Preference for artificial intelligence medicine before and during COVID-19 pandemic: Discrete choice experiment with propensity score matching

Taoran Liu, Winghei Tsang, Yifei Xie, Kang Tian, Fengqiu Huang, Yanhui Chen, Oiying Lau, Guanrui Feng, Jianhao Du, Bojia Chu, Tingyu Shi, Junjie Zhao, Yiming Cai, Xueyan Hu, Babatunde Akinwunmi, Jian Huang, Casper JP Zhang, Wai-Kit Ming

Submitted to: Journal of Medical Internet Research
on: January 07, 2021

Table of Contents

Original Manuscript **5**
Supplementary Files **36**
Figures **37**
Figure 1 ..... **38**
Figure 2 ..... **39**
Figure 3 ..... **40**
Figure 4 ..... **41**
Figure 5 ..... **42**

Preference for artificial intelligence medicine before and during COVID-19 pandemic: Discrete choice experiment with propensity score matching

Taoran Liu[1*] BSc; Winghei Tsang[1*] MBBS; Yifei Xie[2] MBBS; Kang Tian[3] BSc; Fengqiu Huang[1] MPH; Yanhui Chen[1] MBBS; Oiying Lau[1] MBBS; Guanrui Feng[1] MPH; Jianhao Du[1] MPH; Bojia Chu[4] BA, MSc; Tingyu Shi[5] BSc; Junjie Zhao[6] BSc; Yiming Cai[7]; Xueyan Hu[1] BEcon; Babatunde Akinwunmi[8, 9] PhD, MM, MPH, MMSc; Jian Huang[10] PhD, MPH; Casper JP Zhang[11] PhD, MPH; Wai-Kit Ming[1] MD, PhD, MPH, MMSc

[1]Department of Public Health and Preventive Medicine School of Medicine Jinan University Guangzhou CN
[2]International School Jinan University Guangzhou CN
[3]Faculty of Social Sciences University of Southampton Southampton GB
[4]Department of Applied Mathmatics The Hong Kong Polytechnic University Hong Kong HK
[5]University of Southampton Southampton GB
[6]College of Computer Science and Technology Henan Polytechnic University Henan CN
[7]School of Applied Mathematics Beijing normal university (Zhuhai) Zhuhai CN
[8]Department of Obstetrics and Gynecology Brigham and Women's Hospital Boston US
[9]Center for Genomic Medicine (CGM), Massachusetts General Hospital Harvard Medical School Harvard University Boston US
[10]Department of Epidemiology and Biostatistics, School of Public Health Imperial College London London GB
[11]School of Public Health The University of Hong Kong Hong Kong HK
[*]these authors contributed equally

Corresponding Author:
Wai-Kit Ming MD, PhD, MPH, MMSc
Department of Public Health and Preventive Medicine School of Medicine
Jinan University
601 Huangpu W Ave
Tianhe District
Guangzhou
CN

Abstract

Background: Artificial intelligence (AI) has potential uses in relieving the public health pressure caused by the pandemic. In the case of a shortage of medical resources caused by the pandemic, whether people's preference for AI doctors and traditional clinicians has changed is worth exploring.

Objective: We aim to quantify and compare people's preference for AI medicine and traditional clinicians before and during the COVID-19 pandemic to check whether people's preference is affected by the pressure of pandemic.

Methods: The propensity score matching (PSM) method was applied to match two different groups of respondents recruited in 2017 and 2020 with similar demographic characteristics. A total of 2048 respondents (1520 from 2017 and 528 from 2020) completed the questionnaire and were included in the analysis. The Multinomial Logit Model (MNL) and Latent Class Model (LCM) were used to explore people's preferences for different diagnosis methods.

Results: Among these respondents, 84.7% in 2017 and 91.3% in 2020 were confident that AI diagnosis would outperform human clinician diagnoses in the future. Both groups of respondents matched from 2017 and 2020 attached most importance to the attribute 'accuracy', and they prefer the combined diagnosis of AI and human clinicians (2017: odds ratio [OR] 1.645; 95% CI 1.535,1.763, $p < 0.001$; 2020: OR 1.513, 95% CI 1.413, 1.621, $p < 0.001$, Reference level: Clinician). LCM identified three classes with different attribute priorities. In Class 1, the preference for combination diagnosis and accuracy remains constant in 2017 and 2020, and higher accuracy (e.g., 2017 OR for 100% 1.357; 95% CI 1.164, 1.581) is preferred. In Class 2, the 2017 matched data is also very similar to class 2 in 2020, AI combined with human clinicians (2017: OR 1.204, 95% CI 1.039, 1.394, $p = 0.011$; 2020: OR 2.009, 95% CI 1.826, 2.211, $p < 0.001$, Reference level: Clinician) and 20 minutes (2017: OR 1.349, 95% CI 1.065, 1.708, $p < 0.001$; 2020: OR 1.488, 95% CI 1.287, 1.721, $p < 0.001$, Reference level, 0 min) of outpatient waiting time were consistently preferred. In Class 3, the respondents in 2017 and 2020 had different preferences for diagnosis method;

respondents in Class 3 of 2017 prefer clinicians, whereas respondents in Class 3 of 2020 prefer AI diagnosis. As for the latent class segmented according to different sexes, all of the male and female respondent classes from 2017 and 2020 rank accuracy as the most important attribute.

Conclusions: Individual preference for clinical diagnosis between AI and human clinicians were mostly unaffected due to the pandemic. Diagnosis accuracy and expense for diagnosis were of the most important attributes of choice of the type of diagnosis. These findings can provide guidance for policymaking relevant to the development of AI-based healthcare.

(JMIR Preprints 07/01/2021:26997)
DOI: https://doi.org/10.2196/preprints.26997

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Preference for artificial intelligence medicine before and during COVID-19 pandemic: Discrete choice experiment with propensity score matching

Abstract

Background

Artificial intelligence (AI) has potential uses in relieving the public health pressure caused by the pandemic. In the case of a shortage of medical resources caused by the pandemic, whether people's preference for AI doctors and traditional clinicians has changed is worth exploring.

Objective:

We aim to quantify and compare people's preference for AI medicine and traditional clinicians before and during the COVID-19 pandemic to check whether people's preference is affected by the pressure of pandemic

Methods

The propensity score matching (PSM) method was applied to match two different groups of respondents recruited in 2017 and 2020 with similar demographic characteristics. A total of 2048 respondents (1520 from 2017 and 528 from 2020) completed the questionnaire and were included in the analysis. The Multinomial Logit Model (MNL) and Latent Class Model (LCM) were used to explore people's preferences for different diagnosis methods.

Results

Among these respondents, 84.7% in 2017 and 91.3% in 2020 were confident that AI diagnosis would outperform human clinician diagnoses in the future. Both groups of respondents matched from 2017 and 2020 attached most importance to the attribute 'accuracy', and they prefer the combined diagnosis of AI and human clinicians (2017: odds ratio [OR] 1.645; 95% CI 1.535,1.763, $p < 0.001$; 2020: OR 1.513, 95% CI 1.413, 1.621, $p < 0.001$, Reference level: Clinician). LCM identified three classes with different attribute priorities. In Class 1, the preference for combination diagnosis and accuracy remains constant in 2017 and 2020, and higher accuracy (e.g., 2017 OR for 100% 1.357; 95% CI 1.164, 1.581) is preferred. In Class 2, the 2017 matched data is also very similar to class 2 in 2020, AI combined with human clinicians (2017: OR 1.204, 95% CI 1.039, 1.394, $p = 0.011$; 2020: OR 2.009, 95% CI 1.826, 2.211, $p < 0.001$, Reference level: Clinician) and 20 minutes (2017: OR 1.349, 95% CI 1.065, 1.708, $p < 0.001$; 2020: OR 1.488, 95% CI 1.287, 1.721, $p < 0.001$, Reference level, 0 min) of outpatient waiting time were consistently preferred. In Class 3, the respondents in 2017 and 2020 had different preferences for diagnosis method; respondents in Class 3 of 2017 prefer clinicians, whereas respondents in Class 3 of 2020 prefer AI diagnosis. As for the latent class segmented according to different sexes, all of the male and female respondent classes from 2017 and 2020 rank accuracy as the most important attribute.

Conclusions

Individual preference for clinical diagnosis between AI and human clinicians were mostly unaffected due to the pandemic. Diagnosis accuracy and expense for diagnosis were of the most important attributes of choice of the type of diagnosis. These findings can provide guidance for policymaking relevant to the development of AI-based healthcare.

Keywords: propensity score matching, discrete latent traits, patients' preference, artificial intelligence, COVID-19

Introduction

Artificial intelligence (AI) technology, which is also called machine intelligence technology, has been applied in various fields, such as automation, language, and image understanding and analysis and genetic algorithms. Specifically, AI technology can currently perform better than a human on particular tasks and has the potential to substitute some traditional human occupations thanks to continuous advances in medicine, neuroscience, robotic and statistics. Especially in the medical and healthcare field [1], AI technology has gained increasingly widespread applications and a range of future opportunities, for instance, machine learning (ML), consisting of medical big data, electronic health record (EHR), computer vision, natural language processing (NLP), and intelligent robots [2]. In addition, AI technology has largely helped alleviate the masses' demand for clinicians [3].

The currently progressing novel coronavirus disease (COVID-19) has already spread over 217 countries [4] and territories across the world as of November 13, 2020, and brought tremendous threats and challenges to the global public health security system. The COVID-19 outbreak has pushed the medical system and resources of numerous countries to the brink of collapse. AI diagnosis technology, including machine learning, has started to play a role in relieving the burden of the public health system caused by the pandemic and easing the shortage of medical resources. After the commencement of the COVID-19 outbreak, the medical AI team of Alibaba DAMO Academy for Discovery, Adventure, Momentum, and Outlook rapidly developed a set of AI diagnostic technology that can interpret CT images of suspected new COVID-19 cases within 20 s, with an accuracy of 96% [5]. In the fight against the epidemic [6], digital technologies such as cloud computing, artificial intelligence, and blockchain have played a vital role.

Nowadays, AI technology combined with human clinicians, using convolutional neural network (CNN) [7], has greatly improved the diagnosis efficiency and accuracy of diagnosis and significantly reduced the diagnosis time and outpatient queuing. In 2014, the market of AI healthcare applications around the world in 2014 was 663.8 million USD, and is expected to reach 6,662 million USD in 2021 [8]. However, there remains various uncertainty in patients' preference for different diagnostic methods in both developed areas and underdeveloped areas among genders in China. Also, there is neither literature comparing the preference between AI doctors and human clinicians during the COVID-19 pandemic period and usual times nor analysis for patients' decision-making behavior from a range of aspects in different periods.

