Eliminating Geographic Bias Improves Match Results: An Analysis of Program Preferences and Their Impact on Rank Lists and Results

Purushottam A. Nagarkar, M.D.
Jeffrey E. Janis, M.D.
Dallas, Texas; and Columbus, Ohio

Background: Previous studies have demonstrated that programs emphasize *United States Medical Licensing Examination* scores, publications, and geography in creating rank lists. The authors aimed to quantify the importance of geography and to determine how eliminating geographic preferences would affect Match outcomes.
Methods: The Match algorithm was implemented and validated on 6 years of deidentified data from the San Francisco Match (2009 to 2014). A "consensus" ranking was generated for each year—all applicants were ordered into a single list using Markov chain rank aggregation. Each program's rank list was reordered using the consensus list, and a new Match result was simulated. Statistical analysis was carried out with Microsoft Excel.
Results: Variation of program rank lists from the consensus rank list was driven by geography (training in the same medical center or state as the ranking program), "pedigree" (top 25 ranking of applicants' prior training), and foreign medical graduation status. Step 1 scores, publications, and medical school or residency region were not factors. The simulated Match resulted in a slight increase in the match rate. The median normalized number needed to match decreased from 6.7 to 6.5, and 80 percent of applicants had an unchanged or better result compared to the actual Match.
Conclusions: Geography is the primary driver of variation between program rank lists. Removing this variation would result in fewer unfilled positions, no significant change in the average number needed to match, and improved Match outcomes for most applicants. Programs should critically evaluate whether their geographic biases reflect underlying information about applicant quality. (*Plast. Reconstr. Surg.* 142: 82e, 2018.)

Match day is greeted each year by articles about the mysterious Match algorithm[1] and a vague uneasiness about the process.[2–4] A common criticism of the Match is that the participants perceive a lack of control over their fate. Job search systems in other professions rely on one-on-one negotiations and therefore can feel more controllable. The fact that the Match has been shown to be mathematically optimal[5] and economically sound[6] provides cold comfort to participants who worry about turning their future over to an algorithm. There are three aspects of the Match that drive this discomfort. First, you know the factors that influenced your own rank list, but you have no verifiable information regarding the factors influencing everyone else's list. Second, knowing the factors that influence your counterparts does not allow you to reliably predict the result, because of the complex interdependence of the rank lists of all participants. Finally, the Match outcome provides little insight into your performance during the process—there are interviews, and there is a result, with no steps in between. An applicant who matches at her second-choice program has no idea whether she was barely outside the "rank to match" window of her top choice or whether she was ranked dead last. Program and applicant preferences are encoded in rank lists, which are hidden from view.

From the Department of Plastic Surgery, University of Texas Southwestern Medical Center; and the Department of Plastic Surgery, The Ohio State University.
Received for publication May 29, 2017; accepted December 4, 2017.

DOI: 10.1097/PRS.0000000000004485

Disclosure: *The authors have no financial interest to declare in relation to the content of this article.*

Factual knowledge of program preferences is therefore of great value to applicants and potentially even to programs themselves. Furthermore, it is undoubtedly true that the preferences of two programs are likely to be based on different factors, such that their true-preference rank lists may vary even if the same applicants are interviewed by both. The second question, then, is whether applicant factors can predict variations between program preferences.

Our goal with this study was to identify applicant characteristics that predicted the variance in program rank lists. Analyzing this variance required an "average" or baseline against which to measure deviation. This is a thorny problem, because every program ranks only a small subset of all applicants. In his book *The Wisdom of Crowds*, James Surowiecki[7] proposes that the aggregation of decisions made by multiple individuals produces a more optimal consensus than any single individual decision. Each program makes its rank list independently; thus, program rank lists provide an opportunity to apply this concept to the Match. We build here on the work carried out by computer scientists[8] and biochemists[9,10] in the field of data aggregation to create a consensus program rank list, and use it to analyze program variations.

Finally, we also wanted to investigate what would occur if program-specific biases were to disappear—would the outcome of the Match look particularly different? We hypothesized that eliminating program biases would generally result in better outcomes as program behavior more closely began to adhere to the true-preference strategy.

