RESEARCH

A systematic review on machine learning in sellar region diseases: quality and reporting items

Nidan Qiao[1,2]

[1]Department of Neurosurgery, Huashan Hospital, Fudan University, Shanghai, China
[2]Neuroendocrine Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

Correspondence should be addressed to N Qiao: norikaisa@gmail.com

Abstract

Introduction: Machine learning methods in sellar region diseases present a particular challenge because of the complexity and the necessity for reproducibility. This systematic review aims to compile the current literature on sellar region diseases that utilized machine learning methods and to propose a quality assessment tool and reporting checklist for future studies.
Methods: PubMed and Web of Science were searched to identify relevant studies. The quality assessment included five categories: unmet needs, reproducibility, robustness, generalizability and clinical significance.
Results: Seventeen studies were included with the diagnosis of general pituitary neoplasms, acromegaly, Cushing's disease, craniopharyngioma and growth hormone deficiency. 87.5% of the studies arbitrarily chose one or two machine learning models. One study chose ensemble models, and one study compared several models. 43.8% of studies did not provide the platform for model training, and roughly half did not offer parameters or hyperparameters. 62.5% of the studies provided a valid method to avoid over-fitting, but only five reported variations in the validation statistics. Only one study validated the algorithm in a different external database. Four studies reported how to interpret the predictors, and most studies (68.8%) suggested possible clinical applications of the developed algorithm. The workflow of a machine-learning study and the recommended reporting items were also provided based on the results.
Conclusions: Machine learning methods were used to predict diagnosis and posttreatment outcomes in sellar region diseases. Though most studies had substantial unmet need and proposed possible clinical application, replicability, robustness and generalizability were major limits in current studies.

Key Words
- artificial intelligence
- prediction
- pituitary
- growth
- craniopharyngioma

Endocrine Connections (2019) **8**, 952–960

Introduction

Studies using machine learning methods gained popularity in medical researches in recent years. Machine learning methods integrate computer-based algorithms into data analysis to find similar patterns among different samples. The ultimate goal aims at using multiple variables to predict a specific outcome in a particular cohort. There are two types of machine learning algorithms in general: supervised and unsupervised machine learning. In supervised machine learning, both of the predictors and outcome are known; but in unsupervised machine learning, only the predictors are fed into the algorithm.

The most common type of tumors originated in the sellar region included pituitary neoplasm, craniopharyngioma, meningioma and chordoma,

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-19-0156

which overall takes more than 10–15% of tumors in the central nervous system (1). Other non-tumorous sellar region diseases include Rathke's cyst, hypophysitis, hypopituitarism and all the complications due to the treatment of these diseases (2). Machine learning may help to build a more reliable aided diagnostic tool for neuroradiologists and neuropathologists. Better prediction of clinical outcomes in these patients may provide better clinical decision support for either neuroendocrinologists or neurosurgeons. While on the other hand, machine learning methods present a particular challenge because of the complexity in model training and testing. The reproducibility of scientific research has always been of critical importance, which also applies in machine learning studies (3). With the expansion of machine learning in medical studies, the applications in real clinical decision making are booming, which requires both robustness and generalizability (4, 5).

This systematic review aims to compile the current literature on sellar region diseases that utilized different machine learning methods and analyze the reporting items regarding cohort selection, model building and model explanation. Unlike traditional statistical methods, risks of bias and confounding are not the main question of interest in machine learning studies. How to assess the quality of these studies remains unsolved, and a reporting guideline was not available for these studies to follow. This review presents a quality assessment tool and proposes a checklist of reporting items for studies built on machine learning methods.

Methods

Literature for this review was identified by searching PubMed and Web of Science from the date of the first available article to December 1, 2018. The keywords containing 'machine learning' or the algorithms of machine learning were queried with the combination of keywords containing sellar region diseases (Supplementary Table 1, see section on supplementary data given at the end of this article). The search was limited to studies published in English. References in published reviews were manually screened for possible inclusions. The study adheres to the PRISMA guideline, and the checklist was provided in Supplementary Table 2.

Studies were included if they evaluated machine learning algorithms (logistic regression with regulation, linear discriminant analysis, k means, k nearest neighbor, cluster analysis, support vector machine, decision tree-based models and neural networks) for application in prediction of disease originated in the sellar region (both tumorous and non-tumorous diseases). Exclusion criteria were lack of full-text or animal studies.