This study aims to compare the preference for AI diagnosis and traditional human clinicians between two similar demographic groups recruited in 2017 and 2020 to see if people's preference for AI and traditional human clinicians can be affected by the pressure of the COVID-19 pandemic. We used propensity score matching (PSM) to match the two groups of individuals and discrete choice experiment (DCE) to quantify and measure peoples' preference for different diagnosis methods and explore disturbance and impact factors hidden in peoples' decision-making behavior.

Methods

Overview

We designed an online questionnaire to collect demographic information of the participants and investigate the patient preferences for different diagnosis strategies. In brief, the questionnaire hypothesized seven similar scenarios from which the respondents were asked to choose a preferred diagnosis strategy.

The propensity score matching (PSM) method was applied to match two different groups of respondents from 2017 and 2020 with similar demographic characteristics. In addition, models such as multinomial logit models (MNLs) [9, 10] and latent class models (LCMs) [11] were used to evaluate and investigate the different preferences for classes between the two groups of respondents. Also, matched respondents from the 2017 group and 2020 were then compared to explore the heterogeneity or homogeneity in attributes.

Selection of attributes and levels

Individuals would choose different levels of healthcare service combinations under the trade-off. Thus, the patient outpatient queue of The First Affiliated Hospital of Jinan University (Guangzhou Overseas Chinese Hospital) and The First Affiliated Hospital of Sun Yat-sen University was randomly selected and invited to hypothesize that, facing the different alternatives, which diagnosis methods or what attributes have a prominent weight of importance or large impact in their decision.

Thus, combining the patients' hypothesis and relative literature [12–14], six different attributes and their respective levels have been included in our questionnaire experiment: 1) diagnostic method (clinicians; AI plus clinicians; AI); 2) outpatient waiting time before being asked (0, 20, 40, 60, 80, and 100 min); 3) diagnosis time (0, 15, and 30 min); 4) accuracy (ratio of correct diagnosis: 60%, 70%, 80%, 90%, and 100%); 5) follow up after diagnosis (whether the doctor can follow up and follow up at any time, yes or no); 6) diagnostic expenses (RMB 0, 50, 100, 150, 200, and 250). Every attribute and its levels have been presented in Table 1.

Table 1. Attributes hypothesized and their respective levels in a discrete choice experiment.

Label	Attributes description	Levels

nostic methods	The methods of diagnosis that patients choose to receive	Clinicians' diagnosis; AI diagnosis + Clinicians' confirmation; AI diagnosis
atient waiting time	The time that patients wait in the queue before the diagnosing process starts	0min; 20min; 40min; 60min; 80min; 100min
nosis time	The time to wait before obtaining a diagnosis result	0min; 15min; 30min
nostic accuracy	The rate of correct diagnosis	60%; 70%; 80%; 90%; 100%
w-up after diagnosis	Case tracking and follow-up after diagnosis	Yes; No
nostic expenses	The cost of diagnosis	¥ 0; ¥ 50; ¥ 100; ¥ 150; ¥ 200; ¥ 250

DCE instrument design and questionnaire

During the DCE instrument establishment, we applied the fractional factorial design method [15,16] using the Sawtooth lighthouse software (version 9.8.1) to help us obtain the optimal number of treatment scenarios. In practice, it is not always feasible for respondents to make a choice among all of the possible combinations of attributes and levels, which is known as a full factorial design. The full factorial design allows 3240 different combinations ($3 \times 6 \times 3 \times 5 \times 2 \times 6$), which is unreasonable to present them all to respondents. Thus, the fractional factorial method was essential in the design and followed two principles [16-18]: 1) Orthogonality, which means that in DCE, each attribute level should have none or minimal correlation with other attribute levels. 2) Balance, which means that in DCE, each attribute should appear an equal number of times. Following these, we finally set six random questions and one fixed question for each respondent in the DCE.

The DCE questionnaire then contained two parts. The first part required the respondents to fill in their demographic information, such as age (18–20, 21–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60, 61–65, 66–70, 71–75, 76–80, 81–85), sex (male or female) and the educational level (primary school student, primary school graduation, middle school student, middle school graduation, high school student, high school graduation, undergraduate, Bachelor degree, graduate student, Master degree, postgraduate student, PhD degree). The second part contains seven different scenarios, that require the respondent to imagine that they are queuing in outpatient and waiting for a diagnosis, and make a choice between diagnosis strategies. At the end of the questionnaire, respondents are required to estimate how many years (5/10/15/20/30/40/Never) for AI to surpass the human clinicians. An example choice set of different doctors is presented in the S1 Questionnaire.

Data collection

In October 2017 and August 2020, respectively, we had sent our website link that contained the totally same DCE questionnaires to different groups of people across generations through various social media platforms, such as WeChat and QQ. To improve the response rate, we provided incentives (i.e., fit-bit watch and cash in a lottery mode) to respondents who completed the questionnaire.

At the beginning of the questionnaire, a brief background introduction of AI application in medicine (including its advantages and the potential defects compared to traditional clinicians), as well as the purpose of this DCE, were provided. The questionnaire would take only 5–10 min to complete; respondents click 'Agree to take the survey' to start filling the questionnaire. Once the respondents clicked 'Agree to take the survey', then it represents that this respondent has known that he/she willingly participated in this survey research, and his/her privacy would be protected by law.

Propensity score matching mechanism

In our study, propensity score matching (PSM) is a regression method that collects and sets the respondents with a similar basic situation for the treatment group and the control group, and prevalently applied in the study of impact factors and causal effects, such as a medical treatment, a policy decision, or a case study. PSM [19] includes five steps: 1) propensity score estimation; 2) choose matching algorithm; 3) check overlap/common support; 4) matching quality/effect estimation; and 5) sensitivity analysis. The mathematical theory of PSM is primarily based on the Roy-Robin model [20–22]. Our objective to apply the PSM analysis method is to treat our data collected from 2017 as a treatment group, and data collected from 2020 as a control group (S2 Appendix PSM). Next, we match these two group respondents according to their personal information, including age, sex, and educational level. All demographic information was coded as dummy variables; for instance, male respondents were coded as "1", and female respondents were coded as "0".

Algorithm of matching. Among the various matching algorithms, [19] we used the nearest neighbor (NN) [23] algorithm since it is appropriate for identifying individuals in one group that were best matched with the individuals in another group.

Another merit of the NN algorithm is that it can find the controlled individuals according to the treated individuals, which guarantees all the treated individuals are successfully matching; thereby, NN makes the most of the information from the treatment group and the control group. Moreover, we did 1:1 matching, which can effectively reduce the confounding bias [24] and improve research efficiency and credibility.

Statistical analysis

Multinomial logit model. There are various analysis models that can be applied for further DCE statistical analysis, consisting of random effects binary probit and logit, multinomial logit (MNL) model, and mixed logit (MXL) [16, 25]. The theoretical model of DCE is based on the random utility model (S3 Appendix RUM). We assumed the respondents' choice would maximize the utility in each question. The overall utility of the decision-maker consists of a fixed utility and a random utility, which are unobservable. We summarized the interviewee preferences through their

comments, thereby identifying random utilities that were not identified in the question.

We used the MNL model to analyze the preferences of people with different attribute levels. The independent variable only contains attributes related to the healthcare plan and does not contain any information related to the participant. In the MNL model, the respondents picked the healthcare plan depends on the relative importance of the different attributes of the two plans and option none. In queuing time, diagnosis time, and diagnosis cost, our value size of healthcare plan is incrementally arithmetic. We found the maximum likelihood solution for the MNL model of the data.

The results from the MNL model are conditional on the information relating to all the choice options as this attribute is grouped before analysis. In the MNL model, the effect is synonymous with utility. If the value of the coefficient is positive, it means individuals prefer this level of service compared to the others within the same attribute. The MNL introduced is based on a model similar to the logistic regression. The observations are not independent within a block corresponding to the same individual. Instead of having one line per individual like in the classical logit model, there is one row for each level of the attributes of interest, per individual. In this study, for example, there are three types of diagnostic approaches (only doctor/AI combined doctor/AI only), each type of approach having its own characteristics, but an individual can choose only one of three approaches. As part of an MNL model, all three options were presented to each individual, and the individual chooses his preferred option. We presented the odds ratio of respondents' preference between attribute levels.

LC models. We used a latent class model (LCM) [26] to identify different classes of individuals of similar preferences. The purpose of using LCM in our class is to explain the correlation of explicit variables with the least number of potential classes to reach the goal of local independence. LCM first assumes that the null-model is the hypothesized model, and hypothesizes the local independence to exist among explicit variables. Then, LCM increases the number of latent categories from the null-model, and uses maximum-likelihood to estimate various models based on the parameters' limitation, compares and tests the hypothesized model and the observed data, and finds the most appropriate model result. Among different model information evaluation criteria, Akaike's information criteria (AIC) [27] and Bayesian information criterion (BIC) [28] are the most prevalent in selecting LCMs. After the model being determined, each observed data would be classified into appropriate latent classes.

Willingness to pay. Willingness to pay (WTP) is an efficient method that measures how much an individual is willing to sacrifice economically to choose another diagnosis attributes' levels compared to the reference attributes. WTP is used to investigate the homogeneity and heterogeneity of preferences.

Propensity score matching was implemented by STATA 16, and Sawtooth Software Lighthouse Studio 9.8.1 was used to conduct the multinomial logit model and LCMs.

Results

Data collection result

A total of 1520 individuals visited our experimental website in 2017, but only 1317 of them

completed the questionnaire and were included in the analysis. Among the respondents from 2017, ages were from 18 to 85 years, 731 (55.5%) were females, and 1115 (84.7%) believed that AI technology would someday surpass or absolutely substitute human clinicians.

In 2020, only 874 individuals visited our new experimental website, and 528 completed the questionnaire, among which 272 (51.5%) were females, and 91.3% were confident with the AI diagnosis over traditional diagnosis.

