Automated Identification of a Physician’s Primary Patients

THOMAS A. LASKO, MD, MS, STEVEN J. ATLAS, MD, MICHAEL J. BARRY, MD, HENRY C. CHUEH, MD, MS

Abstract  Objective: To develop and validate an automated method for determining the set of patients for whom a given primary care physician holds overall clinical responsibility.

Design: The study included all adult patients (16,185) seen at least once in an ambulatory setting during a three-year period by 18 primary care physicians in ten practices. The physicians indicated whether they considered themselves to be the physician primarily responsible for the overall clinical care of each visiting patient. Statistical models were constructed to predict the physicians’ designations using predictor variables derived from electronically available appointment schedules and demographic information.

Measurements: Predictive accuracy was assessed primarily using the area under the receiver-operating characteristic curve (AUC), and secondarily using positive predictive value (PPV) and sensitivity.

Results: A minimal set of six variables was identified as predictive of the physicians’ designations. The constructed model had a median AUC for individual physicians of 0.92 (interquartile interval: 0.90–0.96), a PPV of 0.94 (interquartile interval: 0.87–0.95), and a sensitivity of 0.95 (interquartile interval: 0.87–0.97).

Conclusion: A statistical model using a minimal set of commonly available electronic data can accurately predict the set of patients for whom a physician holds primary clinical responsibility. Further research examining the generalization of the model to other settings would be valuable.


Determining the set of patients for whom a given physician holds primary clinical responsibility is a fundamental problem of medical research and clinical quality improvement. For primary care physicians (PCPs) this is not a straightforward task, since patients may obtain care from multiple providers for various reasons. Attempts to solve the problem reach back three decades and have been the subject of at least one international conference.

Early motivation for this work was the need to calculate practice-specific event rates in primary care research and epidemiology. These rates required the number of a physician’s patients as their denominator, and so the general problem of identifying those patients has come to be known as the “primary care denominator problem.”

Background

Insurers often address the denominator problem by requiring a patient to enroll with a particular physician when they join the health plan. These enrollment rosters are difficult to maintain with high accuracy, however, and some physicians argue that quality of care ratings should exclude patients who enrolled with them but have never seen them. Furthermore, a quality rating provided by a single insurer using a subset of a physician’s patients is inherently less meaningful as a measure of the physician’s quality of care than a rating using all of the physician’s patients. Quality improvement efforts would therefore benefit from an accurate method to identify all the patients for which a given physician is clinically responsible, regardless of insurer.

For this discussion, we define a primary patient to be one for whom the doctor has the primary role in providing medical care. A more concrete definition (adapted from Kilpatrick) might be that a primary relationship exists when the given patient would first seek primary care from the given doctor if the need arose. We will say that a patient belongs to a physician or practice if there is a primary relationship between them. The set of all primary patients of a particular physician or practice has had several names in the past: practice panel, at-risk population has all been used at one time or another.

We define a coverage patient as one who is seen by a primary care physician but the physician is not currently responsible...
for his or her overall primary care. Examples of this include physicians cross-covering for colleagues, patients obtaining out-of-town medical care, and physicians seeing patients in acute or semiacute care settings. While physicians responsible for the care they provide in these situations, most would not feel that they have assumed overall responsibility for the patient’s primary care.

Our definition of coverage patients also includes former patients, those who used to belong to the physician but have now died, moved, or switched providers. These patients may have been previously considered primary patients, but since the physician is no longer responsible for their primary care, they have become coverage patients.

The perception of whether a patient is a primary or coverage patient may be relative to one’s perspective. Physicians, insurers, and the patients themselves may have different views. This work explicitly takes the physician perspective.

Prior Work
The roster method of tracking the denominator has been done at large scale in several countries, but it has been difficult to overcome the substantial problem of keeping the roster current.\(^{11}\) Anderson et al.\(^{7}\) surveyed an entire Ontario community and used the results to validate a practice roster. The roster had been carefully maintained by a dedicated staff with close social ties to a small, geographically isolated community. The measured roster sensitivity of 89.5% and positive predictive value (PPV) of 85.8% is probably near the limits of what can be expected from manual maintenance methods.