METHODS

Institutional review board approval was not required for this study, as it did not meet the definition of human subject research—all data were deidentified, and did not involve any interaction with the individuals or institutions involved. Six years of deidentified program and applicant data were requested from the San Francisco Match. The National Resident Matching Program previously rejected a similar request, quoting their policy that "individual level data, even de-identified, will not be released." We received 6 years of deidentified data from the San Francisco Match (match years 2009 through 2014). No individually identifiable information about applicants or programs was available to us—all participants had been assigned random alphanumeric codes in the data we received. Applicant characteristics included *United States Medical Licensing Examination* Step 1 scores, number of publications, medical school and residency program(s) attended, and foreign medical graduate status. Program characteristics included only the state in which the program was located.

We used *U.S. News & World Report* rankings of U.S. medical schools,[11] and Doximity rankings[12] of U.S. general surgery; ear, nose, and throat; and oral and maxillofacial surgery residency programs to identify the top 25 reputed medical schools and residency programs in the country in each category. These rankings are not rigorously validated or peer-reviewed—however, although they may not accurately reflect the actual quality of the programs they purport to rank, they do reflect popular opinion on which programs are thought to be among the best. As such, we used these rankings as a proxy for program opinions of applicant "pedigree."

Rank list aggregation was carried out in the statistical computing environment R (R Foundation for Statistical Computing, Vienna, Austria). List aggregation used a space-dependent majority-rule Markov chain algorithm, using the TopKLists package.[9,10] This allowed us to combine all program rank lists from a Match year to create a consensus list of all applicants in that year. Each program's applicants were reranked based on the consensus list to create a new consensus-driven list for the program.

Statistical analysis was carried out using Microsoft Excel (Microsoft Corp., Redmond, Wash.). We identified two subsets of applicants: (1) those who were ranked to match[13] in both the consensus-driven list and the original list ("consensus ranked to match"); and (2) those who were originally ranked to match, but were not ranked to match in the consensus list ("nonconsensus ranked to match").

Binary logistic regression analysis identified whether applicant characteristics could predict which of these two groups the applicant would be. As such, the statistical analysis attempted to explain the variance in program rank lists based on applicant characteristics.

Finally, the Match algorithm (Fig. 1) was implemented in Excel and Visual Basic (Microsoft), and validated against the rank lists and outcomes for each of the years for which data were available. The consensus rank lists were then used to simulate a new Match for each year, and the outcomes were analyzed.