General PSM and logit result

After PSM, 528 respondents out of 1317 respondents recruited in 2017 were matched to the 528 respondents recruited in 2020. The PSM procedure is presented in Figure 1, and the demographic information before and after PSM is presented in Table 2. The general logit model result of respondents of 2017 and 2020 is shown in Table 3, consisting of estimated average preference weights ("effect"), *p*-value, odds ratio (OR) and 95% confidence interval. Generally, individuals from both 2017 and 2020 considered 'accuracy' as the most important attribute (Fig. 2), weighted importance account for 38.53% and 40.55%, respectively, and diagnosis time was considered as the least important attribute, 2.69% and 1.16% respectively. Also, individuals from 2017 and 2020 preferred the combined diagnosis method over AI alone or human clinicians alone according to the OR (2017: OR 1.645, 95% CI 1.535, 1.763; 2020: OR 1.513; 95% CI 1.413,1.621, Reference level: Clinician) (Table 3). In addition, the OR of level diagnosis accuracy increased with the increasing of the rate of accuracy, which indicated that respondents would always prefer the diagnosis method with higher accuracy; for instance, 100% accuracy in 2017 has an OR of 5.043 (95% CI 4.534, 5.609), and 100% accuracy in 2020 has OR of 5.263 (95% CI 4.734, 5.852). The preference characteristics reflected from matched respondents of 2017 are very similar to respondents' preference in 2020.

Overall willingness to pay (WTP). For the WTP of respondents in 2017, respondents were willing to pay 13.991 RMB (approx. 2.239 USD) for the "AI plus clinician" diagnosis method. Also, people were not willing to pay for the increase of outpatient waiting time, but they were willing to pay in exchange for a higher diagnosis rate (1.595 RMB, approx. 0.255 USD for 1% higher accuracy). For the willingness to pay of respondents in 2020, respondents were willing to pay 0.794 RMB, to choose the "AI plus clinician" method rather than the "clinician" method, compared to that in 2017, the intense of WTP of diagnosis method is relatively lower. Similar to respondents in 2017, respondents in 2020 were also not willing to pay for longer outpatient waiting time, but they were willing to pay for higher diagnosis accuracy.

Fig 1. Propensity score matching procedure.

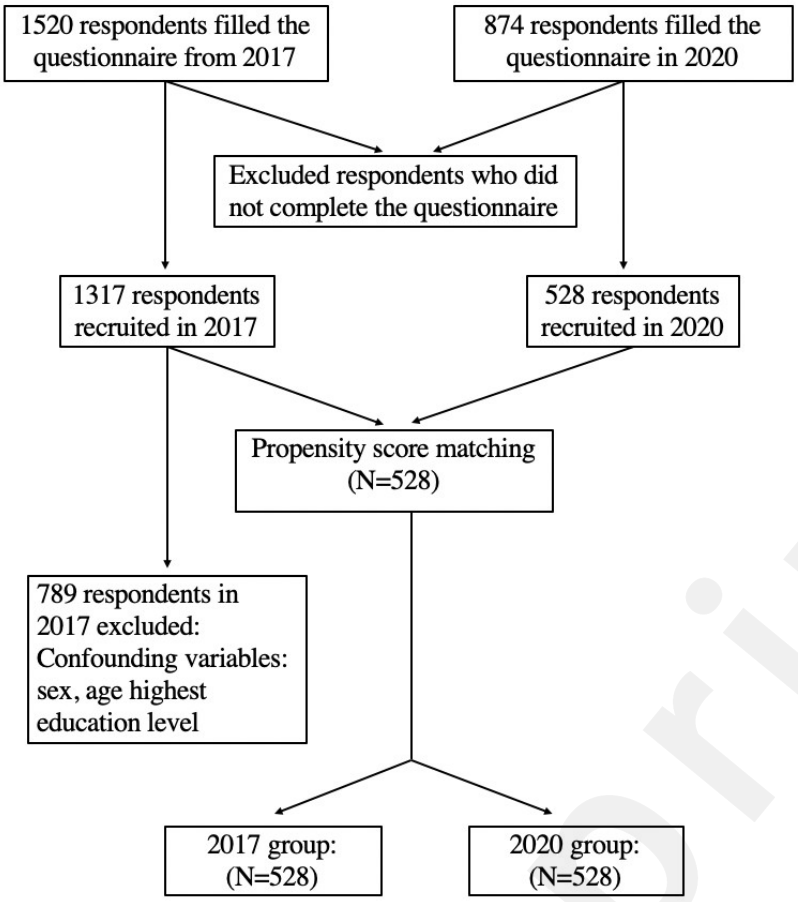


Table 2. Respondents' demographic information of propensity score matching result.

		Non-matched				Propensity score-matched		
seline tching cteristics		2017 (n=1317)	2020 (n=528)	p-value		2017 (n=528)	2020 (n=528)	p-value
x (%)	Male	586 (44.50%)	256 (48.48%)	<0.01		250 (47.35%)	256 (48.48%)	0.966
	Female	731(55.50%)	272 (51.52%)			278 (52.65%)	272 (51.52%)	
ears) (%)	<35	1106 (83.98%)	348 (65.91%)	<0.01		379 (71.78%)	348 (65.91%)	0.688
	≥ 35	211 (16.02%)	180 (34.09%)			149 (28.22%)	180 (34.09%)	
ghest tion level (%)	Primary school - Undergradu ate	1033 (78.44%)	336 (63.64% %)	<0.01		385 (72.92%)	336 (63.64%)	0.125
	Bachelor-PhD	284 (21.56%)	192 (36.36%)			143 (27.08%)	192 (36.36%)	

Table 3. General MNL model result of attribute preference in

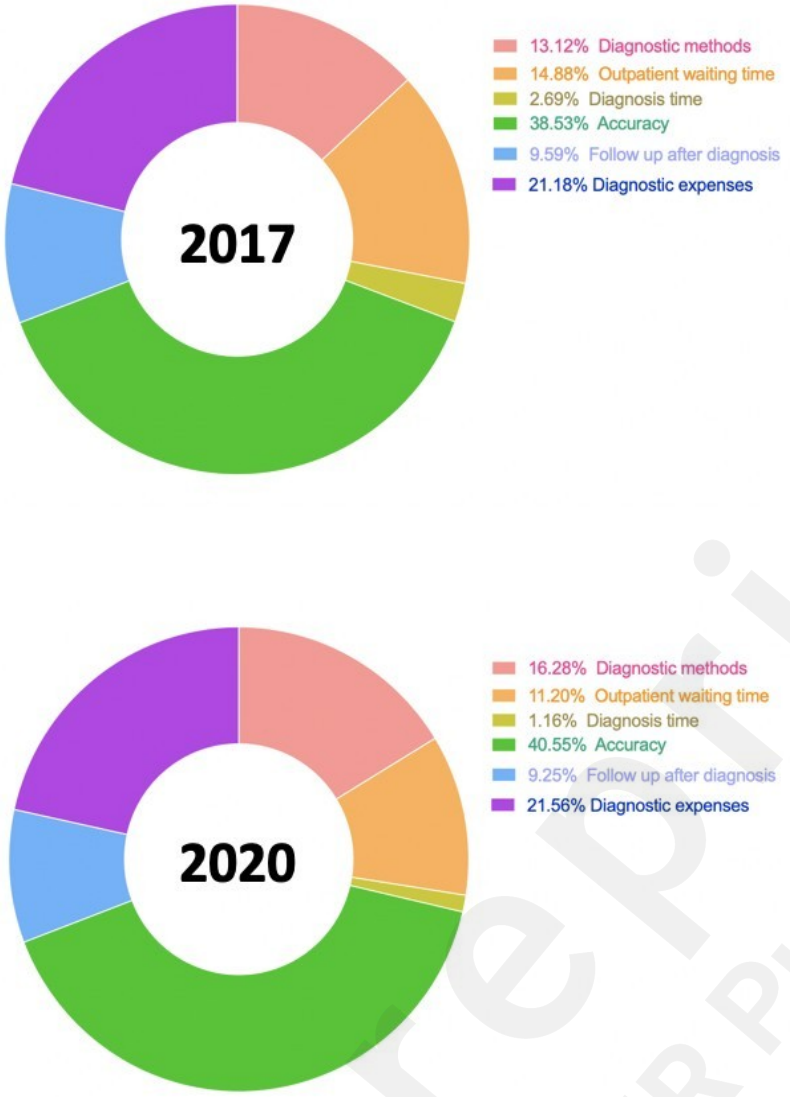
propensity score matching of 2017 and 2020 (N = 528).

	Year 2017					Year 2020			
Levels	Effect	p-value	OR	95% CI		Coefficient	p-value	OR	95% CI
nosis method: Clinician	-0.15	<0.001	Reference			-0.05	0.12	Reference	
sis method: AI+ Clinician	0.35	<0.001	1.64	(1.535 - 1.763)		0.36	<0.001	1.51	(1.413 - 1.621)
sis method: AI	-0.20	<0.001	0.95	(0.885 - 1.016)		-0.31	<0.001	0.78	(0.725 - 0.833)
nt waiting time: 0min	0.31	<0.001	Reference			0.15	0.01	Reference	
nt waiting time: 20min	0.12	0.03	0.82	(0.741 - 0.914)		0.26	<0.001	1.12	(1.013 - 1.245)
nt waiting time: 40nin	-0.03	0.57	0.71	(0.639 - 0.789)		-0.02	0.72	0.85	(0.762 - 0.942)
nt waiting time: 60min	-0.08	0.12	0.67	(0.606 - 0.748)		-0.20	<0.001	0.71	(0.640 - 0.788)
nt waiting time: 80min	-0.31	<0.001	0.54	(0.482 - 0.595)		-0.20	<0.001	0.71	(0.640 - 0.789)
sis time: 0min	0.05	0.19	Reference			-0.02	0.57	Reference	
sis time: 15min	-0.07	0.06	0.89	(0.834 - 0.957)		-0.01	0.83	1.01	(0.946 - 1.084)
sis time: 30min	0.02	0.53	0.98	(0.912 - 1.046)		0.03	0.43	1.05	(0.980 - 1.122)
osis accuracy: 60%	-0.83	<0.001	Reference			-0.83	<0.001	Reference	
osis accuracy: 70%	-0.35	<0.001	1.62	(1.458 - 1.802)		-0.41	<0.001	1.52	(1.365 - 1.684)
osis accuracy: 80%	0.07	0.16	2.47	(2.235 - 2.737)		-0.02	0.72	2.25	(2.033 - 2.487)
osis accuracy: 90%	0.32	<0.001	3.18	(2.867 - 3.526)		0.43	<0.001	3.51	(3.169 - 3.891)
osis accuracy: 100%	0.79	<0.001	5.04	(4.534 - 5.609)		0.83	<0.001	5.26	(4.734 - 5.852)
ow-up after gnosis: With	0.20	<0.001	Reference			0.19	<0.001	Reference	
ow-up after osis: Without	-0.20	<0.001	0.67	(0.620,0.698)		-0.19	<0.001	0.69	(0.656 - 0.715)
is expenses: 0	0.42	<0.001	Reference			0.36	<0.001	Reference	
is expenses:50	0.28	<0.001	0.87	(0.769 - 0.976)		0.23	<0.001	0.88	(0.782 - 0.989)
osis expenses: 100	-0.01	0.82	0.65	(0.576 - 0.730)		0.18	0.00	0.83	(0.738 - 0.935)
osis expenses: 150	0.03	0.66	0.67	(0.599 - 0.760)		-0.06	0.30	0.65	(0.580 - 0.736)
osis expenses: 200	-0.24	<0.001	0.52	(0.459 - 0.585)		-0.19	0.00	0.58	(0.510 - 0.648)
osis expenses: 250	-0.47	<0.001	0.41	(0.363 - 0.465)		-0.52	<0.001	0.41	(0.366 - 0.468)

OR: odds ratio

CI: confidence interval

Fig 2. General estimated weighted importance of attributes for 2017 and 2020.