The roster method is also popular in the United States. Most U.S. insurers and large prepaid group practices require each patient to have a designated PCP. We can find few recent data documenting the accuracy of these designations, but anecdotally many physicians relate the frustrating experience of receiving an insurer’s list of their primary patients and not recognizing many of the listed names. A computational model that could examine the records of all patients visiting a physician and identify the true primary patients would be of tremendous help in maintaining these rosters.

Little published work has focused on developing such a model. Hutchison et al.\(^{12}\) attempted to differentiate primary from coverage patients while testing an algorithm for assigning patients to one of 19 physicians based on billing history. They took advantage of a conversion from a fee-for-service to a capitation payment system (which generated new and presumably accurate rosters), making the provider assignment by using patient visits during one, two, and three years prior to the conversion date. Any patient who had seen a physician in the prior year was counted as a primary patient for that physician. Any patient who had seen a physician in both years two and three prior was also counted as a primary patient. A patient who had visited two or more physicians was counted as a primary patient of the physician visited the most, or in the case of a tie, the most recently. This algorithm, which we will call the Hutchison test, achieved high accuracy for these practices, with an average sensitivity of 95.6% (range, 94.1–97.1) and a positive predictive value of 90.1% (range, 85.0–96.3). The study was not focused on the multi-physician problem, however, and did not report the number of patients who had visited more than one physician during the study period.

Objective
The objective of our study was to develop an accurate and generalizable method of identifying a physician’s primary patients, using data available electronically. Our motivation was to produce a list that physicians would find accurate enough for both quality reporting and clinical use (such as for population-based disease management). Therefore, our reference standard is the physician’s own indication of whether a patient is a primary or coverage patient. To our knowledge, no published studies have used the physician perspective as a reference.

Methods
Candidate List Annotation
The study was approved by the Massachusetts General Hospital (MGH) Institutional Review Board. A convenience sample of 23 primary care physicians representing 12 of the 15 ambulatory practices affiliated with MGH were invited to participate in the study. A total of 18 physicians representing ten practices agreed to participate. All 18 participating physicians completed the study.

Each physician reviewed a list of patients with whom he or she had had any billable contact during the period of 04/01/2000 to 03/31/2003. Patients were seen only in connection with supervising a physician in training were excluded. A total of 18,529 patients were identified.

Each physician annotated each patient record on the list with one of three designations: (1) my patient, indicating a primary relationship; (2) not my patient, indicating a coverage relationship; or (3) my patient with reservations, indicating a primary relationship but with a nonspecific caveat that may be indicative of compliance issues, no-show rates, and other care concerns. Prior to reviewing the records, the physicians participated in discussions outlining what would reasonably constitute each of the three designations. The criteria for each designation were general and conceptual, avoiding any data-centric guidelines that could emerge as artifacts in a statistical model. Specific examples were constructed and circulated (see Appendix 1, available as a JAMIA online supplement at www.jamia.org), and general consensus was achieved on the criteria.

The physicians annotated each patient record using a Web-browser-based tool developed for this project. The tool displayed the patient’s name, past MGH outpatient visits, future scheduled outpatient appointments, the last visit note by the physician reviewing the record, and a link to the patient’s full electronic medical record.

Patient Data Collection
Nonidentifying demographic patient data and systemwide appointment data were retrieved electronically and recorded for each of the 18,529 patients. Appointments that included only consultation with a nurse or a nurse practitioner were excluded. Patients under 18 years of age and patients with missing appointment or demographic data were excluded, leaving a total of 16,185 patients for the 18 physicians. One-third of these (5,358) were randomly set aside as a test set, with the remaining 10,827 used as the training set.

Model Construction
Statistical models were constructed to predict the physician designations given a set of patient variables. The models were constructed using the Averaged One-Dependence Estimator
Algorithm that maximized model performance (described below) was used for all algorithms in this study, with modifications to the source code made locally as needed. A set of 34 candidate input variables for the models was manually constructed from the collected raw patient data (see Appendix 2, available as a JAMIA online supplement at www.jamia.org). These were subjectively identified and constructed based on their likely ability to help predict the physician designations.

