Singh Karandeep (Orcid ID: 0000-0001-8980-2330)
Krumme Alexis (Orcid ID: 0000-0001-9633-861X)
Franklin Jessica M. (Orcid ID: 0000-0002-8941-4116)

A Concept-Wide Association Study to Identify Potential Risk Factors for Non-Adherence Among Prevalent Users of Anti-Hypertensives

Karandeep Singh, MD, MMSc[1]; Niteesh K. Choudhry, MD, PhD[2,3]; Alexis A. Krumme, MS, ScD[2]; Caroline McKay, PhD[4]; Newell E. McElwee, PharmD, MSPH[5]; Joe Kimura, MD[6]; Jessica M. Franklin, PhD[2]

Authors' affiliations:

[1] Departments of Learning Health Sciences and Internal Medicine, University of Michigan Medical School and University of Michigan School of Information, Ann Arbor, MI

[2] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

[3] Center for Healthcare Delivery Sciences, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

[4] Janssen Scientific Affairs, LLC, Horsham, PA

[5] Boehringer Ingelheim, Ridgefield, CT

[6] Atrius Health and Harvard Medical School, Boston, MA

Corresponding Author: Karandeep Singh, MD, MMSc

Address: 1161H NIB, 300 N. Ingalls St., Ann Arbor, MI 48109

Fax: 734-647-3914, Phone: 734-936-1649, E-mail: kdpsingh@umich.edu

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pds.4850

Word count: 3,573

Key words: non-adherence, hypertension, medications, electronic health record

Abstract

Purpose: We sought to determine whether an association study using information contained in clinical notes could identify known and potentially novel risk factors for non-adherence to anti-hypertensive medications.

Methods: We conducted a retrospective concept-wide association study (CWAS) using clinical notes to identify potential risk factors for medication non-adherence, adjusting for age, sex, race, baseline blood pressure, estimated glomerular filtration rate, and a combined comorbidity score. Participants included Medicare beneficiaries 65 years and older receiving care at the Harvard Vanguard Medical Associates network from 2010-2012 and enrolled in a Medicare Advantage program. Concepts were extracted from clinical notes in the year prior to the index prescription date for each patient. We tested associations with the outcome for 5,013 concepts extracted from clinical notes in a derivation cohort (4,382 patients) and accounted for multiple hypothesis testing by using a false discovery rate threshold of less than 5% ($q < 0.05$). We then confirmed the associations in a validation cohort (3,836 patients). Medication non-adherence was defined using a proportion of days covered (PDC) threshold < 0.8 using pharmacy claims data.

Results: We found 415 concepts associated with non-adherence, which we organized into 11 clusters using a hierarchical clustering approach. Volume depletion and overload, assessment of needs at the point of discharge, mood disorders, neurological disorders, complex coordination of care, and documentation of non-compliance were some of the factors associated with non-adherence.

Conclusions: This approach was successful in identifying previously described and potentially new risk factors for anti-hypertensive non-adherence using the clinical narrative.

Introduction

Adherence, defined as the extent to which patients take medications as prescribed,[1] is often poor for anti-hypertensive medications[2] and is associated with worse health outcomes, including increased all-cause mortality,[3] cardiovascular mortality,[4,5] and stroke severity.[6] A U.S. survey of hypertensive adults found that 30.5% of respondents self-identified as being non-adherent to hypertensive medications.[7] Other studies[8,9] have reported the prevalence of non-adherence between 16% and 52%, though the rates depend on how non-adherence is measured.[10]

Adherence is also a complex health behavior, and understanding the reasons why people do not consistently take prescribed medications has been a topic of intense study. Many systematic reviews have been conducted and have identified several factors that account for non-adherence[11–15]. In one of these reviews, Krueger and colleagues[13] grouped the reasons contributing to adherence behaviors into five categories: patient demographic-related factors (*e.g.,* low literacy), psychological and behavioral characteristics, treatment plan issues, disease-related issues (*e.g.,* severity of illness), and healthcare system issues (*e.g.,* relationship with provider, access to care). Psychological and behavioral characteristics linked to non-adherence include a belief that a medication is unimportant or harmful, depression, impaired cognitive function, forgetfulness, anger, stress, anxiety and substance abuse. Treatment plan issues include the experience or fear of side effects, high price, inconvenience, and polypharmacy.

The existing body of literature has two limitations. First, the primary studies linking these factors to non-adherence rely on traditional epidemiological approaches, where a handful of possible exposures are linked to the outcome of non-adherence. Several other factors may be

associated with non-adherence that have never been identified simply because they have not been studied. Second, the body of literature does not provide a way to determine which of these characteristics impact any given patient in a healthcare system without directly attempting to elicit this information from patients. For instance, a patient's prior experience with side effects may contribute to subsequent non-adherence, but determining which patients have experienced side effects is difficult on a large scale.

Though some health exposures that contribute to medication non-adherence may be found in administrative health data such as health insurance claims, many of the contributors are more complex and likely to be found only in the narrative of clinical notes and phone calls. Notes may provide a much richer picture, including symptoms, social issues, and life circumstances that could lead to problems with adherence. A comprehensive review of the electronic health record (EHR) is not feasible on a large scale, but automated extraction of concepts from clinical notes using natural language processing (NLP) software make such an endeavor feasible.

In this paper, we use a concept-wide association study[16] (CWAS) to identify potential risk factors for medication non-adherence. We build off of prior work in which we demonstrated that data from the electronic health record can provide good predictions of future adherence.[17] Unlike traditional epidemiologic studies that typically test a single association, CWAS enables the discovery of new associations through a paradigm of simultaneous testing of multiple associations using pre-specified covariates. This paradigm was first established in the conduct of genome-wide association studies (GWAS) but has been used to discover associations between diseases and environmental exposures in environment-wide association studies

(EWAS)[18,19] and between a single genetic variant and multiple phenotypes using phenome-wide associations studies (PheWAS)[20]. A concept-wide association study (CWAS) is useful to establish relationships between concepts documented in clinical documentation and health outcomes.

