Original Article
Hormonal regulation of male reproduction and hypogonadism

pISSN: 2287-4208 / eISSN: 2287-4690
World J Mens Health 2021 Jan 39(1): 139-146
https://doi.org/10.5534/wjmh.190077

The World Journal of MEN's HEALTH

Optimizing Aging Male Symptom Questionnaire Through Genetic Algorithms Based Machine Learning Techniques

Jin Wook Kim[1], Du Geon Moon[2]

[1]Department of Urology, Chung-Ang University College of Medicine, [2]Department of Urology, Korea University College of Medicine, Seoul, Korea

Purpose: Genetic algorithm (GA) is a machine learning optimization strategy where sample strategies compete for fitness to evolve an optimum solution. This study evolves the Aging Male Symptoms (AMS) with GA to better identify late onset hypogonadism (LOH) with serum testosterone.

Materials and Methods: GA was trained on a training set of standard AMS questionnaire on a nationwide LOH epidemiology study. Random matrices of selectors for particular items were generated. Each generation of was evolved through a fitness function determined by sensitivity. Threshold to determine positive serum testosterone level for LOH was randomized for each competing strategy. After 2,000 runs, with each run producing the best result out of a set of 3,000 randomly generated sets evolved through 300 generations, the best AMS selection matrix was then applied to a separately enrolled validation set to compare outcomes.

Results: Predictability for serum testosterone levels dropped markedly above 3.5 ng/mL during pilot training. Limiting the training to testosterone thresholds between 2.5 and 3.5 ng/mL the GA 93 different strategies. Only a selection of 5 items, determining for a threshold of 20 points and determining for a serum testosterone level of 3.16 ng/mL, showed robust reproducibility within the internal validation set. Applying these conditions to the independent validation set showed sensitivity improved from 0.66 to 0.77, with a specificity of 0.07 to 0.19, respectively.

Conclusions: GA method of selecting questionnaires improved AMS questionnaire significantly. This method can be easily applied to other questionnaires that do not correlate with physiological markers.

Keywords: Hypogonadism; Machine learning; Questionnaire design; Testosterone

INTRODUCTION

The Aging Male Symptom (AMS) questionnaire, though widely used in estimating symptoms for late onset hypogonadism (LOH), has been notoriously inconsistent in its assessment of initial screening or follow-up [1]. Compared to the simpler Androgen Deficiency in the Aging Male (ADAM) questionnaire, it provides a detailed query of clinically relevant points, while also providing a scaled response. Hence, despite the lack of physiological relevance to these questions, AMS continues to enjoy a wide measure of use, despite most studies decrying its lack of specificity [2].

Conventional statistical methods employing various

Received: May 14, 2019 **Revised:** Jul 28, 2019 **Accepted:** Aug 8, 2019 **Published online** Jan 9, 2020
Correspondence to: Jin Wook Kim https://orcid.org/0000-0003-4157-9365
Department of Urology, Chung-Ang University College of Medicine, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea.
Tel: +82-2-6299-1807, **Fax:** +82-2-6298-8351, **E-mail:** jinwook@cau.ac.kr

multivariate regression analyses to identify relevant questionnaires often fall short [3-5]. Regression methods require each component of the multivariate equation to be independent of each other. Even if a cross correlation presents itself through a validation study in an epidemiologic setting to state that all items showed little inter-dependence, one could simply read through the confusing list of all seventeen items to understand the overlapping spectrum of symptoms. Ironically, AMS propounds the complex overlap of its questionnaires by clustering items into symptom groups, further emphasizing that each item was never designed to be independent, and thus, the statistically predictive power of each item on the overall symptom is spread out and weakened.

Furthermore, conventional statistical methods to remove each component from the regression equation is difficult. Removing questions based on statistical relevance in a univariate setting cannot avoid the aforementioned problem of interdependence between each questionnaire. Conversely, including individual questionnaires and comparing various iterations confronts the problem of investigating 2^n possible sets of different strategies for relevance, where n is the number of items in the questionnaire.

