2023, Volume 55, Issue 4, 245-252, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

Early Identification of Family Physicians Using Qualitative Admissions Data

Jacqueline M. Knapke, PhD[a]; Hillary R. Mount, MD[a]; Erin McCabe[b]; Saundra L. Regan, PhD[a]; Barbara Tobias, MD[a]

AUTHOR AFFILIATIONS:

[a] Department of Family and Community Medicine, College of Medicine, University of Cincinnati, Cincinnati, OH

[b] Digital Scholarship Center, University of Cincinnati, Cincinnati, OH

CORRESPONDING AUTHOR:

Jacqueline M. Knapke, Department of Family and Community Medicine, College of Medicine, University of Cincinnati, Cincinnati, OH, jackie.knapke@uc.edu

HOW TO CITE: Knapke JM, Mount HR, McCabe E, Regan SL, Tobias B. Early Identification of Family Physicians Using Qualitative Admissions Data. *Fam Med.* 2023;55(4):245-252.
doi: 10.22454/FamMed.2023.596964

ABSTRACT

Background and Objectives: The medical community has been concerned about the shortage of family physicians for decades. Identification of likely family medicine (FM) student matches early in medical school is an efficient recruitment tool. The objective of this study was to analyze qualitative data from medical school applications to establish themes that differentiate future family physicians from their non-FM counterparts.

Methods: We conducted a qualitative analysis of admissions essays from two groups of 2010-2019 medical school graduates: a study group of students who matched to FM (n=135) and a random sample comparison group of non-FM matches (n=136). We utilized a natural language modeling platform to recognize semantic patterns in the data. This platform generated keywords for each sample, which then guided a more traditional content analysis of the qualitative data for themes.

Results: The two groups shared two themes: emotions and science/academics, but with some differences in thematic emphasis. The study group tended toward more positive emotions and the comparison group tended to utilize more specialized scientific language. The study group exhibited two unique themes: special interests in service and community/people. A secondary theme of religious faith was evident in the FM study group. The comparison group exhibited two unique themes: lab/clinical research and career aspirations.

Conclusions: Aided by machine learning, a novel analytical approach revealed key differences between FM and non-FM student application materials. Findings suggest qualitative application data may contain identifiable thematic differences when comparing students who eventually match into FM residency programs to those who match into other specialties. Assessing student potential for FM could help guide recruitment and mentorship activities.

INTRODUCTION

High-quality primary care is the foundation of a high-quality health care system.[1] Addressing the current and projected increasing shortage of family physicians is critical to ensure comprehensive health care access in the US. In its 2020 report on physician supply and demand, the Association of American Medical Colleges (AAMC) predicted a shortage of between 21,400 and 55,200 primary care physicians in the United States by 2033, due in large part to population growth.[2]

The literature examining factors that influence students' choice of the primary care career is extensive, spanning decades. Two early studies identified important factors that influenced students' choice of primary care, such as income, hours worked, loan repayment, early role models, personal and family factors, and medical school experiences.[3,4] In its 2015 analysis of medical student survey data, the American Academy of Family Physicians (AAFP) found several key factors that influenced students' choice of family medicine (FM): strong mentorship, interest in FM at the start of medical school, perceived support for FM in the greater medical community, and a positive perception of the future of the specialty.[5] An extensive systematic review of the literature from 2003 posited key influences in medical students' specialty choices.[6] A review of 36 articles from 1993-2003 identified more than 10 factors that might affect student choice of family medicine as a specialty. More recently, a systematic review of the literature spanning 1977-2018 summarized results from 75 studies and found 12 primary factors that influence medical students' choice of specialty.[7]

Although the focus of much of the literature in this field is quantitative in nature, utilizing survey or demographic data, some qualitative research has also been done. Qualitative stud-

ies have found several aspects that can impact students' decisions to pursue FM, including faculty mentors, institutional environment, early exposure to FM physicians, high-quality FM clinical experiences, patient interactions, scope of practice, and continuity of care.[8–11] Another study aimed to determine if aspects of student application essays can be correlated to choice of primary care, finding several proportional occurrences that differentiated between FM and non-FM. However, the logistic regression model that included the reviewer agreement on prediction of future primary care practice found only two variables that were statistically significant to students' chosen career practice: interest in basic science research (a negative predictor) and the reviewer's prediction.[12] Some of these factors are associated with characteristics that could be identified on a student's application to medical school, allowing direction of limited resources or opportunities and mentorship toward those students most likely to select FM as their specialty.