Methods

Data Source

Our data was drawn from a population of individuals 65 years and older enrolled in a Medicare Advantage program and receiving care at one of 24 practice sites of Harvard Vanguard Medical Associates (HVMA), a large multi-specialty community-based medical group in eastern and central Massachusetts, between January 2010–December 2012. For these individuals, we linked Medicare claims with structured and unstructured data in the EHR.

From claims data, we extracted data on demographics, all medical encounters, and diagnostic billing codes. In addition, we used medication refill data from pharmacy claims to measure adherence. The HVMA EHR data includes both structured fields, such as demographic characteristics and diagnostic billing codes, and unstructured fields, such as text from progress notes, electronic patient instructions, and patient letters, and telephone encounters (considered collectively as "clinical notes"). Because this network covers primary care, physician specialists, and laboratory testing, the EHR data cover nearly all of the patient's outpatient care encounters and procedures, and contain longitudinal anthropometric and biomarker data.

Prior to receipt and analysis by the research team, clinical notes were de-identified using several steps. Known identifiers such as patient and provider names were searched in the clinical notes and replaced with a series of random letters. All numbers were replaced with "1"s.

The remaining text was de-identified using the MITRE Identification Scrubber Toolkit (MIST), a machine-learning based de-identification tool with a demonstrated F-measure in excess of 0.9.[21]

The Partners Healthcare institutional review board approved this study, and the need for informed consent was waived.

Study Design

We conducted a retrospective cohort study to identify concepts in clinical notes that are associated with anti-hypertensive medication non-adherence. We then grouped the top concepts into clusters based on inter-concept similarity.

We included outpatient adults ≥ 65 years old who filled at least one prescription for an anti-hypertensive during 2011-2012 and were prevalent users of an anti-hypertensive medication. The first fill during this period prior to which patients also had a year of continuous insurance eligibility was considered the index fill. Prevalent users were defined as individuals who had at least one anti-hypertensive filled during the 365 days preceding the index date. We limited our analysis to prevalent users because the rates of non-adherence and factors influencing non-adherence are known to differ among new and prevalent users.[22] Patients were excluded if they had fewer than 112 days of follow-up after the index fill, or if they had fewer than 5 notes in the year preceding the index date. We set a threshold of 5 notes based on prior experience.[16] Patients with fewer than 5 notes are unlikely to have enough information in the EHR to allow accurate judgments about their exposures.

Outcome

We defined the outcome of medication non-adherence based on the proportion of days covered (PDC) in the year following the index fill. PDC is defined as the proportion of days that

the patient had medication available to him, based on a supply diary which strings together adjacent fills using the dates and days' supply of each pharmacy claim for an anti-hypertensive. We set a threshold of PDC < 0.8 to define non-adherence based on prior research linking this threshold to improved cardiac outcomes and use in other quality measures.[23,24] Since patients could fill more than one anti-hypertensive, we considered each anti-hypertensive class that the patient filled separately and averaged the PDC across classes, which is a valid way of measuring medication adherence.[25]

Identification of Concepts in Clinical Notes

Concepts rather than individual words were extracted from clinical notes so that phrases representing the same idea could be grouped together when evaluating their association with the outcome (*e.g.,* "CHF" and "congestive heart failure"). Clinical notes consisted of all unstructured text available in the notes section of the EHR, including notes and phone calls from physicians of all specialties, nurses, case managers, and other care providers. The HVMA EHR is primarily an outpatient record and does not contain admission and inpatient notes. Concepts were coded as binary variables for each patient. Concepts were considered present if they were documented at least once in the clinical notes during the baseline period, and otherwise were considered absent.

Concepts were extracted from clinical notes in the year prior to the index prescription date for each patient (Figure 1) using the National Library of Medicine's MetaMap software[26] (2014 version), which maps phrases to Unified Medical Language System (UMLS) codes known as concept unique identifiers (CUIs). Negated concepts were detected and removed using NegEx algorithm.[27] Extracted concepts were restricted to the Systematized Nomenclature of

Medicine – Clinical Terms (SNOMED-CT) ontology to limit mappings to clinically relevant concepts. Concepts were not limited by semantic type, so all types of concepts contained were extracted, including diagnoses, medications, signs and symptoms, exposures, geographical locations, and actions taken by a physician or patient. Mapping of phrases to multiple concepts was allowed. Concepts with less than 1% patient prevalence were not included in the analysis.

NLP systems may occasionally create erroneous mappings (*e.g.,* the word "Hi" in a phone call note maps to the concept for "Hawaii"). Instead of reporting the name of the concept intended by the UMLS definitions, we report the most common phrase for each concept (i.e., we report "Hi" and not "Hawaii").

Identification of Covariates

We selected covariates that may confound the relationship between the exposures and non-adherence, recognizing that these factors may in fact differ between the tested exposures. We defined baseline systolic and diastolic blood pressure and creatinine values for each lab test as the most recent result on or before the index date, and estimated glomerular filtration rate (eGFR) was computed using the CKD-EPI formula.[28] The combined comorbidity score was computed using claims data in the year prior to the index date.[29] We adjusted for comorbidity because we were concerned that sicker individuals may have lower rates of adherence due to inability to fill prescription. We selected the combined comorbidity score because it combines conditions from the Charlson and Elixhauser measures and was shown to have similar or slightly better performance in predicting mortality as compared to either of the individual comorbidity measures in a Medicare population similar to the population in our study.[29]

Derivation and Validation Cohorts

The overall cohort from which our derivation and validation cohorts were drawn consists of 24 primary care practices. We were concerned that practice patterns and non-adherence may differ between practices of different sizes. Thus, we stratified the assignment of patients to the derivation and validation cohorts at the practice level. Specifically, we ordered the practices based on their number of patients in our dataset. We then assigned practices to the derivation and validation set in alternating order.