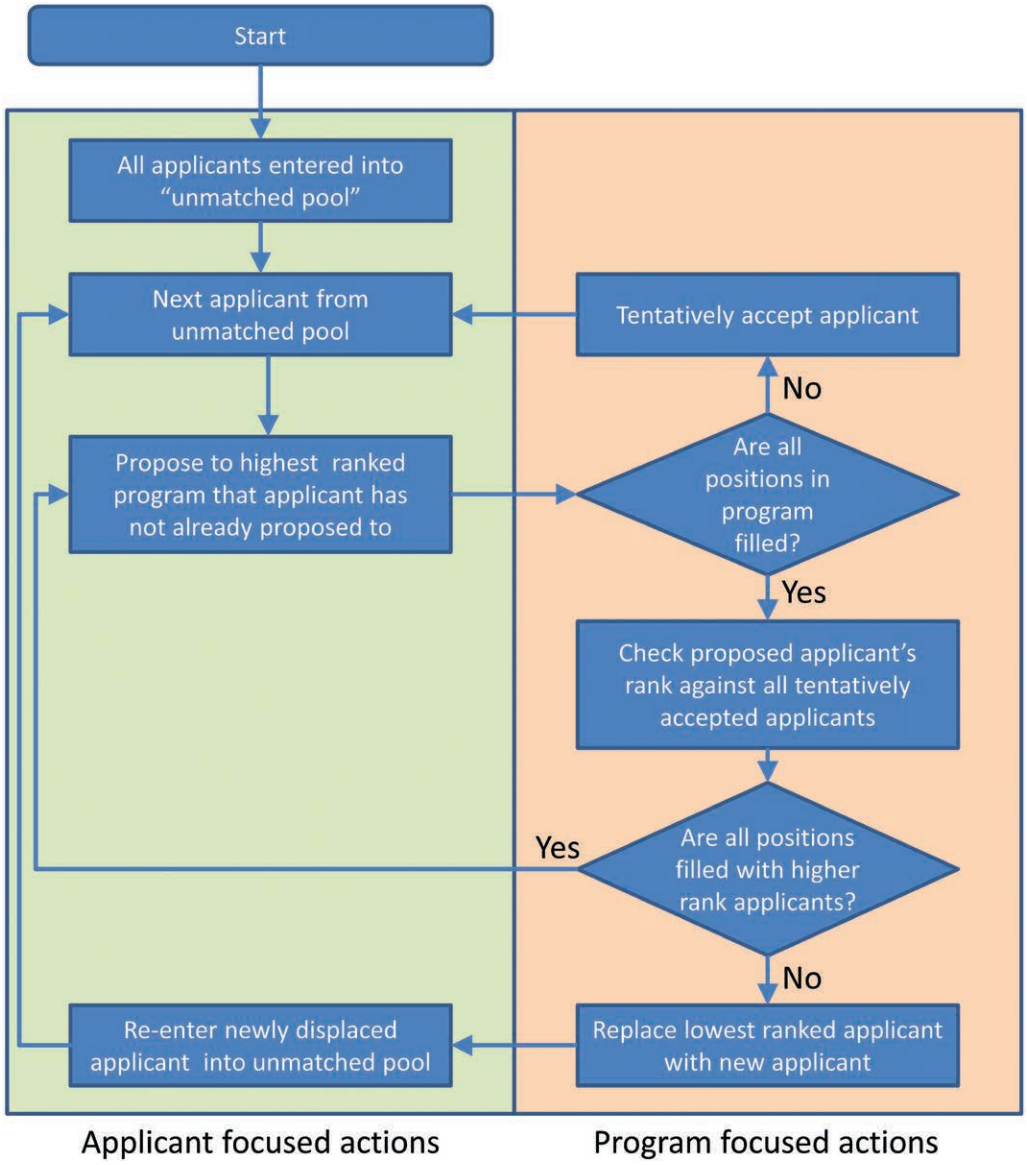

Fig. 1. Flowchart depicting the Gale-Shapley deferred acceptance algorithm used in the Match.

RESULTS

Baseline Data

There were an average of 107 applicants and 51 programs in each Match year. The average match rate for the period of study was 81 percent (range, 73 to 86 percent). From a program perspective, an average of four positions were unmatched per year.

Aggregate Program Preferences

A binary logistic regression model was built to predict ranked-to-match status. The independent variables were all applicant characteristics available: *United States Medical Licensing Examination* Step 1 score, publication count, international medical graduate status, attendance in a residency at the same institution as the program, attendance in a medical school in the same state as the program, residency in the same state as the program, attendance at a top 25 medical school, and at a top 25 residency. The model was statistically significant (chi-square = 80.11, $p < 0.0001$). Step 1 score, publication count, international medical graduate status, medical school in the same state, residency in the same institution as the program, and top 25 residency were all found to be significant factors (Table 1). Applicants who were in a residency at the same institution as the program were 2.5 times more likely to be ranked to match as those who were not. Similarly, each additional publication increased the likelihood of being

Table 1. Binary Logistic Regression Analysis of Program Rankings*

Independent Variable	B	SE	*z* Score	*p*	OR
Step 1 score in top quartile of interviewees?	0.54	0.12	19.15	1.2×10^{-5} †	1.71
Publication count (per publication)	0.05	0.01	18.29	1.9×10^{-5} †	1.05
Residency at same institution as program	0.92	0.27	11.35	7.6×10^{-4}†	2.52
U.S. medical school graduate	0.44	0.18	5.94	0.01†	1.55
Medical school in same state as program	0.35	0.14	6.18	0.01†	1.42
Residency in same state as program	-0.22	0.17	1.74	0.19	0.80
Medical school ranked in top 25?	-0.27	0.18	2.34	0.13	0.76
Residency ranked in top 25?	0.43	0.14	9.35	0.002†	1.53

*Dependent variable is the normalized ranked-to-match status of an applicant, with applicant characteristics as independent variables.
†Statistically significant. Model $\chi^2 = 80.11$, $p < 10^{-13}$.

ranked to match by 5 percent. U.S. graduate status had as powerful an effect as being in a top 25 residency (OR, 1.55 and 1.53, respectively).

Consensus Rank Lists

A binary logistic regression model was built to predict ranked-to-match status in consensus-driven program rank lists, using the same independent variables as in the model above (Table 2). This model was statistically significant ($\chi^2 = 114.25$, $p < 0.0001$), and the only significant variables were the Step 1 score (OR, 2.11) and number of publications (11 percent per publication). International medical graduate status, the applicant's medical school or residency state, and medical school or residency reputation were not significant predictors.