This study undertook a unique methodological approach to analyzing qualitative data found in students' applications to medical school. The design is intentionally hybrid, combining the analytical strengths of machine learning[13] with humanistic enquiry from qualitative analysts and subject matter experts.[14] Once intensive human annotation alone is no longer sustainable, machine learning's aptitude with large-scale data is invaluable.[15] Using a novel approach that was qualitative in nature but guided by machine learning, our study analyzed admissions data from students at one medical school to identify themes that differentiated students who chose FM from students who chose other specialties. We interpreted students' self-described past experiences and future goals to develop a better understanding of the contextual factors that may contribute to the likelihood of their choosing FM.

METHODS

This was a retrospective, longitudinal cohort study. The protocol was reviewed by the University of Cincinnati (UC) Institutional Review Board (IRB) and determined to be not human subjects research (IRB# 2019-1012). Additionally, the protocol received an ancillary Family Educational Rights and Privacy Act (FERPA) review by the UC Office of General Counsel.

Study Sample and Data Set

Using purposeful sampling, we obtained qualitative admissions data for a study group and a comparison group of medical students who graduated between 2010 and 2019. The study group included all students who matched into an FM and/or FM-psychiatry residency program during that time, as well as a randomly selected comparison group of students who matched into a residency program for a specialty other than FM. We aimed for equal numbers in both the study and comparison groups and a sufficient total sample number to achieve saturation during data analysis. We randomly pulled comparison group data for two 5-year periods (2010-2014 and 2015-2019). The Office of Student Affairs and the Office of Admissions and Recruitment at the UC College of Medicine provided deidentified textual application data, which included students' self-described experiences (eg, work, volunteer, research, service) and personal comments (ie, personal essays). The study team had some initial concerns about including specialties such as internal medicine, pediatrics, and psychiatry in the comparison group because they likely present similar primary care personas in their interests and experiences. However, we decided to keep them in the data set in order to have a realistic and representative comparison sample.

Data Analysis

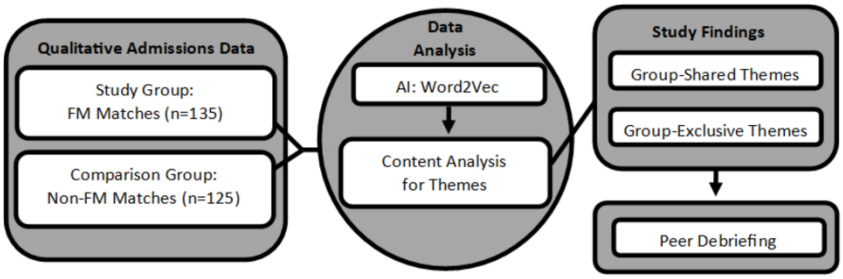
Artificial Intelligence/Word2Vec

Data were first categorized as study group or comparison group and then uploaded onto the UC Digital Scholarship Center's natural language modeling platform called "Model of Models." Using the natural language processing (NLP) technique "Word2Vec" (W2V), this platform modeled our data onto a vector space, based on documents' word embeddings.[16–18] We did not conduct a power analysis as analyses of statistical power have not been the norm in studies that utilize NLP, although new methods are being developed to change this.[19] Our W2V implementation understands words in context, by implementing a narrow word frame (ie, what five words appear before and after the target word). This narrow frame focuses the sense of context and has been shown to work well with smaller data sets like ours.[20] These word embeddings accounted for each word's immediate context by identifying words that frequently co-occurred within a given window. This allowed the platform to map each word onto a vector space that depicted words' relative proximity to one another in terms of their similar context windows. By mapping the word vectors for each group separately, we were then also able to identify words that were both more frequent and, importantly, more unique to each group. We generated two different word lists based on the W2V models for each group, one that generated a list of key terms for each group, as well as one that generated a list of "intersectional" terms, or terms that were frequent in both groups, but with variability in their frequency. These word lists became our codes. The research team did not set a firm threshold for difference in frequency as part of the analysis process because simple word counts would not allow for interpretations of the nuance of text, which were necessary in the qualitative content analysis.