Statistical Analysis

In the derivation cohort, multivariate logistic regression was performed to test the association for each of the concepts with medication non-adherence, adjusting for age, sex, black or Hispanic race, baseline systolic and diastolic blood pressure, estimated glomerular filtration rate, and a combined comorbidity score.

We accounted for multiple hypothesis testing using Storey's method, which controls the false discovery rate, defined as the expected proportion of false positives among all significant hypotheses.[30,31] Using this method, p-values were transformed into q-values. Odds ratios and confidence intervals were not adjusted in any way. Concepts with q-values < 0.05 were reported as potential associations; this equates to a 5% expected proportion of false positives among all concepts declared to have associations. The false discovery rate method was chosen because of its many desirable properties [30]: it explicitly controls the error rate of test conclusions among significant results, scales well in the face of increasing numbers of tests, and has increased power as compared to the Bonferroni method. After identifying potential associations in the derivation cohort, these were considered confirmed if the p-value in the validation cohort was < 0.05 and the effect was in the same direction.

We clustered confirmed predictors into groups using several steps. First, we calculated pairwise Phi correlation coefficients to measure similarity between concepts and then converted this to a distance measure by subtracting from 1. Using this distance measure, we ran a hierarchical clustering algorithm with aggregation using complete linkage. Any number of clusters can be derived from the result of this algorithm (up to the number of observations) by "cutting" the hierarchical cluster tree at varying depths. We measured several cluster stability measures (Silhouette, point biserial correlation, Calinski-Harabasz, Davies-Bouldin, Ray-Turi, Dunn) for all possible clustering solutions between 5 and 20 clusters.[32] We selected an optimal number of clusters based on the cluster stability measures. We reviewed 10 randomly selected sentences for each concept and qualitatively assigned labels to each cluster.

Analyses were performed in R 3.3.2 (Vienna, Austria). Q-values were computed using Storey's *qvalue* R package (available on Bioconductor).[33] Hierarchical clustering was performed using the *stats* package and cluster stability metrics were computed using the *clusterCrit* package.

Results

We identified 8,218 patients who met the inclusion and exclusion criteria (Table 1, Figure 2) from 24 primary care practices, of whom 2,088 (25.4%) were non-adherent to anti-hypertensives (PDC < 0.8). We processed 770,353 notes and extracted 32,693 non-negated concepts from clinical notes. We removed duplicate phrases and considered only the 5,031 concepts with a prevalence of ≥ 1%. The median follow-up period during which we assessed the proportion of days covered was 360 days. We assigned 4,382 patients to the derivation cohort and 3,836 to the validation cohort based on their assigned primary care practice.

Using a false discovery rate threshold of less than 5% ($q < 0.05$), we identified 594 concepts significantly associated with adherence in the derivation cohort, 583 with non-adherence (OR > 1) and 11 with favorable adherence (OR < 1). Of these, 415 concepts had confirmed associations in the validation cohort based on p-value < 0.05 and concordant odds ratios in the two cohorts. All validated concepts were associated with non-adherence (OR > 1). Based on several cluster quality measures (Supplementary Table 1), we grouped the confirmed associations into 11 clusters (Table 2). The odds ratios, confidence intervals, and q-values for individual concepts are provided in Supplementary Table 2. Ten randomly selected sentences from which the cluster descriptions were derived are provided in Supplementary Appendix 1.

Cluster 1 includes concepts related to volume depletion (*e.g.,* hypotension, IV fluids, dehydration, dehydration, tachycardia, hydration, lightheadedness, dry) and volume overload (*e.g.,* 1 pitting edema, pedal edema, lower extremity edema, BNP [beta natriuretic peptide], low salt diet, Diuretic, CXR [chest x-ray], weight gain). Additional concepts in Cluster 1 relate to the evaluation of diagnoses that may mimic volume depletion (*e.g.,* WBC [white blood cell count] and urine culture to work-up infection) or volume overload (*e.g.,* DVT [deep venous thrombosis]). Cluster 2 broadly relates to case management (*e.g.,* Case Manager, Case Management, nurse case manager), assessment of needs at discharge (*e.g.,* mobility in home, skilled nursing facility, medication teaching, assistive device, walker, discharged home, plan of care, HOME ASSESSMENT, rehab, commode, home exercise program, hospice), mood disorders (*e.g.,* mental illness, Anxiety/depression, SSRI [selective serotonin reuptake inhibitor]), and social determinants of health (*e.g.,* Lives with spouse, family support, Social support, upset, afraid, compliance). Cluster 3 captures neurological disorders. Cluster 4

includes items related to coordination of care between the physician's office and the patient (*e.g.,* "back" refers to voicemails left for patients asking them to call back, "Hi" refers to greetings in messages between care providers, "FW" refers to messages forwarded between care providers, "pls" is used as shorthand for "please" in the commonly used phrase "pls call patient", "letter" refers to letters written to patients, "Pool" refers to the pool of staff who answer phone calls for patients, "adv" is shorthand for patients being "advised") and documentation of a variety of symptoms. Cluster 5 is focused on place of residence and related needs (*e.g.,* "hospital bed" in the context of use at home, nursing home, wheelchair, "assisted" used in the context of assisted living), refractoriness to treatment (*e.g.,* refractory, "unresponsive" used in the context of unresponsiveness to treatment), and noncompliance. Cluster 6 refers to management of cardiac arrhythmias with warfarin (*e.g.,* "spontaneous" used in the context of spontaneous development of palpitations/arrhythmias, "prothrombin time" linked to use of warfarin, "accident" includes a note to patients taking warfarin on what to do if they are in a car accident, "remind" includes reminders to patients about the risks of warfarin and need for close monitoring). Cluster 7 refers to patients offered or enrolled in either the Asthma Management Program or COPD Management Program (*e.g.,* "pulmonology" refers to upcoming appointment with pulmonologist, "trained" includes references to "specially trained nurses" that comes from an invitation letter for the Asthma/COPD Management Program, "expiratory" refers to increased expiratory time). Cluster 8 includes need for a translator due to a language barrier and allergies to medications. Cluster 9 refers to patients taking "blood pressure medicine." Cluster 10 relates to counting of cells per high power field (HPF) on urinalysis as well as measurement of ketones. Cluster 11 refers to language taken from plantar fasciitis patient instructions.