Genetic algorithm is a machine learning method best suited to evolve an optimal outcome from near infinite iterations [6,7]. Originally created as an optimization method for complex problems, genetic algorithms create generations of sample strategies that then compete against each other for optimum fitness. Once each generation is complete, all sample strategies are then changed through various methods to create a new generation, where the most fit of the previous generation produce similar but slightly altered offspring strategies (crossover), while poor fit samples are discarded and new random samples are introduced (mutation).

This study applied genetic algorithm on a training set of a previous epidemiological study. The resultant best fit strategy of AMS questionnaires was then applied to a validation set of newly recruited patients.

MATERIALS AND METHODS

1. Pilot training

The matrix of strategies select a different number of questionnaires. Hence to uniformly apply each matrix to determine the outcome, the number of questionnaires for each strategy was multiplied by a factor of either 3 or 4 (out of 5 points), which we termed AMS weight (k), so that each strategy of N questions resulted in an AMS threshold of k×N.

Serum testosterone threshold was also tested against a random cut off value assigned between 2.5 and 5.5 ng/mL. Hence the possible strategies that could possibly evolve did not only compare only against 2^{17} variations of AMS matrices, but also contended with variabilities in threshold of AMS (3 or 4), as well as the continuous threshold of serum testosterone, thus introducing a theoretically infinite amount of complexities. However, the actual variation of serum testosterone threshold is at best equal to m-1, where m is the number of samples in the training set. This amounts to '2 [=either 3 or 4]×(m-1) [=actual possible serum testosterone thresholds]×2^{17}=93,585,408' different strategies to be evaluated.

Therefore, a pilot training was performed to predetermine the level of AMS weight and to limit the range of serum testosterone threshold. An arbitrary iteration of 1,000 cycles were set.

2. Genetic algorithm

Genetic algorithm was written on MATLAB R2019a. The code describes a separate matrix for serum testosterone and seventeen AMS questionnaire items. An arbitrary number of inclusion matrices (gene pool) were constructed initially from randomization to select or exclude each item of the AMS questionnaire. A fitness function was designed to compare and score each strategy to compare with other strategies within its generation. When each generation was complete a variety of crossover and mutation strategies were given based on fitness, with a certain degree allowed for successful sets to pass on their characteristics directly, while some degree of random mutation was also allowed within crossover stages as well. Also, the most unsuccessful portions were entirely randomized from the gene pool.

1) Crossover strategies

Partial crossover strategies and complete crossover strategies were employed randomly on equal ratio. Partial crossover generates two random points between one and seventeen. The Items between each point were switched between adjacent matrices. On the other hand, complete crossover strategies select a single crossover point to switch between strategies.

The higher correlation matrices, arbitrarily set, were preserved, while the rest were crossed over with adjacent matrices either by partial crossovers or complete crossovers. When adjacent matrices were highly similar, with low correlation, these were replaced by random generated new samples. Thus, after successive generations, the sets evolved towards higher correlation.

As with all machine learning strategies, genetic algorithm may also suffer from local minima if only similar crossover strategies were evolved. Thus, when near similar strategies repeated itself, they were replaced by entirely new sets of random mutations, introducing novel strategies to overcome local minima (Fig. 1).

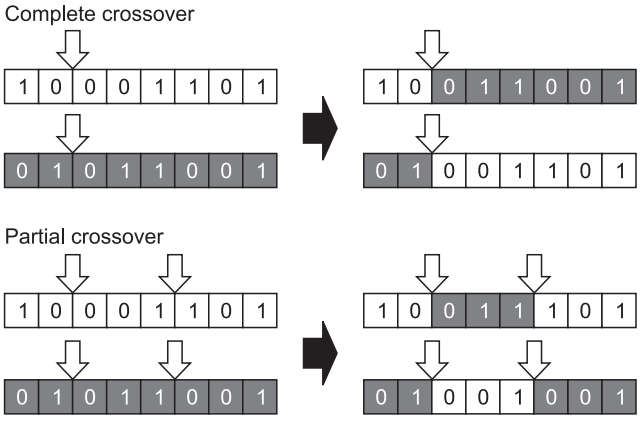

Fig. 1. Methods of crossover mutation.

