Assessing Variation in First-Line Type 2 Diabetes Treatment across eGFR Levels and Providers

Christina X Ji[a]*, Saul Blecker[b], Michael Oberst[a,1], Ming-Chieh Shih[a,2], Leora I Horwitz[b], David Sontag[a]

[a]MIT CSAIL and IMES, 32 Vassar Street, Cambridge, MA 02142

[b]Department of Population Health, NYU Grossman School of Medicine, 550 First Avenue, New York, NY 10016 and Department of Medicine, NYU Grossman School of Medicine, 550 First Avenue, New York, NY 10016

[1]Present address: John Hopkins University, 3400 North Charles Street, Baltimore, MD 21218

[2]Present address: National Tsing Hua University, 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 300044

*Corresponding author: cji@alum.mit.edu, 617-258-0625, 32 Vassar Street 32-G424, Cambridge, MA 02142

Abstract

Objective: The purpose of this study is to assess variation in first-line type 2 diabetes treatment empirically using a large clinical dataset. Since metformin, the guideline-recommended first-line treatment, is contraindicated for severe chronic kidney disease, we examine variation in this treatment decision on two axes—across estimated glomerular filtration rate (eGFR) measurements from the patient and across preferences from the prescribing provider.

Study Design and Setting: Using a large insurance claims dataset, we conducted a retrospective cohort study of patients who were newly initiated on a type 2 diabetes treatment (metformin versus a DPP-4 inhibitor or sulfonylurea). Three years of observation prior to treatment were required, and patients with type 1 or gestational diabetes or without eGFR results were excluded. 1) To test whether the choice of treatment is significantly dependent on eGFR level, we performed a chi-squared test for association between eGFR level and treatment decision. 2) To test whether practice variation exists among providers that cannot be explained by treatment guidelines, we fitted restricted cubic spline models to predict treatment from patient age, eGFR, sex, history of heart failure, and treatment date. Then, we performed a generalized likelihood ratio test (GLRT) to assess whether a model that included provider-specific random effects is a better fit than a model without these random effects.

Results: Among 10,643 eligible patients, the choice of metformin versus a DPP-4 inhibitor or a sulfonylurea was significantly associated with eGFR level ($p < 0.0001$). Among the 2,271 patients seen by 173 providers with at least 10 patients in the cohort, a GLRT found significant variation exists across providers even after accounting for age, eGFR, sex, history of heart failure, and treatment date ($p < 0.0001$).

Conclusion: Our study found significant variation in first-line type 2 diabetes treatments—some that can be explained by treatment guidelines and some that may be due to provider preferences. Further studies can help elucidate whether such variation across providers is appropriate. The data-driven approaches in our study can also be applied to other disease areas to characterize variation in real-world clinical practice and potential opportunities for improvement.

Keywords: practice variation, provider preferences, treatment guidelines, first-line type 2 diabetes treatment, metformin contraindications, observational study

Running Title: Assessing Variation in First-Line Type 2 Diabetes Treatment

Word Count: 2793

What is New?

Key Findings

- We used a large health insurance claims dataset to show that first-line type 2 diabetes treatment is significantly associated with estimated glomerular filtration rate (eGFR).
- We established empirically that significant variation exists in how providers choose between metformin and DPP-4 inhibitors or sulfonylureas for first-line type 2 diabetes treatment, even after accounting for eGFR, age, sex, history of heart failure, and treatment date.

What this adds to what is known?

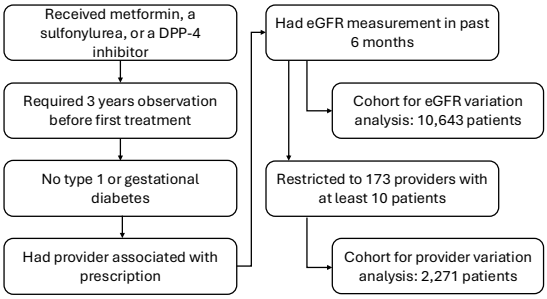
- We proposed a new statistical approach to test for variation in treatment decisions across all providers while accounting for patient characteristics.
- We applied this approach to establish that providers may consider eGFR levels differently when prescribing first-line type 2 diabetes treatments.