The 11 clusters we identified include several factors previously identified by the literature. Our analysis found supporting evidence for 8 of the 12 risk factors for medication non-adherence described in a widely cited review article[1] by Osterberg and Blaschke (Table 3). Krueger and colleagues[13] performed a systematic review that classified factors affecting adherence into 6 categories: patient demographics, family/cultural issues, psychosocial and behavioral characteristics, treatment plan issues, disease-related issues, and healthcare system issues. We found supportive evidence for 5 of these: family/cultural issues (cluster 2), psychological and behavioral factors (cluster 2), treatment plan issues (clusters 5-7 and 9), disease-related issues (cluster 1), and healthcare system-related issues (cluster 2 and 4). We did not find any clusters related to patient demographics but this is not surprising as we adjusted for demographic characteristics in our analysis.

Discussion

This is the first study to use a multiple hypothesis-testing approach utilizing text from the clinical notes to identify potential risk factors for medication non-adherence. Our findings mostly confirm existing knowledge on medication adherence. Our results are important for two reasons. First, existing knowledge has been drawn from multiple clinical studies using carefully assessed exposures and outcomes. That we were able to partially replicate the findings from published literature in a single retrospective cohort study using clinical notes and multiple hypothesis testing is promising because this approach may be useful for evaluating other clinical questions where the published literature is not as rich. Second, the concepts we identified can be directly used in identifying patients at risk for non-adherence. While Osterberg and Blaschke's review identified the presence of psychological problems as a risk factor for non-adherence, our study

provides a mechanism for identifying such patients. Searching the notes for the phrases "Anxiety/depression," "mental illness," "SSRI," and "mood" may be effective ways of identifying such patients. Linking these phrases to clinical decision support would provide a means to flag high-risk patients for targeted interventions.

We also identified some surprising associations. We found that the phrase "Medicare" is linked to non-adherence. While the phrase would seem to imply that a patient is covered by Medicare, a review of the sample sentences (Supplementary Appendix 1) reveals that the concept is often mentioned in the context of a Medicare notice of non-coverage, which is delivered to patients as they near completion of physical rehabilitation. We found the phrase "He" to be linked to be linked to non-adherence, which appears to be used in the clinical documentation to describe male patients (Supplementary Appendix 1). Since we adjusted for sex in our analysis, its significance is likely explained by other contexts in which it was used beyond what we found in our review of sample sentences, as it was found in over 80% of patients. Krueger and colleagues' systematic review found male providers to be linked to favorable adherence, so the use of "he" to describe male providers would not appear to explain this association either. One other surprising finding was that the phrases "Lives with spouse," "Social support," and "family member" were all associated with non-adherence, contrary to previous literature identified by Krueger and colleagues. It is possible that clinicians are more likely to assess and document a patient's living situation if they deem the patient at greater risk for non-adherence. Lastly, the relationship between clusters 9-11 and non-adherence is not clear upon review of the randomly sampled sentences (Supplementary Appendix 1).

The rising adoption of electronic health records and more recent development of data-sharing research networks creates an opportunity to systematically discover predictors of health outcomes from electronic health records. We believe that this approach may be useful especially for the study of rare and understudied health outcomes and behaviors, where a systematic review of the literature may not be as fruitful.

Limitations

As our approach is intended to be hypothesis-generating, caution is needed when interpreting the results because the concepts identified as predictors may not be used consistently in notes despite evaluation of example sentences, may represent erroneous mapping by NLP software, may be confounded, or may represent false positive results (due to 5% false discovery rate). When concepts have multiple meanings or contexts in the notes, we cannot be certain which of these is responsible for the overall association with medication non-adherence, and this may introduce error and limit the interpretability of the analysis. Confounding and exposure misclassification can be particularly difficult to identify in this type of analysis. For example, the association between “COLACE” (cluster 2) and non-adherence may be confounded by “Oxycodone” (also in cluster 2) as stool softeners are commonly prescribed for prevention of opioid-induced constipation. This study draws from a relatively geographically and demographically homogenous population with a low fraction of non-white patients, so its findings may not generalize to other populations. Additionally, our study was limited by residual confounding. The causal mechanisms are likely to differ among the multiple exposures tested in this study and thus the covariates included in our analysis may not fully account for potentially confounders across the breadth of tested exposures.

Conclusions

This approach was successful in identifying previously described and potentially new predictors of anti-hypertensive non-adherence using the clinical narrative as a by-product of routine care delivery.

Acknowledgements: We are grateful to Angela Tong for her assistance with programming and data management.

Study concept and design: Singh, Franklin, Choudhry, McKay, McElwee, Kimura

Acquisition of data: Kimura, Krumme, Franklin, Choudhry

Analysis and interpretation of data: Singh, Franklin, Choudhry, McKay

Drafting of the manuscript: Singh, Franklin

Critical revision of the manuscript for important intellectual content: Singh, Krumme, Franklin, Choudhry, McKay, McElwee, Kimura

Administrative, technical, or material support: Tong

Study supervision: Franklin

Sources of Funding: The research was supported by funding from Merck and by grant 5K12DK111011 from the National Institute of Diabetes and Digestive and Kidney Diseases on which Karandeep Singh is an appointee.

