hCG, human chorionic gonadotrophin; TPR, true positive rate; FPR, false positive rate; PPV, positive predictive value; NPV, negative predictive value; DOR, diagnostic odds ratio.

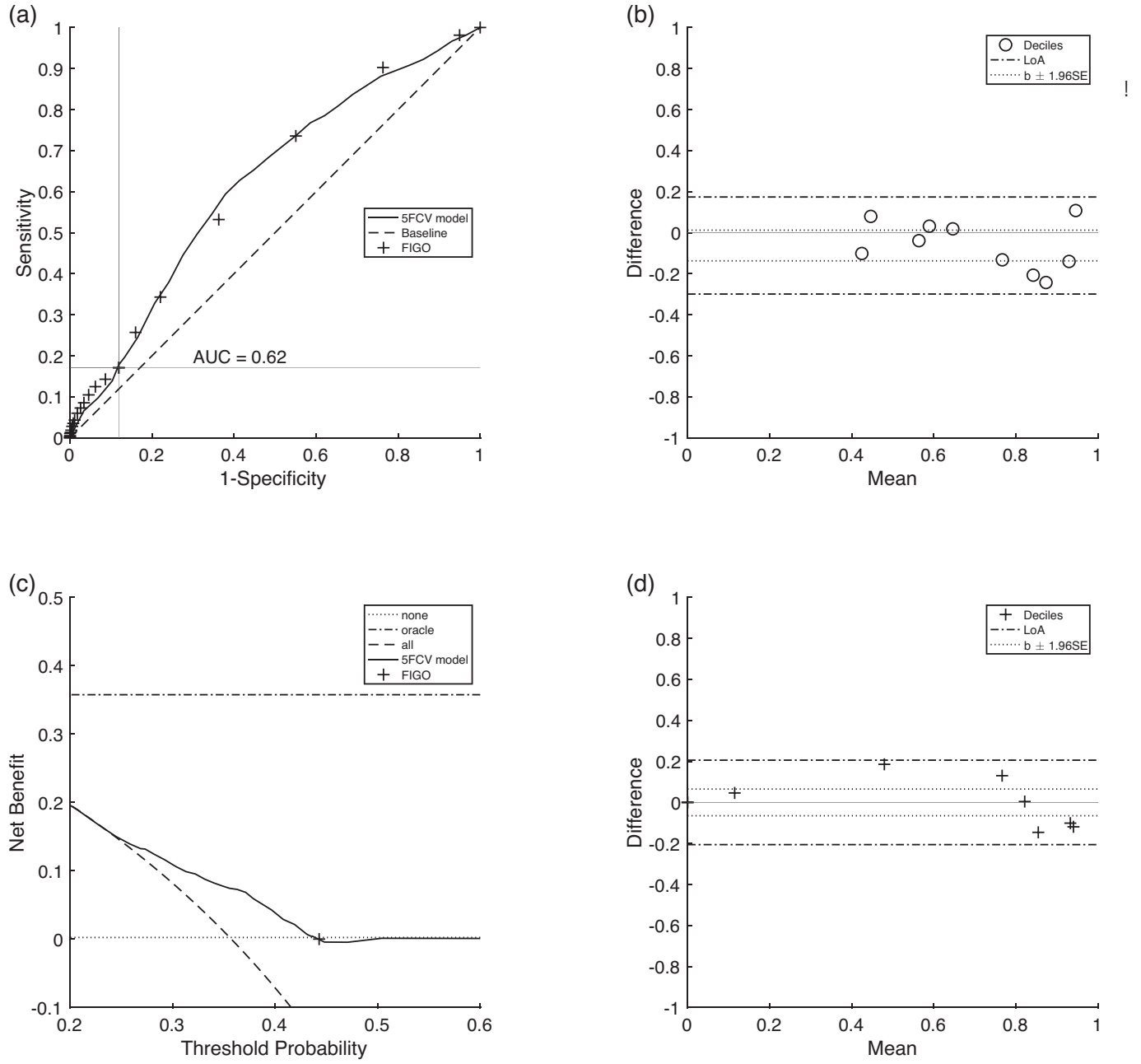

Fig. 2. Training dataset performance of M1: log pre-treatment hCG (raw data) + previous failed chemotherapy (scored data) versus FIGO ($n = 4191$). (a) Receiver Operating Characteristics (ROC) Curve comparing M1 with FIGO. (b) Bland-Altman Calibration plots for M1. Hypothesis tests confirmed non-significance of the least squares slope $p = 0.05$. (c) Decision Curve Analysis (DCA) for M1. Dotted line (treat all patients as low risk (TALR)), the net benefit assuming that no GTN patients will have the outcome (resistance to primary chemotherapy); chained line (oracle), the net benefit associated with a perfect prediction model; dashed line (treat all patients as high-risk (TAHR)), net benefit assuming that all GTN patients will have the outcome; solid line (M1), net benefit when we manage GTN patients according to the predicted risk of the outcome (primary chemotherapy resistance). + represents FIGO performance at a total score = 7 (decision point between low- and high-risk GTN). (d) Bland-Altman Calibration plot for FIGO. Hypothesis tests confirmed non-significance of the least squares slope, p = 0.05. Data shown is the 5FCV result. FIGO, International Federation of Gynecology and Obstetrics scoring system; AUC, area under the curve; CV, cross-validated; LoA, limits of agreement; b, bias; SE, standard error.

score in 1734 patients (41% of the dataset) and of concern, the risk categorisation in 109 patients (2.6% of the dataset), which would have led to different primary treatment. Whilst these may have been data transcription errors, transferring from the paper to computerised format, it is evident that the principle of parsimony holds here: the simplest method is usually the correct one [21]. Scoring systems like FIGO are commonly used in other areas of clinical practice, with the Early Warning System (EWS) as an example, designed to alert clinicians of patient deterioration. However, these systems are highly susceptible to human error, and alarmingly, EWS, studies have shown that approximately one third of scores are incomplete [22] with error rates up to 36% [23,24]. No published data exists for FIGO. Weighted or score-based systems are particularly problematic through incomplete data collection, misassignment of the correct score to the raw data and total score calculation [23]. Given the serious implications upon patient care, automated, computer-based calculators are advocated and becoming increasingly prevalent in clinical practice [25,26].

Streamlined models lend themselves to the development of visual decision aids, such as nomograms (Fig. 5), which ease and enhance the accuracy of the scoring process and provide a visual representation

