The Match: Magic Versus Machines

David A. Ross, MD, PhD

Each spring, at high noon on a Friday in mid-March, thousands of soon-to-be physicians gather with mentors and faculty to have their medical futures revealed before classmates, family, and friends. The event is the culmination of what has been an elaborate, 5-month courtship ritual. The atmosphere is a complex mix of excitement, hope, and fear. The scene is oddly reminiscent of the Sorting Hat from *Harry Potter*.

Participation entails an extraordinary act of faith by all involved: tens of thousands of Type A individuals abdicate control of their future—not to a bit of Hogwarts magic, but to a computer matching program. They are willing to do so because they trust that the computer will render fair outcomes. More darkly, one might say that they are coerced—they participate because the National Resident Match Program effectively has a monopoly on entrance into the graduate medical education system. Regardless, to the extent that participants have faith in the software, it is likely well placed—the algorithm *is* brilliantly designed to optimize outcomes. Its creator, Alvin Roth, was later awarded the Nobel Prize in economics for this and other work with matching algorithms, most notably for organ transplantation.

The Match predated what has now become a broad societal trend to embrace big data and the power of computational approaches. We happily accept Google and Amazon's recommendations for what to read or purchase. As physicians, we are excited by the potential to apply complex data analytics to various medical problems.[1] For all of our success, though, we sometimes forget that a central tenet of these approaches is "garbage-in, garbage out" or, in the words of a *Harvard Business Review* commentary, "If Your Data Is Bad, Your Machine Learning Tools Are Useless."[2] In this regard, the Match may be on shaky footing.

In this issue of the *Journal of Graduate Medical Education*, Hartman et al[3] explore the factors that programs typically use to build their rank order lists—the data *in*. The authors describe some of the problems that arise with the status quo. These problems can include that United States Medical Licensing Examination (USMLE) Step 1 scores, despite being one of the most commonly used metrics to assess applicants, are not good predictors of resident performance; Medical Student Performance Evaluations (MSPE) may obscure comparative data or even suppress important negative information (a devastating betrayal when discovered post-hoc); and applicant personal statements and traditional letters of recommendation may be of limited utility (exclusive of the issue that these documents are plagiarized at surprisingly high rates).[4]

Unfortunately, this may be just the tip of the iceberg. It's not just that these metrics are poor discriminators of future performance; various studies have shown that they may compound group differences—if not overt biases—relating to applicant gender and/or race. USMLE scores show approximately a 1 standard deviation difference based on race, perhaps related to differences in socioeconomic backgrounds (as has been shown with SAT scores).[5–8] Students from groups underrepresented in medicine may receive lower clerkship grades.[8,9] Women may be described differently in letters of recommendation.[10–13] Applicants are described differently in their MSPEs based on race and gender, with white applicants disproportionately described as "excellent," "outstanding," and "best," and black applicants disproportionately described as "competent."[5] Black students are approximately 6 times less likely than other students to be inducted into the Alpha Omega Alpha (AOA) honor society.[14] Even when looking at students *from the same school and with identical grades in all core clerkships*, black students were still 3 times less likely than non-black students to be inducted into AOA.[8] This disparity persisted even after controlling for USMLE scores.[8]

These findings are deeply troubling for our field. They demonstrate that a program that bases decisions on ostensibly "objective" data may actually propagate implicit racial and gender-based biases that are already embedded in the system. Which is to say, it's not just garbage; it's racist and misogynistic garbage.

And yet, as Hartman et al[3] describe, we use these data to build our rank lists because they are the most accessible and most readily quantifiable. By default, they may also serve as surrogate indicators of success. And here the problem only gets worse. There is a principle in economics called Goodhart's law. Loosely described, it says that when an organization implements

DOI: http://dx.doi.org/10.4300/JGME-D-19-00301.1

a new outcome metric, employees will alter their behavior to inflate performance on that metric—even if doing so may undermine the health and productivity of the organization. For example, if a hospital starts tracking length of stay as a key outcome, patients may be discharged more rapidly and potentially prematurely, thereby causing more readmissions or other adverse events. For residency programs, what begins as a matter of convenience can become a self-propagating problem: fetishizing the wrong type of data may cause us to select the wrong applicants.

Of course, the most obvious way that residency programs may judge success in the Match is by how low they go on their rank list to fill all positions. This is also the metric that quintessentially embodies Goodhart's law: if a program is invested in filling a class from within a certain range on their list, they may change the way they rank applicants and favor individuals they believe are more likely to matriculate over those who may be better applicants but have not disclosed interest. It must be said that such behavior directly undermines the purpose—and elegance—of the Match algorithm.

This leads us to perhaps the greatest problem relating to the Match: that programs are forced to create a single rank order list. The mere act of creating such a list implies that applicants *can* be arranged on a single continuous scale, from best to worst. As program directors, we might like to believe that we can reduce all of the data down to a list that roughly correlates—if only probabilistically—with how individuals will perform during training. But even if we *could* use these metrics to identify the "best applicants" who will go on to be the "best residents" (whatever that means)—Is that really what we care about the most?

Many programs pride themselves on having distinct missions beyond training excellent clinicians. These may include addressing the needs of underserved communities, developing leaders in research, or training public policy advocates. The attributes that will predict an applicant's future success in these regards may be independent of—or perhaps even stand in conflict with—their performance on traditional metrics.