What is the implication and what should change now?

- Further work is needed to understand whether the provider variation we discovered for initial type 2 diabetes treatment is appropriate, and if not, how to remove this variation by reaching out to providers or improving treatment guidelines
- Our proposed statistical approach can be applied to determine whether provider variation exists for other diseases after accounting for patient characteristics relevant to the specific disease.

Assessing Variation in First-Line Type 2 Diabetes Treatment across eGFR Levels and Providers

1. Defined cohort with first-line type 2 diabetes treatment



2. Tested for significant treatment variation across eGFR levels

Number of patients in each eGFR category who receive each treatment class as first-line treatment

eGFR level	Metformin	DPP-4i / Sulfonylurea	Total
eGFR < 30	9	55	64
30 ≤ eGFR < 45	92	204	296
45 ≤ eGFR < 60	709	356	1,065
60 ≤ eGFR < 90	3,851	662	4,513
90 ≤ eGFR	4,178	527	4,705
Total	8,839	1,804	10,643

Chi-squared test: Is there significant association between metformin prescription and eGFR level?

Yes, there is significant variation in first-line treatment across eGFR levels ($p < 0.0001$).

3. Tested for significant treatment variation across providers

Fit generalized linear models predicting treatment Y from patient characteristics X with and without provider-specific random effects

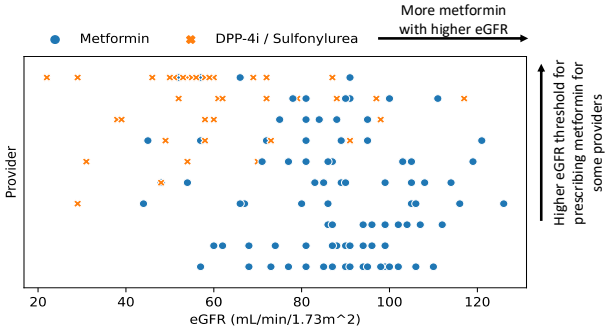
Patient characteristics X:
- Estimated glomerular filtration rate (eGFR)
- Age
- Treatment date
- Heart failure
- Sex

First-line type 2 diabetes treatment Y:
- Metformin vs Sulfonylurea / DPP-4 inhibitor

Generalized likelihood ratio test: Does data have higher likelihood under model with provider-specific random effects compared to model without random effects?

Yes, there is significant variation in first-line treatment across providers ($p < 0.0001$).

4. Visualized provider decisions with respect to eGFR



1. Introduction

Clinical practice guidelines have been developed to promote evidence-based practice in medicine and improve quality of care and clinical outcomes.[1] However, actual practice often deviates from guidelines. This variation may be due to incomplete or unclear guidelines, barriers to adoption, different practice styles of individual providers and health systems, or other factors that influence patient-provider interactions.[2-4] For instance, guidelines may not be applicable for complex patients with multiple co-morbidities. In this case, providers may deviate from the guidelines to provide more appropriate care.[5-7] In other cases though, deviation may be due to poor dissemination of guidelines or lack of resources to support adoption.[2-3] Such variability in guideline compliance may result in suboptimal quality of care or clinical outcomes for patients.

Large clinical datasets offer opportunities to empirically study clinical decision-making and practice variation across providers. In particular, models fit to real-world evidence may be better at capturing decision-making processes that are not covered by or differ from guidelines. The motivation for studying variation with observational data is to understand when variation is captured by guidelines, seems appropriate based on patient characteristics, or may be part of a consistently flawed decision-making process that needs to be remedied.[1,8] In the latter case, we must be careful when flagging a provider or a health system as an outlier that needs attention. A provider who takes on more challenging cases by seeing patients who have more co-morbidities or are more likely to have contraindications for evidence-based care should not be unduly punished. Even if their patients have worse outcomes, much of that variation may still be clinically appropriate.[9] To account for these nuances, we propose first examining variation that can be explained by patient characteristics and then accounting for these characteristics when studying variation across providers. Past works with similar approaches have examined whether the choice of physician has an impact on cancer treatment and hospice enrollment.[10-12]