Latent class model result

After comparing the AIC, BIC, and ABIC of the different potential number of classes, three classes were the most appropriate combination in both matched 2017 respondents and 2020 respondents. The segmented proportion and size of the three classes in the matched 2017 respondents were 43.2% (228 respondents), 42.2% (223 respondents) and 14.6% (77 respondents), respectively, in each class. Meanwhile, in the matched 2020 data, the segmented size means were 44.8% (237 respondents), 48.2% (254 respondents) and 7% (37 respondents), respectively, in each class.

For Class 1 from 2017 matched data (228 respondents), Figure 3 shows that respondents attached the most weighted importance on the attribute of diagnosis methods (32.95%), followed by diagnosis expenses (18.14%). Respondents of class 2 from 2017 were relatively more sensitive to diagnosis accuracy (49.92%) and diagnosis expenses (19.84%). Respondents of class 3 from 2017 had the most preference for diagnosis accuracy (25.66%) and diagnosis expenses (23.21%). For class 1 of respondents from 2020, people preferred the diagnostic expenses (29.99%), followed by diagnosis methods (28.99%). In class 2 of 2020, people chose diagnosis accuracy (52.34%) as the top focus point, followed by diagnosis expenses (14.44%). Regarding class 3 of 2020, people attached the most importance to diagnosis expense (36.21%), followed by diagnosis accuracy (32.84%). It is obvious

to find that three factors that account for the largest proportion of people's preference are diagnosis accuracy, diagnosis expenses, and diagnosis methods. Diagnosis methods occasionally leaped to the most preferred attribute, but typically, accuracy was the most important attribute in people's minds, and diagnosis expenses were steadily maintaining their second important place in people's trade-off.

Table 4. Result of the latent class model in 2017.

		Class 1 (n=228)				Class 2 (n=223)				Class 3 (n=77)			
Attributes	**Variable**	**Coefficient**	***p*-value**	**OR**	**95% CI**	**Coefficient**	***p*-value**	**OR**	**95% CI**	**Coefficient**	***p*-value**	**OR**	**95% CI**
Diagnosis method	**Clinician**	-0.355	<0.001	Reference		0.007	0.923	Reference		0.390	<0.001	Reference	
	AI + Clinician	0.553	<0.001	2.479	(2.240 - 2.743)	0.193	0.011	1.204	(1.039 - 1.394)	0.155	0.167	0.791	(0.636 - 0.984)
	AI	-0.199	<0.001	1.169	(1.060 - 1.289)	-0.200	0.008	0.813	(0.702 - 0.942)	-0.546	<0.001	0.392	(0.308 - 0.500)
Outpatient waiting time	**0min**	0.302	<0.001	Reference		0.365	0.001	Reference		0.564	<0.001	Reference	
	20min	-0.090	0.248	0.676	(0.580 - 0.787)	0.664	<0.001	1.349	(1.065 - 1.708)	0.349	0.028	0.806	(0.594 - 1.095)
	40nin	-0.124	0.112	0.653	(0.561 -	0.174	0.139	0.826	(0.657 - 1.039)	0.135	0.413	0.651	(0.472 -

					0.761)								0.898)
	60min	0.094	0.232	0.812	(0.697 - 0.946)	-0.393	<0.001	0.469	(0.374 - 0.587)	-0.648	<0.001	0.298	(0.207 - 0.429)
	80min	-0.181	0.020	0.617	(0.530 - 0.718)	-0.809	<0.001	0.309	(0.246 - 0.389)	-0.400	0.024	0.381	(0.271 - 0.536)
Diagnosis time	**0min**	0.018	0.724	Reference		0.253	<0.001	Reference		-0.113	0.326	Reference	
	15min	0.004	0.944	0.986	(0.893 - 1.088)	-0.203	0.007	0.633	(0.547 - 0.733)	-0.217	0.068	0.902	(0.717 - 1.135)
	30min	-0.021	0.670	0.961	(0.871 - 1.061)	-0.050	0.501	0.738	(0.639 - 0.854)	0.330	0.004	1.558	(1.252 - 1.938)
Diagnosis accuracy	**60%**	-0.173	0.026	Reference		-2.754	<0.001	Reference		-0.673	<0.001	Reference	
	70%	-0.207	0.009	0.967	(0.829 - 1.128)	-0.790	<0.001	7.129	(5.742 - 8.851)	-0.448	0.016	1.252	(0.877 - 1.787)
	80%	0.252	0.001	1.530	(1.315 - 1.781)	-0.059	0.567	14.804	(12.082 - 18.140)	0.027	0.865	2.014	(1.472 - 2.757)
	90%	-0.004	0.955	1.184	(1.015 -	1.195	<0.001	51.923	(41.340 - 65.216)	0.178	0.275	2.341	(1.705 -

					1.380)								3.214)
	100%	0.132	0.092	1.357	(1.164 - 1.581)	2.408	<0.001	174.651	(129.781 - 235.034)	0.916	<0.001	4.899	(3.631 - 6.611)
Follow-up after diagnosis	**With**	0.183	<0.001	Reference		0.402	<0.001	Reference		0.236	0.002	Reference	
	Without	-0.183	<0.001	0.694	(0.653 - 0.737)	-0.402	<0.001	0.447	(0.406 - 0.493)	-0.236	0.002	0.623	(0.538 - 0.722)
Diagnosis expenses	**□0**	0.201	0.025	Reference		1.114	<0.001	Reference		0.675	<0.001	Reference	
	□50	0.108	0.223	0.912	(0.766 - 1.084)	0.853	<0.001	0.770	(0.592 - 1.002)	0.333	0.068	0.770	(0.592 - 1.002)
	□100	0.017	0.851	0.832	(0.699 - 0.989)	0.031	0.801	0.339	(0.266 - 0.432)	0.045	0.806	0.339	(0.266 - 0.432)
	□150	0.098	0.270	0.902	(0.759 - 1.073)	-0.219	0.082	0.264	(0.206 - 0.337)	0.027	0.882	0.264	(0.206 - 0.337)
	□200	-0.124	0.167	0.723	(0.606 - 0.861)	-0.939	<0.001	0.128	(0.096 - 0.172)	-0.318	0.110	0.128	(0.096 - 0.172)

	□250	-0.299	<0.001	0.607	(0.509 - 0.723)	-0.840	<0.001	0.142	(0.109 - 0.185)	-0.762	<0.001	0.142	(0.109 - 0.185)

Table 5. Result of the latent class model in 2020.

		Class 1 (n=237)				**Class 2 (n=254)**				**Class 3 (n=37)**			
Attributes	**Variable**	**Coefficient**	***p*-value**	**OR**	**95% CI**	**Coefficient**	***p*-value**	**OR**	**95% CI**	**Coefficient**	***p*-value**	**OR**	**95% CI**
Method	**Clinician**	0.090	0.179	Reference		-0.159	<0.001	Reference		-0.220	0.335	Reference	
	AI + Clinician	0.217	0.001	1.135	(0.997 - 1.293)	0.538	<0.001	2.009	(1.826 - 2.211)	0.075	0.733	1.343	(0.877 - 2.059)
	AI	-0.307	<0.001	0.672	(0.587 - 0.769)	-0.379	<0.001	0.803	(0.731 - 0.883)	0.145	0.523	1.442	(0.926 - 2.244)
Outpatient waiting time	**0min**	0.609	<0.001	Reference		-0.021	0.775	Reference		-0.019	0.954	Reference	
	20min	0.197	0.052	0.662	(0.543 - 0.807)	0.377	<0.001	1.488	(1.287 -	0.207	0.504	1.254	(0.686 - 2.291)

									1.721)				
	40nin	-0.119	0.269	0.483	(0.391 - 0.596)	0.024	0.747	1.046	(0.903 - 1.211)	0.451	0.132	1.599	(0.901 - 2.838)
	60min	-0.176	0.087	0.456	(0.373 - 0.557)	-0.282	<0.001	0.770	(0.667 - 0.890)	-0.501	0.166	0.617	(0.308 - 1.238)
	80min	-0.512	<0.001	0.326	(0.268 - 0.397)	-0.098	0.183	0.925	(0.801 - 1.069)	-0.138	0.688	0.887	(0.454 - 1.735)
Diagnosis time	**0min**	-0.021	0.758	Reference		-0.003	0.945	Reference		0.243	0.264	Reference	
	15min	-0.008	0.903	1.013	(0.889 - 1.154)	0.006	0.908	1.009	(0.919 - 1.108)	-0.238	0.330	0.619	(0.386 - 0.992)
	30min	0.029	0.663	1.051	(0.923 - 1.197)	-0.002	0.962	1.001	(0.911 - 1.099)	-0.005	0.982	0.781	(0.508 - 1.200)
Diagnosis accuracy	**60%**	-2.149	<0.001	Reference		-0.272	<0.001	Reference		-1.186	0.021	Reference	
	70%	-1.008	<0.001	3.130	(2.547 - 3.845)	-0.133	0.070	1.149	(0.996 - 1.326)	-0.917	0.047	1.308	(0.544 - 3.143)
	80%	-0.092	0.319	7.820	(6.522 - 9.376)	0.060	0.411	1.394	(1.208 -	0.280	0.396	4.333	(2.285 - 8.215)

