Clinical Recommender Algorithms to Simulate Digital Specialty Consultations

Morteza Noshad[1], Ivana Jankovic[2], Jonathan H. Chen[1,3]
[1]*Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA*
[2]*Division of Endocrinology, Stanford University School of Medicine, Stanford, CA*
[3]*Division of Hospital Medicine, Stanford University, Stanford, CA*

Advances in medical science simultaneously benefit patients while contributing to an overwhelming complexity of medicine with a decision space of thousands of possible diagnoses, tests, and treatment options. Medical expertise becomes the most important scarce healthcare resource, reflected in tens of millions in the US alone with deficient access to specialty care. Combining the growing wealth of electronic medical record data with modern recommender algorithms has the potential to synthesize the clinical community's expertise into an executable format to manage this information overload and improve access to personalized care suggestions. We focus here specifically on outpatient consultations for (Endocrine) specialty expertise, one of the highest demand and most amenable areas for electronic consultation systems. Specifically we develop and evaluate models to predict the clinical orders of these initial specialty referral consultations using an ensemble of feed-forward neural networks as compared to multiple baseline algorithms. As benchmarks closer to the existing standard of care, we used diagnosis-based clinical checklists based on our review of literature and practice guidelines (e.g., Up-to-Date) for each common referral diagnosis as well as existing electronic consult referral guides. Results indicate that such automated algorithms trained on historical data can provide more personalized decision support with greater accuracy than existing benchmarks, with the potential to power fully digital consultation services that could consolidate utilization of scarce medical expertise, improving consistency of quality and access to care for more patients.

Keywords: Electronic Health Records, Collaborative Filtering, Specialty Access, Digital Consultation, Recommender Systems, Neural Networks

1. Introduction

The growing limitations in the scarcest healthcare resource - clinical expertise - is an issue that has long been at the forefront of medicine. This shortage of clinician time is particularly acute in access to medical specialty care. Patients can wait months for outpatient specialty consultation care, which contributes to 20% higher mortality.[1]

Our broader vision beyond specialty consultations is synthesizing massive amounts of medical information to support all decision making. Recommender algorithms could be used for all clinical decision making, while here we focus on specialty consultations as a concrete example point where there clearly is a request for assistance and a gap in availability of that support. We focus on recommending the clinical orders for medications and diagnostic tests from outpatient consultations that any clinician could initiate with adequate support.

This system can consolidate specialty consultation needs and open greater access to effective care for more patients. A key scientific barrier to realizing this vision is the lack of clinically acceptable tools powered by robust methods for collating clinical knowledge, with continuous improvement through clinical experience, crowd-sourcing, and machine learning. Existing tools include electronic consults that allow clinicians to email specialists for advice, but their scale remains constrained by the availability of human clinical experts. Electronic order checklists (order sets) and the e-consult referral guides are in turn limited by the effort to maintain content and adapt to individual patient contexts.[2]

Machine learning approaches are revolutionizing various healthcare areas such as medical imaging,[3] diagnostic models,[4,5] and virtual health assistants[6] by introducing more accurate, low cost, fast, and scalable solutions. Automated diagnostic workflow recommendation is another emerging application of machine learning which has so far mainly been focused on predicting the need for specific medical imaging.[7] However, only a few previous studies have explored the possibility of using machine learning approaches to design a scalable intelligent system that can recommend diagnostic procedures of any type as an alternative to the conventional clinical checklists. Authors in[8] and[9] apply recommender systems based on probabilistic topic modeling and neural networks to predict inpatient clinical order patterns. Other than predicting workflows, recommender systems have also been used for diagnosis in several previous papers.[10,11]