The initial set of candidate variables was algorithmically reduced to a set of ten variables using a simple genetic search algorithm that maximized model performance (described below). The resulting set was then further reduced to a set of six variables by backward elimination, stopping when a statistically significant drop in model performance was measured.

Nine other variable sets were generated in a similar manner using various selection methods, and eight different predictive model types were investigated (Appendix 2). The best performing combination of model type and variable set was the AODE algorithm with the six-variable set described above. The other modeling methods investigated were naïve Bayes, C4.5, logistic regression, bootstrap aggregated C4.5, AdaBoosted C4.5, Artificial Neural Networks, and 1R.

Group models were first constructed using leave-one-out validation at the physician level. That is, 17 training subsets were concatenated, a single model was constructed using 10X cross-validation (described below) on the concatenated set, and then that model was tested on the held-out 18th training subset. This procedure was rotated through all 18 hold-out sets, giving 18 models, one for each physician, representing the combined views of the 17 other physicians regarding the patients of the 18th. Final testing for each model was on the test set for the 18th held-out physician.

Individual models were then constructed using each of the 18 training subsets using 10X cross-validation, with final testing on the corresponding test set.

The 10X cross-validation was performed by randomly partitioning the set into ten equal subsets, training the model using a concatenation of nine of the subsets and testing on the remaining subset. This was repeated for each subset held out in turn as a test set, and the results averaged over all repetitions. The model was then trained on the full original set, with the averaged result used as an estimate of performance.

Statistical Methods

Patients designated as my patient and my patient with reservations were counted as primary patients, and those designated as not my patient were counted as coverage patients. The models’ predictions were in the form of an estimated probability that the physician designated the patient as a primary patient. In order to infer a predicted designation from the probability estimate, the probability estimate was compared to a threshold value, assigning a prediction of primary patient if the probability was greater than the threshold and coverage patient otherwise. The accuracy of the predicted designation depends on the choice of threshold; lowering the threshold tends to increase sensitivity and decrease specificity, and raising the threshold acts in the opposite direction. The appropriate choice of threshold is thus problematic when comparing models of this type.

Comparisons that are independent of threshold choice can be made, however, by using receiver operating characteristic (ROC) curves. These curves plot the sensitivity of a model against its false-positive rate, or (1 – specificity), with each point on the curve corresponding to the model’s performance at a given threshold. A perfect model would have an AUC of 1.0, random guessing would give an AUC of 0.5, and a model that was wrong every time would give an AUC of 0.0. AUCs and confidence intervals were calculated using the nonparametric empirical method and statistical comparison of pairs of curves was performed using the method of DeLong et al., which corrects for the correlation of two curves generated by different models using the same data.

A comparison of the group models was also made with the previously described Hutchison test. This comparison cannot use ROC analysis because the Hutchison test does not produce probabilities, but instead predicts a simple primary or coverage designation. For this case, comparison was made by arbitrarily specifying a threshold of 0.5 for the group model and comparing the resulting sensitivity and PPV to those of the Hutchison test.

Results

Physician Characteristics

About half of the physicians in our sample were women, and there was considerable variation in the physicians’ years in practice, years at MGH, whether they frequently saw colleagues’ patients, the number of clinic hours per week, and the length of time (if any) that their panel has been closed to new patients (Table 1). The diversity of these physicians’ characteristics was reflective of that seen among all providers in their practices (data not shown).

<table>
<thead>
<tr>
<th>Description</th>
<th>Count/Median</th>
<th>Interquartile Interval</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>10 males, 8 females</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years at this institution</td>
<td>9</td>
<td>[7,11.8]</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>Practice style</td>
<td>8 collaborative, 10 solo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinic hours per week</td>
<td>16.2</td>
<td>[8.5,20]</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td>No. of days panel has been closed</td>
<td>328.5</td>
<td>[0,1535.5]</td>
<td>0*</td>
<td>2,646</td>
</tr>
</tbody>
</table>

*Nine of the 18 physicians had open panels (0 Days closed).
Patient Visits to Multiple Physicians
A total of 5,106 (32%) patient records listed visits to exactly one MGH-affiliated primary care physician during the three years (though they may have visited that physician on multiple occasions), with the remaining 11,079 (68%) recording at least one visit to at least one additional primary care physician at an MGH-affiliated practice.