Data obtained from each study were publication characterizes (first author's last name, publication time), cohort selection (sample size, diagnosis), predictors (variables fed into the machine learning models), outcomes (the outcomes as well as the controls, including the distributions between them), model selection (models used in the study, including platforms, packages and parameters), statistics for model performance (methods to evaluated the model, the mean and the variance) and model explanation (any explanation on how important of each predictors and proposed clinical application). Supplements in each study were also reviewed if available.

Quantitative synthesis was inappropriate due to the heterogeneity in outcomes. Summary of included studies was listed in a table and using a narrative approach. The proposed quality assessment (Table 1) of each study consists of five categories: unmet need (limits in current non-machine-learning approach), reproducibility (feature engineering methods, platforms/packages, hyperparameters), robustness (valid methods to overcome over-fit, the stability of results), generalizability (external data validation) and clinical significance (predictors explanation and suggested clinical use). A quality assessment table was provided by listing 'yes' or 'no' of corresponding items in each category.

To provide a clear picture of how to perform a machine learning study, the workflow of a machine learning study was summarized, and notations of terms used in these machine learning studies were provided. The recommended reporting items were also provided based on the results.

Results

After scrutinizing the titles and abstracts generated by the searching strategy, 31 articles left for full-text screening, in which 13 studies were excluded: one study not in English, two duplicated studies, four conference abstracts without full-text and six studies without outcomes in sellar region diseases. Three studies used the same image database such that only the latest published study was included. At last, this systematic review included 16 studies (6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21) (Table 2), with the diagnosis of general pituitary neoplasms, acromegaly, Cushing's disease, craniopharyngioma and growth

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-19-0156

Table 1 Quality assessment of machine learning studies.

Categories	Items	Description	Reported
Unmet need	Limits in current non-machine-learning approach	Low diagnostic accuracy, low human-level prediction accuracy or prolonged diagnostic procedure	Yes/no
Reproducibility	Feature engineering methods	How features were generated before model training	Yes/no
	Platforms/packages	Both platforms and packages should be reported	Yes/no
	Hyperparameters	All hyperparameters which are needed for study replication	Yes/no
Robustness	Valid methods to overcome over-fit	Leave-one-out or k-fold cross-validation or bootstrap	Yes/no
	The stability of results	Calculated variation in the validation statistics	Yes/no
Generalizability	External data validation	Validation in settings different from the research framework	Yes/no
Clinical significance	Predictors explanation	Explanation of the importance of each predictor	Yes/no
	Suggested clinical use	Proposed possible applications in clinical care	Yes/no

hormone deficiency. More than half of the studies were published in the recent 2 years.

The scheme of a machine learning study was summarized in Fig. 1. The process can be categorized into four stages when developing a prediction model. The first step is to bring out the clinical question, which is summarized as 'predicting Outcome using Predictors in a Cohort'. A study should choose the appropriate outcome, predictors and the data source. The data are then pre-processed, which can involve data coding, transformations, imputation and dimension reduction. The training step means how the model (algorithm) finds patterns from the features to match the outcome variable. The trained model should be validated (both internally and externally). Finally, models are explained, and possible clinical applications are provided. The notations of terms used in these machine learning studies were described in Table 3.

Sample size in these studies varied from tens to thousands. The majority of the studies (6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 20) (76.5%) used the diagnosis of a specific disease as the outcome, only four studies (17, 18, 19, 21) tested on the treatment outcome. In the diagnostic studies, three studies (7, 12, 14) used image features to categorize magnetic resonance images (MRIs), two (6, 13) used face photos to predict acromegaly, two (15, 20) predicted growth home deficiency using serum proteins, two (10, 11) used histological spectrum to predict histology diagnosis, one (9) used serum proteins to predict pituitary adenoma and one (8) predicted surgical phase using videos. In studies on treatment outcomes, one study (17) predicted poor early postoperative outcome, one (18) predicted gross-total resection, one (19) predicted response to somatostatin analogs and one (21) predicted growth after growth hormone treatment. All the outcomes were either dichotomized or categorical outcomes except one in the continuous form (21).

Most of the studies (87.5%) arbitrarily chose one or two machine learning models without arguing the reasons. One study (13) chose ensemble models by combining the decisions from multiple models to improve the overall performance. One study (17) compared several models and chose the one with the best performance. With regard to validation methods, five studies (8, 9, 12, 15, 19) used k-fold cross-validation, two studies (14, 20) used bootstrap and three studies (6, 7, 10) used leave-one-out cross-validation. One study (18) used cross-validation but without holdout and five studies (11, 13, 16, 17, 21) did not report the validation method. In studies reporting validation methods, only five (8, 10, 12, 18, 19) reported the variation of the validation statistics.

