# Using Machine Learning to Predict Primary Care and Advance Workforce Research

*Peter Wingrove*[1,2]
*Winston Liaw*, MD, MPH[2,3]
*Jeremy Weiss*, MD, PhD[4]
*Stephen Petterson*, PhD[2]
*John Maier*, MD, PhD[5]
*Andrew Bazemore*, MD, MPH[2]

[1]University of Pittsburgh, School of Medicine, Pittsburgh, Pennsylvania

[2]Robert Graham Center, Washington, DC

[3]University of Houston, College of Medicine, Department of Health Systems and Population Health Sciences, Houston, Texas

[4]Carnegie Mellon University, Pittsburgh, Pennsylvania

[5]University of Pittsburgh, Department of Biomedical Informatics, Pittsburgh, Pennsylvania

**MORE ONLINE**
www.annfammed.org

*Conflicts of interest: authors report none.*

**CORRESPONDING AUTHOR**

Peter Wingrove
361 Darragh Street, #108
Pittsburgh, PA 15213
pmw27@pitt.edu

## ABSTRACT

**PURPOSE** To develop and test a machine-learning–based model to predict primary care and other specialties using Medicare claims data.

**METHODS** We used 2014-2016 prescription and procedure Medicare data to train 3 sets of random forest classifiers (prescription only, procedure only, and combined) to predict specialty. Self-reported specialties were condensed to 27 categories. Physicians were assigned to testing and training cohorts, and random forest models were trained and then applied to 2014-2016 data sets for the testing cohort to generate a series of specialty predictions. Comparing the predicted specialty to self-report, we assessed performance with F1 scores and area under the receiver operating characteristic curve (AUROC) values.

**RESULTS** A total of 564,986 physicians were included. The combined model had a greater aggregate (macro) F1 score (0.876) than the prescription-only (0.745; $P < .01$) or procedure-only (0.821; $P < .01$) model. Mean F1 scores across specialties in the combined model ranged from 0.533 to 0.987. The mean F1 score was 0.920 for primary care. The mean AUROC value for the combined model was 0.992, with values ranging from 0.982 to 0.999. The AUROC value for primary care was 0.982.

**CONCLUSIONS** This novel approach showed high performance and provides a near real-time assessment of current primary care practice. These findings have important implications for primary care workforce research in the absence of accurate data.

*Ann Fam Med* 2020;18:334-340. https://doi.org/10.1370/afm.2550.

## INTRODUCTION

Approximately 1 in 8 Americans works in health care.[1] Translating that into better health depends on the presence of an effective workforce, and many believe the system needs to address shortages and maldistribution.[2-4] In response, Congress established the National Health Care Workforce Commission, though it was never funded.[1]

A primary task of the Commission was to analyze data that would inform responses to threats. For example, organizations have projected increasing shortages of primary care physicians,[4-7] underscoring the need for coordination across agencies and timely, accurate data.[8]

Unfortunately, the data needed are inadequate. Workforce data sets—the American Medical Association's Masterfile and the Centers for Medicare and Medicaid Services' (CMS) National Plan and Provider Enumeration System—have limitations. The Masterfile is a registry that documents medical school, residency, and fellowship training. Whereas training information is accurate, the registry relies on voluntary, self-reported responses for updates.[9] Thus, the Masterfile's accuracy decreases as clinicians age, reduce their hours, or change the type of care they deliver.[7,9] The National Plan and Provider Enumeration System similarly has difficulty reflecting actual practice.[10,11] Congress requires that physicians, regardless of Medicare participation, have unique identifiers—National Provider Identifiers (NPIs). The NPI specialty is self-reported, and there are neither requests for updated information, nor mechanisms to determine whether providers are clinically active.[9] Clinicians are instructed to report changes within 30 days, though there are no penalties for failing to do so.[9]

Even with timely data, misclassification remains a risk. Workforce projections use the most recent residency to categorize specialties. A first problem with this approach is the services might be inconsistent with the residency, eg, family medicine residency graduates might be practicing dermatology. Second, it disregards the contributions of physicians in other specialties and nonphysicians, eg, a rural cardiologist might be practicing primary care.

The method described below overcomes these limitations by evaluating current behavior to infer specialty. Integrating the additional data has the potential to improve accuracy and serve as a check to traditional approaches. Prescription and procedure data are available via the CMS,[12] and technological advances allow us to apply emerging techniques. Machine learning, which develops algorithms to detect patterns, has been used to predict myriad outcomes including cancer survival and myocardial infarctions,[13-17] and has also been applied to Medicare billing data to predict physician specialty and identify fraud; however, this was not restricted to physicians, did not combine specialties performing similar roles, and did not incorporate prescribing data and, as a result, had low accuracy.[18]

The present study combined prescription and procedure data to predict specialty for this purpose. Rather than rely on training, we propose a new method that assesses prescriptions and procedures to determine specialty. The objectives were to describe prescriptions and procedures by specialty, combine data on prescriptions and procedures with machine learning to develop algorithms to predict physician specialties, and test model performance against self-reported specialty.