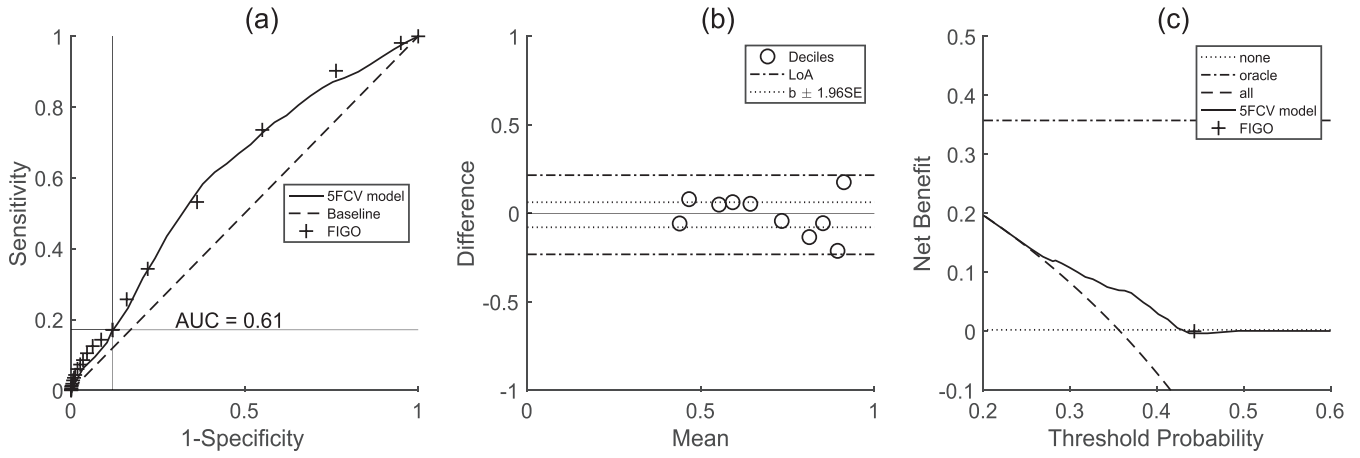

Fig. 3. Training dataset performance of M2: log pre-treatment hCG (raw data) + site of metastases (scored data) (n = 4191). (a) Receiver Operating Characteristics (ROC) Curve comparing M2 with FIGO. (b) Bland-Altman Calibration plots for M2. Hypothesis tests confirmed non-significance of the least squares slope p = 0.05. (c) Decision Curve Analysis (DCA) for M2. Dotted line (treat all patients as low risk (TALR)), the net benefit assuming that no GTN patients will have the outcome (resistance to primary chemotherapy); chained line (oracle), the net benefit associated with a perfect prediction model; dashed line (treat all patients as high-risk (TAHR)), net benefit assuming that all GTN patients will have the outcome; solid line (M2), net benefit when we manage GTN patients according to the predicted risk of the outcome (primary chemotherapy resistance). + represents FIGO performance at a total score = 7 (decision point between low- and high-risk GTN). Data shown is the 5FCV result. FIGO, International Federation of Gynecology and Obstetrics scoring system; AUC, area under the curve; CV, cross-validated; LoA, limits of agreement; b, bias; SE, standard error.

of risk that can be shared with the patient [28]. This is important, given the benefits conferred by involving the patient in their healthcare discussions and decisions [29–31]. Such nomograms can be operated using raw data via an online database, reducing the scope for human error in score assignment and calculation.

There are key cost and efficiency benefits of a simplified model. This could avoid imaging parameters (M1 in this study), obviating cost and time demands for both the health service and patients. However, these imaging investigations are crucial for clinical decision making, such as guiding surgical intervention (hysterectomy, resection of metastases). However, in countries where resources or access to specialist centres is problematic, M1 may be pragmatic, with FIGO performance matched without the need for imaging investigations [33].

Our study has further streamlined FIGO, advancing upon earlier literature, which generated three five-factor models, all of which included imaging variables [14]. Of note, this study trained a LR model upon FIGO decisions (i.e., low- versus high-risk categorisation) and as such, could match FIGO performance exactly (AUC of 0.99–1.00). Dissimilarly, our study trained models upon actual outcome (TR or CR to primary chemotherapy) but selected them to match FIGO performance, which explains the differences in the AUC on ROC curve analysis. Our study used larger patient numbers and more robust modelling analysis that is closely aligned with Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [34], designed to improve the reporting of prediction models. Our study examined not only the significance and variability of the eight FIGO variables using F5CV, but assessed model performance across a wider range of parameters (conventional TPR, FPR, PPV, NPV measures, ROC curve characteristics, calibration analysis, DCA and contingency tables). The model was then validated upon an independent validation cohort, which was not used to select, re-train or re-adjust the models in any way. Moreover, our analysis has revealed the weakness of prior publications that purely relied upon ROC AUC characteristics to compare FIGO performance [14]. Here, we found AUC to provide only a gross