In the quality assessment (Table 4), limits in current non-machine-learning approach were mentioned in most of the studies. During the model training process, only two studies (6, 17) did not provide how the data were transformed into a way that can be fed into the algorithm. But nearly half the studies (43.8%) did not offer the program or the platform for model training and roughly half (43.8%) did not provide hyperparameters which are necessary for the training process. As mentioned above, 62.5% of the studies provided a valid method to combat over-fitting, but only five reported variations in the validation statistics. Only one study (16) validated the algorithm in an external database. Though only four studies (15, 17, 18, 21) reported how to interpret the predictors, most studies (68.8%) suggested possible clinical application of the developed machine learning algorithm.

Based on the results, several recommended reporting items for a machine learning study were proposed (Table 5). Reporting of the background should include results by human intelligence and a summarized research question. In the methods part, it is recommended to report the diagnosis of the cohort;

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-19-0156

Table 2 Summary of studies on sellar region disease using machine learning methods.

	Cohort selection			Outcomes			Performance statistics		
	Sample size	Diagnosis	**Predictors**	Outcomes and controls	Distribution	**Models** (parameters)	Discrimination	CV method	Variation in validation
Learned-Miller 2006 (6)	49	Acromegaly	Parameters from 3D shape of face	Acromegaly/healthy	24:25	SVM (linear or quadratic kernel)	Acc: 85.7%	LOOCV	NA
Kitajima 2009 (7)	43	Sellar mass	Age and 9 MRI features	Pituitary adenoma/ craniopharyngioma/ Rathke's cyst	20:11:12	NN (FC(7)*1)	AUC: 0.990	LOOCV	NA
Lalys 2011 (8)	500	Pituitary adenoma	Features in surgical images	Six surgical phases: nasal incision/retract/tumor removal/column replacement/suture/ nose compress	NA	SVM (linear kernel), HMM	Acc: 87.6%	10-fold CV	s.d.: 2.4%
Hu 2012 (9)	68	Pituitary adenoma	9 serum proteins	NFPA healthy	34:34	Decision tree (Gini index)	Sen: 82.4% Spe: 82.4%	10-fold CV	NA
Steiner 2012 (10)	15	Pituitary adenoma	Spectrum from histology	GH+/GH−/non-tumor cells	1000:1000:1000	k means (k = 10), LDA	Acc: 85.3%	LOOCV	s.d.: 10.5%
Calligaris 2015 (11)	45	Pituitary adenoma	Protein signature in mass spectrometry from histology	ACTH pituitary tumor/GH pituitary tumor/PRL pituitary tumor/pituitary gland	6:9:9:6	SVM	Sen: 83.0% Spe: 93.0%	NA	NA
Paul 2017 (12)	233	Brain tumors	Pixels in MRI images	Meningioma/glioma/ pituitary tumor	208:492:289	CNN ((Cov(64)-Max)*2 + FC(800)*2), NN, SVM	Acc: 94.0%	5-fold CV	s.d.: 4.5%
Kong 2018 (13)	1123	Acromegaly	Features in photos	Acromegaly/healthy	527:596	Ensemble	Acc: 95.5%	NA	NA
Zhang 2018 (14)	112	Pituitary adenoma	Features in MRI images	Null cell adenoma/other subtypes	46:66	SVM (radial kernel)	AUC:0.804 Acc: 81.1%	Bootstrap	NA
Murray 2018 (15)	124	Growth hormone deficiency	Age, sex, IGF1, gene expressions	Growth hormone deficiency/healthy	98:26	RF	AUC:0.990	Out-of-bag (3-fold CV)	NA
Yang 2018 (16)	168	Craniopharyngioma	Expression levels of signature genes	Craniopharyngioma/other brain or brain tumor samples	24:144	SVM (radial kernel)	AUC:0.850	NA	NA
Hollon 2018 (17)	400	Pituitary adenoma	26 patient's characteristics	Poor early postoperative outcome/good	124:276	Elastic net, NB, SVM, RF	Acc: 87.0%	NA	NA
Staartjes 2018 (18)	140	Pituitary adenoma	patient characteristics, MRI features	Gross-total resection/not	95:45	NN (FC(5)*NA)	AUC: 0.96 Acc: 90.9%	5-fold CV without holdout	s.d.:0.08%
Kocak 2018 (19)	47	Acromegaly	Features in MRI images	Response to somatostatin analogs/resistant	24:23	k-NN (k = 5)	Acc: 85.1% AUC: 0.847	10-fold CV	s.d.: 1.5%
Ortea 2018 (20)	30	Growth hormone deficiency	Three serum proteins	Growth hormone deficiency/healthy	15:15	RF, SVM	Acc: 100% AUC: 1.000	Bootstrap	NA
Smyczynska 2018 (21)	272	Growth hormone deficiency with GH treatment	Patient characteristics, GH level, IGF-1 level, GH dose	Height change after GH treatment	0.66 ± 0.57	NN (FC(2)*1)	RMSE: 0.267	NA	NA