## METHODS

### Data Sources

The American Academy of Family Physicians Institutional Review Board approved this study. For this cross-sectional study, we used the 2014-2016 CMS Medicare Fee-For-Service Provider Utilization and Payment Data: Part D Prescriber Public Use Files to identify prescriptions.[19] These data sets include information regarding beneficiaries enrolled in Medicare Part D (70% of all beneficiaries), information about providers (eg, NPI and self-reported specialty), and prescriptions (except for over-the-counter drugs).

To identify procedures, we used the 2014-2016 CMS Medicare Fee-For-Service Provider Utilization and Payment Data: Physician and Other Supplier Public Use Files.[20] In this Medicare Part B data set, procedures were identified with Healthcare Common Procedure Coding System codes. To protect privacy in these data sets, drugs and procedures were not reported by NPI if there were ≤10 claims.

### Variables

To assess the same cohort of physicians, the analysis was restricted to nonpediatric physicians appearing in all 3 years (though they only needed to appear in either the procedure or prescription data sets for a given year). To maintain consistency, physicians were only included if they self-reported the same specialty across all 3 years. We excluded nonphysicians and physician assistants (PAs) and nurse practitioners (NPs) because their subspecialties were not listed. We assigned physicians from specialties with a low number of physicians or for which multiple specialties practice in similar ways into 1 of 27 larger specialties (eg, internal medicine or family medicine were relabeled as primary care). To avoid rare drugs or procedures, we restricted the analysis to the 850 most common prescriptions and 1,500 most common procedure codes and excluded items that did not appear in all 3 years. For each year, we characterized physicians by whether they prescribed or performed each of the 2,350 prescriptions/procedures. We did not account for the number of times they prescribed or performed each.

Physicians were then randomly assigned to 2 groups of the same size (Train and Test). Each physician in the Train and Test groups had a data set of associated prescription/procedure behavior for each of the 3 years.

### Deriving the Algorithm

Random forest is an ensemble learning method that creates decision trees and generates an output based on the class value that appears most frequently, incorporating random variation to generate a lot of trees that are slightly different. This minimizes overfitting and makes the analysis robust to imbalanced data by limiting the pool of possible variables available at each split.[21,22] We selected this method for its conceptual simplicity and favorable statistical properties.[23]

To begin, we trained a separate random forest model (the combined model, consisting of both prescription and procedure data) for each year. Each random forest consisted of 200 trees and had a pool of 100 possible variables at each node. Changes in hyperparameters failed to significantly improve these models over the default settings, with the exception of slightly better performance with more possible variables at each node than the default setting; we selected a value of 100 for simplicity. We chose to run 3 separate models as an alternative to cross-validation. Because the prescription and procedure patterns associated with each specialty should be stable across each year, applying 3 separate random forest models to each year of Test data was a robust way to generate many sets of predictions and assess how consistent the method was at predicting specialty. Though these are imbalanced data, various methods to account for this, including undersampling the larger specialties and weighting the smaller specialties, improved performance for some specialties at the expense of others. Because the goal was accurate prediction for physicians regardless of specialty, we chose to leave the data unbalanced.

### Validating the Algorithm

To assess consistency, we applied each of the 3 random forest models to each of the 3 years of Test data, giving 9 sets of predictions based on the physicians in the Test group. The 9 sets of predictions were compared with self-reported specialty to generate an F1 score (harmonic mean of precision [positive predictive value] and recall [sensitivity]) for each specialty, and a macro F1 score, calculated on the average precision and recall of all specialties. We reported these values as an average across the 9 sets of predictions. We used the 2016 random forest on the 2016 Test data to create sample receiver operating characteristic curves and calculate area under the curve (AUC) values for each specialty.

The F1 score was selected as the primary measure instead of AUC value because of class imbalance. The F1 score is ideal in that it does not take into account true negatives (which will be large no matter what specialty is examined). The F1 score will be low for a given specialty if a significant number of false negatives or false positives occur, and as a result, F1 score can be low for individual specialties even if the model predicts most other specialties well. Because of the large number of true negatives when predicting small specialties, specificity (true negatives/[true negatives + false positives]) can be high even when there are many false positives and precision (true positives/[true positives + false positives]) is low. The high specificity over a large range of sensitivities leads to high AUC values.

### Prescription- and Procedure-Only Subanalyses

We generated 3 additional random forests using only the prescription variables and removing physicians with no prescription data available. We did the same for the procedure variables, removing physicians with no procedure data.