Qualitative Content Analysis

The W2V results guided a traditional content analysis on the full data set. First, we organized the W2V codes into potential themes, without any review of the original data in the application essays and experiences. Next, using an inductive approach that focused on the W2V codes, we carefully read application materials for context and nuance. Early data analysis focused on distilling the essence of experiences being described in the data set and ensuring the key words identified by the topic modeling tool were not misinterpreted, devoid of larger context. Each code was thoroughly examined and if its meaning or attribution to a specific theme was under

FIGURE 1. **Study Methods**



question, we had the benefit of revisiting the original data for clarification. For example, "instrument" was initially a code within the "Science/Academic" theme for both groups. When reviewing the data set for context, we discovered students in the FM group were most often talking about musical instruments, whereas students in the non-FM group were talking about scientific instruments. We also updated the W2V frequency counts to reflect the context of code usage, since frequency of usage was an important factor in determining whether a code should be considered as evidence of a larger theme. After we agreed on a contextualized understanding and frequency count of the model's key terms, we then organized them into final themes. The 6-month analysis process required multiple iterations to reach intercoder agreement. To improve credibility, we employed external member checking, or peer debriefing, of final results.[21,22] We presented the study with results to FM faculty in a division meeting, medical students who had selected FM as their specialty and were taking a summer research course, as well as attendees at three different academic conferences. FM faculty and medical students who had selected FM confirmed that study results aligned with their experiences of interacting with current and future family physicians. Figure 1 provides an overview of our methods.

RESULTS

Our data set included 271 student applications: 136 applications from all graduates who matched to an FM or FM-psychiatry residency program between 2010-2019 and 135 applications from graduates who matched to a random selection of non-FM residency programs during the same time period. Table 1 shows participant demographics and Figure 2 gives a summary of study participants by year of graduation. We pulled comparison group data randomly for two 5-year periods, resulting in 12 comparison applicants each year for 2010-2014 and 15 comparison applicants each year for 2015-2019.

Figure 3 gives a summary of the residency program matches included in the non-FM comparison group, representing 22 specialities. If a student dually matched into both preliminary and advanced residency programs such as internal medicine and anesthesiology, they were coded according to the more specialized program (eg, anesthesiology).

TABLE 1. **Participant Demographics by Group**

		FM Study Group, n (%)	Non-FM Comparison Group, n (%)
Gender	Male	55 (40)	88 (65)
	Female	81 (60)	47 (35)
Age at Application (Yrs)	21	0 (0)	2 (1)
	22	37 (27)	37 (27)
	23	51 (38)	26 (19)
	24	18 (13)	21 (16)
	25	13 (10)	17 (13)
	26	6 (4)	13 (10)
	27	3 (2)	7 (5)
	28 or above	8 (6)	12 (9)
Race	Asian	17 (13)	39 (29)
	Black or African American	11 (8)	11 (8)
	.	.	.
	.	.	.
	Pacific Islander	1 (<1)	0 (0)
	Native American or Native Alaskan	1 (<1)	1 (<1)
	.	.	.
	.	.	.
	White	90 (66)	79 (58.5)
	N/A or other	15 (11)	6 (4)
Pell Grant Recipients	Yes	49 (36)	57 (42)
	No	87 (64)	78 (58)

A small quantity of qualitative data were missing for each group: in the study group, one student's personal essay and experiences could not be obtained and in the comparison group, 10 students' personal essays and experiences could not be obtained due to technical errors. We combined each personal essay and list of experiences into one document per student. Table 2 gives a summary of the full data set by group.

Word2Vec Quantitative Results

The W2V results produced three lists of words: one for each group that included words of high frequency and uniqueness to each group data set, and one list of words used in both groups, with frequency counts reported, allowing for a comparison of intersectionality.

