



Prediction model for testis histology in men with non-obstructive azoospermia: evidence for a limited predictive role of serum follicle-stimulating hormone

Ettore Caroppo^{1,2} · Elisabetta M. Colpi³ · Giuseppe D'Amato¹ · Giacomo Gazzano⁴ · Giovanni M. Colpi^{3,5}

Received: 9 August 2019 / Accepted: 11 October 2019 / Published online: 26 October 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Purpose The present prediction model was intended to verify whether serum FSH level could be predictive of testis histology in patients with non-obstructive azoospermia (NOA).

Methods We evaluated two datasets of patients with NOA: the first (San Paolo dataset) comprising 558 patients, 18–63 years old, the second (Procrea dataset) composed by 143 patients, 26–62 years old; both datasets were combined to obtain a validation set. Multinomial logistic regression was first run with serum FSH and testis volume as independent predictors of testis histology, then, the correctly classified histological subcategories were set as outcome variables of a prediction model in both development and validation sets.

Results Multinomial logistic regression showed that FSH was a significant predictor of testis histology in 58% of cases, although it was unable to correctly classify cases with focal SCO or maturation arrest (MA). A prediction model was then run with hypospermatogenesis (HYPO) and Sertoli-only syndrome (SCO) as outcome variables of a binary logistic regression. FSH significantly predicted both HYPO and SCO, with a sensitivity of 40.9 and 80.7 and a specificity of 84.3 and 46.8 respectively. The model showed a fair discriminative ability (ROC AUC 0.705 and 0.709 respectively) and was adequately calibrated.

Conclusions Supported by a robust statistical analysis, we conclude that serum FSH level cannot be considered a prognostic marker of spermatogenic dysfunction in patients with NOA

Keywords FSH · Testis histology · Testis volume · Non-obstructive azoospermia · Prediction model

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10815-019-01613-8>) contains supplementary material, which is available to authorized users.

✉ Ettore Caroppo

¹ Asl Bari, Department of Maternal and Child Health, Reproductive and IVF Unit, PTA “F Jaia, 70014 Conversano, BA, Italy

² Conversano, Italy

³ Andrology and IVF Unit, Clinica San Carlo, 20037 Paderno Dugnano, MI, Italy

⁴ Division of Anatomic Pathology, Istituto Auxologico Italiano IRCCS, 20100 Milano, Italy

⁵ Andrology Unit, Procrea Institute, 6900 Lugano, Switzerland

Introduction

Serum follicle-stimulating hormone (FSH) level is considered to be clinically useful in the diagnostic evaluation of the infertile men. Since FSH secretion is regulated by a complex interplay between feedback and feedforward inputs within the hypothalamic-pituitary-gonadal axis, the resulting serum levels are expected to reflect both the pituitary and testicular function in physiological and pathological conditions. Indeed, fertile men display significantly lower FSH serum levels compared to infertile men; however, a normal or near-normal serum FSH concentration does not always guarantee normal spermatogenesis [1].

Elevated serum FSH levels are usually found in patients with non-obstructive azoospermia (NOA), but a linear relationship between increasing FSH levels and the degree of spermatogenesis dysfunction has not been demonstrated. There is a consensus that a serum FSH level higher than 7.5 mIU/ml is indicative of spermatogenesis dysfunction [1, 2],

however, azoospermic patients with very high serum FSH levels may harbor isolated areas of spermatogenesis integrity [3]; indeed, studies have demonstrated that FSH levels are poorly predictive (AUC 0.6) of the chance of retrieving testicular sperm in these patients [4]. Sperm retrieval, however, may be affected by a number of confoundings, including the surgeon skill and experience and the time and attention dedicated to the search for sperm in the testicular specimens; still, successful sperm retrieval (SSR) is a nearly unpredictable event, particularly in the microTESE setting.

A question that remains to be answered is to what extent serum FSH level is related to spermatogenesis in patients with NOA. Since testis histology has been found to be a reliable marker of spermatogenesis, few previous studies have been sought to determine the relationship between serum FSH level and testis histology, but unfortunately, they provided conflicting results mostly due to their small sample size [5, 6].

In the present study, we have built a prediction model with validation and redevelopment in order to evaluate whether serum FSH level could be predictive of testis histology and, therefore, of spermatogenesis in patients with NOA.