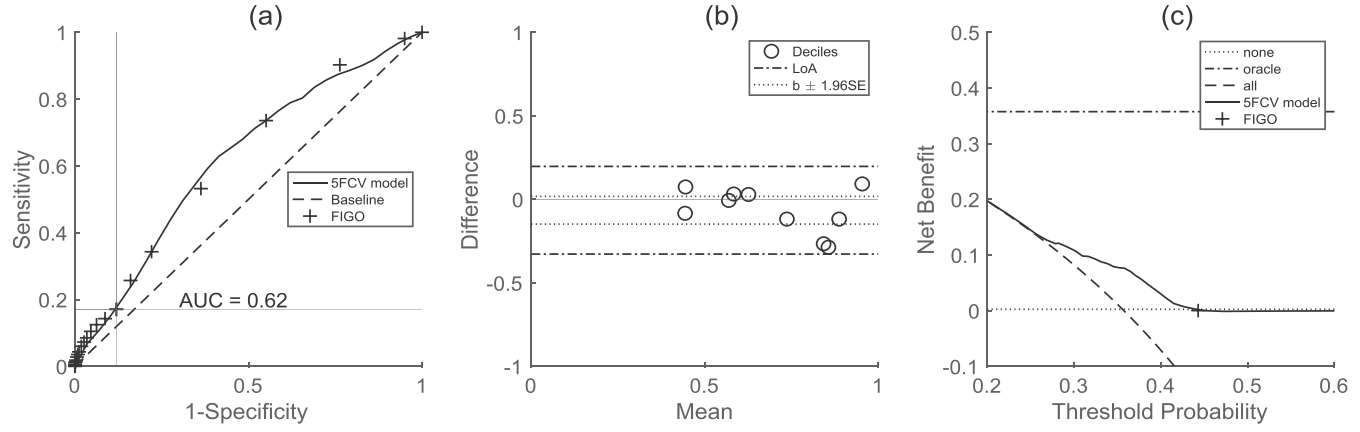

Fig. 4. Training dataset performance of M3: log pre-treatment hCG (raw data) + number of metastases (scored data) (*n* = 4191). (a) Receiver Operating Characteristics (ROC) Curve comparing M3 with FIGO. (b) Bland-Altman Calibration plots for M3. Hypothesis tests confirmed non-significance of the least squares slope p = 0.05. (c) Decision Curve Analysis (DCA) for M3. Dotted line (treat all patients as low risk (TALR)), the net benefit assuming that no GTN patients will have the outcome (resistance to primary chemotherapy); chained line (oracle), the net benefit associated with a perfect prediction model; dashed line (treat all patients as high-risk (TAHR)), net benefit assuming that all GTN patients will have the outcome; solid line (M3), net benefit when we manage GTN patients according to the predicted risk of the outcome (primary chemotherapy resistance). + represents FIGO performance at a total score = 7 (decision point between low- and high-risk GTN). Data shown is the 5FCV result. FIGO, International Federation of Gynecology and Obstetrics scoring system; AUC, area under the curve; CV, cross-validated; LoA, limits of agreement; b, bias; SE, standard error.