FIGURE 2. **Research Groups by Graduation Year**

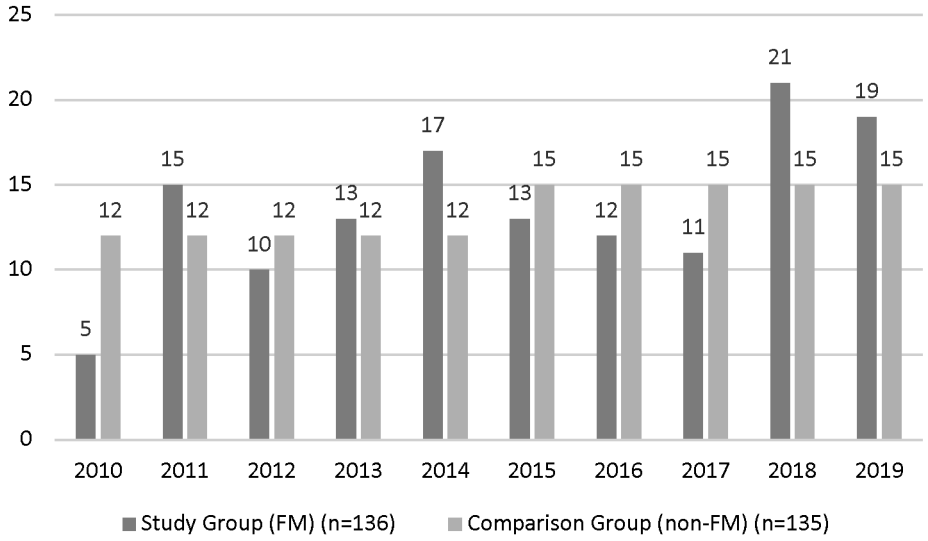
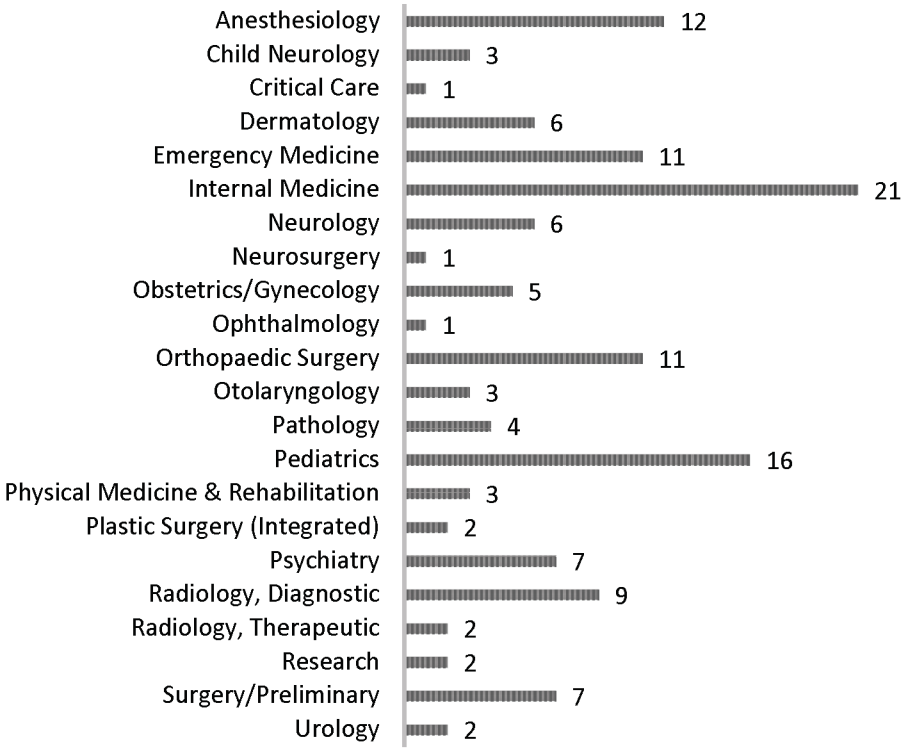