Metformin is the guideline-recommended initial treatment for type 2 diabetes. It is effective at improving glycemic control and promoting weight loss without increasing risk of hypoglycemia.[13] The medication is also cost-effective. American Diabetes Association (ADA) guidelines recommend starting metformin unless "there are contraindications."[14] However, the guidelines do not list most contraindications and only suggest, rather than explicitly recommend, that metformin is safe in chronic kidney disease (CKD) so long as eGFR is at least 30 mL/min/1.73m^2.[14] For patients with mild renal impairment, defined as eGFR levels between 30 and 60, metformin only started being recommended in 2016 after several large-scale cohort studies found that metformin did not increase the risk of lactic acidosis in these patients.[15-17] Because there is uncertainty around the specific eGFR threshold at which the risk of metformin in CKD outweighs the benefits in diabetes, the decision to prescribe another medication because metformin is contraindicated falls on the individual provider, leading to practice variation.

Using a large insurance database, we examined how providers decided to initiate metformin versus another diabetes medication for patients with type 2 diabetes. First, we assessed whether there is significant dependence between the choice to initiate metformin and eGFR levels. This allowed us to verify that the patient characteristics used to define treatment guidelines indeed play a role in real-world decision-making. Second, we modelled how eGFR

and other factors influenced the likelihood of prescribing metformin in general and for each individual provider. We then used these models to assess whether significant variation exists across providers. This part of our analysis established the existence of variation that cannot be explained by treatment guidelines and relevant patient characteristics. We hope this work will encourage future studies that examine how this variation affects patient outcomes and, if needed, how to issue better guidance to reduce harmful variation.

2. Methods

2.1. Dataset and Cohort

We performed a retrospective cohort study using insurance claims data from a large insurance provider in the northeast United States, spanning from 2012 to 2021. The database includes laboratory test results when the insurance company has a contract with the laboratory center filling the claim. Similar insurance claims datasets have been used before for prediction of many health outcomes, including onset of type 2 diabetes, abnormal eGFR results, hospitalization, and mortality.[18-20]

Patients who received an initial medication fill for metformin, a sulfonylurea, or a dipeptidyl peptidase-4 (DPP-4) inhibitor were included in the cohort. Patients who were prescribed multiple treatment classes on the first, did not have exactly one provider associated with the prescription, were not observed for at least 3 years prior to the initial medication fill, or did not have an eGFR measurement in the past 6 months were excluded. See Appendix A for definitions of the treatment classes and number of patients after applying each exclusion criterion. For the analysis on provider-based variability, we limited our cohort to patients who were seen by providers who had at least 10 patients who met the inclusion and exclusion criteria. Treatment decisions were attributed to the provider listed on the insurance claim for the medication fill.

The outcome of interest was a binary variable for whether the patient was prescribed metformin versus a sulfonylurea or a DPP-4 inhibitor. Our primary exposure was eGFR, which was based on the last measurement obtained in the 6 months prior to the initial medication fill. As described in Appendix A, the appropriate race-adjusted eGFR concept was prioritized when available, and patients without eGFR measurements were excluded. Other variables included age, sex, and the presence of a diagnosis for heart failure. These variables were chosen as they may contribute to variation in decision making related to use of metformin.[21-22] We also included treatment date as a covariate since metformin usage increased over time, particularly when metformin was no longer contraindicated for patients with moderate renal failure in 2016.[15-17]

2.2. Statistical Approach to Test for Variation

Because first-line diabetes treatment guidelines are heavily dependent on eGFR, we first verified that the general treatment policy observed in real-world data aligned with guidelines. We created the following 5 eGFR categories based on the definitions for the different stages of CKD: below 30, 30-44, 45-59, 60-89, and at least 90.[8] We calculated the number of patients in each

eGFR category who were prescribed each medication class. These counts were used in a chi-squared test to determine if there was a difference in rates of metformin use by eGFR category.