In this work, we address the problem of predicting outpatient specialty workflows. Specifically, our objective is to predict which tests and procedures would be ordered at the first specialty visit for a patient referred by a primary care physician (PCP), based on their unique medical records. This prediction could provide automated decision support and recommendations at primary care visits or specialist pre-visit screenings to allow diagnostic procedures to be completed while the patient is awaiting their in-person specialist visit. As opposed to manually-created medical checklists, which are mainly based on diagnosis (e.g., common laboratory and imaging tests a clinician can order to evaluate diabetes), the proposed data-driven algorithm utilizes the patient's previous laboratory results, diagnosis codes, and the most recent procedures as input and recommends follow-up laboratory orders and procedures. The proposed recommender model offers several key benefits including scalability to answer unlimited queries on-demand; maintainability through automated statistical learning; adaptability to respond to evolving clinical practices; and personalizability of individual suggestions with greater accuracy than manually-authored checklists. We categorized the input EHR data into three groups: diagnostic data, including the diagnosis codes and laboratory results; procedures ordered by the referring PCP; and the specialist being referred to (recognized by their ID). This grouping of the data allows us to use appropriate base models for each of the input data categories and process them separately. The first base model is a neural network based multi-label classifier with diagnostic data as input and specialty procedures as labels. The second model is a collaborative filtering AutoEncoder (AE) with the PCP and specialty procedures as input and output, respectively. The designed collaborative filtering AutoEncoder is similar to the the deep learning based collaborative models proposed in.[12,13] The predictions from the base models are then fed into an ensemble neural network to improve the predictions from each of the base learners. Unlike traditional ensemble methods that use the ratings from base learners to improve predictions,[14] the proposed approach leverages the specialist ID number as side information to personalize the recommendations both for the patient and speciality provider. Here, we develop and measure the potential advantages and tradeoffs of the proposed method compared to clinical checklists and several other baselines.

2. Cohort and Data Description

In this work, we address the prediction of future clinical diagnostic steps for the outpatients referred to the Stanford Health Care Endocrinology Clinic between January 2008 and December 2018. To have adequate access to the patients' clinical records, we only considered those referred by a PCP within Stanford Health Care Alliance network, which totally includes 6511 patients (67% Female, mean Age 53.2 years, min Age 16, max Age 89). We aimed to predict the procedures (primarily laboratory and imaging tests) the endocrinologist would order at the first in-person visit. Because the procedures ordered could depend on the time window between the referral and the first specialist visit, we restricted the cohort to only those patients with a first specialist visit within 4 months after referral.

For each patient in our cohort we used electronic health record (EHR) data to extract all laboratory results within two months before the referral as well as the procedures ordered by the referring PCP. We further included the receiving specialist's identity (33 unique specialists) as side information to allow the model to personalize predictions per specialist as well, as the Endocrinology clinic directors noted that even different specialists within the same clinic had subtly different preferences for diagnostic evaluation approaches to common conditions.

3. Proposed Method

The proposed method is an ensemble model that takes the patient's clinical information and the specialist ID as input and predicts the future procedures. In order to feed the data into the model and train the base and ensemble models, we needed to pre-process the data to the appropriate format.

3.1. *Data Pre-Processing*

The defined cohort included 6511 patients and, within the defined cohort, there were 2993 unique laboratory tests, 2158 unique procedures, and 11810 unique diagnosis codes. Given that it would not be practical to train a model with several thousand data dimensions and output labels using only 6511 samples, we restricted each data category to only the most frequent types. Specifically, we only considered the top 100 most common laboratory tests and the top 60 procedures. We also restricted the diagnosis codes to the top 10 most prevalent codes related to endocrinology: *Diabetes mellitus Type I or II, Hypercalcemia, Hyperlipidemia, Hypothyroidism, Hyperthyroidism, Osteopenia, Thyroid cancer, Thyroid nodule, and Obesity.*

The raw laboratory results in the EHR data are mainly continuous data, which we converted into one-hot encoded format using the clinical laboratory defined "normal range" for each value. Thus, each laboratory value was embedded into a three dimensional binary vector, where the first dimension represents whether the laboratory value is available for the patient and the second and third dimensions indicate whether the laboratory value is low or high (in case of a normal result both are 0). Thus, if a patient has any missing clinical information, the one hot encoding approach appropriately considers it the the encoded data format. Finally, the samples were randomly shuffled and split into the train and test sets with 80% and 20% of the entire sample sizes, respectively.