FIGURE 3. **Comparison Group Specialty Matches (n=135)**



 https://doi.org/10.22454/FamMed.2023.596964

TABLE 2. Summary of Qualitative Data Set

	FM Study Group	Non-FM Comparison Group	Total
Students	135	125	260
Data set page count	585	437	1,022
Data set word count	267,658	260,274	527,932

Qualitative Findings

A W2V-guided content analysis revealed several thematic differences between the two groups. Themes were either shared between the two groups but presented in ways that were particular to each, or themes were unique to a specific group. Two themes emerged from both groups, but in different ways: emotions and scientific interests. Two themes were unique to the FM study group: an orientation toward service and a special interest in communities and people. One additional theme of religious faith was found only in the FM study group, but we classified this as a secondary theme, attributable to UC's Midwestern location. Two themes were unique to the non-FM comparison group: research interests and strong career aspirations.

Themes Shared Between Groups

The two themes that were shared across both groups were emotions and scientific experiences and interests. However, the two groups presented distinctive profiles within these two themes. FM matches tended to use more emotional language overall, and more positive emotional language, whereas non-FM matches used less emotional language overall, but also more negative emotional language. For example, some of the key words associated with the emotions theme in the FM group were "compassion," "happiness," "joy," "care," and "encourage." Conversely, some of the key words associated with the emotions theme in the non-FM group were "fear," "failure," "difficult," "stress," and "suffer." Similarly, both groups wrote extensively about both their past academic experiences and future scholarly interests, particularly as they related to their scientific studies. However, the FM study group used scientific terms less frequently overall, and those terms tended to be more general, such as "chart," "diagnosis," "symptom," and "sick." The non-FM comparison group used scientific terms more frequently, and those terms tended to be more specialized, such as "pharma," "robot," "instrument," "transplant," and "interpret."

Group-Specific Themes

Each group also presented its own unique themes. These are topic areas that certainly appeared in the alternative group, but not with enough frequency or emphasis to be characterized as a theme. In the FM study group, students wrote extensively about two key interests: a commitment to service, and a desire to deeply engage people, both at the community level and the individual level. In expressing their commitment to service, students in the FM study group used key terms such as "income," "disabled," "underserved," "outreach," "urban," "afford," "disparity," "need," and of course, "service." In discussing the importance of communities and people both in their past experiences and in their future goals, students in the FM study group used words such as "community," "club," "group," "member," "participate," "children," and "people." Finally, religious faith emerged as a secondary theme in the FM study group. FM matches described the importance of their religious faith using the following key terms: "ministry" (45 in FM/12 in non-FM), "Bible" (41/22), "Christian" (58/30), "mission" (84/54), "Xavier" (University, 56/20), and "[University of] Notre Dame" (46/15). We attribute this secondary theme to UC's location in a Midwestern city with a strong Catholic identity.[23] Again, the non-FM comparison group also employed some of these words in their application documents, but not so frequently that service, community/people, or faith could be considered themes in the non-FM comparison group.

We identified themes unique to the non-FM comparison group as well, particularly around their experiences and interests in research and their career aspirations. Students in the non-FM group described lab and clinical research experiences and interests, using key terms like "tissue," "device," "mice," "protocol," "sequence," "assay," "instrument," "experiment," and "abstract." They also expressed stronger career aspirations overall in both frequency of code usage and strength of language employed, characterizing their ambitions using terms such as "aspire," "intellectual," "cure," "career," "success," "goal," and "reward/ing." Although the FM study group also described research experiences and career aspirations, the frequency and emphasis with which they discussed these two topics was not such that these topics could be classified as primary themes in the study group. Table 3 provides a summary of these findings, including representative quotations and word counts for each code from the data set to illustrate each theme.

Themes from the study group's application data portrayed medical students who care about individuals and their stories, reflect positively on their life experiences, desire to be part of a compassionate community, want to contribute to overcoming health inequities, and view medicine as a profession of both science and service. Themes from the comparison group's application data portray medical students who exhibit greater interest in achieving their career aspirations and enjoy the intellectual challenge of medicine, bring a wealth of past research and lab experiences to their medical studies, and are motivated and excited by the possibilities of scientific discovery in medicine.