									1.608)				
	90%	1.027	<0.001	23.951	(19.729 - 29.078)	0.119	0.114	1.478	(1.276 - 1.711)	0.549	0.080	5.666	(3.117 - 10.301)
	100%	2.223	<0.001	79.203	(61.423 - 102.129)	0.226	0.002	1.645	(1.423 - 1.901)	1.274	<0.001	11.709	(6.759 - 20.285)
Follow-up after diagnosis	**With**	0.540	<0.001	Reference		0.066	0.025	Reference		0.260	0.108	Reference	
	Without	-0.540	<0.001	0.340	(0.310 - 0.373)	-0.066	0.025	0.876	(0.827 - 0.928)	-0.260	0.108	0.594	(0.436 - 0.810)
Diagnosis expenses	**□0**	0.466	<0.001	Reference		0.425	<0.001	Reference		0.946	0.006	Reference	
	□50	0.359	0.002	0.912	(0.766 - 1.084)	0.136	0.104	0.749	(0.636 - 0.882)	1.162	<0.001	1.241	(0.669 - 2.303)
	□100	0.273	0.019	0.832	(0.699 - 0.989)	0.213	0.012	0.809	(0.686 - 0.955)	-0.180	0.645	0.324	(0.152 - 0.694)
	□150	-0.010	0.934	0.902	(0.759 - 1.073)	-0.104	0.215	0.589	(0.500 - 0.694)	-0.154	0.684	0.333	(0.159 - 0.696)

	□200	-0.347	0.004	0.723	(0.606 - 0.861)	-0.145	0.084	0.565	(0.480 - 0.666)	-1.551	0.015	0.082	(0.025 - 0.272)
	□250	-0.740	<0.001	0.607	(0.509 - 0.723)	-0.524	<0.001	0.387	(0.327 - 0.458)	-0.222	0.586	0.311	(0.141 - 0.687)

Table 6. Willingness to pay of respondents in 2017.

Attribute		**Willingness to pay**							
		Overall N=528 (100%)		Class 1 n=228 (43.2%□		Class 2 n=223 (42.2%□		Class 3 n=77 (14.6%)	
		RMB (□)	USD ($)	RMB (□)	USD ($)	RMB (□)	USD ($)	RMB (□)	USD ($)
Diagnosis method	AI + Clinician	-13.991	-2.239	-3.026	-0.484	-0.220	-0.035	0.307	0.049
	AI	1.499	0.240	-0.520	-0.083	0.245	0.039	1.223	0.196
Outpatient waiting time		8.919	1.427	0.617	0.099	0.964	0.154	0.532	0.085
Diagnosis time		-0.568	-0.091	0.073	0.012	0.069	0.011	-0.436	-0.070

Diagnosis accuracy	-1.137	-0.182	-0.441	-0.071	-2.854	-0.457	-1.198	-0.192
Follow-up after diagnosis	11.324	1.812	1.219	0.195	0.953	0.153	0.618	0.099
Diagnosis expenses	Reference		Reference		Reference		Reference	

1 RMB=0.16 USD

Table 7. Willingness to pay of respondents in 2020.

Attribute		**Willingness to pay**							
		Overall N=528 (100%)		Class 1 n=237 (44.8%)		Class 2 n=254 (48.2%)		Class 3 n=37 (7%)	
		RMB (¥)	USD ($)	RMB (¥)	USD ($)	RMB (¥)	USD ($)	RMB (¥)	USD ($)
Diagnosis method	AI + Clinician	-0.794	-0.127	-0.171	-0.027	-1.327	-0.212	-1.307	-0.209
	AI	0.483	0.077	0.536	0.086	0.417	0.067	-1.620	-0.259
Outpatient waiting time		0.377	0.060	0.700	0.112	0.187	0.030	0.612	0.098
Diagnosis time		-0.054	-0.009	-0.040	-0.006	0.004	0.001	0.058	0.009

Diagnosis accuracy	-1.595	-0.255	-2.997	-0.479	-0.443	-0.071	-5.645	-0.903
Follow-up after diagnosis	0.726	0.116	1.455	0.233	0.252	0.040	2.306	0.369
Diagnosis expenses	Reference		Reference		Reference		Reference	

1 RMB=0.16 USD

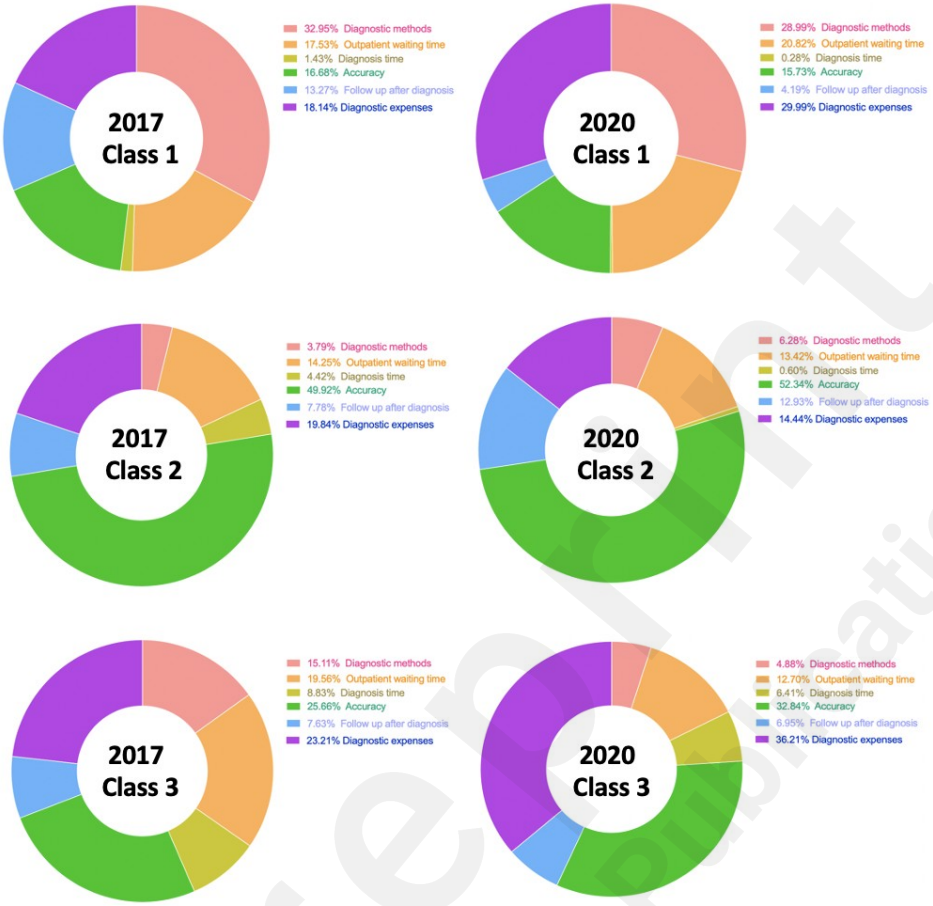


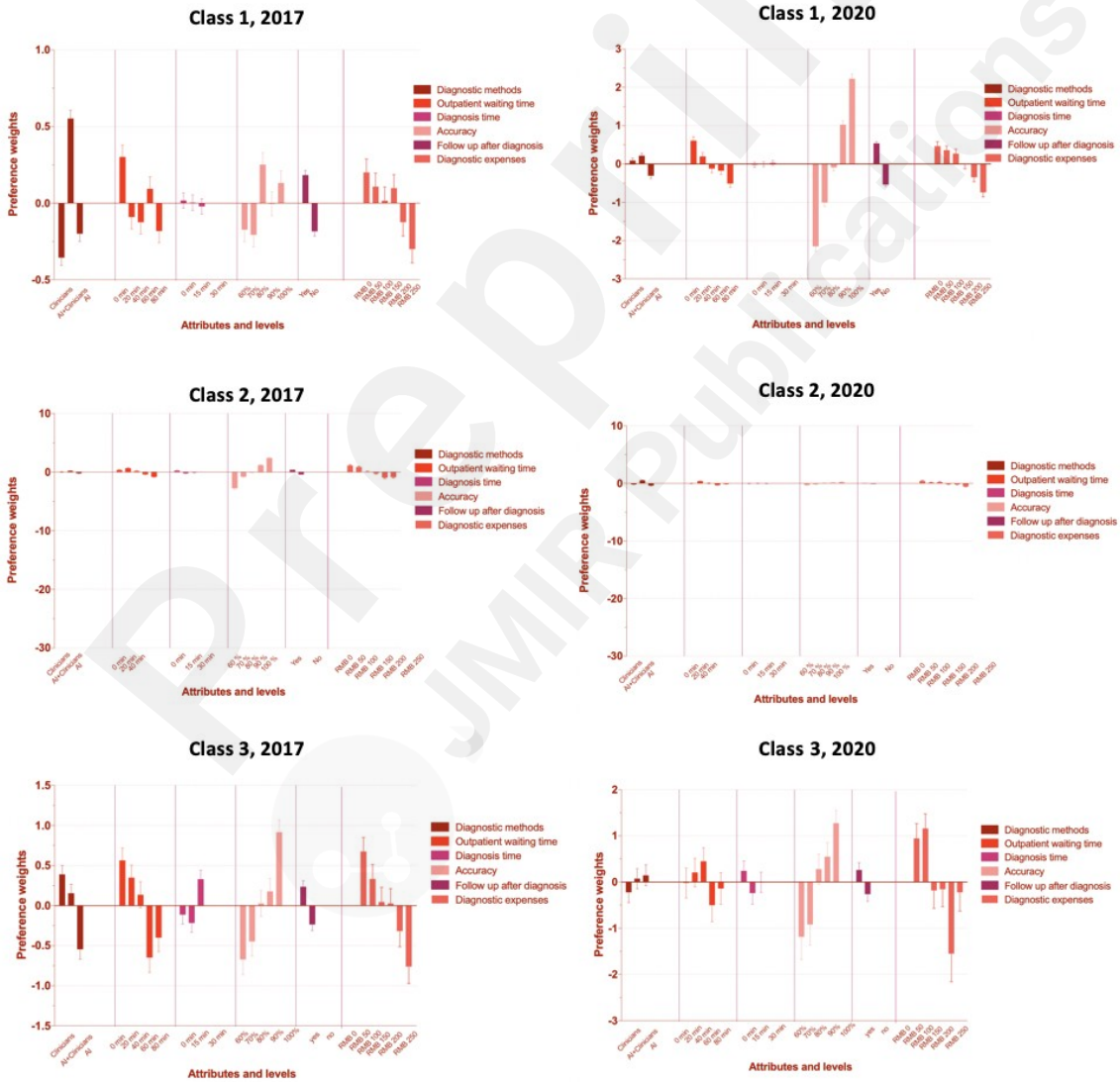
Fig 3. Weighted importance of latent class model in 2017 and 2020.