We used the 3 prescription-only models to generate 9 predictions (eg, the 2016 prescription-only model can generate predictions using the 2014, 2015, and 2016 Test data sets) based on the Test data sets, looking only at variables for prescriptions. We did the same for the 3 procedure-only models. We then generated an F1 score for each specialty and macro F1 scores for the prescription-only and procedure-only sets of predictions.

### Statistical Analysis

We used 2-sided paired *t* tests to assess whether the performance of the combined method differed from the prescription-only or procedure-only method, by specialty as well as macro F1 score. Data are presented as mean (%) or mean (95% CI). We considered $P < .05$ to be statistically significant.

### Aggregate Analysis

We summed the predicted number of physicians in each specialty for the 9 predictions generated by the combined random forests, averaged the counts, and compared them to the specialty distribution of the Test set to assess if the overall predicted physician counts were in line with the actual Test set counts. To assess model consistency at the individual physician level, we looked at 2016 data for physicians in the Test set and used the 3 combined (2014-2016) models to generate 3 predictions. We defined model agreement as all 3 models predicting the same specialty. We focused on a single year of prescribing and procedural data for physicians in the Test set because even though we excluded physicians who did not self-report a consistent specialty across all 3 years, it was still possible that a physician's actual specialty had changed year to year. Choosing to apply the 2014-2016 random forest models to just the 2016 Test data set removed the possibility that the model was inconsistent when a physician in the Test data set showed behavior that changed across the years; prediction disagreement in terms of their 2014, 2015, and 2016 specialty might have reflected the model working as intended. We then categorized physicians according to whether their self-reported specialties did or did not match the predictions.

Statistical analyses were performed with Stata version 15.0 (StataCorp, LLC). The random forest models were run with the ranger package in R, and AUC was calculated in R with the pROC package.[24]

### National Provider Identifiers

Despite its flaws, self-report via NPI is an appropriate reference standard. First, it effectively deals with the concern that historical training differs from current practice by divorcing specialty categorization from residency training. This would remain an issue if we used the American Medical Association's Masterfile. Second, by only including those physicians who appeared in the prescribing or procedural data set, we excluded those not clinically active. Our models are based on the aggregate behavior of a large number of physicians, and we hypothesized that they are not meaningfully influenced by the small number of physicians with inaccurate self-reported specialty.

## RESULTS

We included 564,986 physicians (n = 282,493 in the Train and Test groups). A breakdown by specialty for the Train and Test sets is shown in Table 1. The smallest specialty was allergy/immunology, comprising 0.6% of the physicians in both data sets, and the largest was primary care, comprising 35.6% and 35.9% of the Train and Test sets, respectively. Using prescription data only, approximately 40% of physicians identified as primary care compared with approximately 34% using procedure data only (Supplemental Table 1). Psychiatrists exhibited a similar pattern, with more appearing in the prescription data set than the procedure data set. The inverse was true for specialists who routinely perform procedures.

Primary care physicians prescribed the greatest mean number of unique drugs (61.4), more than 50% more than the next greatest group (cardiologists, 38.1) (Table 1). Radiologists had the greatest mean number of unique procedure codes (35.7).

Comparing the combined and procedure-only predictions, the combined model was significantly better for 18 (66.7%) specialties, worse for 8 (29.6%), and no different for 1 (3.7%) (Table 2; see Supplemental Table 2 for recall, negative predictive, and positive predictive values). Comparing the combined to prescription-only predictions, 19 (70.4%) were significantly better, 6 (22.2%) were worse, and 2 (7.4%) were no different. Macro F1 scores also showed statistically significant differences; the combined model (0.876) was more than 0.05 greater than the procedure-only model (0.821) and more than 0.10 greater than the prescription-only model (0.745).

With respect to the overall robustness of the combined model, 22 specialties (81.5%) had mean F1 scores > 0.80, and 15 (55.6%) had sc0ores > 0.90 (Table 2). The 3 worst specialties were plastic surgery (0.533), physical medicine and rehabilitation (0.586), and neurosurgery (0.650), and the combined model was significantly better than the procedure-only and prescription-only models for all 3 of these specialties. No specialty had a score of < 0.500 for the combined model. The F1 score for the combined model for primary care was 0.920.

These performance characteristics translated to high AUC values (Supplemental Table 3); 22 specialties (81.5%) had AUC values > 0.99. The lowest AUC was for primary care (0.982).