To assess whether providers have different treatment policies with respect to eGFR, we needed to account for how eGFR and other factors affected the likelihood of prescribing metformin. This is important for two reasons: 1) Some providers see more challenging patients.[9] For example, patients with severe CKD are more likely to seek out specialists than general practitioners for treatment. Modeling patient features accounts for these differences in patient populations across providers. 2) We are interested not just in how providers prescribe the treatments at different rates but in how providers differ in how they take these factors into account when making treatment decisions. Thus, these features need to be included when predicting the likelihood of prescribing metformin.

To account for these features when modeling the treatment decisions, we first found the best model in each of the following two model families: 1) A generalized linear model with restricted cubic spline features for eGFR, age, and treatment date and binary indicators for heart failure and sex. The cubic spline features allow us to model non-linear relationships.[23] For details on how the restricted cubic spline features are defined, see Appendix B. 2) A generalized linear model with the previously mentioned features, random intercepts, and random slopes for the eGFR features. The random intercepts and random slopes account for differences in how providers account for eGFR and weigh the risks and benefits of each treatment when making decisions.

Then, we used the best models in these two families to examine whether there was variation across all providers. We assessed this by performing a generalized likelihood ratio test (GLRT) for whether model family 2 is a better fit for the observed samples than model family 1.[24-25] By evaluating the likelihood only at the observed samples, this test focused on decisions for the types of patients who were actually seen by each provider. As an additional analysis, we also examined whether any individual provider deviated significantly from the average treatment policy. For this part, we conducted separate GLRTs using only the samples from patients seen by a particular provider. We used the Benjamini-Hochberg procedure to keep the expected false discovery rate at 5%.[26]

The method we propose joins a large body of work on identifying outlying providers described in Appendix C, including mixture models of provider effects, hierarchical and Bayesian approaches to modeling random effects, and Markov chain Monte Carlo simulations of normal behavior.[27-34] Unlike these prior works, our method tests whether variation exists across all providers rather than whether a single provider is an outlier. This is particularly useful when there are few samples per provider. While there may be insufficient power to identify individual providers as outliers, the total number of samples across all providers may provide sufficient power to identify whether variation exists across all providers.

To visualize how the treatment decisions made by each provider vary with eGFR, we first plotted the treatments observed for patients in the dataset against each eGFR value, with the decisions for each provider in a separate row. Then, we plotted the treatment policies learned by

the models fit for the GLRT. The model predicts the likelihood of prescribing metformin given eGFR level, age, sex, treatment date, and history of heart failure. We held all features besides eGFR constant, so the policy shown is for a 50-year-old female given treatment on 2014-05-25 with no history of heart failure. These features were chosen arbitrarily to fall within the observed population. To create this second plot, we varied the eGFR value from the minimum to the maximum observed (3 to 155) and plotted the predicted likelihood of prescribing metformin in general and for each provider. Unlike the observed decisions in the first plot, the second figure shows the policy a provider may follow at any eGFR value.

Data was extracted using SQL from a postgres database. Models were fit using R. All other statistical procedures were carried out using Python.

3. Results

3.1. Descriptive Statistics and Variation across eGFR levels

A total of 10,643 patients met the inclusion and exclusion criteria. Metformin was prescribed to 83.0% of these patients, and the other 17.0% were prescribed a sulfonylurea or a DPP-4 inhibitor. The mean age was 60.6 (standard deviation 12.9). 52.8% of patients were male. 6.7% of patients had heart failure in the past 730 days. Metformin was prescribed for only 63.9% of patients who had heart failure, compared to 84.4% of patients who did not have heart failure. The mean eGFR level was 84.0, with a standard deviation of 20.5.