Acc, accuracy; ACTH, adrenocorticotropic hormone; AUC, area under curve; BoVW, bag-of-visual-word; CNN, convolutional neural network; Cov, convolutional layer; CV, cross-validation; FC, fully-connected neural network; GH, growth hormone; HMM, hidden Markov model; IGF1, insulin-like growth hormone 1; LDA, linear discriminant analysis; LOOCV, leave-one-out cross-validation; Max, max pooling layer; MRI, magnetic resonance image; NA, not available; NB, naïve Bayesian; NFPA, non-functional pituitary adenoma; NN, neural network; PRL, prolactin; RF, random forest; RMSE, root mean square error; s.d., standard deviation; Sen, sensitivity; Spe, specificity; SVM, support vector machine.

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-19-0156

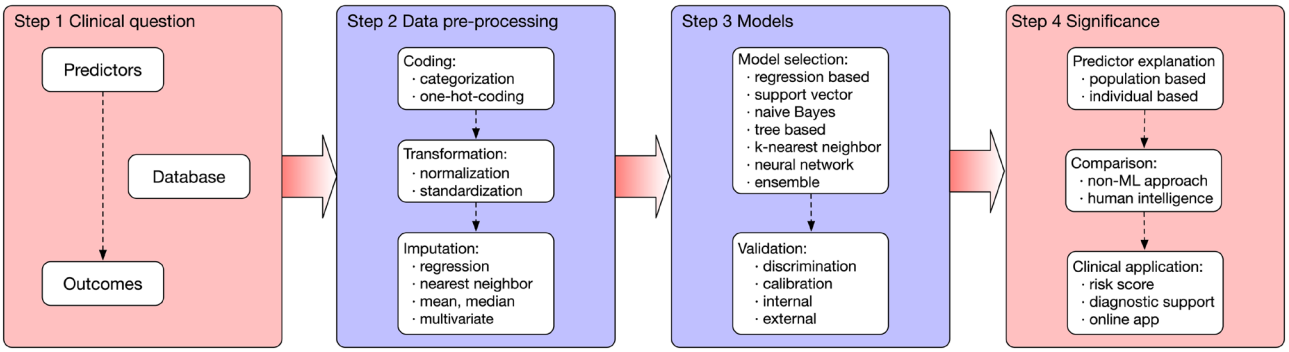

Figure 1
The scheme of a machine learning study. The process can be categorized into four steps: a good clinical question; pre-processed data; training and validation of the model and significance in clinical applications.

Table 3 Notations of special machine learning terms.

Terms	Explanations
Unsupervised learning	A subgroup of machine leaning models with the purpose of finding similarities among samples where no outcomes are available
Supervised learning	A subgroup of machine leaning models with both predictors and outcomes, and the purpose is to learn the mapping function from the predictors to the outcomes
Feature	Predictors in a machine learning algorithm
Categorization	Transforming a continuous variable into a categorical variable
One-hot encoding	Using a vector (all the elements of the vector are 0 except one) to re-code a categorical variable
Standardization	Rescale data to a specific range, e.g., dividing by mean or dividing by standard deviation
Normalization	Transforming unnormalized data into normalized data, e.g., logarithm transformation
Over-fit	The established model corresponds too exactly to the training dataset, and may therefore fail to predict future unseen observations
Imputation	Assigning the value of a missing data, e.g., using the mean of the existing data
Dimension reduction	Representing the original data with lesser dimensions
Training	The learning process of the data pattern by a model
LASSO	Least Absolute Shrinkage and Selection Operator: A regression analysis method that performs both variable selection and regularization
SVM	Support Vector Machine: Finding the best hyperplane to separate data in a high dimensional space
Naïve Bayes	A simple probabilistic classifier based on Bayes' theorem
kNN	k Nearest Neighbor: Classification of a sample according to the distance to other samples in the multidimensional space
Neural network	A family of models inspired by biological neural networks
Tree	A tree-like graph model of decisions and their possible consequences
Ensemble	Combining several different models, calculating predictions from these models and then those predictions are used as weighted inputs into another regression model for the ultimate prediction
Parameters	Coefficients of a model formula that need to be learned from the data
Hyperparameters	All the configuration variables of a model which are often set manually by the practitioner
Validation	Calculating performance of a trained model in a separated dataset
Discrimination	The ability of a model to separate individuals in multiple classes
Calibration	How well a model's predicted probabilities concur with the actual probabilities
Cross-validation	First, the data is partitioned into k (5 or 10) equally sized parts randomly with one part as the validation dataset and others as the training dataset. This process is repeated for k times with each of the subsamples used exactly once as the validation dataset
Leave-one-out	Leaving one sample out each time and training the model on the remaining samples. The process is repeated multiple times till all the samples are "leave-outed" once
Bootstrapping	Randomly sampled data from the whole original data (patients can be sampled multiple times) can be used to create new data. Training and validation are based on the new data, and the resampling process is repeated multiple times
Robust	The stability of a model in cross-validation or in sensitivity analysis
Feature importance	How much the accuracy decreases when the feature is excluded