Materials and methods

Study population

We retrospectively evaluated two datasets of patients: the first (San Paolo dataset) comprising 558 patients, 18–63 years old, with non-obstructive azoospermia who referred to Andro-Urology and IVF Unit, San Paolo Hospital, Milan, Italy, to undergo surgery aimed at retrieving testicular sperm with either TESE or microTESE and the second composed of 143 patients with NOA, 26–62 years old, referred to the Andrology Unit, Procrea Institute Lugano, Switzerland, to undergo microTESE. Both cohorts had been previously evaluated: San Paolo cohort had been evaluated in a retrospective cohort study [7] and in a diagnostic accuracy study aiming at individuating possible predictors of sperm retrieval success [8], while Procrea cohort had been utilized for a diagnostic accuracy study aiming at determining predictive factors of seminiferous tubules caliber [9]. Both datasets contained de-identified, non-coded data.

The two cohorts of patients had been evaluated by the same urologist, who also had performed all the surgical procedures. The criteria for the diagnosis of NOA were the same as those applied in previous studies. Briefly, patients were defined to have NOA if semen analysis showed azoospermia without spermatozoa in the pellet, ejaculate volume and pH were higher than 1 mL and 7.2, respectively, and no sign of obstruction of the seminal tract was detected on physical examination, scrotal and/or transrectal ultrasound, when applicable. Exclusion criteria were obstructive azoospermia, history of

testicular torsion, mumps, or bacterial orchitis. In order to render the two cohorts homogenous for testis histology patterns, six patients from the Procrea dataset, two with testicular cancer and four with complete hyalinosis (histology patterns not included in San Paolo dataset), were excluded from the present analysis.

FSH serum level and testicular volume (measured by ultrasound) values were available for all the included patients. Testis volume was tested as candidate predictor in the multinomial logistic regression, but since it was not found to be predictive of testis histology (see below), it was not included in the prediction model.

Testis histology

All the histopathology specimens from both datasets had been examined by a single experienced pathologist (GG). Patients undergoing cTESE had a fragment of subcapsular parenchyma removed soon after the incision of the albuginea, while in patients undergoing microTESE a random biopsy sample representative of the overall appearance of the testicular parenchyma was taken. In both cases, fragments were fixed in Bouin's solution and sent to the pathologist. Histological analysis was conducted by examining at least 100 different tubule sections. The histopathological results were defined as: (i) Sertoli cell-only syndrome (SCO) when the seminiferous tubules were exclusively populated by Sertoli cells; (ii) focal SCO when tubules with SCO were interspersed with rare tubule containing germ cells; (iii) maturation arrest (MA), characterized with a complete arrest of the spermatogenetic maturation sequence at spermatogonia/spermatocytes level; (iv) hypospermatogenesis (HYPO), when all stages of spermatogenesis were present but reduced to a varying degree.

IRB approval

This is a secondary analysis of already published data. For both datasets, local IRB approval for the retrospective analysis of de-identified data had been obtained before conducting the original research. The present study was performed on de-identified, non-coded data.

Statistical analysis

The comparison of data as stratified according to testis histology was evaluated by non-parametric statistical tests (two-tailed Mann-Whitney *U* test, Chi-square test or Kruskal-Wallis test as appropriate) since the variables were not normally distributed according to one-sample Kolmogorov–Smirnov test.

San Paolo set was used to run a multinomial logistic regression with testis histology as dependent categorical variable with 3 degrees of freedom, in order to verify whether FSH

and testis volume could be used as covariate/predictors in a prediction model. Once the required assumptions were checked (assumption of linearity was confirmed by the Box-Tidwell test and multicollinearity was excluded by running a linear regression with multicollinearity diagnostic test), we looked at the correct predictions and selected SCO and HYPO as the two more representative categories (see Results for details). We then created two one-versus-all dummy binary classifiers, DHYPO and DSCO, in order to run binary logistic regression, compute sensitivity and specificity, and quantify the predictive ability of each model.

Model building

Data from San Paolo dataset (development set) were used for building a model to compute the probability to detect HYPO or SCO on the basis of serum FSH level in patients with NOA. The dependent/outcome variables were the two dummy binary variables DHYPO and DSCO, where any observation of HYPO or SCO, respectively, was coded as 1, while the others were coded as 0.

The assumption of the linear relationship between the predictor (continuous) and its log odds was verified by the Box-Tidwell test; then, the predictor was tested for multicollinearity. No missing data were found. Three outliers were detected by visual inspection of box plots and by computing the Cox distance but, since they did not affect the results and were judged to be “real life” values, we did not remove them from the dataset.