As demonstrated in Table 1, both categories of medications were prescribed across the range of eGFR values. However, there was a noticeable and significant increase in the prevalence of metformin prescriptions as eGFR values increased ($p<0.0001$ from chi-squared test). Only 14.1% of patients with eGFR less than 30 were prescribed metformin. This rate increased to 31.1% among patients with eGFR 30-44 and 66.6% among patients with eGFR 45-59. For patients whose eGFR measurements indicated no sign of kidney damage, metformin prescription rates were 85.3% among patients with eGFR 60-89 and 88.8% among patients with eGFR ≥ 90. See Table 1 in Appendix A for the number of patients in each category.

3.2. Variation across Providers

There were 173 providers who saw at least 10 patients in the dataset. Among these 173 providers, the mean number of patients per provider was 13.1 with a range of 10 to 41. The total number of patients in this restricted cohort was 2,271. The mean per-provider metformin prescription rate was 82.2% with a range from 22.7% to 100.0%. The maximum likelihood model with and without random effects both had 4 knots each for eGFR, age, and treatment date. The GLRT comparing the data likelihood under the model with random effects and the model without random effects indicated significant variation metformin usage across providers even when accounting for eGFR, age, sex, prior heart failure, and treatment date (G-statistic 387.3, $p<0.0001$).

While there was variation across all providers, we were not able to identify any single provider as an outlier because each provider only had a few samples. In the tests we performed

with only a single provider's patients, we found that no particular provider significantly deviates from the average treatment policy. The smallest p-value for a single provider was 0.059.

Figure 1 illustrates that metformin use increased as eGFR increases. However, there was heterogeneity among providers. Most of this variation among providers occurred at low eGFR values. Some providers still prescribed metformin when eGFR was low, while other providers did not use metformin even when eGFR was high. Providers at the top of the figure prescribed no metformin. Those at the bottom prescribed entirely metformin.

The black line in Figure 2 that depicts the model learned without provider-specific random effects illustrates how the metformin prescription probability started around 20% for small eGFR values and increased to around 90% as eGFR levels started to indicate normal kidney function. The blue lines in Figure 2 depict the policies learned for each provider. Again, when eGFR was low, some providers still prescribed metformin with high probability, while other providers were less inclined to give metformin. Other values for patient age, sex, and treatment date yielded similar plots.

4. Discussion

Using a large insurance claims dataset, we studied how the choice of first-line treatment for type 2 diabetes varies with different levels of kidney damage and with prescribing provider. The 83.0% metformin prescription rate in our cohort is fairly consistent with the 58-77% metformin use rate found in prior studies.[35-38] We also observed that metformin was prescribed significantly more frequently among patients with higher eGFR levels, as would be expected based on treatment guidelines. However, eGFR level, age, sex, history of heart failure and treatment date did not fully explain variation in metformin prescription. We also found significant variation across providers. This suggests that if a patient sees different providers, they may receive different care, be prescribed different treatments, and ultimately have different outcomes.

The degree of provider level variation is particularly interesting because of its novelty and the potential for interventions to improve patient care. To our knowledge, this is the first study that uses a large observational dataset to empirically assess how first-line type 2 diabetes treatment varies across providers while accounting for patient characteristics. Using the observational dataset allowed us to evaluate variation with the decisions that were actually made. Most prior research in this area relied on self-reported interviews and surveys, which may not always reflect clinical practice.[39] One study demonstrated that providers are not always aware of the appropriate timing for and contraindications to metformin.[40] Another survey found that only a third of providers are aware that the eGFR cutoff typically used for prescribing metformin is 30.[21] Other work discussed in Appendix C focused on process-of-care indicators, such as whether annual assessments were performed.[41-44]

Because this study demonstrated that variation exists across providers who prescribe first-line type 2 diabetes treatments, it raises the question of whether this variation needs to be addressed. If the variation is inappropriate, how can the negative effects be mitigated? A non-specific intervention would be to provide brief education about first-line diabetes treatments to

providers whose treatment decisions do not align with the guidelines related to eGFR and metformin. More data-driven and individualized interventions such as reviewing provider practice patterns to identify areas for improvement or imposing financial penalties for unwarranted variation would require accurate assessment of the performance of individual clinicians, identification of outlying providers, and good intervention design.[45-50] For such interventions, there must be sufficient evidence based on decisions for a large number of patients that the provider is indeed making poor treatment decisions to avoid incorrectly penalizing providers.[51-54]