From the aspect of the odds ratio, our research finds that in 2017 (Table 4), respondents of class 1 and 2 preferred the combined diagnosis method (class 1: OR 2.479, 95% CI 0.997, 2.743; class 2: OR 1.204, 95% CI 1.039, 1.394) over other two methods except class 3. People preferred the “0 min” outpatient waiting time (except class 2), “0 min” diagnosis time (except class 3), and “0 RMB” diagnosis expenses across all classes. Also, all classes in 2017 preferred higher diagnosis accuracy (e.g., level “100%” in class 3: OR 4.899, 95% CI 3.631, 6.611). All classes’ people thought follow-up after diagnosis was important.

Similarly, in 2020 (Table 5), people in class 1 and class 2 preferred the combined diagnosis method (class 1: OR 1.135, 95% CI 0.997, 1.293; class 2: OR 2.009, 95%

CI 1.826, 2.211) except class 3. People in class 2 preferred “20 min” outpatient waiting time (OR 1.488, 95% CI 1.287, 1.721). Also, very similar to the respondent in classes of 2017, people preferred higher accuracy across all classes. Follow-up after diagnosis was still important across all classes. The strength of preferences mentioned above is visually presented in Figure 4, which is quantified by preference weights (coefficient) of each attributes’ level.

Fig 4. Latent class preference weights in 2017 and 2020.



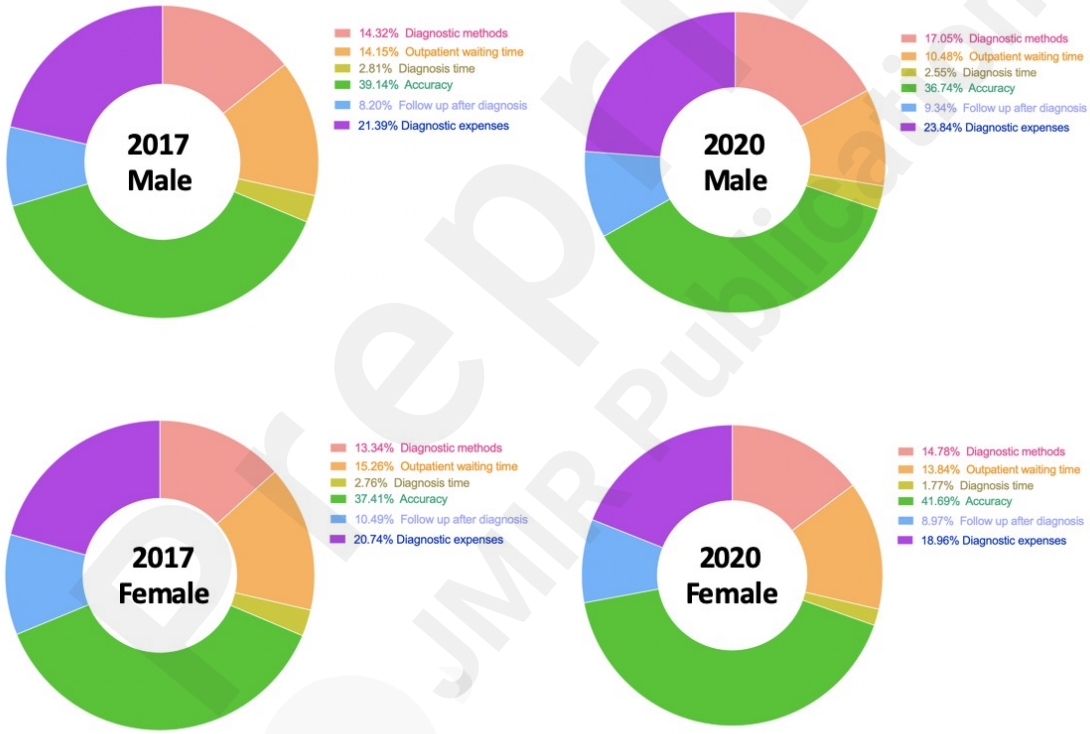
Regarding WTP, we found that people’s WTP was highly consistent with the

situation of OR. People in 2017 (Table 6) were willing to pay for the combined diagnosis method, except class 3. Also, people were always not willing to pay for a longer diagnosis time, except for class 3. All classes of people in 2017 were willing to pay for higher diagnosis accuracy and follow-up after diagnosis.

In 2020 (Table 7), all classes of people are willing to pay for the combined method, except class 3, in which people are willing to pay more for the AI diagnosis method. All classes of people were willing to pay for shorter outpatient waiting times, higher diagnosis accuracy, and follow-up after diagnosis.

For the latent class segmented, according to different sexes, the male class in 2017 (Fig 5) attached the most importance on diagnosis accuracy (39.14%) followed by diagnosis expenses (21.39%). The female class in 2017 also thought that diagnosis accuracy and diagnosis expenses were the most expensive, 37,41% and 20.74%, respectively. In 2020, the male class thought the diagnosis accuracy was the most important, with an accounting proportion of 36.74%, followed by diagnosis expenses (23.84%). Also, the female class of 2020 ranked diagnosis accuracy (41.69%) as the first attribute in a trade-off, followed by diagnosis expenses (18.96%). The LCM of the male and female classes in 2017 and 2020 show that there was no obvious preference heterogeneity among male and female respondents, among different periods.

Fig 5. Weighted importance of latent class model in 2017 and 2020 with male and female respondents.



Discussion

In this study, we collected information on the preference for AI-based diagnosis from two different groups of individuals from 2017 and 2020 (during the COVID-19 pandemic). By using the PSM method, we matched the respondent from the two groups with a similar distribution of age, sex and educational level. We did not find

differences between respondents' preferences in 2017 and 2020 comparing the two demographically similar groups of individuals. Accuracy and the expense for diagnosis play an essential role in people's choices.

The DCE questionnaire always depends on respondent rate; that is to say, people actively clicking the website link and completing the questionnaire is essential for the expansion of sample size and the whole research. With the PSM method, we can easily compare and check whether people's preferences change in peacetime and in unusual times such as the COVID-19 pandemic.

Different models have been applied in our paper, the MNL model and the LC model, and both of these two models have various advantages and drawbacks in quantifying respondents' preference. From the general PSM logit model, our research found that people's preference for diagnostic accuracy remains substantial in their trade-off regardless of which diagnosis methods or which periods. Moreover, diagnosis expense, following the preference for accuracy, also held huge, weighted importance in people's decision-making in both 2017 and 2020. We considered that the accessibility and availability of medical resources are still big problems in China, especially in some rural areas of China, because of the insufficient distribution of the medical insurance system [29, 30] and too low per capita income.

In general, we found that people's preferences for different diagnoses are largely similar, indicating that people's preferences for different diagnoses and the trade-off are not affected much by the factors caused by the COVID-19 pandemic. However, from our LC model, our research still found some very slight heterogeneity of individual preference among different groups of respondents and different sexes, which are unobservable in the logit model. Although the weighted importance of accuracy was still approximately consistent across all of the classes, it might not be the most important attribute in people's trade-off anymore. In our research, class 1 of the respondents from 2017 and 2020 considered diagnosis expense is the most important factor in their decision-making, followed by the methods of diagnosis. From the LC model result of different sexes, male respondents from 2017 or 2020 held that accuracy of diagnosis the most important when choosing a diagnosis

strategy. As to the attribute levels, we found that the majority of respondents preferred the combination diagnosis method of “AI plus human clinicians” confirmation to a single diagnosis mode (only “AI” or only “human clinicians”). It is understandable since the majority of people believed that the accuracy would be improved through mix the different modes of diagnosis. Also, it is noticeable that part of the respondents preferred a longer time for diagnosis and outpatient queuing. Though there is no relative literature to demonstrate that the diagnosis time and outpatient time are negatively or positively correlated to the accuracy, it is possible that some patients would rather spend some time waiting for a doctor for a more accurate result than receive a result of the quicker diagnosis. Accessibility or the price of the AI service is an important issue, especially in rural areas or impoverished areas. Therefore, before pricing the AI technology service, it is advisable to survey and analyze the assignable funds of residents. For the residents in rural areas, governments can consider adding the AI diagnosis to health insurance or relevant subsidy projects. The other factor that should be considered is the accuracy of the AI diagnosis, since only a product or service with high accuracy can be promoted and adverted. In the early stage of an AI technology service entering into a market, relevant users could consider combining the AI technology with human wise. Therefore, in the future, AI diagnosis developers should focus on working on improving the accuracy of the diagnosis, while trying to reduce the diagnosis expenses to an acceptable range of patients.

There are some shortcomings and limitations in our research, especially in the data collecting procedure. It is clear that the sample size could limit the power of our analysis. Our sample might not be representative of the entire Chinese population. A further limitation is that the deployment and distribution of AI technology medical service is non-obvious, especially in rural areas [31], and an uneducated group of residence. Thus, for AI popularizing, there are still many obstacles to be overcome, and there is still a long way to go in popularizing conceptual projects.

Conclusion

Our study shows that, in general, the preferences of respondents in 2017 and 2020 did not show significant differences; people's preference for artificial intelligence versus clinical diagnosis was unaffected by the COVID-19 pandemic. However, preference for higher diagnostic accuracy and low diagnosis expense was evident regardless of diagnosis methods, waiting time, and follow-up service.

In summary, affordability and accuracy are the two principal factors to be considered to promote AI-based healthcare. A combination of AI-based healthcare and professionals will be more easily accepted by the general public in AI development.

Declaration of interest

The authors declare no conflicts of interest.

Funding

No funding received in this work.

Reference

1. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial Intelligence in Health Care: Bibliometric Analysis. Journal of Medical Internet Research. 2020;22(7):e18228–e18228.