Probability was set as $p < 0.5$ for entry and $p > 0.1$ for removal. Decision threshold was left as 0.5 for DSCO model, while for the DHYPO model, due to imbalanced dataset, the threshold set arbitrarily at 0.5 returned a classification matrix with 0 true positive and 0 false positive. This typically occurs with imbalanced data: when most of the observations are in one class, the model always predicts that every case belongs to that class in order to achieve high accuracy (the accuracy paradox). We then modeled four additional logistic regressions with reduced classification threshold (0.4 to 0.1) to individuate the model with the highest Matthews Correlation Coefficient (MCC), which is a measure of the quality of binary classification that takes into account true and false positives and negatives. The model with the threshold being set at 0.3 performed best in terms of MCC both in development and in redevelopment set. In the model with SCO as dependent variable, the Box-Tidwell test revealed a non-linear relationship between the predictor (FSH) and the logit of the outcome, therefore we transformed FSH to better fit the data using a polynomial $FSH + FSH^2 + FSH^3$.

For model internal validation, we focused on discrimination and calibration. Discrimination (ability of the model to distinguish between cases with and without the outcome) was assessed by area under curve (AUC) estimates derived from

receiving operator characteristic (ROC) curve; AUC less than 0.60 reflects poor discrimination, 0.60 to 0.75 possibly helpful discrimination, more than 0.75 clearly useful discrimination [10]. Calibration (correspondence between the predicted and observed probabilities) was assessed by the Hosmer and Lemeshow test and visually by comparing predicted probabilities and observed probabilities based on the covariate (FSH) expressed as deciles.

Model validation and redevelopment

Data from Procrea dataset were checked to be used for external validation. Since the sample size was not large enough, we combined this dataset with San Paolo dataset and a new prediction model was developed on the combined data set with the same outcome variables DHYPO and DSCO. The resulting models can be considered validation and model redevelopment according to the TRIPOD statements [11]. Outliers were managed as for the development set. There were only two missing data that were managed with pairwise deletion.

For model redevelopment, once the assumptions were verified, we used the same criteria of model development (e.g. decision threshold). For model validation, discrimination was assessed by AUC estimates derived from ROC curve; while calibration was assessed by the Hosmer and Lemeshow test and visually by comparing the predicted probabilities as obtained from the development set with the observed probabilities computed on the validation set.

Sample size

The minimum sample size required for the analysis was determined according to the rule of 10 events per predictor (EPV). Binary logistic regression with HYPO as outcome/dependent variable required a minimum sample size of 110 patients, as the smaller of the number of subjects who experienced or did not experience the outcome was 110 and the number of candidate predictors was 1; binary logistic regression with SCO as outcome/dependent variable required a minimum sample size of 257 patients, given 257 the smaller number of patients who experienced or did not experience the outcome and 1 the number of candidate predictors.

All computations were performed using IBM SPSS for Windows (Chicago, IL, USA) and by STATA StataCorp LLC, (College Station, TX, USA).

Results

The basal characteristics of patients from the two datasets are displayed in Table 1. Both datasets were comparable for all studied parameters. SCO was the most frequent histological

Table 1 Baseline characteristics of patients with NOA in the two datasets

	Development set	Validation set
N, subjects	558	695
FSH (mIU/ml)	18.6 (17.1) [0.6–127]	19.1 (15.6) [0.6–127]
Testis volume (ml)	7.2 (5) [0.9–21.3]	7.3 (5) [0.9–21.3]
Testis histology patterns		
Sertoly-only syndrome	301 (53.9%)	389 (56%)
Focal SCO	75 (13.4%)	83 (11.9%)
Maturative arrest	72 (12.9%)	88 (12.7%)
Hypospermatogenesis	110 (19.7%)	135 (19.4%)

Data are expressed as median (IQR) with range in squared brackets, or as count with percentages in parentheses

pattern in patients with NOA. The comparison of FSH and testis volume as stratified according to testis histology is displayed in Table 2. FSH level was significantly higher and testis volume was significantly lower in patients with SCO or FSCO compared with those with MA or HYPO ($p < 0.0001$), while no difference was observed between patients with SCO and FSCO as well as between those with MA and HYPO in both datasets (subgroups analysis by two-sided Mann-Whitney U test, Table 2)

Multinomial logistic regression

A multinomial logistic regression was performed to model the relationship between the predictors and membership in the four subgroup of testis histology patterns (SCO, FSCO, MA, HYPO) (Fig. 1). The traditional $\alpha < 0.05$ criterion of statistical significance was employed for all tests. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) estimated that the model fitted values were close to the true expected values (AIC 1215.558, BIC 1254.47, -2 Log Likelihood 1197.558). Addition of the predictor to a model that contained only the intercept significantly improved the fit between model and data, $\chi^2(6) = 107.298$, $p < 0.0001$.

Significant unique contributions were made by FSH only (Supplementary Table 1). This was confirmed by the likelihood ratio test, which showed that the variable has a significant overall effect on the outcome (-2 Log Likelihood of Reduced Model: intercept 1203.61, FSH 1271.29, testis volume 1201.95; Chi-square: intercept 6053, $p = 0.109$, FSH 73.737, $p < 0.0001$, testis volume 4392, $p = 0.222$). The logistic model resulted in 58.2% correct prediction. Correct predictions were more frequent for SCO (95%) than for HYPO (35.5%), while 0% correct prediction was obtained for FSCO and MA (Supplementary Table 4).