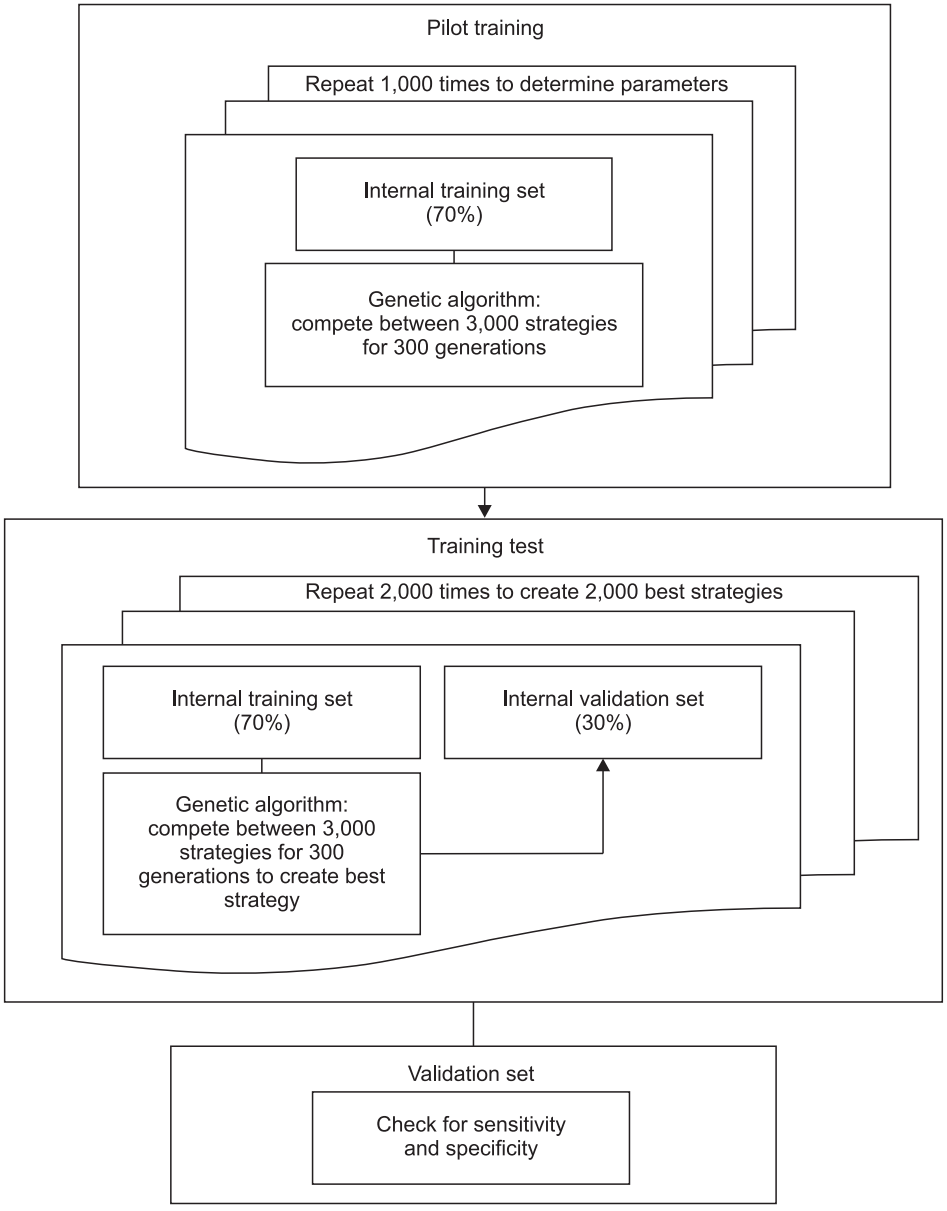

Fig. 2. Overall outline of analysis.

2) Generations and gene pool

Genetic algorithm scale is defined by the number of gene pools created and the number of generations these must evolve through. 3,000 strategies (not considering the variety produced through mutation) were evolved through 300 generations, and each such genetic algorithm cycle was reiterated 2,000 times. Appropriate strategies to minimize either serum testosterone threshold range or weight of AMS questionnaire was determined through the pilot study (Fig. 2).