These models also generated relatively accurate predictions

**Table 1. Prescriptions and Procedures and Comparison of Train and Test Data Sets, by Specialty**

| Specialty | Drugs Prescribed, Mean No. | Procedure Codes, Mean No. | Train, n (%) | Test, n (%) |
|---|---|---|---|---|
| Allergy/immunology | 12.9 | 8.7 | 1,611 (0.6) | 1,625 (0.6) |
| Anesthesiology | 16.4 | 7.4 | 16,087 (5.7) | 16,110 (5.7) |
| Cardiology | 38.1 | 21.3 | 11,465 (4.1) | 11,170 (4.0) |
| Dermatology | 12.8 | 17.3 | 5,609 (2.0) | 5,498 (1.9) |
| Emergency medicine | 8.6 | 5.5 | 18,689 (6.6) | 18,663 (6.6) |
| Endocrinology | 31.7 | 9.7 | 2,376 (0.8) | 2,497 (0.9) |
| Gastroenterology | 16.2 | 13.2 | 5,999 (2.1) | 5,960 (2.1) |
| Hematology-oncology | 19.2 | 18.5 | 5,638 (2.0) | 5,572 (2.0) |
| Infectious disease | 21.8 | 7.0 | 2,337 (0.8) | 2,328 (0.8) |
| Nephrology | 36.6 | 13.3 | 3,735 (1.3) | 3,691 (1.3) |
| Neurology | 27.4 | 9.4 | 6,431 (2.3) | 6,217 (2.2) |
| Neurosurgery | 5.1 | 8.8 | 1,976 (0.7) | 2,008 (0.7) |
| Obstetrics and gynecology | 5.4 | 4.6 | 11,361 (4.0) | 11,505 (4.1) |
| Ophthalmology | 13.7 | 13.2 | 8,837 (3.1) | 8,755 (3.1) |
| Orthopedic surgery | 5.9 | 13.5 | 10,980 (3.9) | 11,095 (3.9) |
| Otolaryngology | 9.9 | 10.5 | 4,322 (1.5) | 4,262 (1.5) |
| Pathology | 5.6 | 10.2 | 4,682 (1.7) | 4,831 (1.7) |
| Physical medicine and rehabilitation | 14.5 | 10.6 | 3,610 (1.3) | 3,438 (1.2) |
| Plastic surgery | 3.1 | 5.7 | 1,864 (0.7) | 1,795 (0.6) |
| Primary care | 61.4 | 11.7 | 100,682 (35.6) | 101,498 (35.9) |
| Psychiatry | 20.9 | 4.0 | 15,075 (5.3) | 14,974 (5.3) |
| Pulmonology | 22.8 | 12.7 | 5,282 (1.9) | 5,395 (1.9) |
| Radiation oncology | 3.5 | 14.5 | 1,926 (0.7) | 1,903 (0.7) |
| Radiology | 4.3 | 35.7 | 11,840 (4.2) | 11,816 (4.2) |
| Rheumatology | 33.4 | 15.1 | 1,975 (0.7) | 2,030 (0.7) |
| Surgery | 5.1 | 9.2 | 13,536 (4.8) | 13,278 (4.7) |
| Urology | 17.3 | 20.1 | 4,568 (1.6) | 4,579 (1.6) |
| **Total** | | | 282,493 (100) | 282,493 (100) |

Note: Prescribing data are from the 2014-2016 Centers for Medicare and Medicaid Services (CMS) Medicare Fee-For-Service Provider Utilization and Payment Data: Part D Prescriber Public Use Files.[19] Procedure data are from 2014-2016 CMS Medicare Fee-For-Service Provider Utilization and Payment Data: Physician and Other Supplier Public Use Files.[20]