Model development

The results of the two models for the prediction of HYPO or SCO in both model development and redevelopment are summarized in Fig. 2 and Supplementary Table 2. According to the results of the binary logistic regression, FSH level significantly predicted the HYPO pattern (odds ratio 0.933, 95% CI 0.912–0.955) with a sensitivity of 40.9% and a specificity of 84.3%, PPV 39.1, NPV 85.3, positive likelihood ratio (LR+) 2.62, negative LR (LR-) 0.7 Accuracy was not taken into account, not being an adequate metric when imbalanced set are evaluated. The resulting prediction equation is the following: $\text{Log}(\text{HYPO}) = -0.199 - 0.069 \text{FSH}$. The mean observed probability of detecting HYPO in patients with NOA was 19.7%, with a minimum of 3 and a maximum of 46%; the predictive probabilities were similar (mean 19.7%, range 0.01–44%; Supplementary Table 3); the probabilities of finding HYPO decreased with increased FSH levels. The model had a possibly helpful discriminative capacity (AUC 0.705) and showed a good internal calibration (Hosmer and Lemeshow test: Chi-square 11.087, $df 8$, $p = 0.197$).

In the model with SCO as dependent variable, FSH level significantly predicted the SCO pattern (odds ratio 1303, 95%CI 1139–1490) with a diagnostic accuracy 68.5%, sensitivity of 80.7% and specificity of 54%, PPV 67.3, NPV 70.5, LR+ 1.76, LR- 0.36. The resulting prediction equation is the following: $\text{Log}(\text{SCO}) = -2.424 + 0.265 \text{FSH} - 0.007 \text{squared}$

Table 2 FSH and testis volume values as stratified according to testis histology in the development and validation sets

		SCO	FSCO	MA	HYPO
Development set	FSH [#] (mIU/ml)	23.3 (18) [1.10–127] ¹	19.8 (17.7) [4.3–47.1] ¹	9.5 (13.08) [0.6–45.1]	12.1 (12.5) [2.4–56]
	Testis volume [#] (ml)	6.4 (3.57) [0.9–21.3] ¹	7.3 (4.4) [1.8–15] ²	9.4 (5.4) [1.4–18]	8 (5.05) [2.16–20]
Validation set	FSH [#] (mIU/ml)	22.7 (15.8) [1.1–127] ¹	21 (17.7) [4.3–47.1] ¹	9.6 (12.6) [0.6–45.1]	12.5 (12.6) [1.8–56]
	Testis volume [#] (ml)	6.9 (3.9) [0.9–21.3] ¹	7.3 (4.4) [1.8–15] ²	9 (5.5) [1.4–18]	8 (4.7) [2.16–20]

Data are expressed as median (IQR) with range in squared brackets, or as count with percentages in parentheses

[#] Kruskal-Wallis test, $p < 0.0001$

Subgroup analysis by two-sided Mann-Whitney U test

¹ $p < 0.0001$ vs MA and HYPO

² $p < 0.01$ vs MA and HYPO

Fig. 1 Multinomial logistic regression with testis histology as dependent categorical variable with 3 degrees of freedom, and FSH as independent continuous covariate. Adjusted predictions are displayed with 95% CI