DISCUSSION

Using novel analysis methods that integrated machine learning and traditional content analysis, our study discovered linguistic differences between future FM and future non-FM physicians using qualitative medical school application data. Although medical school applicants in the study sample shared some

TABLE 3. Themes With Selected Codes and Quotations From the Study and Comparison Groups

FM Study Group Themes with Selected Codes (Code Counts for FM/Non-FM)	FM Selected Quotes	Non-FM Comparison Group Themes with Selected Codes (Code Counts for Non-FM/FM)	Non-FM Selected Quotes
Emotions: More emotional language overall and more positive emotions **Codes:** Compassion (119/62) Happiness (43/5) Joy (36/8) Care (563/250) Encourage (103/44)	"From this, I saw not only the potential of medicine to alleviate illness but also the power of compassion to heal the human being." "I found the most joy in knowing something about the patients I encountered, in trading stories and in being connected." "What I lacked, however, was knowledge and training to affect [individuals'] stories of health and happiness, their most vulnerable stories, for the better."	**Emotions:** Less emotional language overall and more negative emotions **Codes:** Difficult (114/80) Stress (80/51) Suffer (72/50) Fail (61/38)	"I experienced the stressful atmosphere of the hospital during long overnight shifts and the difficulty in juggling numerous patients, including uncooperative and unresponsive patients." "I learned how important it is to not only treat the patient, but also to provide support for the family as they suffer alongside their loved one."
Science/Academic: Less frequent use of scientific terms overall and terms used were more general **Codes:** Chart (47/15) Diagnosis (36/10) Symptom (34/8) Sick (34/7)	"Reviewing charts gave me insight into the scientific aspect of clinical practice." "In her practice I saw how she used her knowledge of all the things I studied in biology, chemistry, and anatomy to unravel a patient's symptoms, find a diagnosis, and treat the condition effectively."	**Science/Academic:** More frequent use of scientific terms overall and terms used were more specialized **Codes:** Pharma (56/40) Cardiac (99/70) Robot (32/3) Instrument (31/20) Spinal (30/17)	"I also learned about the process behind cancer diagnosis and treatment plan prescription, including the interpretation of pathology reports and PET scans." "My . . . rotation, in the cardiac cath lab, provided me with the opportunity to assist in and observe minor heart procedures such as catheterizations, angiograms, angioplasties, as well as emergency procedures, such as codes."
Special Interest in Service Codes: Service (486/258 Income (35/8) Disabled (65/15) Underserved (37/11) Outreach (34/21) Urban (34/14) Afford (31/4) Disparity (27/2) Need (382/133)	"My personal experiences not only sparked a strong interest in healthcare, but also cemented my commitment to community service." "It is deeply gratifying to have a positive influence on children from underserved communities." "I began to realize that I could do my part to help eliminate the systemic healthcare disparities I saw growing up and bring culturally competent care to minority communities."	**Research/Labs Codes:** Tissue (45/22) Device (40/24) Mice (54/25) Sequence (33/14) Assay (32/12) Instrument (31/13) Abstract (29/17)	"Often I'd use protocols outlined in recently published literature to synthesize compounds for investigation which taught me the value of careful documentation, interpretation, and reporting of results." "As I advanced, I learned more technical skills such as how to grow and purify proteins, conduct gel electrophoresis, assay, screen drugs, extract DNA from cells via midi prep kits, and use the Beckman 2000 and VPrep pipetting robots."
Community/People Codes: Community (482/263) Club (212/122) Group (415/300) Participate (263/176) Children (549/303) People (421/180)	"Helping others in my community provided a sense of service and accomplishment that I wanted to incorporate into my life and future career choice." "Because I understand how it feels to be vulnerable, I am empathetic toward people in need and one of my primary goals is to treat people with dignity." " . . . I think, for me, it all comes down to simply helping people."	**Career Aspirations Codes:** Aspire (47/30) Intellectual (42/31) Career (344/268) Success (164/108) Goal (188/144) Reward/ing (117/73)	"I perceive a record or accomplishment as inspiration to better myself. I ask myself to do more, and to do what others haven't done before. I anticipate and expect to be successful because I hold myself to a higher standard." "I desire and embrace intellectual challenge because I've come to realize the satisfaction it can yield." "My goal is [to] gain exposure to a wide variety of specialties and identify those that I should consider as a possible career path."

themes between them, such as emotions and scientific interests, these shared themes exhibited important differences in the content analysis.