3.2. *Ensemble Model*

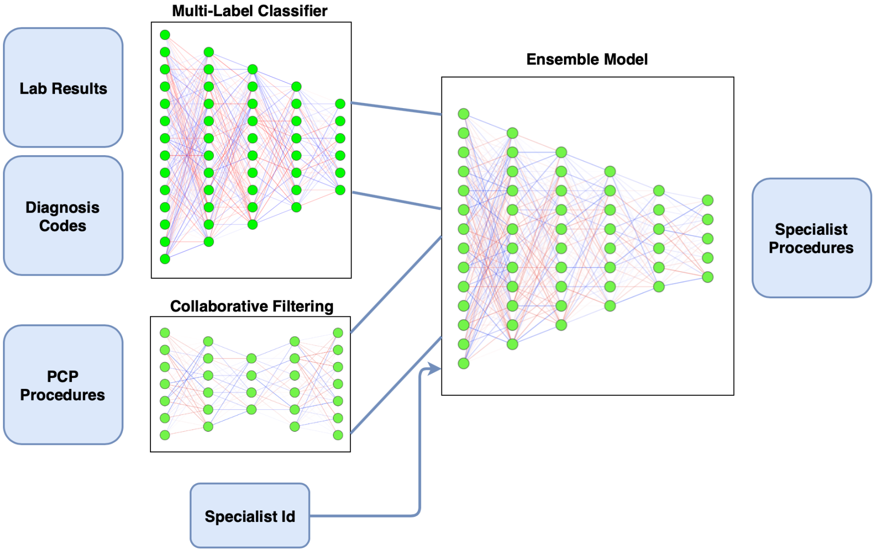


Fig. 1: The proposed model consists of two base models which were trained separately and an ensemble model that combines the prediction results from the base models using the trained neural network. The first base model is a neural network based multi-label classifier with diagnostic data as input and specialty procedures as labels. We refer to this network as diagnostic model (abbreviated as DM). The second base model is an AutoEncoder (AE) based collaborative filtering architecture with the PCP and specialty procedures as input and output. The predictions from the base models are then fed into an ensemble neural network which includes the specialist ID as side information to get the final predicted specialty procedures.

We leverage neural network based models to assess the performance potential beyond classical methods based on their flexibility in terms of model architecture and number of training coefficients. The proposed model consists of two base models which are trained separately and an ensemble model that combines the prediction results from the base models using the trained neural network (Figure 1). The first base model is a neural network based multi-label classifier with diagnostic data as input and specialty procedures as labels. A multi-class classifier has the advantage of being able to predict multiple labels (specialty procedures) with the same model and neural network is one of the most efficient methods to implement that. The neural network consists of 5 fully connected layers with the dimensions $310-200-100-80-60$ and rectified linear unit (ReLU) activations. The network is trained using stochatic gradient descent

(SGD) with a learning rate of 0.001 and mean square error (MSE) loss function. We trained the network for 400 epochs with a batch size of 256. After each layer, a dropout regularization with $p = 0.3$ is used to prevent overfitting. We refer to this network as the Diagnostic Model (DM). The second base model is an AutoEncoder (AE) based collaborative filtering architecture with the PCP and specialty procedures as input and output. The AE consists of 5 fully connected layers with dimensions $60 - 60 - 40 - 60 - 60$. The predictions from the base models are then fed into an ensemble neural network which includes the specialist ID as side information to get the final predicted specialty procedures. The ensemble neural network consists of 6 fully connected layers with the dimensions $130 - 200 - 150 - 100 - 80 - 60$ and each output neuron represents the score for a procedure ID. For all of the neural network based methods we performed several hyperparameter optimizations. The scores are normalized within the range $[0, 1]$ and could be interpreted as an uncalibrated probability that the corresponding procedure is ordered by the specialist. Based on the predicted scores for the procedures we can take two different recommendation approaches. The first method applies a fixed threshold where procedures are recommended only if their respective scores are above the threshold. This will result in variable numbers of procedures being recommended for different patients. In the second approach the algorithm always recommends the top k procedures. Thus, in this approach only the order of the scores are important not their values. We ultiamtely converged on recommendations based on a fixed score threshold that resulted in superior performance.