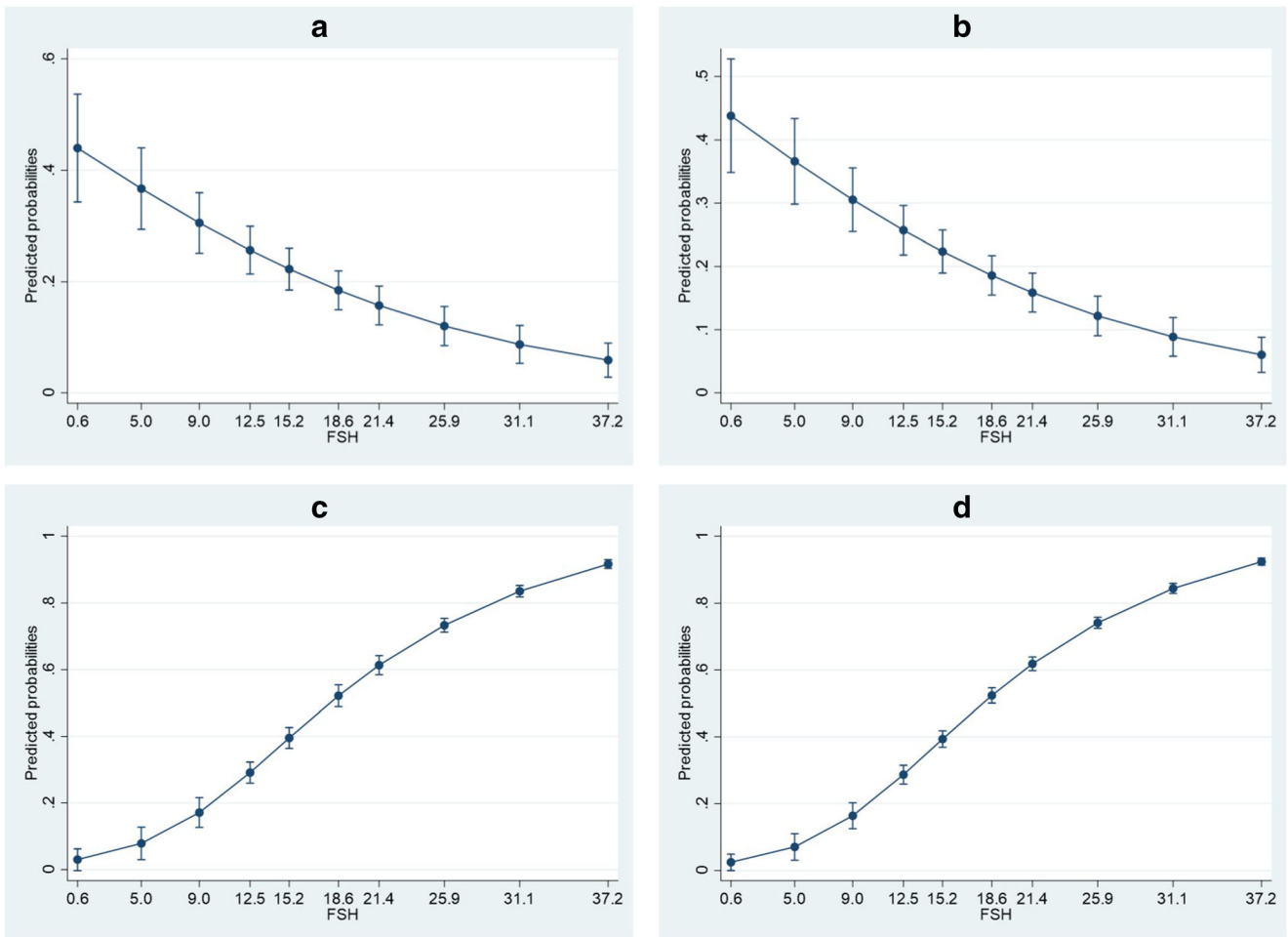
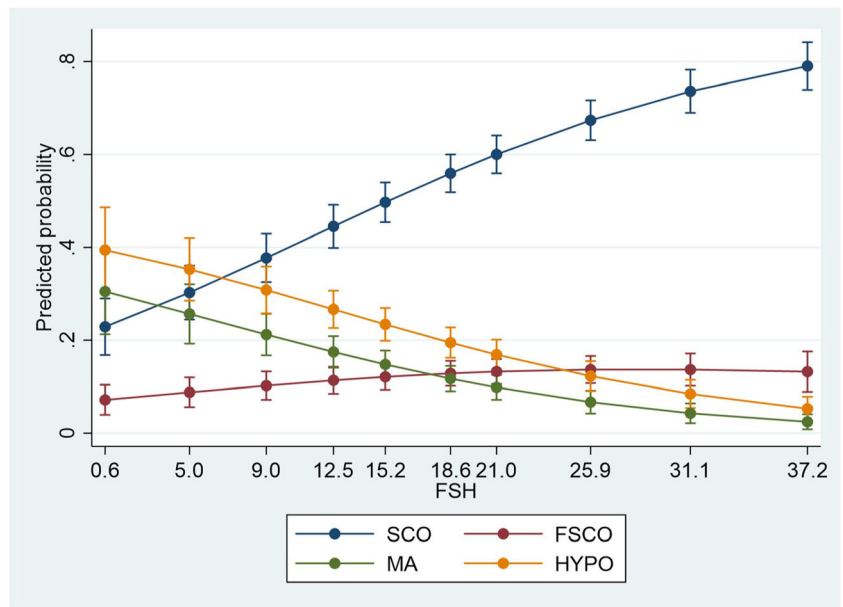


Fig. 2 Binary logistic regression plots. **a** Development set with HYPO as outcome variable. **b** Validation/redevelopment set with HYPO as outcome variable. **c** Development set with SCO as outcome variable. **d** Validation/redevelopment set with SCO as outcome variable. X axis

displays the lowest limits of FSH deciles; FSH deciles ranges are displayed in Supplementary Table 3. Adjusted predictions are displayed with 95% CI