Candidate List Annotations
There was a median of 779 patients per physician candidate list (range, 218–2,244). A total of 11,178 (69.1%) patients were designated as a primary patient and 5,007 (30.9%) as a coverage patient. The median fraction of primary patients for all physicians was 0.83 (range, 0.15–0.96).

Predictive Variables
The six variables in the final variable set are listed below in descending order of predictive value. Their distributions are presented in Figure 1. References to “all physicians” in these descriptions mean all primary care physicians in all MGH-affiliated practices.

Waiting fraction: the total number of days waited for appointments with the given physician, divided by the total waited for all physicians combined.

Visit difference: the total number of visits that a patient has made to the given physician minus the total to all other physicians combined.

Days since last visit: the number of days since the last visit to the given physician.

Future difference: the total number of appointments scheduled for future visits with the given physician, minus the total for all other physicians combined.

Idle ratio: the number of days since the last visit to the given physician, divided by the number of days since the first visit.

Practice style: indicates whether the given physician practices using a collaborative style where the physician frequently covers patients of other physicians (and perhaps vice versa) or a solo style where the physician infrequently covers the patients of others. This was a subjective attribute assigned prospectively.

Model Accuracy
The AODE method tied or outperformed the other modeling methods on all candidate variable sets (Appendix 2). The AUC (using cross-validation) of the final AODE model on the full training set of all patients was 0.948 (standard deviation = 0.006). The final model produced a test-set AUC of at least 0.90 for 14 of the 18 physicians (see Table 2 available as a JAMIA online supplement at www.jamia.org). There were no significant differences between the performance of any group model and its corresponding individual model.

Comparison to Prior Work
The AODE model was more accurate than the Hutchison test (see Table 3, available as a JAMIA online supplement at www.jamia.org). For the data set as a whole, the PPV of the Hutchison test was 0.011 lower (95% confidence interval [CI]: −0.024 to 0.001) than the AODE model (0.918 versus 0.929), and the sensitivity was 0.084 lower (95% CI: −0.198 to −0.074) than AODE (0.851 versus 0.935). For individual physicians, the median sensitivity of the AODE model was 0.95 (interquartile interval: 0.87–0.97) and the median PPV was 0.94 (interquartile interval: 0.87–0.95), whereas the Hutchison test had median sensitivity of 0.85 (interquartile interval: 0.81–0.90) and a median PPV of 0.92 (interquartile interval: 0.86–0.98).

Discussion
Our goal was to identify an accurate and generalizable method for enumerating the primary patients of a given physician, using data available electronically. We plan to use these patient lists for quality improvement efforts and population-based disease management. In our study population, we found that most patients saw multiple primary care physicians and therefore require some method to determine which physician(s) are actually responsible for their overall care. Our AODE-based model was able to achieve a high degree of accuracy in making these determinations, with a test-set AUC above 0.90 for 14 of the 18 physicians.

The group model performed equivalently to each physician’s individual model, giving us confidence in the generalizability of the group model to other physicians at MGH. Since the reference standard for all models was the physicians’ own assessment of whether the patient belongs to MGH. Since the reference standard for all models was the physicians’ own assessment of whether the patient belongs to the practice, the ability of the group model to perform equivalently to the individual models implies that the physicians are in general agreement about the types of patients that should be classified as belonging to their practice.

As discussed above, our models produce predictions in the form of an estimated probability that a given patient is a primary patient. The aspects of a model’s performance that are more relevant to clinicians, however, are its sensitivity and
PPV. If we imagine that the model was selecting patients for a doctor’s roster, then the model’s sensitivity represents the fraction of primary patients that made it to the roster, and PPV represents the fraction of the roster that consists of true primary patients. A perfect algorithm would put all of a doctor’s primary patients on the roster (sensitivity = 1.0), and no one else (PPV = 1.0).