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-19-0156

Table 4 Quality assessment of machine learning studies in sellar region disease.

	Unmet need	Reproducibility			Robustness		Generalizability	Clinical significance	
	Limits in current non-machine-learning approach	Feature engineering	Platforms, packages	Hyperparameters	Valid methods for over-fitting	Stability of results	External data validation	Predictors explanation	Suggested clinical use
Learned-Miller 2006 (6)	Yes	No	Yes	No	Yes	No	No	No	Yes
Kitajima 2009 (7)	Yes	Yes	No	Yes	Yes	No	No	No	Yes
Lalys 2011 (8)	No	Yes	No	Yes	Yes	Yes	No	No	Yes
Hu 2012 (9)	No	NA	Yes	Yes	Yes	No	No	No	Yes
Steiner 2012 (10)	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes
Calligaris 2015 (11)	Yes	NA	No	No	No	No	No	No	Yes
Paul 2017 (12)	Yes	Yes	No	Yes	Yes	Yes	No	No	No
Kong 2018 (13)	Yes	Yes	No	Yes	No	No	No	No	Yes
Zhang 2018 (14)	Yes	Yes	Yes	No	Yes	No	No	No	Yes
Murray 2018 (15)	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes
Yang 2018 (16)	Yes	Yes	Yes	Yes	No	No	Yes	No	No
Hollon 2018 (17)	No	No	Yes	No	No	No	No	Yes	No
Staartjes 2018 (18)	Yes	Yes	Yes	No	No	Yes	No	Yes	Yes
Kocak 2018 (19)	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Ortea 2018 (20)	Yes	NA	Yes	No	Yes	No	No	No	Yes
Smyczynska 2018 (21)	Yes	Yes	No	Yes	No	No	No	Yes	No

NA, no need.

locations and period of the included patients; how the control group was determined; all the variables as predictors; the data coding, data transformation methods; missing data imputation methods; any censoring data. The methods part should also include the reason for choosing a specific model; the platform and the package for model building (recommended in Supplementary Table 3) and all the hyperparameters in the model if applicable. Reporting of the results should include the rate of binary outcome or the distribution of categorical or continuous outcome; the appropriate validation statistic based on the clinical question; the 95% confidence interval by cross-validation or bootstrap and whether an external validation was obtained. Reporting of the discussion should include the reason if arbitrarily chosen cut-off value; the clinical meaning of the discrimination or calibration statistics; explanation of the model (provide coefficients or feature importance if possible); discussion on how the model will be integrated into clinical care.

Table 5 A proposed reporting checklist of future studies using machine learning.

Reporting of background should include
Results of human intelligence or non-machine-learning approach
A summarized research question
Reporting of method should include
Diagnoses of the cohort
Locations and time span of the patients included
How the control group was determined
All the variables as predictors
Data coding and data transformation methods
Missing data imputation methods
Any censoring data
The reason for choosing a specific model
The platform and the package for model building
All the hyperparameters in the model if applicable
Reporting of results should include
The rate of binary outcome or the distribution of categorical or continuous outcome
The appropriate validation statistic based on the clinical question
95% confidence interval of validation statistic by cross-validation or bootstrapping
Whether an external validation was obtained
Reporting of the discussion should include
The reason if arbitrarily chosen cut-off value
Clinical meaning of the discrimination or calibration statistics
Explanation of the model (provide coefficients or feature importance if possible);
Discussion on how the model will be integrated in clinical care

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-19-0156

Discussion

The review summarized studies on sellar region disease with machine learning methods about cohort selections, predictors, outcomes, model buildings and validation methods. A quality assessment tool was proposed in these aspects: unmet needs, reproducibility, robustness, generalizability and clinical significance. A reporting checklist from the introduction to the discussion was also provided for future studies.