The results of this study should be interpreted and applied carefully. One limitation is that the data originated from a selective group of individuals at one institution when compared to the at-large population, therefore the language used is also distilled to a similar group of individuals who have completed similar education prior to applying to medical school. An important next step in the research process is to enlarge the data set to capture a cross-section of medical students from different sizes and types of medical schools from every major region of the United States in order to see if the same themes are present in a broader data sample. Although novel, the analysis process also presents a limitation in that the content analysis was guided by W2V results, therefore additional themes may have been missed by the machine that would have been noted by a straightforward content analysis.

Findings from this study may have important implications for the recruitment of future family physicians because they

provide a foundation for the thematic differences in medical school application data when comparing students who eventually matched to FM residency programs and students who matched to other specialties. Medical school applications are numerous, but machine learning analysis of admissions data could help prospectively identify students whose applications are thematically similar to past FM matches. Such students could be supported through mentorship, connection with other students interested in FM to allow exploration of the field of FM earlier in the medical school experience, and/or with longitudinal shadowing opportunities, factors that have been shown to impact FM residency choice.[4] They may also benefit from opportunities to participate in community service events since community and service were two themes in the FM data set.

Previous studies of factors that influence medical students' choice of specialty have primarily used quantitative methods, although some qualitative studies do exist. Our results do not suggest a qualitative approach is superior, but rather complementary to quantitative surveys or demographic analyses. Our results echo previous findings in the literature regarding important factors, such as values[6] (service and community, in our FM data set), interest in basic science research[12] (research/labs, in our non-FM data set), and prestige[24] and career opportunities[7] associated with a specialty (research interest and career aspirations, in our non-FM data set). Most importantly, previous studies have shown the impact that faculty role models, mentorship, interest in FM at the start of medical school, and a positive perception of the future of FM can have on students' choice of FM.[5–8,24] Building on this evidence, our findings provide the necessary foundation for understanding the characteristics and interests of future family physicians while they are still in medical school. In a novel analysis of qualitative data derived from medical students' application documents, this study utilized machine learning to identify key differences between students who matched into FM residency programs and students who matched into non-FM residency programs.

Acknowledgments

The authors acknowledge Barbara Gadzinski in University of Cincinnati Office of Student Affairs and Andrea (Andi) Oakes in Office of Information Technology for their generosity of time and effort in providing access to student admissions data. Additionally, the natural language processing tool used for this project was developed by UC's Digital Scholarship Center through a grant from the Andrew W. Mellon Foundation, Public Knowledge program, 2005-07903. The authors thank James Lee, Ezra Edgerton, and Andrew Boylan for their development of the platform and support of this work.

Presentations

Parts of this study were presented at the following conferences:

- Cracking the Code: How to Recognize a Future Family Medicine Physician. Family Medicine Education Consortium (FMEC) Annual Meeting, Pittsburgh, PA, October 8-10, 2021.
- Use of Machine Learning to Guide Qualitative Data Analysis of Medical School Admissions Data. International Congress of Qualitative Inquiry (ICQI), Urbana-Champaign, IL (virtual), May 2021.
- Pipeline Tracking: Use of Artificial Intelligence To Identify Potential FM Students. Society of Teachers of Family Medicine, New Orleans, LA, May 2021.