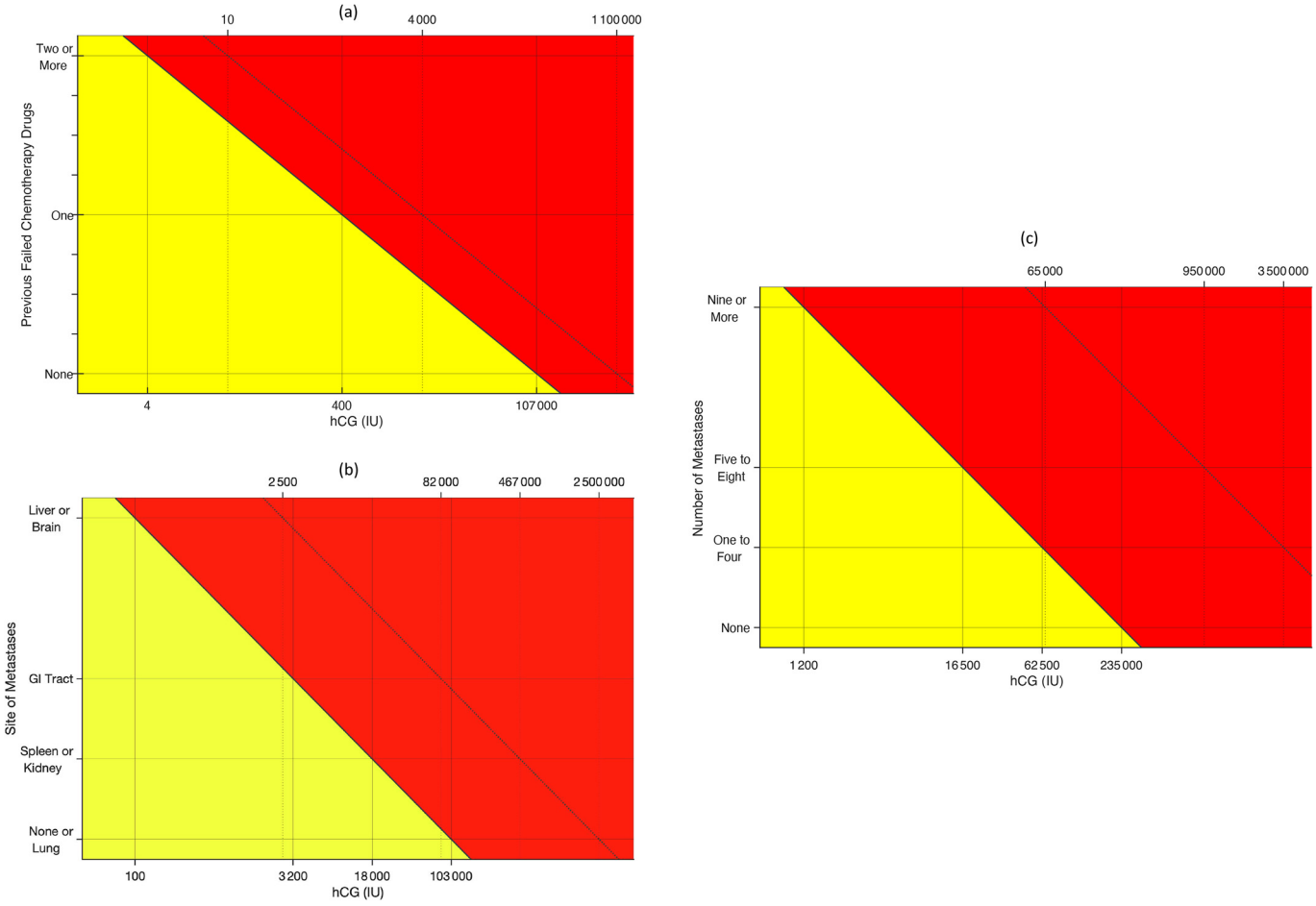

Fig. 5. Nomograms developed from the estimated model parameters. (a) M1: log pre-treatment hCG (raw data) + previous failed chemotherapy (scored data). (b) M2: log pre-treatment hCG (raw data) + site of metastases (scored data). (c) M3:log pre-treatment hCG (raw data) + number of metastases (scored data). The solid line delineates the boundary between low- (yellow) and high-risk (red) treatment, whilst the area to the right of the dotted line represents ultra-high-risk GTN patients.
hCG, human chorionic gonadotrophin; IU, international unit.

performance measure, with small variation in AUC across all explored models, despite considerable differences in other performance measures, confirming that ROC shape, especially on the left-hand-side, is crucial. Taken together, in addition to the 50 repeats of the 5FCV process performed within the model selection process, these techniques reduce the likelihood of model overfit and dependency upon a particular spilt of the data. This gives confidence that model performance should be replicated when confronted with a new cohort [35]. Indeed, the results of our validation analysis would certainly support this.

Limitations of this study include the inherent bias introduced through the retrospective nature of the study and change in imaging parameters over the study period, although all patients were rescored according to the FIGO 2000 criteria. The different hCG assays used between the centres is a further limitation, yet the current FIGO system operates internationally with different hCG assays, so this variation must be incorporated into any streamlined model. The small number of ultra-high risk GTN (total FIGO score ≥ 13) patients [4,5,36], in both the training ($n = 139$) and validation ($n = 5$) datasets limits the applicability of the model to this subgroup. The relatively small numbers within the validation cohort is a further limitation, yet due to the rare nature of GTN, accruing larger numbers would take several years. Like FIGO, a limitation of the models and nomograms herein concern decisions at the borderline between low- and high-risk treatment groups. A degree of uncertainty exists, and as