3. Patients

The training set was acquired from 1,335 patients from a prospective study performed in 2014 [8]. In summary, all regions and provinces of South Korea. General checkup centers were randomly selected from the Health Insurance Review and Assessment service registry, weighing for regional population distribution by province. All participants were randomly selected through a multilevel stratified random sampling from each administrative district from January to November 2011. Only participants between 40 and 80 years of age were enrolled. For the purpose of a single visit completion of participation, only participants who visited for routine biennial physical checkup before 11:00 a.m. were eligible for inclusion to standardize measurements for serum testosterone. Within this training set, for each iteration of a thousand, patients were randomly assigned a 70 to 30 ratio to either the internal training set or the internal validation set. For each internal training set, an optimization of genetic algorithm was performed and then applied to the internal validation set. Excluding missing data provided 1,335 volunteers.

The validation set was enrolled from population of healthy volunteers between March to December 2018 at a single institute. The training set was set for 120 volunteers (Table 1).

All patients were excluded for 5α-reductase inhibitor use, phosphodiesterase 5 inhibitor use, hormone replacement therapy or any alternative solution, herbal or experimental, to improve sexual function within the last 3 months.

4. Ethics statement

The present study protocol was reviewed and approved by the Institutional Review Board of Chung-Ang University Hospital (Reg. No. C2015130). Informed consent was submitted by all subjects when they were enrolled.

RESULTS

1. Pilot training

Pilot training test was performed at 1,000 iterations to determine serum testosterone threshold range and AMS questionnaire weight. As shown in the scatterplot (Fig. 3), there was no difference between AMS weights 3 and 4. However, there was a significant drop-off of sensitivity when serum testosterone thresholds were raised above 3.5 ng/mL.

Table 1. Characteristics of validation population

Characteristic	Mean	SD
Age (y)	60.80	9.14
Serum testosterone (ng/mL)	3.49	1.12
Free testosterone (pg/mL)	79.85	33.26
Smoking state		
Quit smoking (%)	53.3	-
Still smoking (%)	25.8	-
Never smoked (%)	20.8	-
Height (cm)	168.63	5.53
Weight (kg)	72.90	9.87
Systolic BP (mmHg)	131.14	13.01
Diastolic BP (mmHg)	78.77	11.00
Waist (cm)	90.65	6.98
Hip (cm)	98.94	5.83

SD: standard deviation, BP: blood pressure.

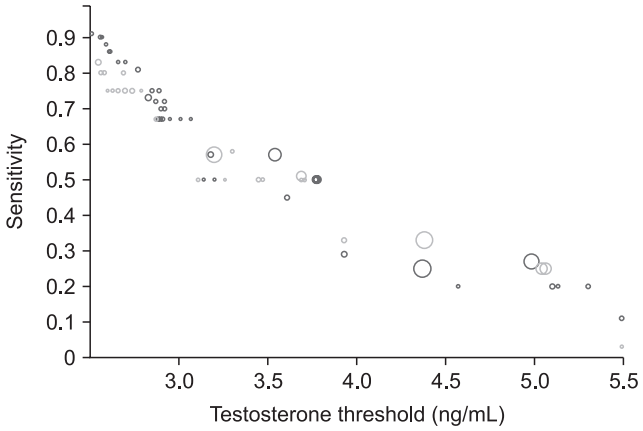

Fig. 3. Serum testosterone threshold *vs.* sensitivity displayed during 1,000 pilot training cycles; darker circles showing Aging Male Symptoms (AMS) weight at 3, lighter circles showing AMS weight at 4. There was no difference between applying AMS weight, however, sensitivity drop-off was noticeable above threshold of 3.5 ng/mL.

2. Genetic algorithm training and internal validation

With AMS weight randomly assigned either 3 or 4, and serum testosterone threshold set randomly between 2.5 to 3.5 ng/mL, the training set of 1,335 volunteers were developed through 3,000 gene pool samples evolved to 300 generations 2,000 times to produce 2,000 best strategies and their applied internal validation results.