Disclosures: NKC reports grants from CVS Caremark during the conduct of the study and from Sanofi; AstraZeneca; Medisafe; the National Heart, Lung, and Blood Institute; Merck; and Pharmaceutical Research and Manufacturers of America. NKC and JMF are co-principal investigators with funding from Merck on this study. CM is a former employee of Merck and owns stock in the company. CM is a current employee of Janssen, a relationship that began after the study ended. NEM is a former employee of Pfizer and Merck and owns stock in both companies. NEM is a current employee of Boehringer Ingelheim, a relationship that began after this study ended. JK is an employee of Atrius Health, where the study was conducted. KS and AAK have no relevant disclosures to report.

References

1. Osterberg L, Blaschke T. Adherence to Medication. *N Engl J Med*. 2005;353(5):487-497. doi:10.1056/NEJMra050100
2. Cohen DL, Townsend RR. Medication Non-Adherence: A Bigger Problem Than Physicians Assume. *J Clin Hypertens (Greenwich)*. May 2016. doi:10.1111/jch.12838
3. Molnar MZ, Gosmanova EO, Sumida K, et al. Predialysis Cardiovascular Disease Medication Adherence and Mortality After Transition to Dialysis. *Am J Kidney Dis*. April 2016. doi:10.1053/j.ajkd.2016.02.051
4. Herttua K, Martikainen P, Batty GD, Kivimäki M. Poor Adherence to Statin and Antihypertensive Therapies as Risk Factors for Fatal Stroke. *J Am Coll Cardiol*. 2016;67(13):1507-1515. doi:10.1016/j.jacc.2016.01.044
5. Kim S, Shin DW, Yun JM, et al. Medication Adherence and the Risk of Cardiovascular Mortality and Hospitalization Among Patients With Newly Prescribed Antihypertensive Medications. *Hypertension*. 2016;67(3):506-512. doi:10.1161/HYPERTENSIONAHA.115.06731
6. Lee KB, Lee J-Y, Choi N, et al. Association between insufficient medication of antihypertensives and the severity of acute ischemic stroke. *Clin Hypertens*. 2015;22:11. doi:10.1186/s40885-016-0047-8
7. Tong X, Chu EK, Fang J, Wall HK, Ayala C. Nonadherence to Antihypertensive Medication Among Hypertensive Adults in the United States—HealthStyles, 2010. *J Clin Hypertens (Greenwich)*. February 2016. doi:10.1111/jch.12786
8. Schmieder RE, Ott C, Schmid A, et al. Adherence to Antihypertensive Medication in Treatment-Resistant Hypertension Undergoing Renal Denervation. *J Am Heart Assoc*. 2016;5(2). doi:10.1161/JAHA.115.002343
9. Van Wijk BL, Klungel OH, Heerdink ER, de Boer A. Rate and determinants of 10-year persistence with antihypertensive drugs. *J Hypertens*. 2005;23(11):2101-2107.
10. Maeng DD, Snyder RC, Medico CJ, Mold WM, Maneval JE. Unused medications and disposal patterns at home: Findings from a Medicare patient survey and claims data. *J Am Pharm Assoc (2003)*. 2016;56(1):41-46.e6. doi:10.1016/j.japh.2015.11.006
11. Gellad WF, Grenard J, McGlynn EA. *A Review of Barriers to Medication Adherence: A Framework for Driving Policy Options*.; 2009.
12. Gellad WF, Grenard JL, Marcum ZA. A systematic review of barriers to medication adherence in the elderly: looking beyond cost and regimen complexity. *Am J Geriatr Pharmacother*. 2011;9(1):11-23. doi:10.1016/j.amjopharm.2011.02.004
13. Krueger KP, Berger BA, Felkey B. Medication adherence and persistence: a comprehensive review. *Adv Ther*. 22(4):313-356.
14. Bowry ADK, Shrank WH, Lee JL, Stedman M, Choudhry NK. A Systematic Review of Adherence to Cardiovascular Medications in Resource-Limited Settings. *J Gen Intern*

Med. 2011;26(12):1479-1491. doi:10.1007/s11606-011-1825-3

15. Marshall IJ, Wolfe CDA, McKevitt C. Lay perspectives on hypertension and drug adherence: systematic review of qualitative research. *BMJ*. 2012;345.
16. Singh K, Betensky RA, Wright A, Curhan GC, Bates DW, Waikar SS. A Concept-Wide Association Study of Clinical Notes to Discover New Predictors of Kidney Failure. *Clin J Am Soc Nephrol*. 2016;11(12):2150-2158. doi:10.2215/CJN.02420316
17. Franklin JM, Gopalakrishnan C, Krumme AA, et al. The relative benefits of claims and electronic health record data for predicting medication adherence trajectory. *Am Heart J*. November . doi:10.1016/j.ahj.2017.09.019
18. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS One*. 2010;5(5). doi:10.1371/journal.pone.0010746
19. Patel CJ, Ioannidis JPA. Studying the elusive environment in large scale. *JAMA*. 2014;311(21):2173-2174. doi:10.1001/jama.2014.4129
20. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31(12):1102-1111. doi:10.1038/nbt.2749
21. Deleger L, Molnar K, Savova G, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *J Am Med Inform Assoc*. 2013;20(1):84-94. doi:10.1136/amiajnl-2012-001012
22. Blackburn DF, Swidrovich J, Lemstra M. Non-adherence in type 2 diabetes: practical considerations for interpreting the literature. *Patient Prefer Adherence*. 2013;7:183-189. doi:10.2147/PPA.S30613
23. Choudhry NK, Glynn RJ, Avorn J, et al. Untangling the relationship between medication adherence and post–myocardial infarction outcomes. *Am Heart J*. 2014;167(1):51-58.e5. doi:10.1016/j.ahj.2013.09.014
24. Trends in Part C & D Star Rating Measure Cut Points. https://www.cms.gov/medicare/prescription-drug-coverage/prescriptiondrugcovgenin/downloads/2014-trends-in-part-c-and-d-star-rating-measure-cut-points.Pdf. Accessed April 21, 2017.
25. Choudhry NK, Shrank WH, Levin RL, et al. Measuring concurrent adherence to multiple related medications. *Am J Manag Care*. 2009;15(7):457-464.
26. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. 2001:17-21. doi:D010001275 [pii]
27. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34(5):301-310. doi:10.1006/jbin.2001.1029
28. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med*. 2009;150(9):604-612. doi:10.7326/0003-4819-150-9-200905050-00006

29. Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J Clin Epidemiol*. 2011;64(7):749-759. doi:10.1016/j.jclinepi.2010.10.004

30. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J Clin Epidemiol*. 2014;67(8):850-857. doi:10.1016/j.jclinepi.2014.03.012

31. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann Stat*. 2003;31(6):2013-2035.

32. Desgraupes B. Clustering Indices. https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf. Published 2013. Accessed January 31, 2018.

33. Bass AJ, Dabney A, Robinson Maintainer John Storey DD. Q-value estimation for false discovery rate control. http://qvalue.princeton.edu. Published 2016.

Tables

Table 1. Baseline characteristics of the patients included in the study.

	No. (%) or Mean (SD)		
Characteristic	**Derivation cohort (12 practices)**	**Validation cohort (12 practices)**	**p-value**
Number of patients	4,382	3,836	
Age, years	76.6 (5.4)	76.0 (5.4)	< 0.001
Sex			
Female	2,637 (60.2%)	2,189 (57.1%)	0.005
Male	1,745 (39.8%)	1,647 (42.9%)	
Race			
White	3,970 (90.6%)	3,101 (80.8%)	
Black	154 (3.5%)	359 (9.4%)	< 0.001
Hispanic	28 (0.6%)	42 (1.1%)	
Other/Unknown	230 (5.2%)	335 (8.7%)	
Baseline systolic blood pressure (mmHg)	130.3 (14.8)	130.0 (14.5)	0.366
Baseline diastolic blood pressure (mmHg)	72.8 (9.2)	74.0 (9.3)	< 0.001
Baseline eGFR, ml/min/1.73m^2			
≥ 90	480 (11.0%)	460 (12.0%)	
60 – 89	2,451 (55.9%)	2,268 (59.1%)	0.001
30 – 59	1,328 (30.3%)	1,000 (26.1%)	
15 – 29	104 (2.4%)	92 (2.4%)	
< 15	19 (0.4%)	16 (0.4%)	
Combined comorbidity score, median (IQR)	0 (-1 – 2)	0 (-1 – 2)	0.430
Follow-up time after index date, years, median (IQR)	360 (360 – 360)	360 (360 – 360)	0.763
No. of notes in the year prior to index date, median (IQR)	74 (43 – 120)	76.5 (47 – 118)	0.015

Table 2. Concepts associated with non-adherence, clustered by similarity and sorted within each cluster by odds ratio (highest to lowest).

Cluster	Concepts	Pooled OR in derivation cohort (95% CI)*	Pooled OR in validation cohort (95% CI)*
1. Volume depletion and overload	hypotension, 1 pitting edema, stopped, initiation, IV fluids, notified, pass, He, pedal edema, Discharge Date, Left back, sacral, period, lower extremity edema, was, maintained, speak, emergency room, left anterior, u weeks, stabilized, BNP, 1 36, little, feeling better, hold, Epic, 6/6, dehydration, machine, rapid, Bloating, awake, WBC, Monday, excessive, wanted, tachycardia, DVT, coordinating, compression stockings, when, 3/4, accurate, pain Oral, place, hydration, stockings, slowly, higher dose, some, leg edema, causes, r/t, trace, right hand, beta blocker, signs, NT, night, fracture, documentation, reason, standing, monitor, u/s, primary, 5/5, leg pain, G, behavioral, air, until, extensive, lightheadedness, longer, sounds, clinic, decreased, heart, all, related, book, low salt diet, difficulty, overnight, moderate, improving, but, Tuesday, protocol, afternoon, urine culture, after, weakness, denied, Resp, tired, post, live, easily, exertion, dry, Diuretic, physical activity, knee, consult, first, progressive, MRI, cut back, increased, support, culture, sitting, increasing, right arm, understanding, imaging, Wednesday, CXR, energy, mL, fine, EKG, talk, Final, cooking, position, bowel movements, over, weight gain, placement, mood, only, hard, ultrasound, move, hand, hearing, immediately	1.36 [1.19, 1.53]	1.37 [1.19, 1.55]
2. Case management, Assessment of needs at discharge, Mood disorders, Social determinants of health	necrosis, ICU, homebound, knowledgeable, Transfers, Case Manager, home services, Lives with spouse, Case Management, mobility in home, skilled nursing facility, family support, involved patient's, transport, nurse case manager, residence, supervision, continent, following procedure, Social support, returned home, Family Name, Toileting, initial, referral to, Leaving home, DME, medical facility, Social Services, regular diet, coordination, toilet, medication teaching, Admissions care, Cognition, mental illness, Ambulatory, nursing, assistive device, Lower body, walker, attending physician, bathing, discharged home, plan of care, Oxycodone, HOME ASSESSMENT, Upper body, ambulance, home visit, DNR, steps 1, independent bathing, nurse, development, rehab, Coagulation, Date appointment, SENNA, poor, Ability to lift, conscious level, close supervision, retention, Anxiety/depression, Hospital Patient, assist, SSRI, Elder, ENVIRONMENTAL, elevator, Short term, level 1, care taker, commode, hospital, Handicap, Medicare, referred to, self care, independent Walking, referral, admitted, secondary to, rehabilitation, Gait, medication changes, house, dependent, physical therapist, hospitalization, d/c, d, home exercise program, bed, Action, hospice, coming to, independent, Patient reports, steady, stand, wean, activity, hospitalized, ADLs, COLACE, dressing, oriented, mobility, mom, Nutrition, pain management, supply, cane, bowel, Packet, upset, emergency, incontinence, willing, afraid, Medication review, compliance, Medical Center, DOB, refused, bladder, p, family member, 1B, maximum	1.53 [1.34, 1.72]	1.56 [1.35, 1.77]
3. Neurological disorders	left upper extremity, unsteady gait, attentive, gait abnormality, head CT, neurologist, XE, confusion, tap, cerebellar, unsteady, unaware, difficulty walking, memory, once per day, mental status, small vessel, visual fields	1.57 [1.32, 1.81]	1.49 [1.24, 1.74]
4. Coordination of care between physician's	back, heartbeat, failure, dry mouth, transferred, number, used, give, treatment, chest pain, pills, Device, shortness of breath, sleeping, REVIEW OF SYSTEMS, oxygen, morning, upset stomach, prepared,	1.32 [1.16, 1.48]	1.34 [1.17, 1.51]