But as we noted, the sensitivity and specificity (and consequently also the PPV) of a model depend on the threshold we use to convert the model’s estimated probability to an actual predicted designation of primary patient or coverage patient. One of the benefits of a probabilistic model is that the threshold can be adjusted to suit our particular needs. If the goal is to ensure that as few coverage patients as possible wind up on the physician’s list, then we choose a high threshold to increase the model’s PPV. If, conversely, we wish to ensure that a minimal number of primary patients get left off the list, then we can choose a low threshold to increase the model’s sensitivity. Although there is always a trade-off between sensitivity and specificity (which roughly translates to a trade-off between sensitivity and PPV), a model with a higher AUC will suffer a smaller trade-off.

To our knowledge, the Hutchison test is the only reported model available for comparison, and the group model shows a relevant performance advantage over it. For our comparison, we chose an arbitrary threshold of 0.5, producing a model with equivalent PPV to the Hutchison test (0.92) but substantially higher sensitivity (0.85 versus 0.93), with a 95% CI of the difference of −0.07 to −0.19. In other words, both methods would include on average eight incorrect patients per 100 on the physician’s list, but the Hutchison test would have left off 15 primary patients per 100, whereas the group model would have missed only seven. Experimenting with the threshold, we found the 0.5 point to be near optimum for most physicians and that changing the threshold generally tended to lower the sensitivity without substantially changing the PPV (data not shown). This implies that the value of the more complicated AODE model over the simpler Hutchison test may be in capturing more true positives rather than reducing the number of false positives.

The candidate variables constructed as possible inputs to the model were limited to those that could generalize to other institutions. Five of the six final variables can be constructed using data that would be available to most institutions with electronic billing and scheduling systems. The sixth variable, which encodes whether the physician practices in a collaborative or solo style, is known to the physician and can be self-identified.

We avoided variables such as the “primary care physician of record” field in our registration database because while most institutions have such a field, its accuracy may vary across institutions. Although partially accurate variables are certainly useful in statistical models, information about the accuracy of the variable gets built into the structure of the model. If that accuracy changes across institutions, the generalizability of the model would suffer.

Limitations and Future Work
Three of the 18 physicians designated small numbers of patients as coverage patients, which limited the power of the study in their cases.

The performance of the final model may vary with patient population, practice patterns, or other factors, and further testing on practices from other institutions will be necessary to validate its wider generalizability. The fact that we included practices with a wide variety of practice volume, extent of cross-coverage, and physician demographics to construct the model gives us hope that it will generalize well to other adult primary care practices. Comparison with models for pediatric or specialist practices, which may have very different patient visit patterns, would be especially interesting.

We deliberately used a physician perspective of the patient’s designation because many quality initiatives are directed toward primary care providers and their set of patients. Patients may have a different view of who their primary physician is, and comparing our model with others built using the patient’s perspective as the gold standard would be interesting. However, collecting this information from patients would be logistically challenging.

Finally, the usefulness of our results may be limited by the technical nature of our modeling methods, which may restrict their accessibility to institutions with specialized technical staff. Further research should assess whether and to what degree accuracy is lost by simplifying the methods to those that could be easily implemented. We are encouraged by the performance of the logistic regression model, which produced AUC values that were within a half percentage point of the AODE results on most of our candidate datasets (Appendix 2). The difference between our final model and that produced by logistic regression is statistically significant, although the loss in practical terms may be acceptable in many circumstances. Since logistic regression models are relatively simple to implement, they represent a promising direction for future work.

Summary
We have developed and validated a mathematical model that can accurately identify the patients of a particular primary care physician using a minimal data set available electronically at most institutions. The reference standard for the model was the physicians’ assessments of their own clinical responsibility to the patient. A single model was accurate across a set of physicians with varying characteristics, indicating that the physicians generally agreed with each other about the types of patients for which they feel primary clinical responsibility. This model has the potential to improve the accuracy of quality-of-care assessments and to facilitate practice-based epidemiologic research and population-based disease management.

References