Though machine learning methods have the potential advantage of increasing the prediction power, researchers should always focus on the clinical questions. The unmet needs in current practice either in diagnosis or in posttreatment prediction were the drivers for expanding the use of this new method. In particular cases, results of human-level intelligence (13) should be tested in scenarios where the predictions are majorly dependent on clinicians' subjective judgments in current standard care, for example, in studies in predicting gross-total resection rate after pituitary adenoma surgery (18). In both studies, human-level intelligence results by either physicians' judgment or conventional prediction technique were provided. In particular situations, if the diagnostic process needs too much human labor, it was also a good argument for the application of machine learning methods (22, 23).

In general, machine learning studies were retrospective observational studies, and the predictors were usually all the variables which have been recorded. On the other hand, features can be generated by transforming data already collected using specific methods (standardization, normalization, centralization) (24). These methods should be reported in the method part for study replication. There were also a few feature selection methods (25), and most of them were based on maximizing the validation statistics. But we should be bear in mind that feature selection can either improve the robustness or have the potential to harm the generalizability.

Unsupervised leaning models are usually not used in clinical studies, because the purpose of these approaches is to find similarities among samples where no outcomes are available, for example, genomic grouping. In selecting specific supervised machine learning algorithms, no common rules apply. Because there was no guarantee that a certain algorithm performs the best in all kind of data. In general, neural networks performs better than other models in image data, and tree-based models perform better in tabular data.

Platforms, packages, parameters and hyperparameters were other critical issues for study replication, but only half of the studies provided this information. Algorithms like logistic regression with regulation, linear discriminant analysis, k means and k nearest neighbor are relatively easy to implement and do not require many hyperparameters. Support vector machine, decision tree-based models and neural networks are more complicated and need tons of hyperparameters during training. Proper reporting was necessary for study replication using these models.

Leave-one-out holds one sample out each time and trains the model on the remaining samples. Similarly, k-fold cross-validation (k = 5 or 10 in general) holds 1/5th or 1/10th of the samples out each time and trains the model on the remaining samples (9, 19, 22). Bootstrap samples patients from the whole original data randomly to create new data in which a model was trained, and the resampling process is repeated multiple times (14, 20). It was not recommended to randomly split the data into two parts (training and testing) because it may have a big chance to achieve a relatively 'good' testing data such that biasing the model performance to the better direction.

Calibration seems not so important in sellar region disease in this systematic review. When the research question is to predict the classification, it is not important whether the predicted probabilities deviate to the real probabilities because the goal is to discriminate the predicted values between the two classes. On the other hand, in situations when predicting the probability of a specific class (e.g. mortality risk) is important, the predicted probabilities should be calibrated to avoid deviating too much from the actual probabilities (26). Calibration is usually measured by Hosmer–Lemeshow goodness-of-fit test or by calibration belt plotting the distribution of real probability versus the predicted probability (27).

Generalizability is another major concern in machine learning studies. The population to be generalized should have similar characteristics distribution and outcome proportion. If a model is to be truly applied in the clinical setting, it should be validated in another database. Recent food drug administration approved aided diagnostic tool for diabetic retinopathy diagnosis, atrial fibrillation detection and other diseases all require validation in the external dataset (4).

Sometimes clinicians want to know which factors drive the model for the prediction in the whole population or in a particular patient, which highlights the importance of model explanation. On the population level, this can be solved by looking at coefficients of each variable in logistic regressions or calculating feature importance in

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-19-0156

tree-based models or neural networks. But sometimes individual-level explanation may be more important, which necessitate the interpretation of each variable in each sample. This procedure can be calculated by SHAP score, which means the contribution of each variable to the final prediction value (28). But both explanations only tell why the model performs like the way it functions, but not anything about how we can improve our clinical practice, which is one major limit in machine learning methods.

Clinical applications include multiple aspects. Developing a smartphone application for acromegaly detection may help to increase the diagnostic rate of acromegaly (13). Using histological spectrum to differentiate different tumor types may help quicker and more accurate intra-operative diagnosis (11). Predicting somatostatin analog sensitivity can guide future clinical trials by recruiting patients more sensitive to the medications (19). Precise prediction of postoperative adverse events may help to alarm surgeons to pay more attention to those patients who have a higher likelihood of developing these events (17). Web-based online real-time prediction can also help increase physician–physician or physician–patient communication (29).