FSH + 0 cubic FSH. The mean calculated probability of detecting SCO in patients with NOA was 53.9%, with a minimum of 18 and a maximum of 79%; the predictive probabilities were similar (mean 53.9%, range 9–100%; Supplementary Table 3); the probabilities of finding SCO increased with increased FSH levels. The model had a possibly helpful discriminative capacity in the development set (AUC 0.709) and showed a good internal calibration (Hosmer and Lemeshow test: Chi-square 9322, df 8, $p = 0.316$). Calibration plots for both HYPO and SCO as outcome variable are displayed in Fig. 3

Validation and model redevelopment

The calculated probabilities of both HYPO and SCO were similar to those of the development model (mean 19.4%, range 0–46% and 55.9%, range 0–81% for HYPO and SCO respectively; Table 5). Both models had a possibly helpful

discriminative capacity (0.69 and 0.702 for HYPO and SCO respectively) and were well calibrated according to Hosmer and Lemeshow test (Chi-square 10.302, $p = 0.244$ and 12.427, $p = 0.133$ for HYPO and SCO respectively). Calibration plots are displayed in Fig. 3

The model redevelopment confirmed that FSH predicts HYPO with a sensitivity of 38.55 and specificity of 85.5% ($\text{Log}(\text{HYPO}) = -0.208 - 0.068 \text{ FSH}$), and SCO with a sensitivity of 84.32%, specificity 46.8%, accuracy 67.9% ($\text{Log}(\text{SCO}) = -2.507 + 0.290 \text{ FSH} - 0.008 \text{ squared FSH} + 0 \text{ cubic FSH}$).

Discussion

The present study demonstrates the limited ability of serum FSH in predicting the severity of spermatogenesis impairment in patients with NOA. As demonstrated by multinomial