**Table 2. F1 Scores for Random Forests, by Specialty and Type of Training Data**

| | F1 Score, Mean (95% CI) | | |
|---|---|---|---|
| Specialty | Combined | Prescription Only | Procedure Only |
| Allergy/immunology | 0.912<br>(0.906-0.917) | 0.860[a]<br>(0.855-0.865) | 0.903[a]<br>(0.901-0.905) |
| Anesthesiology | 0.951<br>(0.950-0.952) | 0.624[a]<br>(0.615-0.632) | 0.963[a]<br>(0.962-0.964) |
| Cardiology | 0.938<br>(0.935-0.940) | 0.917[a]<br>(0.913-0.919) | 0.929[a]<br>(0.927-0.932) |
| Dermatology | 0.966<br>(0.964-0.968) | 0.940[a]<br>(0.937-0.943) | 0.964[b]<br>(0.963-0.965) |
| Emergency medicine | 0.897<br>(0.894-0.899) | 0.716[a]<br>(0.708-0.723) | 0.914[a]<br>(0.911-0.916) |
| Endocrinology | 0.865<br>(0.858-0.871) | 0.869[a]<br>(0.863-0.875) | 0.623[a]<br>(0.615-0.630) |
| Gastroenterology | 0.923<br>(0.922-0.924) | 0.901[a]<br>(0.897-0.905) | 0.900[a]<br>(0.897-0.903) |
| Hematology-oncology | 0.874<br>(0.872-0.876) | 0.872<br>(0.869-0.876) | 0.677[a]<br>(0.673-0.680) |
| Infectious disease | 0.745<br>(0.740-0.750) | 0.758[a]<br>(0.754-0.763) | 0.474[a]<br>(0.462-0.486) |
| Nephrology | 0.885<br>(0.882-0.887) | 0.866[a]<br>(0.864-0.868) | 0.882[a]<br>(0.879-0.884) |
| Neurology | 0.885<br>(0.883-0.887) | 0.895[a]<br>(0.892-0.897) | 0.732[a]<br>(0.725-0.739) |
| Neurosurgery | 0.650<br>(0.645-0.656) | 0.377[a]<br>(0.364-0.389) | 0.631[a]<br>(0.626-0.635) |
| Obstetrics and gynecology | 0.920<br>(0.917-0.923) | 0.928[a]<br>(0.927-0.929) | 0.868[a]<br>(0.865-0.870) |
| Ophthalmology | 0.982<br>(0.982-0.982) | 0.975[a]<br>(0.974-0.976) | 0.987[a]<br>(0.987-0.987) |
| Orthopedic surgery | 0.884<br>(0.880-0.888) | 0.760[a]<br>(0.756-0.765) | 0.901[a]<br>(0.900-0.903) |
| Otolaryngology | 0.932<br>(0.927-0.937) | 0.874[a]<br>(0.868-0.880) | 0.951[a]<br>(0.949-0.952) |
| Pathology | 0.987<br>(0.986-0.987) | 0.005[a]<br>(0-0.012) | 0.990[a]<br>(0.990-0.990) |
| Physical medicine and rehabilitation | 0.586<br>(0.583-0.589) | 0.380[a]<br>(0.374-0.387) | 0.492[a]<br>(0.483-0.500) |
| Plastic surgery | 0.533<br>(0.527-0.539) | 0.314[a]<br>(0.287-0.341) | 0.383[a]<br>(0.378-0.388) |
| Primary care | 0.920<br>(0.917-0.922) | 0.911[a]<br>(0.910-0.912) | 0.878[a]<br>(0.876-0.880) |
| Psychiatry | 0.930<br>(0.929-0.932) | 0.938[a]<br>(0.936-0.940) | 0.740[a]<br>(0.734-0.745) |
| Pulmonology | 0.836<br>(0.834-0.837) | 0.843[a]<br>(0.839-0.848) | 0.818[a]<br>(0.814-0.822) |
| Radiation oncology | 0.939<br>(0.933-0.945) | 0.691[a]<br>(0.680-0.702) | 0.976[a]<br>(0.975-0.977) |
| Radiology | 0.979<br>(0.977-0.980) | 0.275[a]<br>(0.272-0.279) | 0.984[a]<br>(0.983-0.985) |
| Rheumatology | 0.916<br>(0.913-0.919) | 0.916<br>(0.913-0.918) | 0.726[a]<br>(0.719-0.733) |
| Surgery | 0.774<br>(0.767-0.781) | 0.624[a]<br>(0.614-0.634) | 0.735[a]<br>(0.731-0.740) |
| Urology | 0.962<br>(0.958-0.965) | 0.950[a]<br>(0.947-0.953) | 0.962<br>(0.961-0.963) |
| Macro F1 | 0.876<br>(0.874-0.878) | 0.745[a]<br>(0.741-0.748) | 0.821[a]<br>(0.820-0.823) |

Note: Prescribing data are from 2014-2016 Centers for Medicare and Medicaid Services (CMS) Medicare Fee-For-Service Provider Utilization and Payment Data: Part D Prescriber Public Use Files.[19] Procedure data from 2014-2016 CMS Medicare Fee-For-Service Provider Utilization and Payment Data: Physician and Other Supplier Public Use Files.[20]

[a] $P$ <.01 (paired $t$ test vs combined).

[b] $P$ <.05 (paired $t$ test vs combined).

for specialty counts (Table 3). Nineteen (70.4%) of the predicted counts for specialties were within 5% of the actual counts. The models underestimated the number of physicians in several specialties, including infectious disease, neurosurgery, physical medicine and rehabilitation, and plastic surgery. In contrast, the model overestimated the number of physicians practicing primary care by 3.7%.

With respect to consistency, the 3 models predicted the same specialty for 97.0% of physicians, when applied to the same year of Test prescription and procedure data (2016) (Table 4). Among these, 89.4% were consistently predicted as the specialty that matched their self-report, whereas 7.6% were consistently predicted as a nonmatching specialty. These values were 98.3%, 92.6%, and 5.8%, respectively, for primary care.