Embracing the multidimensionality of applicants requires an act of courage: it may entail selecting individuals who score lower on traditional metrics with the confidence that their strengths and future potential outweigh the risks. For example, it may mean ranking highly someone with a history of USMLE Step 1 failure but exceptional commitment to working with underserved communities. Or it may mean ranking someone with superb research credentials but worse clinical performance in medical school. In both cases, the program may expect that the resident will struggle in the short term but still believe that they can persevere and have greater impact in the end. From a game theory perspective, making these choices is a losing bet: the conventional candidate may be more likely to succeed and, even if they struggle, no one will question the decision to accept them; if the unconventional candidate struggles, complaints and second-guessing will abound—and, even if they succeed, the positive payoff may be so remote from training as to appear irrelevant.

Embracing a more holistic perspective may also alleviate some of the intrinsic stress of the Match. The more we hone in on our unique goals, the more we may recognize that we are competing less with perceived rival programs than we think. For example, when I debrief with colleagues at the end of each recruitment season, I am constantly reminded: we're looking for different things in applicants. Moreover, whether we successfully recruit a talented new class may have less to do with our actions or our perceived program quality so much as the state of the field as a whole. In a good year, with many strong applicants applying in a particular specialty, all programs will do relatively well; conversely, in an "off" year, all programs will struggle together.

This leads to a critical point: if we view the Match as a zero-sum game, with other programs as adversaries, then we allow ourselves to be divided. As the field of medicine faces more and more external threats (eg, with evolving systems of care), our real goal ought to be to better ally and advocate together. Especially apropos to this point is the suggestion that pressure to score well on the USMLE may be a significant contributor to burnout, depression, and suicidality in medical students. Is it worth it to us to force students to struggle for a score that has little or no predictive value in residency?

When the students of Hogwarts step up to be sorted, the Sorting Hat looks deep into their character to determine the best fit. With rare exceptions, all of the students are qualified and all are expected to succeed. Each will achieve basic competencies and have the opportunity to thrive in their own unique way. By in large, the same is true with the Match. The fact that a student is graduating from medical school is proof of a level of skill and accomplishment. While we tend to focus on the computer part of the Match, it can never be better than its inputs—the data that we use to generate our rank lists. There will always be a role for traditional metrics, but we would be wise to temper our enthusiasm for them and embrace a more holistic and nuanced approach. We should stop thinking and talking about "best applicants" and focus more on how we can identify those who best

align with each of our program's unique missions. And maybe try for a little magic.

References

1. Mukherjee S. A.I. versus M.D.: What happens when diagnosis is automated? *The New Yorker*. March 27, 2017. https://www.newyorker.com/magazine/2017/04/03/ai-versus-md. Accessed April 26, 2019.
2. Redman TC. If your data is bad, your machine learning tools are useless. *Harvard Business Review*. April 2, 2018. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless. Accessed April 26, 2019.
3. Hartman ND, Lefebvre CW, Manthey DE. A narrative review of the evidence supporting factors used by residency program directors to select applicants for interviews. *J Grad Med Educ*. 2019;11(3):268–273.
4. Maruca-Sullivan PE, Lane CE, Moore EZ, Ross DA. Plagiarised letters of recommendation submitted for the National Resident Matching Program. *Med Educ*. 2018;52(6):632–640. doi:10.1111/medu.13546.
5. Ross DA, Boatright D, Nunez-Smith M, Jordan A, Chekroud A, Moore EZ. Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. *PLoS One*. 2017;12(8):e0181659. doi:10.1371/journal.pone.0181659.
6. Edmond MB, Deschenes JL, Eckler M, Wenzel RP. Racial bias in using USMLE step 1 scores to grant internal medicine residency interviews. *Acad Med*. 2001;76(12):1253–1256.
7. Cooper RA. Impact of trends in primary, secondary, and postsecondary education on applications to medical school. II: considerations of race, ethnicity, and income. *Acad Med*. 2003;78(9):864–876.
8. Wijesekera T, Kim M, Moore E, Sorenson O, Ross D. All other things being equal: exploring racial and gender disparities in medical school honor society membership. *Acad Med*. 2019;94(4):562–569. doi:10.1097/ACM.0000000000002463.
9. Lee KB, Vaishnavi SN, Lau SK, Andriole DA, Jeffe DB. "Making the grade:" noncognitive predictors of medical students' clinical clerkship grades. *J Natl Med Assoc*. 2007;99(10):1138–1150.
10. Dutt K, Pfaff DL, Bernstein AF, Dillard JS, Block CJ. Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nature Geoscience*. 2016;9:805–808.
11. Madera JM, Hebl MR, Martin RC. Gender and letters of recommendation for academia: agentic and communal differences. *J Appl Psychol*. 2009;94(6):1591–1599. doi:10.1037/a0016539.
12. Trix F, Psenka C. Exploring the color of glass: letters of recommendation for female and male medical faculty. *Discourse & Society*. 2003;14(2):191–220.
13. Watson C. Sex-linked differences in letters of recommendation. *Women and Language*. 1987;10(2):26.
14. Boatright D, Ross D, O'Connor P, Moore E, Nunez-Smith M. Racial disparities in medical student membership in the Alpha Omega Alpha Honor Society. *JAMA Intern Med*. 2017;177(5):659–665. doi:10.1001/jamainternmed.2016.9623.

David A. Ross, MD, PhD, is Associate Professor, Department of Psychiatry, Yale School of Medicine.

Corresponding author: David A. Ross, MD, PhD, Yale School of Medicine, 300 George Street, Suite 901, New Haven, CT 06511, david.a.ross@yale.edu