**Methods**

*1. Design*

This is a randomized, controlled, multicenter study conducted in three hospitals in Chongqing, China, including four SRT training sites certified by the Chongqing Municipal Health Commission.

*2.* *Participants and setting*

A total of 228 residents studying at these four SRT training sites between January 2017 and December 2020 were invited to participate in the study (30 residents were ultimately excluded). The inclusion criteria were as follows: active residents employed in four training sites who were willing to participate in the training program and assessment procedure. The following exclusion criteria were used: residents who had previously participated in similar curriculum and residents who completed less than 50% of the training program.

The sample size and power analysis were calculated using “Compare 2 Means: 2-Sample, 2-Sided Equality” tool in the “Power and Sample Size” website (<http://powerandsamplesize.com/Calculators/>) according to our [preliminary](javascript:;) [experiment](javascript:;) results. Briefly, this calculator uses the following formulas to compute sample size and power, respectively:



Where:

* κ=nA/nB is the matching ratio
* σ is standard deviation
* Φ is the standard Normal distribution function
* Φ−1is the standard Normal quantile function
* α is Type I error
* β is Type II error, meaning 1 − β is power

Seven attending doctors and 6 senior doctors were invited as directors. All of the residents had completed basic theoretical courses and clinical clerkships in a medical college.

*3. Ethical consideration*

The present study was approved by the Ethics Committee of Xinqiao Hospital, Army Medical University. Participants, who had been previously informed about the project, voluntarily participated this project and agreed with the randomization. Meanwhile participant could quit the project at any time during the research. Written informed consent was obtained from all participants. Meanwhile all the researchers signed a confidentiality agreement regarding the participants’ data.

*4. Intervention and control groups*

The students and directors were randomized into two groups: the control (LBL) group and the PAL group. There were 99 residents in each group. The program was completed in 2 weeks.

In the *control group*, i.e., the PBL group, the program was teacher-centered, and the lecture-based teaching method was used for theoretical knowledge teaching. The directors taught epidemiology, definitions, classifications, pathophysiology, diagnosis and differential diagnosis, and treatment of VD-related diseases. For clinical practice teaching, different typical cases of VD-related patients were chosen and taught by the directors.

In the *PAL group*, the 99 students were randomized into 33 subgroups (3 students in one subgroup). One student in the subgroup was selected as the tutor by other members. All of the selected student tutors received routine tutor training that comprised three individual 3-h sessions led by directors, as previously reported [[10](#_ENREF_10)]. After that, the students learned knowledge of VD-related diseases, and then the theoretical knowledge was taught by the student tutor, followed by a group discussion. For clinical practice, the same VD-related patient cases in the control group were chosen. The students performed a medical history inquiry and physical examination by themselves, followed by a group discussion under a student tutor's organization. Finally, each group worked as a team to come up with a diagnosis and treatment plan. The entire process was supervised and commented on by one director.

*5.* *Instruments and data collection*

*5.1* *Examination performance*

Examination performance was evaluated before and after teaching.

Mastery of theoretical knowledge was assessed through a written examination paper consisting of 40 single best answer questions that contained four parts (“*Concept and classification of VD-related diseases*,” “*Symptoms of different VD-related diseases*,” “*Signs of different VD-related diseases*”, and “*Treatment of different VD-related diseases*”, with 2 points awarded for each question and 10 questions for each part) and 2 open-ended questions (“*Diagnosis and differential diagnosis*”, with 10 points for each question). All the questions were chosen from our examination database of neurology. The examination questions in the database were classified into different difficulty levels, and questions were randomly chosen from different levels and constituted a final examination paper for all of the participants enrolled in this research. Details of the paper were shown in supplementary materials.

Clinical skills were also examined by analyzing the case of a standardized patient. The evaluation was performed in five parts: “*Medical history inquiry*,” “*Physical examination*,” “*Diagnosis and treatment plan*,” “*Communication skills and care for patients,*” and “*Overall evaluation*” (20 points for each part). The rating scale used for clinical skills evaluation of VD-related diseases was given in supplementary materials.

Examination was performed within one week before and after teaching. The examination performance was evaluated by the teaching [secretaries](javascript:;) (they are also clinical doctors) blinded to the group assignments.

*5.2* *Questionnaire*

*5.2.1* *Student perceptions and satisfaction*

The students completed a researcher-made questionnaire (3-point scale with 1 = improved and 3 = not improved) asking about their perceptions of the teaching process from 7 aspects. This questionnaire included items about “*Study activeness*,” “*Mastery of theoretical knowledge*,” “*Efficiency of learning*,” “*Application of clinical knowledge*,” “*C**ommunication skills*,” “*Self-study ability*,” and “*Self-confidence and teamwork skills*.” The items and scopes of the questionnaire were chosen according to previous report [[1](#_ENREF_1), [26](#_ENREF_26)] and our own teaching experience.

Additionally, we investigated satisfaction and the risk of burnout using different teaching methods with a researcher-made questionnaire (3-point scale with 1 = disagree and 3 = disagree).

Questionnaire was completed within one week after teaching.

*5.2.2 Validity and Reliability*

The responses on the scale in the questionnaire were collected, and the validity and reliability were analyzed using SPSS 18.0 software.

*6. Statistical Methods*

The basic characteristics of each cohort are presented as the mean ± SD. For continuous variables (such as score), an independent-samples *t*-test was used to evaluate statistical significance between groups, while a paired-samples *t*-test was used to analyze the paired data (pretest and posttest). For categorical variables (such as sex), chi-square analysis was used to evaluate statistical significance. The statistical method used for variable with three categories was also chi-square analysis in Descriptive Statistics → Crosstabs. If any cells had expected count less than 5, Fisher’s Exact Test would be applied. *P*-values less than 0.05 were considered significant. Analyses were performed with SPSS 18.0 software.