such, we do not propose that our guidance is any more prescriptive than FIGO. Neither FIGO nor our models consider the response to second-line single-agent chemotherapy, hence individual patient preference and clinician experience remain of vital importance. Borderline low-risk patients should continue to be counselled regarding their likelihood of responding to single-agent chemotherapy and be offered a choice between this approach versus a primary, multi-agent regimen. Despite our large dataset, 487 patients received second-line single-agent chemotherapy, of which only $n = 51$ were TR. Accounting for second-line outcomes in a modelling strategy such as this would not only deviate from the study purpose, to match FIGO, but would prove problematic at the model building stage owing to the very low TR prevalence (10%) and small cohort size. While recognised techniques such as over-sampling could be employed to reduce variability in model estimation, cross-validated results could not be relied upon to infer prospective performance. Furthermore, there exists no independent validation set to test real-life performance.

The trade-off between over- and under-treatment must always be at the forefront of changing practice, given that GTN is an inherently curable condition [2,36], involving a young patient population with a desire for future pregnancies. Overtreatment and the consequences of this should be avoided, particularly given that ~50% of low-risk patients resistant to first-line single-agent chemotherapy achieve remission with a second-line single-agent [12]. In this regard, our streamlined models are reassuring (Supplementary Table 7, online only) and could quickly and efficiently be tested worldwide, using an online nomogram for determining treatment strategy (low- or high-risk) (Fig. 5). Given the importance of imaging investigations to guiding treatment decisions, we

recommend that M2 and M3, involving hCG and the site or number of metastases respectively, are subject to future validation. Further testing of the models and nomograms using data from worldwide GTN centres is now required to test their functionality and reproducibility across different populations and hCG assays.