4. Experiment Design

The problem of predicting the specialty procedures using the laboratory test results, diagnosis codes, and PCP procedures is, in general, a multi-label classification problem and recommender system methods cannot be directly applied. However, we can split the clinical data into two major groups such that we can separately apply a multi-label classification model to the first group (laboratory test results and diagnosis codes) and a collaborative filtering model to the second group (PCP procedures), which is of the same type as the output labels (Specialty procedures). We compared the results to two standard collaborative filtering methods, i.e. singular value decomposition (SVD) and probabilistic matrix factorization (PMF). We compared the performance of the proposed ensemble method to each of the base models, i.e., the diagnostic model (DM) and AutoEncoder (AE), as well as the collaborative filtering methods SVD and PMF, and conventional clinical checklists and referral guides. Because of the long wait time for new appointments and to ensure appropriateness of consults, clinical checklists are sometimes provided to schedulers in subspecialty clinics to ensure that relevant laboratory results will be available before the initial patient visit. For example, an endocrinologist may request that a patient referred for osteoporosis not be scheduled until a bone density scan, a chemistry panel, and a vitamin D level are provided to the clinic. Alternatively, diagnosis-related checklists may represent an "order panel" of laboratory results and procedures that one would commonly order for a given medical condition to improve clinical care. Checklists represent a general list of laboratory results or procedures that an individual clinician or subspecialty clinic has decided are relevant to most patients with that diagnosis, but are often manually curated, not personalized to the individual patient, and do not update automatically as clinical evidence changes. For the clinical checklist baseline comparisons, we looked to existing clinical order sets available in the electronic medical record system designed for specific referral diagnoses and respective electronic consult referral guides that summarized initial recommended steps for common referral diagnoses. In addition, our two board-certified clinical authors (IJ and JHC) retrieved and reviewed clinical practice guideline references (e.g., UpToDate.com) to draft diagnostic workup checklists for each of the main referral diagnoses that were confirmed by consensus agreement. We further compared the results to an multi-label classifier based on aggregate neural networks (ANN) with 6 fully connected layers which utilizes all the laboratory test results, diagnosis codes, PCP procedures, and specialist ID as unified input to predict the specialist-ordered procedures. Incorporating the specialist ID as side information reflects the further real-world practice we observed when interviewing multiple specialists and clinic directors who indicated that even within specialist groups, different individual physicians sometimes still have idiosyncratic preferences for expected diagnostic workup sequences and lists for common conditions treated.

5. Results

5.1. *Specialty Referrals Efficiency*

To estimate the potential needs and impact on clinical practice and patient access, we extracted several statistics on the current utilization of (Endocrine) specialty referrals in the Stanford Healthcare system in 2017, summarized int Table 1 below.

Metric	Description
5,675	Referral Orders to Specialty
51%	Referrals Orders followed by Specialty New Patient visit within 12 months
$67 + / - 66$	Days between Referral to completed New Patient visit (avg +/- stdev)
4,796	New Patient visits (including external referrals and direct appointments)
92%	New Patient visits with only orders that could be done in advance
60%	New Patient visits with Only Diagnostic Tests
50%	New Only Diagnostic Test visits with follow-up visit within 12 months
$96 + / - 85$	Days between New Only Diagnostic Test and follow-up visit (avg +/- stdev)

Table 1: Specialty referrals metrics

Outpatient visit metrics for Endocrinology specialty visits referred from primary care visits in 2017 at Stanford Healthcare in Table 1 highlight many missed opportunities for clinical improvement. Half of referrals are lost to follow-up, never completed within 12 months, indicating large gaps in access to and completion of desired specialty care. The average waiting time for in-person consultations is over 2 months. Yet clinical orders that result from over 90 percent of specialty New Patient visits could have been done in advance (i.e., NOT specialty injections, chemotherapy, or procedures like biopsies). The majority of New Patient visits only result in diagnostic test orders, without any medications or interventions at all. Half of these New Diagnostic Test Only visits require another visit within another 12 months.