**Table 3. Predicted vs Actual Counts of Physicians, by Specialty, for Combined Random Forest Models**

| Specialty | Predicted | Actual | Predicted/ Actual, % |
|---|---|---|---|
| Allergy/immunology | 1,590 | 1,625 | 97.8 |
| Anesthesiology | 16,100 | 16,110 | 99.9 |
| Cardiology | 10,790 | 11,170 | 96.6 |
| Dermatology | 5,569 | 5,498 | 101.3 |
| Emergency medicine | 17,886 | 18,663 | 95.8 |
| Endocrinology | 2,330 | 2,497 | 93.3 |
| Gastroenterology | 6,064 | 5,960 | 101.7 |
| Hematology-oncology | 5,468 | 5,572 | 98.1 |
| Infectious disease | 1,704 | 2,328 | 73.2 |
| Nephrology | 3,702 | 3,691 | 100.3 |
| Neurology | 5,747 | 6,217 | 92.4 |
| Neurosurgery | 1,342 | 2,008 | 66.8 |
| Obstetrics and gynecology | 11,425 | 11,505 | 99.3 |
| Ophthalmology | 8,758 | 8,755 | 100.0 |
| Orthopedic surgery | 11,008 | 11,095 | 99.2 |
| Otolaryngology | 4,021 | 4,262 | 94.3 |
| Pathology | 4,790 | 4,831 | 99.2 |
| Physical medicine and rehabilitation | 2,154 | 3,438 | 62.7 |
| Plastic surgery | 1,419 | 1,795 | 79.1 |
| Primary care | 105,225 | 101,498 | 103.7 |
| Psychiatry | 14,912 | 14,974 | 99.6 |
| Pulmonology | 5,525 | 5,395 | 102.4 |
| Radiation oncology | 1,881 | 1,903 | 98.8 |
| Radiology | 11,547 | 11,816 | 97.7 |
| Rheumatology | 2,086 | 2,030 | 102.8 |
| Surgery | 14,949 | 13,278 | 112.6 |
| Urology | 4,498 | 4,579 | 98.2 |

Note: The predicted counts are based on the combined models and averaged all 9 sets of predictions. The actual counts are the number of physicians by specialty in the Test data set. Values are rounded to the nearest integer.

## DISCUSSION

In this study, we developed high-performing models to predict specialties. With noted exceptions, these models exhibited high F1 scores and AUC values, especially in comparison to earlier work.[18]

For several specialties, including neurosurgery and physical medicine and rehabilitation, the models' performance was suboptimal. We hypothesize that these specialties have high overlap with other specialties, making classification difficult. This finding was not true for primary care, suggesting that the constellation of procedures and prescriptions is also important. Whereas primary care shares prescriptions and procedures with a broad range of specialties, few share its breadth.

Our method has implications for primary care workforce studies. For example, this approach can be used to identify primary care PAs/NPs, who do not have mandated residencies and have eluded classification.[25] Workforce projections have been hampered by these limitations. For example, across 40 state workforce assessments, 60% did not include PAs/NPs, citing inadequate data as justification for their exclusion.[26] To capture the contribution of PAs/NPs, researchers have relied on surveys and state licensing data,[27,28] which have response rates of 20% to 30%.[29,30]

Our approach also enhances the accuracy and granularity of projections. As noted, workforce projections rely on training though this might not reflect current practice.[5-7] Our approach provides a near real-time assessment of behavior. This subtle distinction might affect residencies created and policies supported. This method also allows for identification of physicians not easily categorized such as those providing HIV care.[31]