Although machine learning approach provided additional prediction power comparing to conventional regression models, several concerns in applying this approach were listed as follows: (1) the superiority of prediction power were not guaranteed in every case; (2) machine learning approach is more data consuming and time consuming, thus is less efficient than conventional models; (3) different platforms, different packages and multiple hyperparameters of machine learning approach restrict its replicability among different research groups. Current gaps of knowledge still exist in how to correctly explain the machine learning models either in the global level or in the individual level.

Conclusion

Machine learning methods were used to predict diagnosis and posttreatment outcomes in sellar region diseases. Though most studies had substantial unmet needs and proposed possible clinical application, replicability robustness assessed by variations in the validation statistics and generalizability evaluated by the external database, were major limits in current studies. Population-level and individual-level predictors explanation are also directions for future improvements.

Supplementary data
This is linked to the online version of the paper at https://doi.org/10.1530/EC-19-0156.

Declaration of interest
The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding
Dr Qiao is supported by 2018 Milstein Medical Asian American Partnership Foundation translational medicine fellowship. This study is supported by Shanghai Committee of Science and Technology, China (grant NO. 17JC1402100 and 17YF1426700).

Acknowledgement
Research involving human participants and/or animals: this article does not contain any studies with human participants performed by any of the authors.

References

1 Bresson D, Herman P, Polivka M & Froelich S. Sellar lesions/pathology. *Otolaryngologic Clinics of North America* 2016 **49** 63–93. (https://doi.org/10.1016/j.otc.2015.09.004)
2 Freda PU & Post KD. Differential diagnosis of sellar masses. *Endocrinology and Metabolism Clinics of North America* 1999 **28** 81–117, vi. (https://doi.org/10.1016/S0889-8529(05)70058-X)
3 Goodman SN, Fanelli D & Ioannidis JPA. What does research reproducibility mean? *Science Translational Medicine* 2016 **8**:341ps12. (https://doi.org/10.1126/scitranslmed.aaf5027)
4 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 2019 **25** 44–56. (https://doi.org/10.1038/s41591-018-0300-7)
5 Erickson BJ, Korfiatis P, Akkus Z & Kline TL. Machine learning for medical imaging. *RadioGraphics* 2017 **37** 505–515. (https://doi.org/10.1148/rg.2017160130)
6 Learned-Miller E, Lu Q, Paisley A, Trainer P, Blanz V, Dedden K & Miller R. Detecting acromegaly: screening for disease with a morphable model. *Medical Image Computing and Computer-Assisted Intervention* 2006 **9** 495–503. (https://doi.org/10.1007/11866763_61)
7 Kitajima M, Hirai T, Katsuragawa S, Okuda T, Fukuoka H, Sasao A, Akter M, Awai K, Nakayama Y, Ikeda R, *et al*. Differentiation of common large sellar-suprasellar masses effect of artificial neural network on radiologists' diagnosis performance. *Academic Radiology* 2009 **16** 313–320. (https://doi.org/10.1016/j.acra.2008.09.015)
8 Lalys F, Riffaud L, Morandi X & Jannin P. Surgical phases detection from microscope videos by combining SVM and HMM. *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*. MCV 2010. Lecture Notes in Computer Science **6533** 54–62. (https://doi.org/10.1007/978-3-642-18421-5_6)
9 Hu X, Zhang P, Shang A, Li Q, Xia Y, Jia G, Liu W, Xiao X & He D. A primary proteomic analysis of serum from patients with nonfunctioning pituitary adenoma. *Journal of International Medical Research* 2012 **40** 95–104. (https://doi.org/10.1177/147323001204000110.
10 Steiner G, Mackenroth L, Geiger KD, Stelling A, Pinzer T, Uckermann O, Sablinskas V, Schackert G, Koch E & Kirsch M. Label-free differentiation of human pituitary adenomas by FT-IR spectroscopic imaging. *Analytical and Bioanalytical Chemistry* 2012 **403** 727–735. (https://doi.org/10.1007/s00216-012-5824-y)

11 Calligaris D, Feldman DR, Norton I, Olubiyi O, Changelian AN, Machaidze R, Vestal ML, Laws ER, Dunn IF, Santagata S, *et al*. MALDI mass spectrometry imaging analysis of pituitary adenomas for near-real-time tumor delineation. *PNAS* 2015 **112** 9978–9983. (https://doi.org/10.1073/pnas.1423101112)

12 Paul JS, Plassard AJ, Landman BA & Fabbri D. Deep learning for brain tumor classification. *Proceedings SPIE* 2017 **10137** 1013710–1013717. (https://doi.org/10.1117/12.2254195)