The above implies that more than 30 percent of New Patient specialty visits have *modifiable* delays where a pre-visit digital consultation guide to complete initial (diagnostic) clinical orders could at least consolidate two in-person specialty care visits into a single visit, if not eliminating the need for an in-person follow-up completely. Besides sparing the immediate patient another multi-month wait for a follow-up visit, freeing up low value clinic visit spots opens access for *all* critical patient visits. Extrapolating to the nationwide shortage of (Endocrinology) specialists, the above implies effective use and distribution of the types of tools proposed here could enable access to care for hundreds of thousands more patients every year in the US alone. This is consistent with our surveys of specialists who estimate about half of their New Patient visits do not have appropriate initial clinical workup completed.

5.2. *Model Performance*

By varying the score threshold for each of the prediction methods (to convert predicted scores into binary predictions) for each procedure order, we can obtain different performance metrics including precision (positive predictive value, the fraction of predicted procedure orders the specialist actually ordered) and recall (sensitivity, the fraction of orders the specialist actually ordered that were predicted). Therefore, the methods are evaluated in terms of precision, recall, and area under the receiver operating curve (AUROC) metrics. Figure 2 represents the precision-recall graph of the proposed ensemble method compared to the base models (diagnostic model and AutoEncoder), collaborative filtering methods (SVD and PMF), the aggregate neural network model (ANN), clinical checklists and referral guides. As shown the ensemble model achieves a better precision-recall trade-off compared to other models, the clinical checklist and referral guides. More than a quarter of specialty referrals do not clearly map to a single classic diagnosis, which means that simple static guides will not be adequately personalized and viable in these and many more cases.

Precision at different fixed values of recall are represented in Figure 3. The ensemble method achieves a better precision-recall trade-off compared to the other models. The methods are also compared in terms of AUROC in Figure 4. The ensemble method achieves the highest AUROC of 0.80 compared to the other methods.

Figure 5 shows an example of model inputs and outputs. The patient was referred for hyperthyroidism, presumably for a low thyroid stimulating hormone (TSH). The true specialist orders were all related to thyroid conditions, except for vitamin D, which was presumably repeated due to theprevious low laboratory test values. The predicted specialist orders with a predicted score above a fixed threshold of 0.20 captured a portion of the thyroid-related laboratory tests as well as the repeat vitamin D, a metabolic panel (appropriate for a general review of the patient's body chemistry context) and a hemoglobin A1c (not as relevant for the primary thyroid disorder, but likely reflects the Endocrinologists attention to other common disorders like diabetes). Finally, we compare the performance of the ensemble method using two selection approaches based on the predicted scores (discussed in Section 3.2). As shown in Figure 6, the selection method based on a fixed threshold (η) performs better than the selection method based on the fixed k.

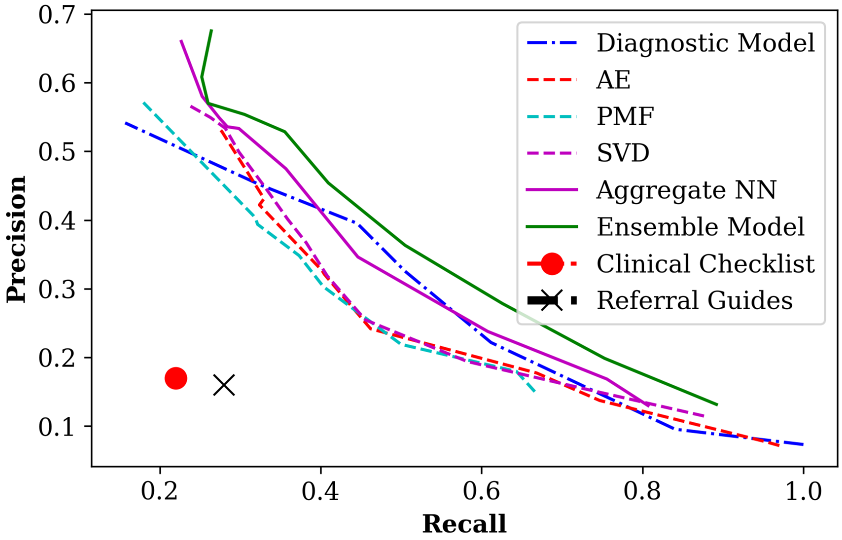
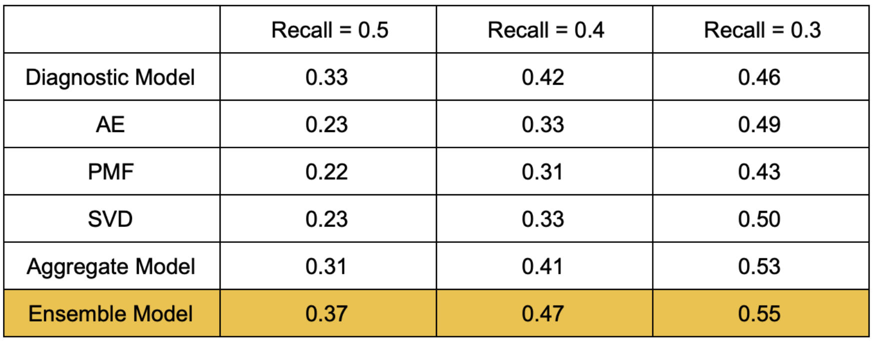