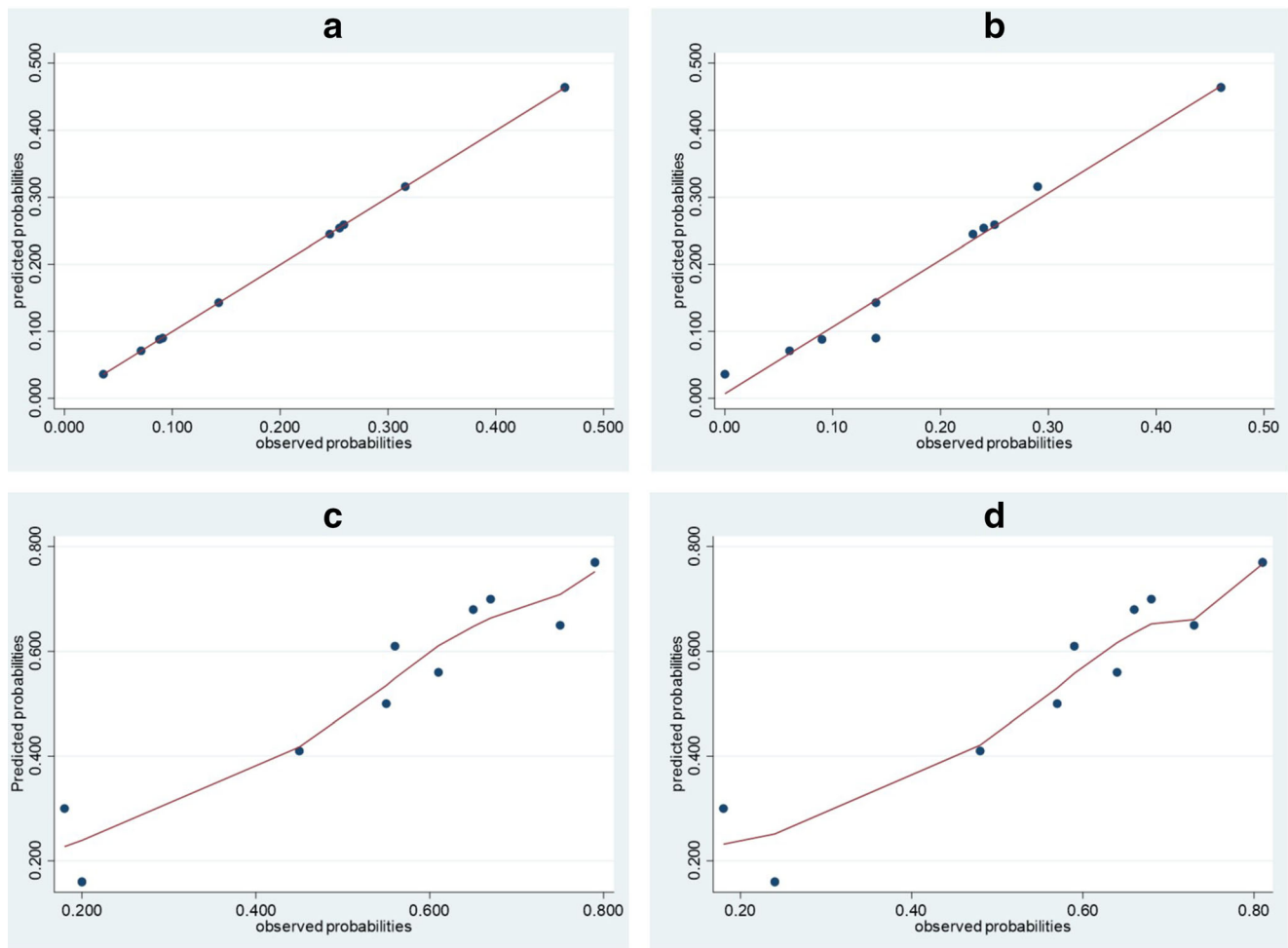


Fig. 3 Relationship between observed and predicted probabilities based on predictor (FSH) deciles. **a** Development set with HYPO as outcome variable. **b** Validation set with HYPO as outcome variable. **c** Development set with SCO as outcome variable. **d** Validation set with SCO as outcome variable. X axis displays the lowest limits of FSH

deciles; FSH deciles ranges are displayed in Supplementary Table 3. Adjusted predictions are displayed with 95% CI. Since the relationship between observed and predicted probabilities with SCO as outcome variable was not linear, Lowess curve fitting was used

logistic regression analysis performed on San Paolo dataset, FSH could correctly classify the histopathological patterns in only 58% of cases; relevantly, FSH was not able to discriminate between cases with SCO and FSCO, as well as between those with MA and HYPO.

Following the results of the multinomial logistic regression analysis demonstrating that FSH was a significant predictor of HYPO and SCO, we built a prediction model to evaluate to what extent serum FSH level could predict these two histological patterns, which have been associated with the highest and lowest chances of successful sperm retrieval respectively [12]. According to the prediction equation, the relationship between FSH and HYPO was negative, so the probability of finding this histopathological pattern decreases with the increase of FSH value. This was confirmed by the diagnostic accuracy measures: the sensitivity of 40.9 and a specificity of 84.3 imply that FSH can be better used, although to a certain extent, to exclude HYPO rather than to predict it. Conversely, increased FSH levels were associated with increased probability of finding SCO (sensitivity 80.7%); yet, lower FSH levels were relatively unable to correctly exclude it (specificity 54%).

Taken together, these results suggest that the severity of spermatogenesis dysfunction in men with NOA cannot be fully predicted by FSH levels. Such a finding is not difficult to explain. It has been classically demonstrated that FSH level correlates with the number of spermatogonia and, to a lesser extent, primary spermatocytes [13]; indeed, treatment with testosterone or 17 beta estradiol benzoate, which reduces the number of germ cells, led to a consequent significant rise in serum FSH levels in a rat model [14]. It is reasonable to expect, therefore, that patients that harbor focal areas of intact or less severely compromised spermatogenesis surrounded by tubules with complete SCO would display serum FSH levels that are indistinguishable from those with complete SCO syndrome, given that the overall number of spermatogonia is expected to be very small in both cases; similarly, the presence of a comparable number of spermatogonia in patients with MA and HYPO may justify the finding of similar serum FSH levels in both histotypes. Other factors that, to a lesser extent, may explain the variability of FSH levels in NOA patients include age [15, 16] and single nucleotide polymorphisms (SNPs) of FSH β subunit (FSHB) and FSH receptor, that may be associated with reduced FSH levels due to decreased FSHB promoter activity [17] and with increased FSH levels due to reduced FSH signal transduction [18] respectively.

Unfortunately, we could not adjust the data for the age of patients, as we used two datasets with de-identified data that did not fully provide the age of all patients. This, together with the retrospective design, represents a limitation of the present study.

In conclusion, the results of the present study suggest that FSH is unable to fully predict the histopathological patterns in patients with NOA; as a consequence, serum FSH level, while having been demonstrated to be of help for the differential diagnosis of NOA, cannot be used as prognostic markers of the severity of spermatogenesis dysfunction in these patients.

Compliance with ethical standards

Conflict of interest Caroppo E. reports personal fees from Ferring Pharmaceutical, outside of the submitted work. Colpi EM, D'Amato G, Gazzano G, and Colpi GM has nothing to disclose.