**Table 4. Model Agreement and Specialty Match Using 2016 Data**

| Specialty | Count | Models Predicting the Same Specialty, % | Specialty Match, %[a] | Specialty Mismatch, %[b] |
|---|---|---|---|---|
| Allergy/immunology | 1,625 | 97.1 | 89.6 | 7.5 |
| Anesthesiology | 16,110 | 97.9 | 94.3 | 3.6 |
| Cardiology | 11,170 | 96.9 | 90.4 | 6.5 |
| Dermatology | 5,498 | 98.8 | 96.7 | 2.1 |
| Emergency medicine | 18,663 | 98.3 | 87.0 | 11.3 |
| Endocrinology | 2,497 | 95.8 | 83.3 | 12.5 |
| Gastroenterology | 5,960 | 97.2 | 92.4 | 4.8 |
| Hematology-oncology | 5,572 | 94.9 | 84.9 | 10.0 |
| Infectious disease | 2,328 | 91.1 | 61.2 | 29.9 |
| Nephrology | 3,691 | 96.7 | 86.9 | 9.8 |
| Neurology | 6,217 | 94.5 | 83.1 | 11.4 |
| Neurosurgery | 2,008 | 80.6 | 48.3 | 32.3 |
| Obstetrics and gynecology | 11,505 | 96.7 | 90.6 | 6.1 |
| Ophthalmology | 8,755 | 99.1 | 97.9 | 1.2 |
| Orthopedic surgery | 11,095 | 94.6 | 86.1 | 8.5 |
| Otolaryngology | 4,262 | 96.8 | 89.5 | 7.3 |
| Pathology | 4,831 | 99.3 | 97.8 | 1.5 |
| Physical medicine and rehabilitation | 3,438 | 83.2 | 41.6 | 41.6 |
| Plastic surgery | 1,795 | 80.7 | 42.2 | 38.5 |
| Primary care | 101,498 | 98.3 | 92.6 | 5.7 |
| Psychiatry | 14,974 | 97.9 | 92.1 | 5.8 |
| Pulmonology | 5,395 | 96.1 | 83.2 | 12.9 |
| Radiation Oncology | 1,903 | 95.9 | 91.0 | 4.9 |
| Radiology | 11,816 | 99.1 | 96.4 | 2.7 |
| Rheumatology | 2,030 | 97.6 | 91.7 | 5.9 |
| Surgery | 13,278 | 91.7 | 77.7 | 14.0 |
| Urology | 4,579 | 97.3 | 94.5 | 2.8 |
| **Overall** | 282,493 | 97.0[c] | 89.4[c] | 7.6[c] |

For this analysis, we applied the 2014, 2015, and 2016 combined random forests to 2016 Test data, for a total of 3 predictions based on prescribing and procedure data for a single year. Model agreement is defined as all 3 models predicting the same specialty.

[a] All 3 models predicted the self-reported specialty.

[b] All 3 models predicted a specialty that differed from the self-reported category.

[c] Mean across all specialties weighted by number in each specialty.

There are several limitations to the study. First, we excluded physicians not billing Medicare, only participating in Medicare Advantage, or only providing pediatric care. Physicians had to prescribe drugs or perform procedures >10 times to appear in the data set. A national all-payer claims database would overcome these limitations. Second, we evaluated a single technique in this analysis. Whereas random forest models are broadly used, it is possible that other techniques or changes to parameters might improve accuracy.[32] Third, we only included physicians appearing in 3 consecutive years. These analyses need to be repeated with a cohort that involves physicians with less longitudinal data to determine if results are similar. Fourth, we were unable to understand the motivations behind scope deviations, eg, a family physician could practice differently because of unique disease patterns in their service area. Understanding these motivations via a qualitative approach would provide additional context. Finally, we used self-reported specialty for training and testing. As mentioned, this database does not have a penalty for out-of-date information, though physicians are instructed to report changes.[9]

In summary, we report a novel method for identifying primary care physicians. These models exhibit high performance, and because they identify the practice patterns of specialties, they can be used to identify primary care PAs and NPs. By assessing current practice rather than historical training, this approach has the potential to change how the primary care workforce is tracked.

**To read or post commentaries in response to this article, see it online at https://www.AnnFamMed.org/content/18/4/334.**

**Key words:** biostatistical methods; workforce; Medicare

Submitted Feb 22, 2019; revised, Nov 27, 2019; accepted Jan 6, 2020.

**Prior presentation:** 2017 North American Primary Care Research Group Annual Meeting; November 17-21, 2017; Montreal, Canada.