Fig. 2: Precision-Recall graph of the proposed ensemble method compared to the base models (diagnostic model and AutoEncoder), collaborative filtering methods (SVD and PMF), and the aggregate neural network model (ANN). Notably, all methods substantially outperform credible real-world standard of care benchmarks of static clinical checklists and referral guides.



	Recall = 0.5	Recall = 0.4	Recall = 0.3
Diagnostic Model	0.33	0.42	0.46
AE	0.23	0.33	0.49
PMF	0.22	0.31	0.43
SVD	0.23	0.33	0.50
Aggregate Model	0.31	0.41	0.53
Ensemble Model	0.37	0.47	0.55

Fig. 3: Precision at fixed recall for the proposed ensemble method compared to the base models (diagnostic model and AE), collaborative filtering methods (SVD and PMF), the aggregate neural network model (ANN). The ensemble model achieves a better precision-recall trade-off compared to other models and the clinical checklist. Again, the ensemble model achieves a better precision-recall trade-off compared to other models.

6. Discussion

The generalizability of the proposed model to more diverse types of patients with different conditions depends on several key assumptions. As mentioned in 3.1, due to the model's learning limitations with respect to the number of patients, we only included a portion of the laboratory tests, diagnosis codes, and procedures as features and labels in our data, which degrades the performance of the recommendation model. Further, the recommended items

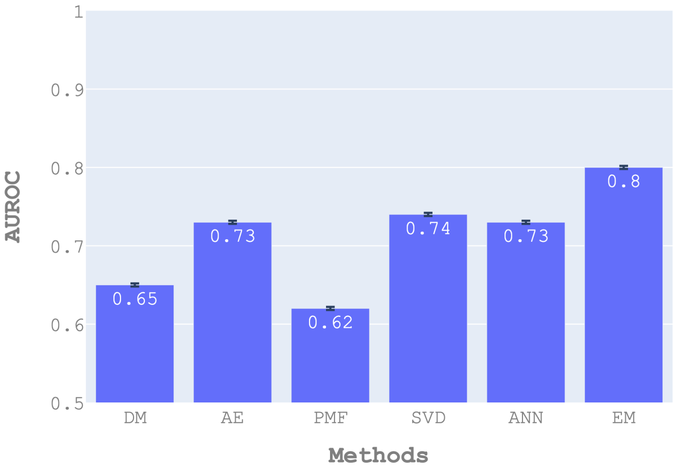
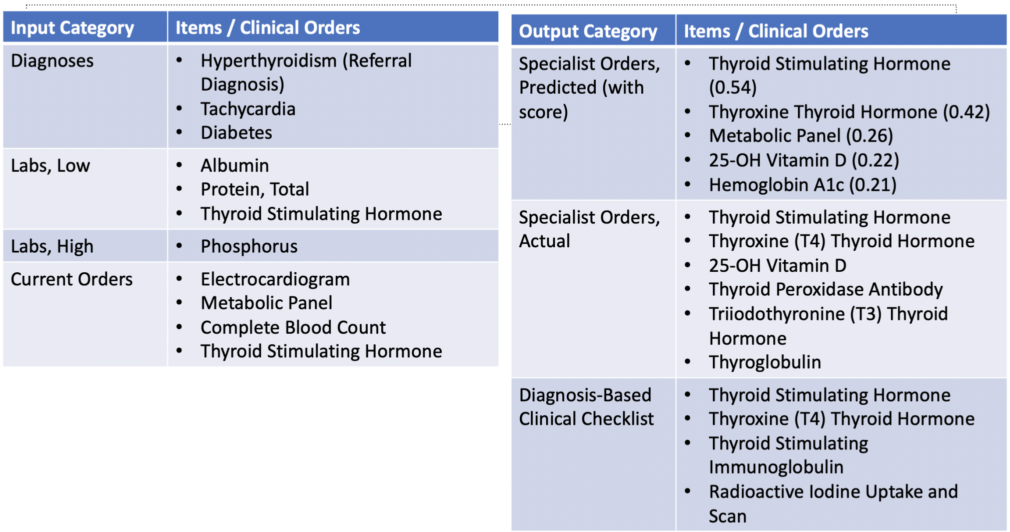