Drivers of Variation from the Consensus Rank List

Fifty-seven percent of ranked-to-match applicants would not have been ranked-to-match in a consensus rank list. Another binary logistic regression model was built to predict whether a ranked-to-match applicant would have remained ranked-to-match in a consensus rank list. To account for the smaller sample size in this model, we included attendance at a medical school or residency in the same region as the program as a variable. The other independent variables chosen were the same as in the previous logistic regression models (Table 3). This model was statistically significant ($\chi^2 = 29.31$, $p = 0.0003$). Step 1 score, number of publications, attendance at a residency in the same region as the program, and attendance at a top 25 residency were significant variables. Both Step 1 score and number of publications were negative predictors for being a nonconsensus ranked-to-match applicant (OR, 0.57; and decrease in likelihood of 10 percent per publication, respectively). In other words, having a top quartile Step 1 score or a high number of publications was likely to result in being a consensus ranked-to-match applicant. By contrast, residency geography or ranking predicted nonconsensus ranked-to-match status (OR, 1.73 and 2.11, respectively).

Changes in Match Outcome with Consensus Rank Lists

Our model reproduced the actual Match results for each of the available years with 100 percent accuracy, validating our algorithm implementation. We simulated each year's Match after reordering program rank lists to reflect the consensus rank list for that year. In these simulations, there were three fewer unfilled positions per year and, consequently, three more applicants matched into positions each year. The average rank on his or her list at which an applicant matched was unchanged. Forty-two percent of applicants would have matched at the same program as in their actual outcome, but 39 percent would have matched at a higher ranked program and 19 percent would have matched at a lower ranked program. From a program standpoint, the normalized number needed to match (i.e., number of ranks needed per position offered) would have decreased from 6.7 to 6.5. Twenty-five percent of programs would have had the same result as in their actual outcome, 48 percent would have improved their result, and 27 percent would have matched to a lower ranked applicant.

DISCUSSION

Our results show that programs, unsurprisingly, value objective criteria such as Step 1 scores and research productivity, and subjective criteria such as the reputation of the applicant's training program. However, they somewhat surprisingly value geographic familiarity with the applicant, preferring applicants who graduated from medical school or residency training in the same state.

The problem of aggregating multiple ranked lists into a single consensus list is one that has been studied for some time in computer science

Table 2. Binary Logistic Regression Analysis of Consensus Program Rankings*

Independent Variable	B	SE	*z* Score	*p*	OR
Step 1 score in top quartile of interviewees?	0.76	0.12	39.58	3.2×10^{-10}†	2.14
Publication count (per publication)	0.11	0.01	84.10	4.7×10^{-20}†	1.11
Residency at same institution as program	0.16	0.35	0.21	0.64	1.18
U.S. medical school graduate	0.13	0.16	0.68	0.41	1.14
Medical school in same state as program	0.05	0.16	0.11	0.74	1.05
Residency in same state as program	-0.23	0.17	1.82	0.18	0.80
Medical school ranked in top 25?	-0.35	0.18	3.63	0.06	0.71
Residency ranked in top 25?	-0.10	0.15	0.43	0.51	0.90

*Dependent variable is the normalized ranked-to-match status in the consensus rank lists, with applicant characteristics as independent variables.
†Statistically significant. Model $\chi^2 = 114.25$, $p < 10^{-20}$.

Table 3. Binary Logistic Regression Analysis of Ranked-to-Match Applicants*

Independent Variable	B	SE	*z* Score	*p*	OR
Step 1 score in top quartile of interviewees?	-0.57	0.25	5.20	0.02†	0.57
Publication count (per publication)	-0.10	0.03	14.14	0.0002†	0.90
Residency at same institution as program	0.39	0.53	0.54	0.46	1.48
US medical school graduate	-0.06	0.38	0.02	0.88	0.94
Medical school in same region as program	-0.26	0.25	1.11	0.29	0.77
Residency in same region as program	0.55	0.28	3.77	0.05†	1.73
Medical school ranked in top 25?	-0.17	0.34	0.27	0.61	0.84
Residency ranked in top 25?	0.74	0.30	6.06	0.01	2.11

*Dependent variable is nonconsensus ranked-to-match status.
†Statistically significant. Model $\chi^2 = 29.3$, $p < 0.0003$.