Funding

The research team would like to thank Weston Park Cancer Charity, Sheffield, U.K. for funding Dr. Victoria Parker's salary as a Clinical Research Fellow with large grants (CA154 and CA184). MJS acknowledges support of the Imperial Biomedical Research Centre funded by the National Institute of Health Research (NIHR) and Experimental Cancer Medicine Centre (ECMC) supported by NIHR and Cancer Research UK. The funders had no role in study design, data collection, analysis, interpretation or writing of the report.

Author contributions

VLP and RFH conceived and designed the study. VLP collected and assembled data. VLP and RFH performed the data analysis. VLP, XA, KS, BWH, AAP, JEP, MCW, JAT, MJS and RFH interpreted the data. VLP wrote the manuscript, with editorial input from BWH, AAP, JEP, MCW, JAT, MJS and RFH. All authors approved the final version of the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the following individuals: Julie Ford and Tracey Byne (STDC) for providing administrative support; Laura Ellis, David Drew (STDC) and Terry Tin (CCTDC) for facilitating exploration and interrogation of the specialist databases; medical students Emily Press, Freya Rhodes, Scarlett Strickland and Adam Temple (The University of Sheffield) for their involvement in raw data collection; medical students Adam Foster and Bryony Cushen (The University of Sheffield) for their involvement in database cleaning.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygyno.2023.11.017.