Fig. 4: AUROC of the proposed ensemble method (EM) compared to the base models (diagnostic model and AE), collaborative filtering methods (SVD and PMF), the aggregate neural network model (ANN). Error bars show the 95% confidence interval computed using bootstrapped resampling.



Input Category	Items / Clinical Orders
Diagnoses	• Hyperthyroidism (Referral Diagnosis) • Tachycardia • Diabetes
Labs, Low	• Albumin • Protein, Total • Thyroid Stimulating Hormone
Labs, High	• Phosphorus
Current Orders	• Electrocardiogram • Metabolic Panel • Complete Blood Count • Thyroid Stimulating Hormone

Output Category	Items / Clinical Orders
Specialist Orders, Predicted (with score)	• Thyroid Stimulating Hormone (0.54) • Thyroxine Thyroid Hormone (0.42) • Metabolic Panel (0.26) • 25-OH Vitamin D (0.22) • Hemoglobin A1c (0.21)
Specialist Orders, Actual	• Thyroid Stimulating Hormone • Thyroxine (T4) Thyroid Hormone • 25-OH Vitamin D • Thyroid Peroxidase Antibody • Triiodothyronine (T3) Thyroid Hormone • Thyroglobulin
Diagnosis-Based Clinical Checklist	• Thyroid Stimulating Hormone • Thyroxine (T4) Thyroid Hormone • Thyroid Stimulating Immunoglobulin • Radioactive Iodine Uptake and Scan

Fig. 5: Example Inputs and Outputs for referral clinical order predictions. Adapted from a real-world patient's information available at the time of referral and the predicted specialist's clinical orders vs. actual specialist orders vs. simple checklists (order sets) based only on the primary referral diagnosis. In this example, the ensemble model predicts 5 candidate clinical orders that exceed a score threshold, of which 2 were actually ordered by the specialist who ordered 6 total items. (Precision = 2/5 = 40%, Recall = 2/6 = 33%).

based on the prediction model are learned based on specialists' preferences; they are not necessarily correct or incorrect orders. Gold standards for "correctness" are largely elusive