and more recently in genomics and proteomics. Building on a significant body of work carried out in these fields, we used the concept of the Markov chain to carry out rank list aggregation.[14] A Markov chain is a discrete statistical model in which the future state of the model is dependent only on its current state. The transition matrix of a Markov chain is the probability distribution that governs the transition from the current state to the next state. We represented the set of applicants in a year as a Markov chain, and constructed the transition matrix (i.e., the likelihood that two applicants would switch relative positions on the consensus rank list) based on the relative ranks of the applicants in a majority of the individual program lists. In the stationary distribution of this Markov chain, a higher probability for an applicant would signify a higher rank, and therefore the stationary distribution is the consensus rank list.

The consensus ranking can be expected to smooth out program-specific variations. As expected, in the binary logistic regression model used to analyze the consensus list, geography was no longer found to be relevant, and only the objective criteria in our data (i.e., Step 1 scores and publications) were relevant. When analyzing deviations of actual rank lists from the consensus list, we found that the nonobjective criteria (residency reputation and geographic familiarity) were likely to boost an applicant into rank-to-match status at the expense of objective criteria (e.g., Step 1 scores or publication counts).

Perhaps the correct interpretation of these results is that programs are upgrading the ranks of applicants who have a geographic connection to their state, or perhaps they are negatively ranking applicants who *do not* have a such a geographic connection. Either interpretation of the data is valuable. This geographic bias, whether it is positive or negative, may not be conscious, and simply being aware of it could help programs be more attentive to how they build their rank lists. Similarly, being aware of this bias may allow applicants to be judicious in their choice of interviews. For example, applicants now know that having a geographic connection to a program may allow one to overcome a poor Step 1 score or a lack of publications, and vice versa, that applying to a program with which one has no geographic connection has a lower chance of success unless one has a very strong application on the objective criteria.

Finally, we aimed to investigate how removing these program biases would affect the Match. To do this, we constructed a functional version of the San Francisco Match, implementing the Gale-Shapley applicant-proposing algorithm. The model simulated the actual Match results with 100 percent accuracy, which, incidentally, confirms that the San Francisco Match uses the

applicant-proposing algorithm, and not the program-proposing one. Our Match simulation results using consensus rank lists were instructive. First, there were three fewer unfilled positions per year on average in the simulations. This is, in our opinion, a small but significant number that would have an enormous positive impact on the affected programs and applicants. Second, from the applicants' standpoint, the outcome was a qualified improvement: 39 percent of applicants had a better outcome in the simulation, whereas only 19 percent had a worse outcome. Because the average rank at which applicants matched did not change, this tells us that the drop for applicants who had worse outcomes was greater than the improvement for those who had better outcomes. Intuitively this makes sense—if program biases were causing some objectively poor candidates to have inflated ranks, a smoothing out of these biases could cause these candidates to fall precipitously. From the program standpoint, the new Match outcome was an improvement for 48 percent of programs, compared with just 27 percent that did worse. These two perspectives—the applicant's and the program's viewpoints—suggest that program biases allow some programs (the 27 percent) to "steal" good applicants that would otherwise have matched at programs that were higher on their (the applicants') lists. This aligns with what we know about the Match—not following a true-preference strategy can produce worse outcomes for both programs and applicants.

As always, statistical analyses such as this are unable to account for the purely subjective aspects of interviews, personal interactions, and reference letters. Certainly, an applicant can have an unrepresentatively bad interview day and end up with a poor rank (or vice versa, have a good day and an undeservedly high rank), and this could be the source of some variations in rank. However, because we used several years of data with thousands of rankings, we believe such incidental biases should not rise to the level of statistical significance. As such, we believe the program biases we have identified are real. There may, of course, be benign reasons for these biases (e.g., programs that do not share a geographic connection with an applicant may not know the applicant's references very well, or may not have had as many opportunities to interact with the applicant before the interview). For example, in the basic program preference regression, the strongest predictor of being ranked to match was being in residency training at the same institution as the ranking program. This points to familiarity as being very important for ranking. However, note that attending medical school in the same state was an independent predictor, whereas residency training in the same state was not. It is hard to believe, therefore, that familiarity with applicant references is the only reason for this finding of geographic bias.