office and patient, documentation of symptoms	level, causing, Hi, mixed, natural, diarrhea, specialist, moist, pls, edema, abdominal pain, responsibility, repair, evidence of, complications, sleepy, affected, slight, combination, quality, course, gastritis, taper, discharge, letter, every hours, arm, fall, FW, discomfort, relieved, M, Pool, middle, sleep, appetite, approximately, findings, scattered, direct, cancel, meeting, sclerae, drop, qhs, cases, amount, technique, ARTHRALGIA, nausea, Magnesium, treat, vomiting, bedtime, consistent, proximal, adv, learn, finished, Oral mucosa, Advanced, swelling, interested, Oropharynx, complaint, tid, midline		
5. Home health needs, Refractory to treatment, noncompliance	hospital bed, denial, PICC line, noncompliance, unresponsive, DECUBITI, writer, GH, refractory, nursing home, wheelchair, assisted	1.91 [1.56, 2.25]	1.72 [1.35, 2.09]
6. Management of cardiac arrhythmias with warfarin	spontaneous, lateral malleolus, H 2, prothrombin time, during, taking medications, communicate, accident, remind, noon, planning, adjustment	1.46 [1.23, 1.68]	1.49 [1.25, 1.73]
7. Asthma/COPD** management program	pulmonology, responsive, trained, expiratory, offered, approach	1.65 [1.40, 1.90]	1.42 [1.16, 1.67]
8. Language barrier, presence of drug allergies	Translator, On admission, take blood pressure, Allergy-drug	1.56 [1.28, 1.83]	1.52 [1.22, 1.82]
9. Blood pressure medication	blood medicine	1.99 1.56, 2.42]	1.61 [1.16, 2.05]
10. Urinalysis	HPF, KETONE	1.49 [1.22, 1.75]	1.55 [1.26, 1.84]
11. Plantar fasciitis instructions	ball of foot	1.87 [1.40, 2.34]	1.91 [1.46, 2.37]

Note: q-value < 0.05 for all concepts. Additional details for each concept in Supplementary Appendix 1. OR = odds ratio. CI = confidence interval. Prior to analysis, notes were de-identified by converting all numbers to “1” followed by use of de-identification software.
* Pooled OR and 95% CI is the mean OR and 95% CI for the individual concepts in each cluster.
** COPD = chronic obstructive pulmonary disease

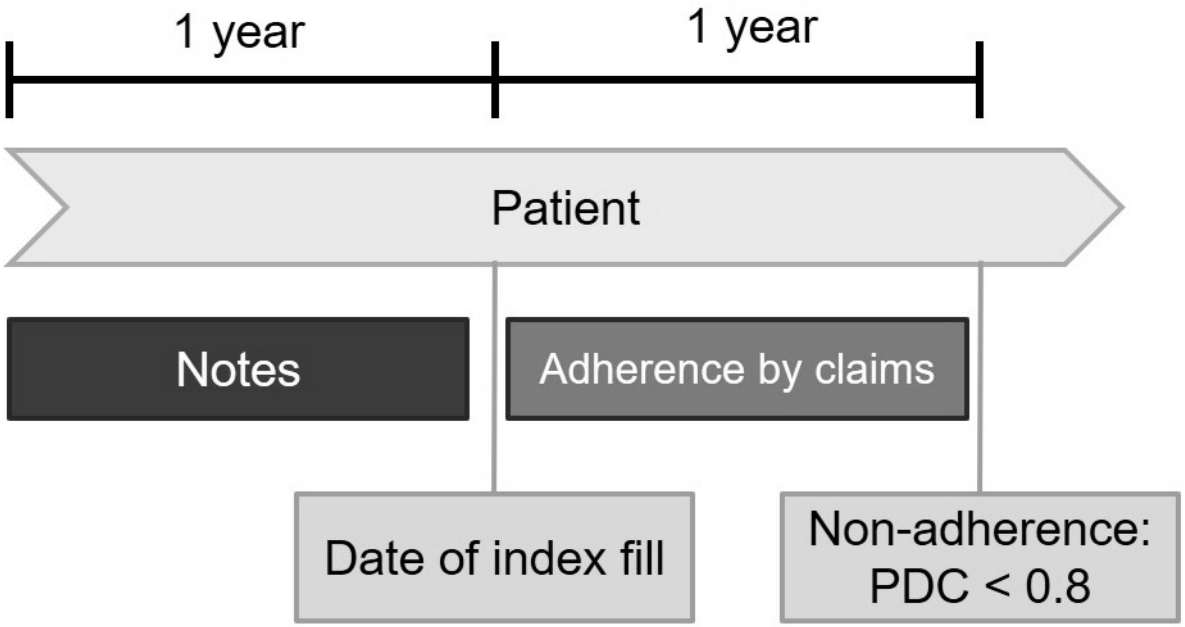
Table 3. Comparing predictors identified by a review article[1] to our notes-driven approach.