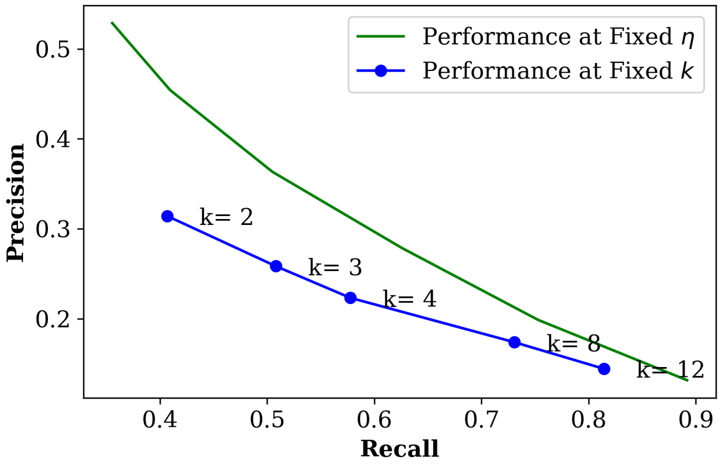


Fig. 6: Precision-Recall performance comparison of the ensemble method using different threshold approaches to selecting the top predictions. A fixed number (k) can used to select the top 5, or top 10, or other top k items to predict, or a score threshold to select all top items with a predicted score above that threshold. The curves illustrate the natural tradeoff between precision vs. recall with varying threshold values, and moreso, indicates that score-based thresholding will consistently outperform fixed count selections.

for medical care, but our prior studies have already assessed this in different ways such as alignment with clinical practice guidelines,[9] risk-adjusted patient outcome scores,[?] and expert panel consensus of clinical appropriateness,[15] where we largely find that the "normative" clinical behavior we predict already naturally aligns with such external measures of correctness.

7. Conclusion

In this work, we addressed the problem of predicting outpatient specialty diagnostic workups, specifically the orders expected to result from adult Endocrinology referrals. We proposed a data-driven model that recommends follow-up procedure orders based on patients' clinical information. Several evaluations illustrate that the proposed method can outperform conventional clinical checklist and baseline methods.

References

1. Julia C Prentice and Steven D Pizer. Delayed access to health care and mortality. *Health services research*, 42(2):644–662, 2007.
2. B Middleton, DF Sittig, and A Wright. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearbook of medical informatics*, 25(S 01):S103–S116, 2016.
3. Maryellen L Giger. Machine learning in medical imaging. *Journal of the American College of Radiology*, 15(3):512–520, 2018.
4. Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.
5. Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised repre-

sentation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

6. Patrick Kenny, Thomas Parsons, Jonathan Gratch, and Albert Rizzo. Virtual humans for assisted health care. In *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*, pages 1–4, 2008.
7. Selin Merdan, Khurshid Ghani, and Brian Denton. Integrating machine learning and optimization methods for imaging of patients with prostate cancer. In *Machine Learning for Healthcare Conference*, pages 119–136, 2018.
8. SK Lakshmanaprabu, Sachi Nandan Mohanty, Sujatha Krishnamoorthy, J Uthayakumar, K Shankar, et al. Online clinical decision support system using optimal deep neural networks. *Applied Soft Computing*, 81:105487, 2019.
9. Jonathan H Chen, Mary K Goldstein, Steven M Asch, Lester Mackey, and Russ B Altman. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *Journal of the American Medical Informatics Association*, 24(3):472–480, 2017.
10. Maytiyanin Komkhao and Wolfgang A Halang. Recommender systems in telemedicine. *IFAC Proceedings Volumes*, 46(28):28–33, 2013.
11. Fang Hao and Rachael Hageman Blair. A comparative study: classification vs. user-based collaborative filtering for clinical prediction. *BMC medical research methodology*, 16(1):172, 2016.
12. Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.
13. Oleksii Kuchaiev and Boris Ginsburg. Training deep autoencoders for collaborative filtering. *arXiv preprint arXiv:1708.01715*, 2017.
14. Mohammad Moghimi, Serge J Belongie, Mohammad J Saberian, Jian Yang, Nuno Vasconcelos, and Li-Jia Li. Boosted convolutional neural networks. In *BMVC*, volume 5, page 6, 2016.
15. Andre Kumar, Rachael C Aikens, Jason Hom, Lisa Shieh, Jonathan Chiang, David Morales, Divya Saini, Mark Musen, Michael Baiocchi, Russ Altman, et al. Orderrex clinical user testing: a randomized trial of recommender system decision support on simulated cases. *Journal of the American Medical Informatics Association*, 27(12):1850–1859, 2020.