Ethical approval The present study did not involve human or animal participants, being a retrospective analysis of de-identified data from two separate databases. For this type of study formal consent is not required.

IRB approval This is a secondary analysis of already published data. For both datasets, local IRB approval for the retrospective analysis of de-identified data had been obtained before conducting the original research. The present study was performed on de-identified, non-coded data.

References

1. Practice Committee of the American Society for Reproductive Medicine in collaboration with the Society for Male Reproduction and Urology. Evaluation of the azoospermic male: a committee opinion. *Fertil Steril*. 2018;109:777–82.
2. Schoor RA, Elhanbly S, Niederberger CS, Ross LS. The role of testicular biopsy in the modern management of male infertility. *J Urol*. 2002;167:197–200.
3. Ramasamy R, Lin K, Veeck G, Rosenwaks Z, Palermo GD, Schlegel PN. High serum FSH levels in men with nonobstructive azoospermia does not affect success of microdissection testicular sperm extraction. *Fertil Steril*. 2009;92:590–3.
4. Li H, Chen LP, Yang J, Li MC, Chen RB, Lan RZ, et al. Predictive value of FSH, testicular volume, and histopathological findings for the sperm retrieval rate of microdissection TESE in nonobstructive azoospermia: a meta-analysis. *Asian J Androl*. 2018;20:30–6.
5. Bergman M, Behre HM, Nieschlag E. Serum FSH and testicular morphology in male infertility. *Clin Endocrinol*. 1994;40:133–6.
6. Von Eckardstein S, Simoni M, Bergmann M, Weinbauer GF, Gassner P, Schepers AG, et al. Serum inhibin B in combination with serum follicle-stimulating hormone is a more sensitive marker than FSH alone for impaired spermatogenesis in men, but cannot predict the presence of sperm in testicular tissue samples. *J Clin Endocrinol Metab*. 1999;84:2496–50.
7. Caroppo E, Colpi EM, Castiglioni M, Gazzano G, Vaccalluzzo L, D'Amato G, et al. Role of FSH level, testicular volume and histology in choosing the appropriate surgical technique of sperm retrievals in patients with non-obstructive azoospermia. *Hum Reprod*. 2015;30:i132–3.
8. Caroppo E, Colpi EM, Gazzano G, Vaccalluzzo L, Scropo FI, D'Amato G, et al. Testicular histology may predict the successful sperm retrieval in patients with non-obstructive azoospermia undergoing conventional TESE: a diagnostic accuracy study. *J Assist Reprod Genet*. 2017;34:149–54.
9. Caroppo E, Colpi EM, Gazzano G, Vaccalluzzo L, Piatti E, D'Amato G, et al. The seminiferous tubule caliber pattern as evaluated at high magnification during microdissection testicular sperm

- extraction predicts sperm retrieval in patients with non-obstructive azoospermia. *Andrology*. 2019;7:8–14.
10. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models users' guide to the medical literature. *JAMA*. 2017;318:1377–84.
 11. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyeberg 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–W73.
 12. Flannigan R, Bach PV, Schlegel PN. Microdissection testicular sperm extraction. *Transl Androl Urol*. 2017;6:745–52.
 13. De Kretser DM, Burger HG, Hudson B. The relationship between germinal cells and serum FSH levels in males with infertility. *J Clin Endocrinol Metab*. 1974;38:787–93.
 14. Walczak-Jedrzejowska R, Marchlewska K, Oszukowska E, Filipiak E, Slowikowska-Hilczner J, Kula K. Estradiol and testosterone inhibit rat seminiferous tubule development in a hormone-specific way. *Reprod Biol*. 2013;13:243–50.
 15. Grunewald S, Glander HJ, Paasch U, Kratzsch J. Age-dependent inhibin B concentration in relation to FSH and semen sample qualities: a study in 2448 men. *Reproduction*. 2013;145:237–44.
 16. Yang Q, Huang YP, Wang HX, Hu K, Wang YX, Huang YR, et al. Follicle-stimulating hormone as a predictor for sperm retrieval rate in patients with nonobstructive azoospermia: a systematic review and meta-analysis. *Asian J Androl*. 2015;17:281–4.
 17. Grigorova M, Punab M, Ausmees K, Laan M. FSHB promoter polymorphism within evolutionary conserved element is associated with serum FSH level in men. *Hum Reprod*. 2008;23:2160–6.
 18. Tamburino L, La Vignera S, Tomaselli V, Condorelli RA, Cannarella R, Mongioì LM, et al. The -29G/A FSH receptor gene polymorphism is associated with higher FSH and LH levels in normozoospermic men. *J Assist Reprod Genet*. 2017;34:1289–94.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.