Predictors in review article	Related phrases identified using a notes-driven approach	Clusters
Presence of psychological problems, particularly depression	mental illness, Anxiety/depression, SSRI, mood	1, 2
Presence of cognitive impairment	Cognition, confusion, memory, mental status	2, 3
Treatment of asymptomatic disease	--	--
Inadequate follow-up or discharge planning	--	--
Side effects of medication	Hypotension, dehydration, tachycardia, beta blocker, Diuretic, dry mouth, Allergy-drug	1, 4, 8
Patient's lack of belief in benefit of treatment	refused	2
Patient's lack of insight into the illness	unaware	3
Poor provider-patient relationship	--	--
Presence of barriers to care or medications	upset, afraid, noncompliance, Translator	2, 5, 8
Missed appointments	cancel	4
Complexity of treatment	medication teaching, Medication review	2
Cost of medication, copayment, both	--	--

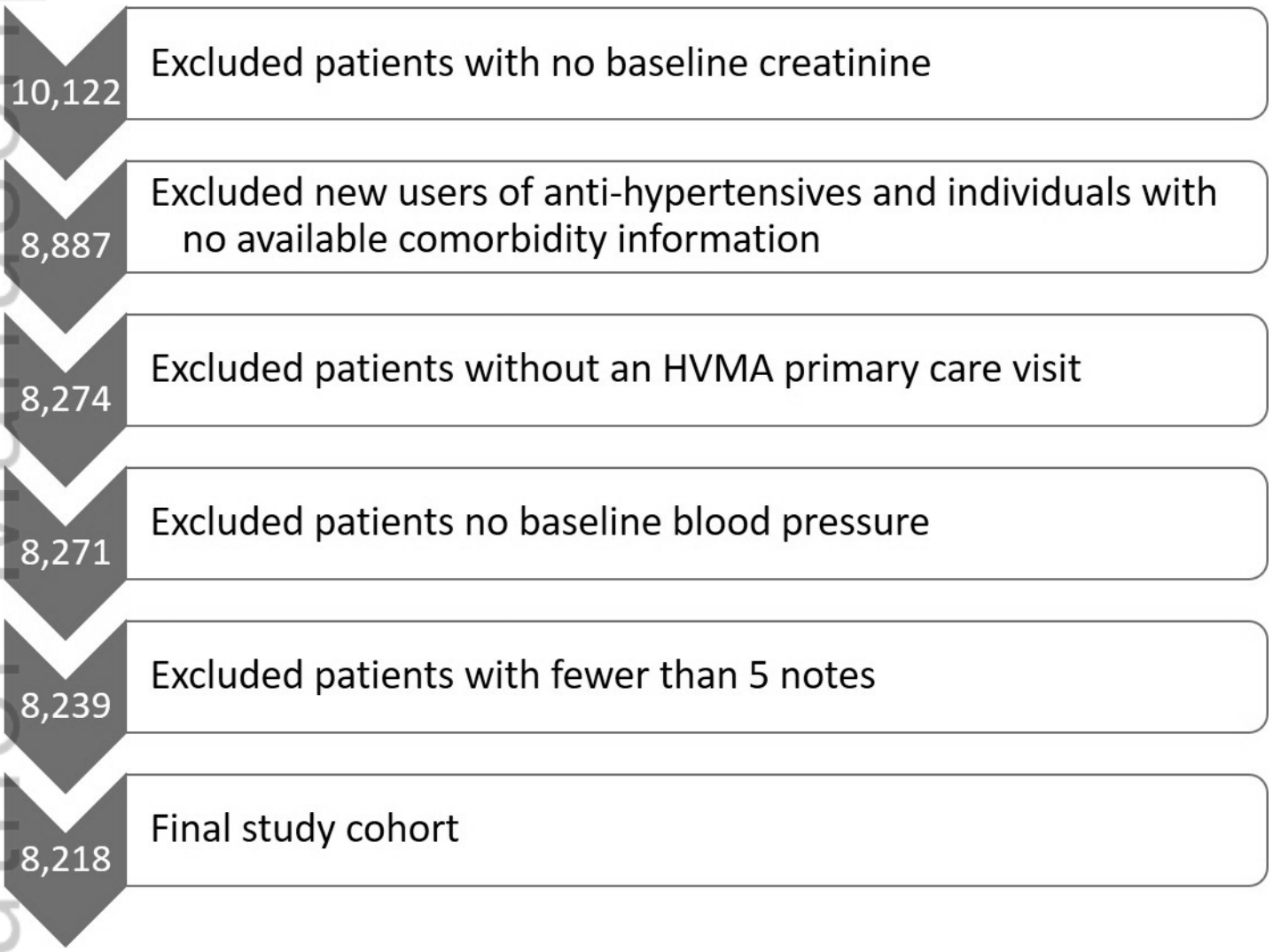
Figure Legends

Figure 1. Analysis of clinical notes in relation to index fill and outcome assessment.

Figure 2. Flow diagram of patient selection for study cohort.



PDS_4850_f1.jpg



PDS_4850_f2.jpg