2. Vaira L, Bochicchio MA, Conte M, Casaluci FM, Melpignano A. MamaBot: a System based on ML and NLP for supporting Women and Families during Pregnancy. Proceedings of the 22nd International Database Engineering & Applications Symposium. 2018; p. 273–277.

3. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):5829945–5829945.

4. WHO Coronavirus Disease (COVID-19) Dashboard. WHO. Updated on Nov 12, 2020. Available from: https://covid19.who.int/table

5. L P. AI-based COVID-19 CT report becomes a national collection.; 2020. Available from: WWW.CHINANEWS.COM.

6. Yassine HM, Shah Z. How could artificial intelligence aid in the fight against coronavirus? Expert Review of Anti-infective Therapy. 2020;18(6):493–497.

7. Zandieh SO, Yoon-Flannery K, Kuperman GJ, Langsam DJ, Hyman D, Kaushal R. Challenges to EHR Implementation in Electronic- Versus Paper-based Office Practices. Journal of General Internal Medicine. 2008;23(6):755–761.

8. Sullivan F. From $600 M to $6 Billion, Artificial Intelligence Systems Poised for Dramatic Market Expansion in Healthcare; 2016. Available from: https://ww2.frost.com/news/press-releases/600-m-6-billion-artificial- intelligence-systems-poised-dramatic-market-expansion-healthcare/.

9. Mcfadden D. Regression-based specification tests for the multinomial logit model. Journal of econometrics. 1987;34(1-2):63–82.

10. Anas A. Discrete choice theory, information theory and the multinomial logit and gravity models. Transportation Research Part B: Methodological. 1983;17(1):13–23.

11. Greene WH, Hensher DA. A latent class model for discrete choice analysis: con trasts with mixed logit. Transportation Research Part B: Methodological. 2003;37(8):681–698.

12. Mekonnen AB, Enqusselasie F. Patients' preferences for attributes related to health care services at hospitals in Amhara Region, northern Ethiopia: a discrete choice experiment. Patient preference and adherence. Patient Preference and

Adherence. 2015.

13. Alizadeh A, Eftekhaari TE, Mousavi SH. Patient preferences for hospital quality in Bandar Abbas using a discrete choice experiment. Life Science Journal. 2012;9(4):1882–1886.

14. Ryan M. Using conjoint analysis to take account of patient preferences and go be yond health outcomes: an application to in vitro fertilisation. Soc Sci Med. 1999;48(4):535–546.

15. Gunst RF, Mason RL. Fractional factorial design. Wiley Interdisciplinary Re. views: Computational Statistics. 2009;1(2):234–244.

16. Organization WWH, USAID; CapacityPlus; THE WORLD BANK. How to Conduct a Discrete Choice Experiment for Health Workforce Recruitment andRetention in Remote and Rural Areas. In: A User Guide with Case Studies. World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland: WHO Press; 2012.Available from: https://www.who.int/hrh/resources/DCE_UserGuide_WEB.pdf?ua=1.

17. Gunst RF, Mason RL. Fractional factorial design. Wiley Interdisciplinary Re views: Computational Statistics. 2009;1(2):234–244.

18. Bailey RA. Balance, Orthogonality and Efficiency Factors in Factorial Design. Journal of the Royal Statistical Society: Series B (Methodological). 1985;47(3):453–458.

19. Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys. 2008;22(1):31-72.

20. Rubin DB. Estimating causal effects of treatments in randomized and nonran domized studies. Journal of Educational Psychology. 1974;66(5):688–688.

21. Roy AD. Some thoughts on the distribution of earnings. Oxford Economic Papers. 1951;3(2):135-146.

22. Castaño-Muñoz J, Duart JM, Sancho-Vinuesa T. The Internet in face-to-face. higher education: Can interactive learning improve academic achievement? British Journal of Educational Technology. 2014;45(1):149–159.

23. Cost S, Salzberg S. A weighted nearest neighbor algorithm for learning with sym bolic features. Machine Learning. 1993;10(1):57–78.

24. Austin PC. Variance estimation when using inverse probability of treatment. weighting (IPTW) with survival analysis. Statistics in Medicine. 2016;35(30):5642–5655.

25. Hensher DA, Greene WH. The mixed logit model: the state of practice. Transpor. tation. 2003;30(2):133–176.

26. Greene WH, Hensher DA. A latent class model for discrete choice analysis: con. trasts with mixed logit. Transportation Research Part B: Methodological. 2003;37(8):681–698.

27. Sakamoto Y, Ishiguro M, Kitagawa G. Akaike information criterion statis-tics. Reidel D, editor. Dordrecht, The Netherlands; 1986.

28. Weakliem DL. A critique of the Bayesian information criterion for model selection. Sociological Methods & Research. 1999 Feb;27(3):359-97.

29. Hu R, Dong S, Zhao Y, Hu H, Li Z. Assessing potential spatial accessibility of. health services in rural China: a case study of Donghai county. International Journal for Equity in Health. 2013;12(1):35–35.

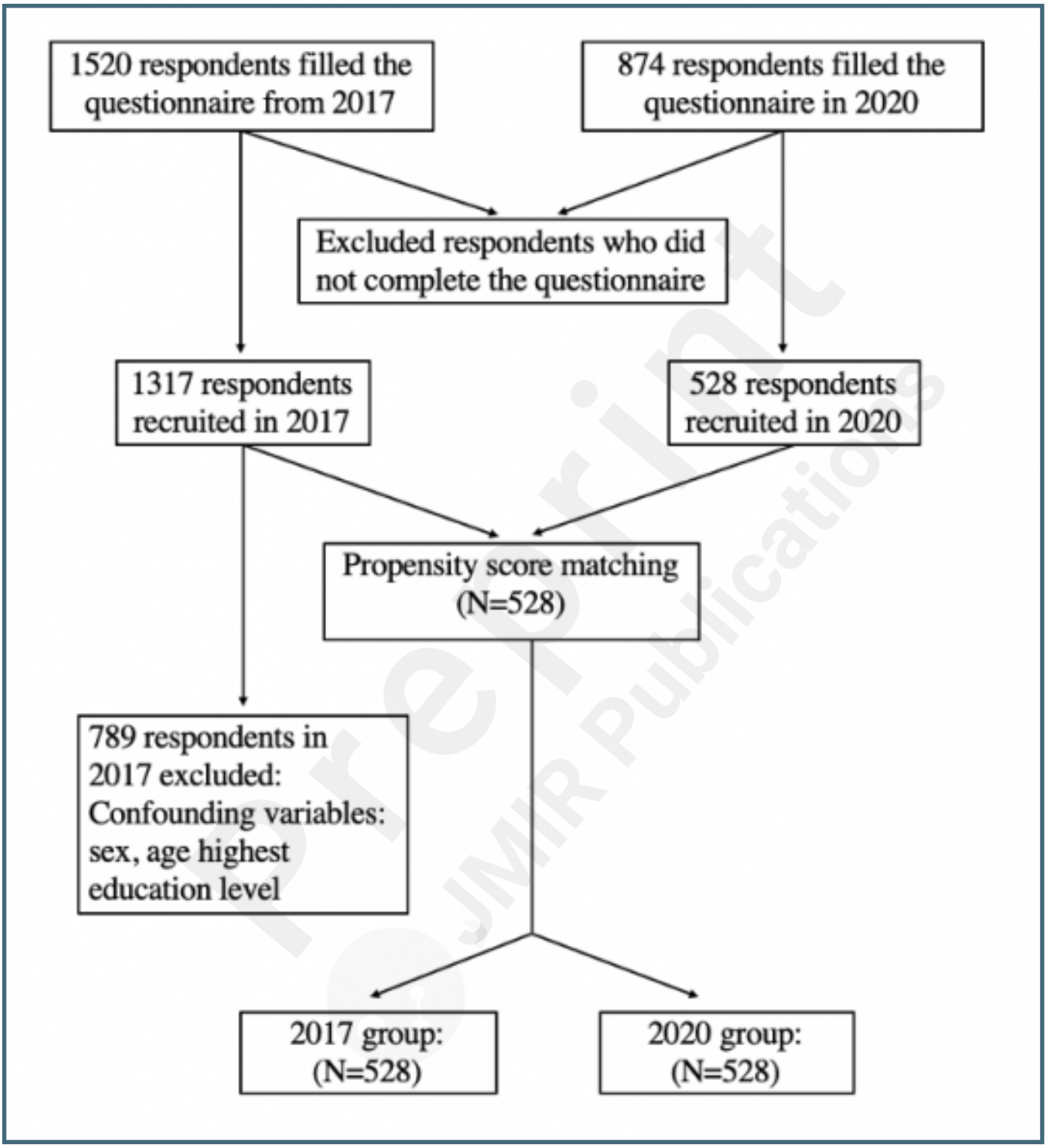
30. Wang X, Yang H, Duan Z, Pan J. Spatial accessibility of primary health care in. China: a case study in Sichuan Province. Social Science & Medicine. 2018; 209: 14–24.

31. Guo J, Li B. The application of medical artificial intelligence technology in rural. areas of developing countries. Health equity. 2018;2(1):174–181.