**Supplemental materials:** available at https://www.AnnFamMed.org/content/18/4/334/suppl/DC1/.

## References

1. Buerhaus PI, Retchin SM. The dormant National Health Care Workforce Commission needs congressional funding to fulfill its promise. *Health Aff (Millwood)*. 2013;32(11):2021-2024.
2. Committee on the Governance and Financing of Graduate Medical Education; Board on Health Care Services; Institute of Medicine. *Graduate Medical Education That Meets the Nation's Health Needs*. Eden J, Berwick D, Wilensky G, eds. National Academies Press; 2014. https://pubmed.ncbi.nlm.nih.gov/25340242/. Accessed May 19, 2020.
3. Chen C, Petterson S, Phillips RL, Mullan F, Bazemore A, O'Donnell SD. Toward graduate medical education (GME) accountability: measuring the outcomes of GME institutions. *Acad Med*. 2013;88(9):1267-1280.
4. Council on Graduate Medical Education. *Towards the Development of a National Strategic Plan for Graduate Medical Education. 23rd Report.* https://www.hrsa.gov/sites/default/files/hrsa/advisory-committees/graduate-medical-edu/reports/April2017.pdf. Published 2017. Accessed May 19, 2020.
5. Association of American Medical Colleges. *The Complexities of Physician Supply and Demand: Projections from 2013 to 2025*. https://www.kff.org/wp-content/uploads/sites/3/2015/03/ihsreportdownload.pdf. Published 2015. Accessed Jun 26, 2016.
6. Duchovny N, Trachtman S, Werble E; Congressional Budget Office. *Projecting Demand for the Services of Primary Care Doctors*. https://www.cbo.gov/system/files/115th-congress-2017-2018/workingpaper/52748-workingpaper.pdf. Published 2017. Accessed May 19, 2020.
7. Petterson SM, Liaw WR, Tran C, Bazemore AW. Estimating the residency expansion required to avoid projected primary care physician shortages by 2035. *Ann Fam Med*. 2015;13(2):107-114.
8. Starfield B, Shi L, Macinko J. Contribution of primary care to health systems and health. *Milbank Q*. 2005;83(3):457-502.
9. Bindman AB. Using the National Provider Identifier for health care workforce evaluation. *Medicare Medicaid Res Rev*. 2013;3(3):E1-E10: mmrr.003.03.b03.
10. Will KK, Williams J, Hilton G, Wilson L, Geyer H. Perceived efficacy and utility of postgraduate physician assistant training programs. *JAAPA*. 2016;29(3):46-48.
11. American Nurses Association. *Nurse Practitioner Perspective on Education and Post-graduate Training*. https://www.nursingworld.org/practice-policy/nursing-excellence/official-position-statements/id/nurse-practitioner-perspective-on-education/. Published 2014. Accessed May 19, 2020.
12. Centers for Medicare and Medicaid Services. *Medicare Provider Utilization and Payment Data: Part D Prescriber*. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html. Published May, 2017. Updated Nov 2019. Accessed May 19, 2020.
13. Gupta S, Tran T, Luo W, et al. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open*. 2014;4(3):e004007.
14. Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D. Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records. *AI Mag*. 2012;33(4):33-45.
15. Zhai H, Brady P, Li Q, et al. Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children. *Resuscitation*. 2014;85(8):1065-1071.
16. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12(4):e0174944.
17. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.
18. Bauder RA, Khoshgoftaar TM, Richter AN, Herland M. Predicting medical provider specialties to detect anomalous insurance claims. *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, San Jose, CA; 2016. pp. 784-790.
19. Centers for Medicare and Medicaid Services. *Medicare Fee-For-Service Provider Utilization & Payment Data Part D Prescriber Public Use File: A Methodological Overview*. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Prescriber_Methods.pdf. Published 2019. Accessed Jun 26, 2020.
20. Centers for Medicare and Medicaid Services. *Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier Public Use File: A Methodological Overview*. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf. Published 2014. Updated 2019. Accessed May 19, 2020.
21. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2/3:18-22.
22. Khoshgoftaar TM, Golawala M, Van Hulse J. An empirical study of learning from imbalanced data using random forest. *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. Patras, Greece; 2007. pp. 310-317.
23. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
24. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw*. 2017;77(1):1-17.
25. Wiltse Nicely KL, Fairman J. Postgraduate nurse practitioner residency programs: supporting transition to practice. *Acad Med*. 2015;90(6):707-709.
26. Morgan P, Strand De Oliveira J, Short NM. Physician assistants and nurse practitioners: a missing component in state workforce assessments. *J Interprof Care*. 2011;25(4):252-257.
27. Doescher MP, Andrilla CH, Skillman SM, Morgan P, Kaplan L. The contribution of physicians, physician assistants, and nurse practitioners toward rural primary care: findings from a 13-state survey. *Med Care*. 2014;52(6):549-556.
28. Spetz J, Fraher E, Li Y, Bates T. How many nurse practitioners provide primary care? It depends on how you count them. *Med Care Res Rev*. 2015;72(3):359-375.
29. American Academy of Physician Assistants. *Physician Assistant Census Report: Results From the 2010 AAPA Census*. https://www.aapa.org/wp-content/uploads/2016/12/2010_AAPA_Census_Report.pdf. Published 2011. Accessed May 19, 2020.
30. US Department of Health and Human Services, Health Resources and Services Administration, National Center for Health Workforce Analysis. *Projecting the Supply and Demand for Primary Care Practitioners Through 2020*. https://bhw.hrsa.gov/sites/default/files/bhw/nchwa/projectingprimarycare.pdf. Published 2013. Accessed May 19, 2020.
31. Gilman B, Bouchery E, Barrett K, et al. *HIV Clinician Workforce Study: Final Report*. Cambridge, MA: Mathematica Policy Research; 2013.
32. Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: a survey and results of new tests. *Pattern Recognition*. 2011;44(2):330-349.