There are other weaknesses of this study. Once again, it is restricted to data from the San Francisco Match, because we did not have access to National Resident Matching Program data. Analyzing National Resident Matching Program data would be very valuable, especially because independent programs are increasingly converting to an integrated model. Nevertheless, we believe that the preferences our analysis reveals are likely shared by integrated programs—after all, the program directors and faculty in decision-making positions at integrated programs are not so very different from those in independent programs.

Another weakness is the reliance on *U.S. News & World Report* and Doximity data to identify the top 25 medical schools and residencies. We certainly do not endorse either ranking system, but we do think they provide useful proxies for the common wisdom regarding medical school and residency reputation. The fact that our analysis shows that this ranking is in fact an independently predictive variable of applicant rank is a point in support of the validity of the ranking. We also presume that the applicant characteristics in the San Francisco Match are valid. For example, the number of publications reported could be exaggerated by applicants, which would introduce potential error to the statistical model.

Finally, our analysis is blind to post-Match outcomes (i.e., training outcomes for matched applicants). Data on resident performance, in-service scores, board pass rates, research productivity in plastic surgery, career success, and so forth would add a very interesting dimension to this analysis.

CONCLUSIONS

Plastic surgery independent programs consistently value Step 1 scores and publication counts, but they also provide an implicit boost to applicants who share a geographic connection with the program. The variation of a program's list from the consensus of the group is driven almost entirely by geography. Furthermore, removing this variation results in unchanged or improved outcomes for 80 percent of applicants and no significant change in the number needed to match for programs. Applicants would do well to critically examine the objective strength of their

applications when applying to or interviewing with programs with which they share no geographic bonds. Programs would do equally well to engage in some introspection when assigning rankings to assess whether the bias we have identified here is a conscious and considered decision. Eliminating the geographic bias would result in fewer unfilled positions and arguably better outcomes for applicants and programs alike.

Purushottam A. Nagarkar, M.D.
Department of Plastic Surgery
University of Texas Southwestern Medical Center
1801 Inwood Road, 4th Floor
Dallas, Texas 75390
purushottam.nagarkar@utsouthwestern.edu

REFERENCES

1. Palmer B. Unmatched: The system for placing medical school graduates in residency programs is inefficient and ugly. Available at: http://www.slate.com/articles/health_and_science/medical_examiner/2015/04/match_day_for_medical_residency_the_scramble_foreign_doctors_and_a_shortage.html. Accessed September 10, 2017.
2. Sindhu K. Match day is coming up: Here's how medical students game the residency system. Available at: https://www.statnews.com/2017/03/08/match-day-residency-medical-student/. Accessed September 10, 2017.
3. Barry-Jester L. Another 34,000 people are about to put their future in the hands of an algorithm. Available at: https://fivethirtyeight.com/features/another-34000-people-are-about-to-put-their-future-in-the-hands-of-an-algorithm/. Accessed September 10, 2017.
4. Roy A. How a Nobel economist ruined the residency matching system for newly minted M.D.'s. Available at: https://www.forbes.com/sites/theapothecary/2014/04/15/how-a-nobel-economist-ruined-the-residency-matching-system-for-newly-minted-m-d-s/#365d36d05585. Accessed September 10, 2017.
5. Gale D, Shapley L. College admissions and the stability of marriage. *Am Math Mon.* 1962;69:9–15.
6. Roth AE, Peranson E. The redesign of the matching market for American physicians: Some engineering aspects of economic design. *Am Econ Rev.* 1999;89:748–780.
7. Surowiecki J. *The Wisdom of Crowds.* New York: Doubleday; 2005.
8. Fagin R, Kumar R, Sivakumar D. Comparing top k lists. *SIAM J Discrete Mathematics* 2003;17:134–160
9. Lin S, Ding J. Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA Studies. *Biometrics* 2009;65:9–18.
10. Lin S. Space oriented rank-based data integration. *Stat Appl Genet Mol Biol.* 2010;9:Article 20.
11. *U.S. News & World Report.* Best medical schools: Research. Available at: http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-medical-schools/research-rankings. Accessed September 10, 2017.
12. Doximity. Residency navigator 2017–2018. Available at: https://residency.doximity.com/. Accessed September 10, 2017.
13. Malafa MM, Nagarkar PA, Janis JE. Insights from the San Francisco Match rank list data: How many interviews does it take to match? *Ann Plast Surg.* 2014;72:584–588.
14. Meyn S, Tweedie RL. *Markov Chains and Stochastic Stability.* 2nd Ed. New York: Cambridge University Press; 2009.