References

[1] FIGO: current FIGO staging for cancer of the vagina, fallopian tube, ovary, and gestational trophoblastic neoplasia, Int. J. Gynaecol. Obstet. 105 (2009) 3–4.
[2] M.J. Seckl, N.J. Sebire, R.A. Fisher, et al., Gestational trophoblastic disease: ESMO clinical practice guidelines for diagnosis, treatment and follow-up, Ann. Oncol. 24 (Suppl. 6) (2013) vi39-50.
[3] B.W. Hancock, J. Tidy, Placental site trophoblastic tumour and epithelioid trophoblastic tumour, Best Pract. Res. Clin. Obstet. Gynaecol. 74 (2020 Jul) 131–148.
[4] C. Lok, N. van Trommel, L. Massuger, et al., Practical clinical guidelines of the EOTTD for treatment and referral of gestational trophoblastic disease, Eur. J. Cancer 130 (2020) 228–240, Available from: https://www.esgo.org/network/eottd/.
[5] J.A. Tidy, M. Seckl, B. Hancock, on behalf of the Royal College of Obstetricians and Gynaecologists, Management of Gestational Trophoblastic Disease: green-top guideline no. 38 - June 2020, BJOG 128 (2021) e1-e27.
[6] V.L. Parker, A.A. Pacey, J.E. Palmer, et al., Classification systems in gestational trophoblastic neoplasia - sentiment or evidenced based? Cancer Treat. Rev. 56 (2017) 47–57.
[7] FIGO Oncology Committee, FIGO staging for gestational trophoblastic neoplasia 2000. FIGO Oncology committee, Int. J. Gynaecol. Obstet. Ireland (2002) 285–287.
[8] B.W. Hancock, E.M. Welch, A.M. Gillespie, et al., A retrospective comparison of current and proposed staging and scoring systems for persistent gestational trophoblastic disease, Int. J. Gynecol. Cancer 10 (2000) 318–322.
[9] M.J. Seckl, N.J. Sebire, R.S. Berkowitz, Gestational trophoblastic disease, Lancet 376 (2010) 717–729.
[10] Winter M.C, Singh K: Gestational Trophoblastic Neoplasia: A Guide to Management at Weston Park Hospital, Sheffield Teaching Hospitals NHS Foundation Trust, 2022.
[11] M.C. Winter, J.A. Tidy, A. Hills, et al., Risk adapted single-agent dactinomycin or carboplatin for second-line treatment of methotrexate resistant low-risk gestational trophoblastic neoplasia, Gynecol. Oncol. 143 (2016) 565–570.
[12] V.L. Parker, M.C. Winter, E. Whitby, et al., Computed tomography chest imaging offers no advantage over chest X-ray in the initial assessment of gestational trophoblastic neoplasia, Br. J. Cancer 124 (2021) 1066–1071.
[13] V.L. Parker, B.W. Hancock, A.A. Pacey, et al., PREDICT-1: Improving the FIGO Prognostic Scoring System for Gestational Trophoblastic Neoplasia - Is it Possible? 2021.
[14] Y.K. Eysbouts, P.B. Ottevanger, L. Massuger, et al., Can the FIGO 2000 scoring system for gestational trophoblastic neoplasia (GTN) be simplified? A new retrospective analysis from a nationwide data-set, Ann. Oncol. 28 (8) (2017 Aug 1) 1856–1861.
[15] V.L. Parker, M.C. Winter, J.A. Tidy, et al., PREDICT-GTN 1: can we improve the FIGO scoring system in gestational trophoblastic neoplasia? Int. J. Cancer 152 (2023) 986–997.
[16] M. Stone, Cross-Validatory choice and assessment of statistical predictions, J. R. Stat. Soc. B. Methodol. 36 (1974) 111–133.
[17] A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests, BMJ i6 (2016).
[18] A.J. Vickers, B. van Calster, E.W. Steyerberg, A simple, step-by-step guide to interpreting decision curve analysis, Diagnost. Prognost. Res. 3 (2019) 18.
[19] A.J. Vickers, Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers, Am. Stat. 62 (2008) 314–320.
[20] J.M. Bland, D.G. Altman, Comparing methods of measurement: why plotting difference against standard method is misleading, Lancet 346 (1995) 1085–1087.
[21] J. Laird, The law of parsimony, Monist 29 (1919) 321–344.
[22] D.A. Clifton, L. Clifton, D.-M. Sandu, et al., 'Errors' and omissions in paper-based early warning scores: the association with changes in vital signs—a database analysis, BMJ Open 5 (2015), e007376.
[23] M. Edwards, H. McKay, C. Van Leuvan, et al., Modified early warning scores: inaccurate summation or inaccurate assignment of score? Crit. Care 14 (2010) P257.
[24] A.F. Smith, R.J. Oakey, Incidence and significance of errors in a patient 'track and trigger' system during an epidemic of Legionnaires' disease: retrospective casenote analysis, Anaesthesia 61 (2006) 222–228.
[25] J.M. Lockwood, J. Thomas, S. Martin, et al., AutoPEWS: automating pediatric early warning score calculation improves accuracy without sacrificing predictive ability, Pediatric Quality Safety 5 (2020), e274.
[26] M.A. Mohammed, R. Hayton, G. Clements, et al., Improving accuracy and efficiency of early warning scores in acute care, Br. J. Nurs. 18 (2009) 18–24.
[28] V.P. Balachandran, M. Gonen, J.J. Smith, et al., Nomograms in oncology: more than meets the eye, Lancet Oncol. 16 (2015) e173–e180.
[29] A. Coulter, A. Collins, King's Fund C, Making Shared Decision-Making a Reality: No Decision about me, without me, King's Fund, London, 2011.
[30] NICE, Shared Decision Making Collaborative: A Consensus Statement, 2016.
[31] M. Stewart, J. Belle Brown, W. Weston, et al., Patient-Centered Medicine: Transforming the Clinical Method, Sage Publications, Thousand Oaks, 1995.
[33] V.L. Parker, M.J. Seckl, B.W. Hancock, Global differences in management and treatment: a critical appraisal from a UK perspective, Gestat, Trophobl. Disease Chapter 21, 1-33 (2021). Available from: https://isstd.org/gtd-book.html
[34] K.G.M. Moons, D.G. Altman, J.B. Reitsma, et al., Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration, Ann. Intern. Med. 162 (2015) W1.
[35] J.M.G. Taylor, D.P. Ankerst, R.R. Andridge, Validation of biomarker-based risk prediction models, Clin. Cancer Res. 14 (2008) 5977–5983.
[36] H.Y.S. Ngan, M.J. Seckl, R.S. Berkowitz, et al., Update on the diagnosis and management of gestational trophoblastic disease, Int. J. Gynaecol. Obstet. 143 (Suppl. 2) (2018) 79–85.