Overall 93 different strategies for determining AMS were devised through machine learning, with an overall sensitivity of 0.67 and a specificity of 0.41. However, within internal validation these outcomes dropped off significantly to a sensitivity of 0.56 and a specificity of 0.06. When these items were divided by the robustness of the overall matrix to maintain a persistent 0.6 sensitivity in the internal validation set, only a few items remained to prove viable (items, 4, 8, 12, 14, and 17).

Independent t-test showed serum testosterone for robust outcomes was 3.16±0.22 *vs.* 2.74±0.16 ng/mL ($p<0.001$), while chi-square test showed robust sets preferred AMS weight of 4 than 3 ($p<0.001$).

Thus, based on these recommendations, items 4, 8, 12, 14, and 17, determining for an AMS score of 20 (AMS weight 4×5 items) or above to predict for serum testosterone of 3.16 ng/mL was used. A matrix composing of only these items, solving for serum testosterone 3.16 ng/mL produced predicted a sensitivity of 0.90, and a specificity of 0.26 for the entire training set (Table 2, Fig. 4).

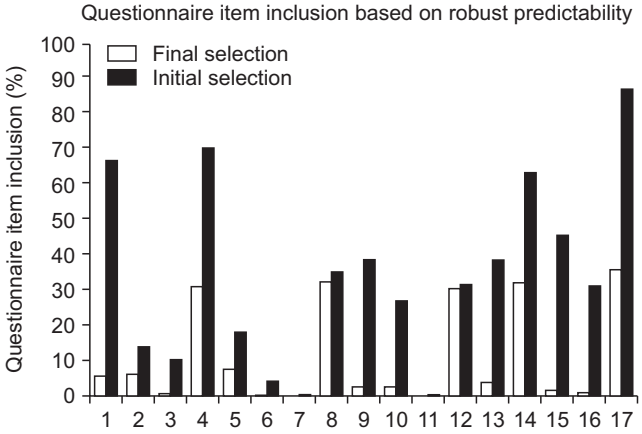

Fig. 4. Results of genetic algorithm show initial selections which were contrasted against internal validation to discriminate final selection matrix.

Table 2. Comparison of questionnaire item selection

Item No.	Item	Original AMS	Linear regression		Genetic algorithm		Genetic algorithm +robustness	
			p-value	Selection	Frequency	Selection	Frequency	Selection
1	Decline in your feeling of general well-being	O	0.45	-	0.30	O	0.06	-
2	Joint pain and muscular ache	O	0.80	-	0.10	-	0.06	-
3	Excessive sweating	O	0.48	-	0.05	-	0.01	-
4	Sleep problems	O	0.88	-	0.47	O	0.31	O
5	Increased need for sleep, often feeling tired	O	0.19	O	0.12	-	0.08	-
6	Irritability	O	0.74	-	0.02	-	0.00	-
7	Nervousness	O	0.94	-	0.00	-	0.00	-
8	Anxiety	O	0.78	-	0.34	O	0.33	O
9	Physical exhaustion/lacking vitality	O	0.95	-	0.17	O	0.03	-
10	Decrease in muscular strength	O	0.28	-	0.13	O	0.03	-
11	Depressive mood	O	0.84	-	0.00	-	0.00	-
12	Feeling that you have passed your peak	O	0.95	-	0.31	O	0.31	O
13	Feeling burnt out, having hit rock-bottom	O	0.15	O	0.18	O	0.04	-
14	Decrease in beard growth	O	0.79	-	0.45	O	0.33	O
15	Decrease in ability/frequency to perform sexually	O	0.61	-	0.20	O	0.02	-
16	Decrease in the number of morning erections	O	0.59	-	0.13	O	0.01	-
17	Decrease in sexual desire/libido	O	0.23	-	0.57	O	0.36	O
	Sensitivity	0.66		0.58		0.68		0.77
	Specificity	0.07		0.37		0.12		0.19

AMS: Aging Male Symptoms.

3. Validation

Validation test for 1,335 patients of the epidemiological study showed a sensitivity of 0.77, and a specificity of 0.19, improving over a sensitivity of 0.66, and a specificity of 0.07 when considering the entire AMS set as a matrix for equal serum testosterone cut off points of 3.16 ng/mL.