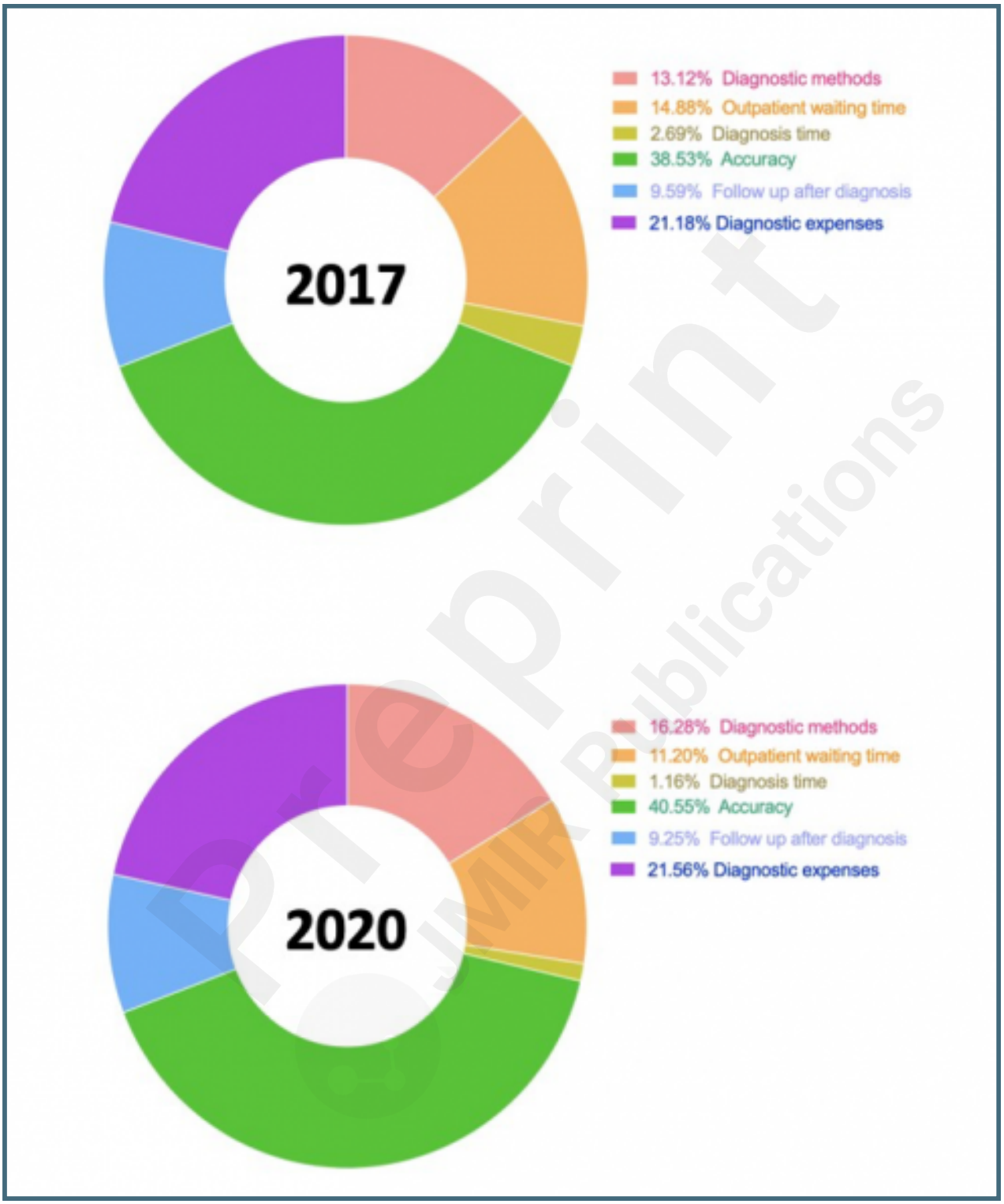
Supplementary Files

Figures

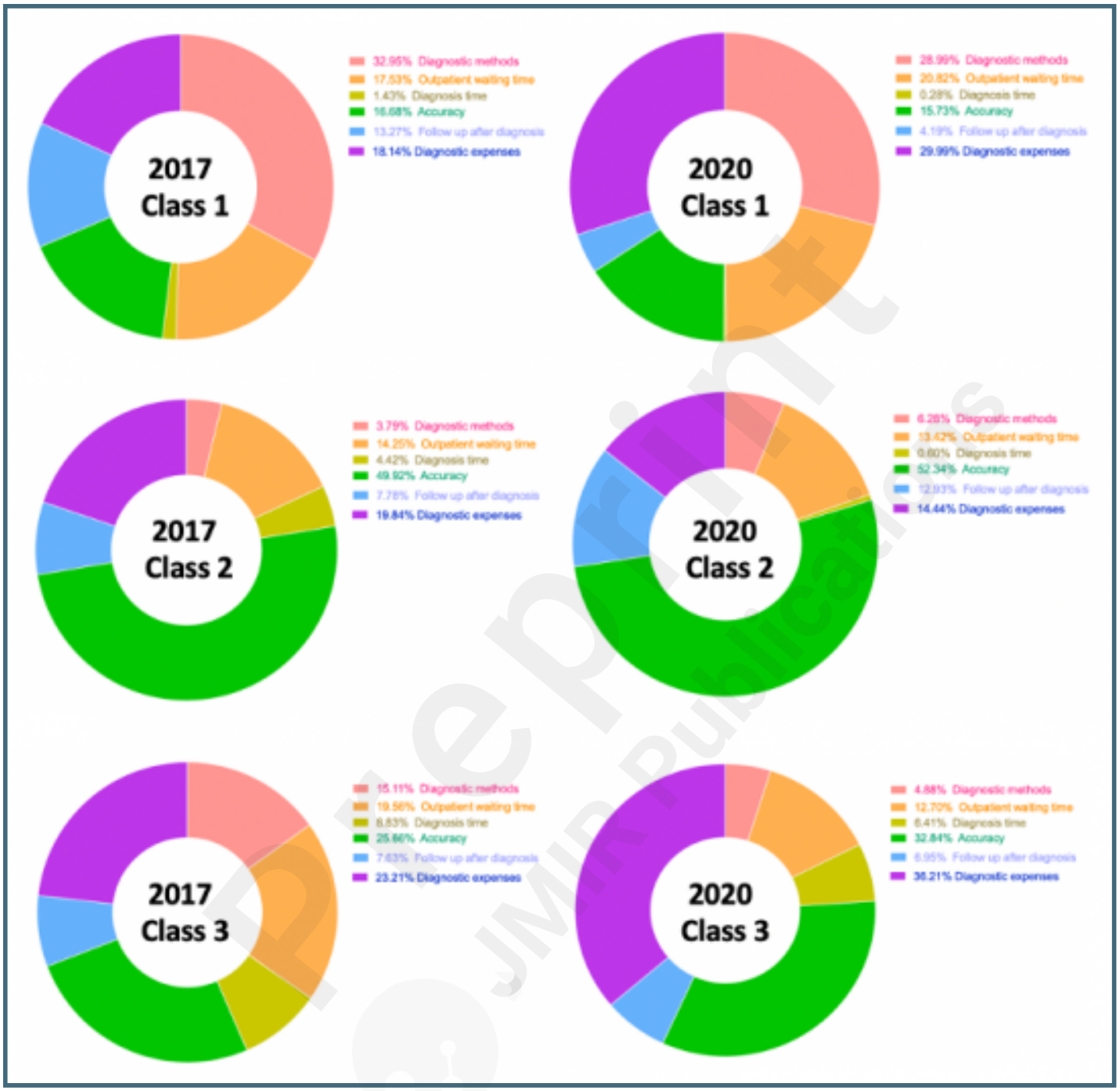
Propensity score matching procedure.



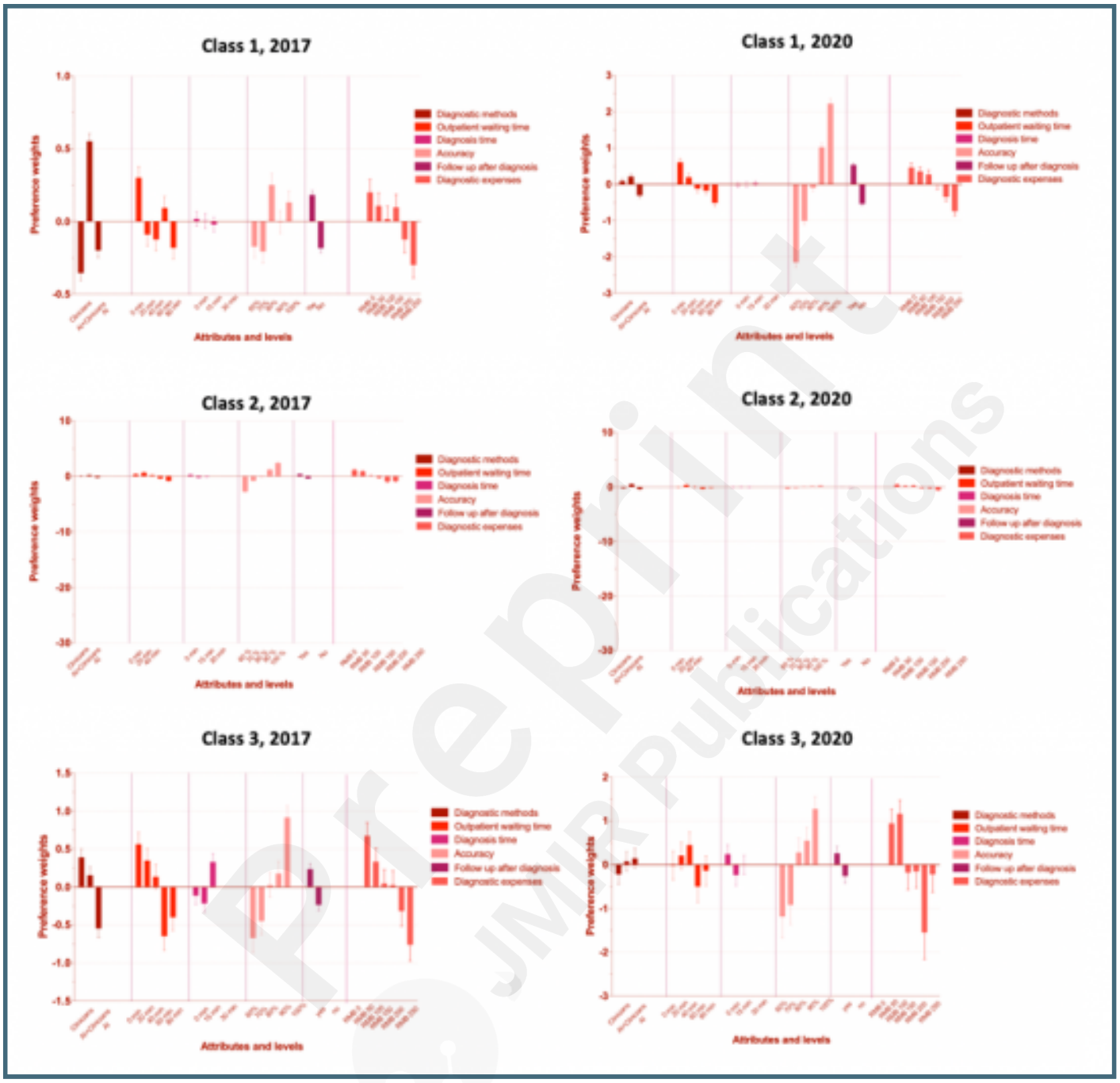
General estimated weighted importance of attributes for 2017 and 2020.



Weighted importance of latent class model in 2017 and 2020.



Latent class preference weights in 2017 and 2020.



Weighted importance of latent class model in 2017 and 2020 with male and female respondents.