13 Kong X, Gong S, Su L, Howard N & Kong Y. Automatic detection of acromegaly From facial photographs using machine learning methods. *EBioMedicine* 2018 **27** 94–102. (https://doi.org/10.1016/j.ebiom.2017.12.015)

14 Zhang S, Song G, Zang Y, Jia J, Wang C, Li C, Tian J & Di Dong ZY. Non-invasive radiomics approach potentially predicts non-functioning pituitary adenomas subtypes before surgery. *European Radiology* 2018 **28** 1–10. (https://doi.org/10.1007/s00330-017-5180-6)

15 Murray PG, Stevens A, De Leonibus C, Koledova E, Chatelain P & Clayton PE. Transcriptomics and machine learning predict diagnosis and severity of growth hormone deficiency. *JCI Insight* 2018 **3** 992–914. (https://doi.org/10.1172/jci.insight.93247)

16 Yang J, Hou Z, Wang C, Wang H & Zhang H. Gene expression profiles reveal key genes for early diagnosis and treatment of adamantinomatous craniopharyngioma. *Cancer Gene Therapy* 2018 **25** 227–239. (https://doi.org/10.1038/s41417-018-0015-4)

17 Hollon TC, Parikh A, Pandian B, Tarpeh J, Orringer DA, Barkan AL, McKean EL & Sullivan SE. A machine learning approach to predict early outcomes after pituitary adenoma surgery. *Neurosurgical Focus* 2018 **45** E8. (https://doi.org/10.3171/2018.8.FOCUS18268)

18 Staartjes VE, Serra C, Muscas G, Maldaner N, Akeret K, van Niftrik CHB, Fierstra J, Holzmann D & Regli L. Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: a pilot study. *Neurosurgical Focus* 2018 **45** E12. (https://doi.org/10.3171/2018.8.FOCUS18243)

19 Kocak B, Durmaz ES, Kadioglu P, Korkmaz OP, Comunoglu N, Tanriover N, Kocer N, Islak C & Kizilkilic O. Predicting response to somatostatin analogues in acromegaly: machine learning-based high-dimensional quantitative texture analysis on T2-weighted MRI. *European Radiology* 2018 **20** 1–9. (https://doi.org/10.1007/s00330-018-5876-2)

20 Ortea I, Ruíz I, Cañete R, Caballero-Villarraso J & Cañete MD. Identification of candidate serum biomarkers of childhood-onset growth hormone deficiency using SWATH-MS and feature selection. *Journal of Proteomics* 2018 **175** 1–32. (https://doi.org/10.1016/j.jprot.2018.01.003.

21 Smyczyńska U, Smyczyńska J, Hilczer M, Stawerska R, Tadeusiewicz R & Lewinski A. Pre-treatment growth and IGF-I deficiency as main predictors of response to growth hormone therapy in neural models. *Endocrine Connections* 2018 **7** 239–249. (https://doi.org/10.1530/EC-17-0277)

22 Qiao N. Using deep learning for the classification of images generated by multifocal visual evoked potential. *Frontiers in Neurology* 2018 **9** 638. (https://doi.org/10.3389/fneur.2018.00638)

23 Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY, *et al*. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017 **318** 2211–2223. (https://doi.org/10.1001/jama.2017.18152)

24 Jamshidi A, Pelletier JP & Martel-Pelletier J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nature Reviews. Rheumatology* 2019 **15** 49–60. (https://doi.org/10.1038/s41584-018-0130-5)

25 Wang L, Wang Y & Chang Q. Feature selection methods for big data bioinformatics: a survey from the search perspective. *Methods* 2016 **111** 21–31. (https://doi.org/10.1016/j.ymeth.2016.08.014.

26 Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, McGinn T & Guyatt G. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* 2017 **318** 1377–1384. (https://doi.org/10.1001/jama.2017.12126)

27 Nattino G, Finazzi S & Bertolini G. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. *Statistics in Medicine* 2014 **33** 2390–2407. (https://doi.org/10.1002/sim.6100.

28 Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, King-Wai Low D, Newman SF, Kim J, *et al*. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2018 **2** 749–760. (https://doi.org/10.1038/s41551-018-0304-0)

29 Karhade AV, Thio QCBS, Ogink PT, Shah AA, Bono CM, Oh KS, Saylor PJ, Schoenfeld AJ, Shin JH, Harris MB, *et al*. Development of machine learning algorithms for prediction of 30-day mortality After surgery for spinal metastasis. *Neurosurgery* 2018 **35** 2419. (https://doi.org/10.1093/neuros/nyy469)

Received in final form 4 June 2019
Accepted 11 June 2019
Accepted Preprint published online 11 June 2019