DISCUSSION

The initial machine learning process to train the algorithm best suited to correlate with serum testosterone was performed against the original epidemiologic data. The data from 1,895 participants were cleaned for missing values, leaving 1,335 participants. Initially we attempted to take advantage of the benefit of using an AMS score over simpler identifiers, such as the ADAM score, *i.e.*, a gradually scaled response providing, hopefully, a full spectrum of responses. However, the correlation function showed weak reproducibility primarily through oversimplifying the outcome.

Hence, sensitivity was chosen as the fitness function, while expanding the complexity of the genetic algorithm to inspect the entire spectrum of serum testosterone thresholds. Interestingly, the initial internal training provided a wide array of questionnaires fit for validation, but was most successful to show that a significant portion of the items, such as 2, 3, 6, 7, 11 had no function in relevance to serum testosterone.

Additionally, through rigorous internal validation, we were also able to discern 'false positive' questionnaires, such as items, 1, 9, 10, 13, 15, and 16, which showed high fitness during training but sharply lost relevance when applied to internal validation. Through this process, only a select number of items showed sharp distinction and were successful in identifying low serum testosterone in the independent validation set.

It is noteworthy to see that these select items do not overlap with items selected through conventional methods, such as linear regression (Table 2). Despite the robustness for selecting matrices best fitting the outcome, frequentist methods could not thoroughly predict all possible combinations that could be mired from deviations from the prerequisites of variable independence and normality [9]. For example, items such as 4 and 5, both describing some symptom of sleep cannot possibly be independent of each other, yet a multitude of investigations concerning the AMS often present outcomes where investigators treat these as independent variables in analysis. Genetic algorithms can overcome these pitfalls by dissociating the process of selection from the assumption of normality and independence [6,7,10]. The algorithm itself only compares and develops the outcomes based on fitness alone; if codependent factors are included, it does not claim to make judgement that these factors independently affect the overall outcome. Indeed, if such codependence would hamper evolving the fitness function, the genetic algorithm would most likely either eliminate it entirely from the process until such time that randomness reintroduces such factors into the equation again, or stall into a local minima, again, in which randomization would rescue it from over fitting.

Ultimately, the overall outcome shows only relatively distinct items included in the final questionnaire. One single item for sleep (item 4), one single item for mood (item 8), one single item for vitality (item 12), one single item that had been generally ignored (item 14), and one single item most central to the AMS (item 17). Also, it is interesting to note that, during the entire process, the threshold for serum testosterone had been left randomized, only limiting the focus when predictability went out of hand, yet ultimately the threshold had settled down to 3.16 ng/mL as the most deterministic level, which can be converted to 11 nmol/L, a widely accepted level for LOH [4,11,12].

Overall, one cannot say that a validation sensitivity of 0.77 is very high, especially considering a poor specificity of only 0.19. Recently Lu et al [13], approached LOH diagnosis through machine learning techniques employing a decision tree method to incorporate various serologic criteria and compared these to questionnaires. The results showed higher predictive value in serologic criteria alone.

However, dismissing questionnaires in diagnosing LOH entirely runs the fundamental problem of approaching a disease without clarifying its symptoms. How can a disease be independent of presenting symptoms? Without clarification of what centrally constitutes the constellation of presentations of a symptom leaves us with no symptoms and only a series of equations describing serology.

Hence, any attempt to salvage these questionnaires through any means is worthwhile. While AMS suffers from any credible criteria, it does not suffer from any lack of symptoms queried. AMS provides a wide range

of symptoms, whether they are specifically related to the diagnosis of LOH.

Finally, in clarification, the questionnaire used in this study was the Korean translation of the AMS questionnaire, previously used in several nationwide studies [8,14,15]. There is poor validation information regarding the AMS, as questions of its utility had already arose by the time several nationwide studies had been conducted, and no future prospects of formal validation seems likely. Perhaps a refinement of the questionnaire may help improve these matters.