REFERENCES

1. National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Care Services; Committee on Implementing High-Quality Primary Care. Implementing High-Quality Primary Care: Rebuilding the Foundation of Health Care, . National Academies Press; 2021. doi: https://doi.org/10.17226/25983
2. IHS Markit Ltd. The complexities of physician supply and demand: projections from 2018 to 2033. 2020. https://www.aamc.org/system/files/2020-06/stratcomm-aamc-physician-workforce-projections-june-2020.pdf, Accessed January 24, 2022.
3. Rosenthal MP, Diamond JJ, Rabinowitz HK. Influence of income, hours worked, and loan repayment on medical students' decision to pursue a primary care career. *JAMA.* 1994;271(12):914-917.
4. Xu G, Rattner SL, Veloski JJ, Hojat M, Fields SK, Barzansky B. A national study of the factors influencing men and women physicians' choices of primary care specialties. *Acad Med.* 1995;70(5):398-404.
5. Kost A, Bentley A, Phillips J, Kelly C, Prunuske J, Morley CP. Graduating medical student perspectives on factors influencing specialty choice: an AAFP national survey. *Fam Med.* 2019;51(2):129-136.
6. Senf JH, Campos-Outcalt D, Kutob R. Factors related to the choice of family medicine: a reassessment and literature review. *J Am Board Fam Pract.* 2003;16(6):502-512.
7. Yang Y, Li J, Wu X. Factors influencing subspecialty choice among medical students: a systematic review and meta-analysis. *BMJ Open.* 2019;9(3):22097-22097.
8. Jordan J, Brown JB, Russell G. Choosing family medicine: what influences medical students? . *Can Fam Phys.* 2003;49:1131-1137. .
9. Scott I, Wright B, Brenneis F, Brett-Maclean P, Mccaffrey L. Why would I choose a career in family medicine? reflections of medical students at 3 universities. *Can Fam Phys..* 2007;53(11):1956-1957.
10. Mutha S, Takayama JI, O'neil EH. Insights into medical students' career choices based on third- and fourth-year students' focus-group discussions. *Acad Med.* 1997;72(7):635-640.
11. Petchey R, Williams J, Baker M. Ending up a GP': a qualitative study of junior doctors' perceptions of general practice as a career. *Fam Pract.* 1997;14(3):194-198.
12. Hull AL, Glover PB, Acheson LS. Medical school applicants' essays as predictors of primary care career choice. *Acad Med.* 1996;71(1):37-39.
13. Arbabi A, Adams DR, Fidler S, Brudno M. Identifying clinical terms in medical text using ontology-guided machine

learning. *JMIR Med Inform.* 2019;7(2):12596-12596.
14. Powers-Fletcher MV, Mccabe EE, Luken S. Convergence in viral outbreak research: using natural language processing to define network bridges in the bench-bedside-population paradigm. *Harv Data Sci Rev.* 2021.
15. Harrison CJ, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction to natural language processing. *BMC Med Res Methodol.* 2021;21(1):158-158.
16. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. *Proceedings of the 2010 LREC Workshop on New Challenges for NLP Frameworks.* 2010.
17. Jang B, Kim I, Kim JW. Word2vec convolutional neural networks for classification of news articles and tweets. *PLoS One.* 2019;14(8):220976-220976.
18. Xia C, He T, Li W, Qin Z, Zou Z. Similarity analysis of law documents based on word2vec. *Inst Electrical and Electronics Engineers*;2019:354-357.
19. Card D, Henderson P, Khandelwal U, Jia R, Mahowald K, Jurafsky D. With little power comes great responsibility. *Assoc Computational Linguistics.* 2022;2020:9263-9274.
20. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. *Neural Inform Processing Syst.* 2013;2:3111-3119.
21. Lincoln YS, Guba EG, Pilotta JJ. Naturalistic inquiry. *Int J Intercult Relat.* 1985;9(4):438-439.
22. Stahl NA, King JR. Expanding approaches for research: understanding and using trustworthiness in qualitative research. *J Dev Educ.* 2020;44(1):26-28.
23. Wartman S. Area Catholics by the numbers.. *The Cincinnati Enquirer.* 2015. https://www.cincinnati.com/story/news/2015/09/18/area-catholics-numbers/72410868/.
24. Alavi M, Ho T, Stisher C. Factors that influence student choice in family medicine: a national focus group. *Fam Med.* 2019;51(2):143-148.