CONCLUSIONS

Genetic algorithm method of selecting questionnaires improved AMS questionnaire significantly. This method can be easily applied to other questionnaires that do not correlate with physiological markers.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1C1B5076536).

The authors thank Dr. Hyun-joon Kim, Samsung Advanced Institute of Technology, Device Lab, for his technical assistance for this study.

Conflict of Interest

The authors have nothing to disclose.

Author Contribution

Conceptualization: JWK, DGM. Data curation: JWK, DGM. Formal analysis: JWK. Funding acquisition: JWK. Investigation: JWK, DGM. Methodology: JWK, DGM. Project administration: JWK, DGM. Resources: JWK, DGM. Software: JWK. Supervision: JWK, DGM. Validation: JWK, DGM. Visualization: JWK. Writing–original draft: JWK. Writing–review & editing: JWK, DGM.

Data Sharing Statement

The data analyzed for this study have been deposited in HARVARD Dataverse and are available at https://doi.org/10.7910/DVN/9V2I0P.

REFERENCES

1. Dobs AS. The role of accurate testosterone testing in the treatment and management of male hypogonadism. Steroids 2008;73:1305-10.
2. Taniguchi H, Kawa G, Kinoshita H, Matsuda T. Change of aging males' symptoms (AMS) rating scale in Japanese late-onset hypogonadism (LOH) patients administered androgen replacement therapy. J Men Health 2011;8:S67-70.
3. Kang S, Park HJ, Park NC. Serum total testosterone level and identification of late-onset hypogonadism: a community-based study. Korean J Urol 2013;54:619-23.
4. Lunenfeld B, Saad F, Hoesl CE. ISA, ISSAM and EAU recommendations for the investigation, treatment and monitoring of late-onset hypogonadism in males: scientific background and rationale. Aging Male 2005;8:59-74.
5. Kong XB, Guan HT, Li HG, Zhou Y, Xiong CL. The ageing males' symptoms scale for Chinese men: reliability,validation and applicability of the Chinese version. Andrology 2014;2:856-61.
6. Brown EC, Sumichrast RT. Impact of the replacement heuristic in a grouping genetic algorithm. Comput Oper Res 2003;30:1575-93.
7. Mantzaris D, Anastassopoulos G, Adamopoulos A. Genetic algorithm pruning of probabilistic neural networks in medical disease estimation. Neural Netw 2011;24:831-5.
8. Moon du G, Kim JW, Kim JJ, Park KS, Park JK, Park NC, et al. Prevalence of symptoms and associated comorbidities of testosterone deficiency syndrome in the Korean general population. J Sex Med 2014;11:583-94.
9. Wilcox RR. Some practical reasons for reconsidering the Kolmogorov-Smirnov test. Br J Math Stat Psychol 1997;50:9-20.
10. Eisenbarth H, Lilienfeld SO, Yarkoni T. Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory-Revised (PPI-R). Psychol Assess 2015;27:194-202.
11. Morales A, Lunenfeld B; International Society for the Study of the Aging Male. Investigation, treatment and monitoring of late-onset hypogonadism in males. Official recommendations of ISSAM. International Society for the Study of the Aging Male. Aging Male 2002;5:74-86.
12. Wang C, Nieschlag E, Swerdloff R, Behre HM, Hellstrom WJ, Gooren LJ, et al. ISA, ISSAM, EAU, EAA and ASA recommendations: investigation, treatment and monitoring of late-onset hypogonadism in males. Int J Impot Res 2009;21:1-8.
13. Lu T, Hu YH, Tsai CF, Liu SP, Chen PL. Applying machine learning techniques to the identification of late-onset hypogonadism in elderly men. Springerplus 2016;5:729.
14. Park DS, Kim TB, Ku JH, Kim SW, Paick JS. Correlation be-

tween androgen deficiency on the aging males questionnaire and the aging males' symptoms Scale and their relationship with serum testosterone. Korean J Urol 2008;49:1035-40.
15. Ryu JK, Cho KS, Kim SJ, Oh KJ, Kam SC, Seo KK, et al. Korean Society for Sexual Medicine and Andrology (KSSMA) Guideline on erectile dysfunction. World J Mens Health 2013; 31:83-102.