Our study should be interpreted in the context of its limitations. First, this study was limited to a single insurance provider whose beneficiaries are primarily located in the northeast United States. As a result, the findings may not be generalizable. Second, while our models included several patient characteristics, they may not have accounted for all patient-level factors that contribute to treatment decisions. Third, while we chose to evaluate the performance of individual providers, variation may also be studied at the level of provider groups or health care facilities.[55]

Finally, the statistical approach in this study is more broadly applicable to treatment decisions for other diseases as well. The models in our method can be built with relevant patient characteristics for those diseases. All code is publicly available at https://github.com/clinicalml/t2dm_provider_variation_analysis.

Acknowledgments

We would like to thank Rebecca Boiarsky and Justin Lim for setting up the databases used in this study. We would also like to thank James Denyer, Aaron Smith-McLallen, Stephanie Gervasi, and the rest of the data science group at Independence Blue Cross for providing the dataset.

Funding and Competing Interests

This work was supported in part by Independence Blue Cross, Office of Naval Research Award No. N00014-21-1-2807, and the LEAP program from the Ministry of Science and Technology in Taiwan. There are no competing interests. The insurance claims dataset was provided by Independence Blue Cross. This research was ruled exempt by MIT's IRB (protocol E-4025).

References

1. Mercuri M, Gafni A. Examining the role of the physician as a source of variation: Are physician-related variations necessarily unwarranted? *J Eval Clin Pract*. 2018;24(1):145-151. doi:10.1111/jep.12770
2. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA*. 1999;282(15):1458-65. doi:10.1001/jama.282.15.1458
3. Grol R, Dalhuijsen J, Thomas S, et al. Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. *BMJ*. 1998;317(7162):858-861.
4. Pronovost PJ. Enhancing physicians' use of clinical guidelines. *JAMA.* 2013;*310*(23):2501-2502.
5. Giugliano D, Esposito K. Clinical inertia as a clinical safeguard. *JAMA*. 2011;305(15):1591-2. doi:10.1001/jama.2011.490
6. Safford MM, Shewchuk R, Qu H, et al. Reasons for not intensifying medications: differentiating "clinical inertia" from appropriate care. *J Gen Intern Med*. 2007;22(12):1648-55. doi:10.1007/s11606-007-0433-8
7. Lebeau JP, Cadwallader JS, Aubin-Auger I, et al. The concept and definition of therapeutic inertia in hypertension in primary care: a qualitative systematic review. *BMC Fam Pract*. 2014;15:130. doi:10.1186/1471-2296-15-130
8. Shashar S, Codish S, Ellen M, et al. Determinants of medical practice variation among primary care physicians: protocol for a three phase study. *JMIR Res Protoc*. 2020;9(10):e18673. doi:10.2196/18673
9. Landon BE, O'malley AJ, Keegan T. Can choice of the sample population affect perceived performance: implications for performance assessment. *J Gen Intern Med*. 2010;25(2):104-109. doi: 10.1007/s11606-009-1153-z.
10. Lipitz-Snyderman A, Sima CS, Atoria CL, et al. Physician-driven variation in nonrecommended services among older adults diagnosed with cancer. *JAMA Intern Med.* 2016;176(10):1541-1548. doi: 10.1001/jamainternmed.2016.4426.
11. Hawlet ST, Hofer TP, Janz NK, et al. Correlates of between-surgeon variation in breast cancer treatments. *Med Care.* 2006 Jul;44(7):609-16. doi: 10.1097/01.mlr.0000215893.01968.f1.

12. Obermeyer Z, Powers BW, Makar M, et al. Physician characteristics strongly predict patient enrollment in hospice. *Health Affairs.* 2015;34(6):993-1000. doi: 10.1377/hlthaff.2014.1055.
13. Choi JG, Winn AN, Skandari MR, et al. First-line therapy for type 2 diabetes with sodium-glucose cotransporter-2 inhibitors and glucagon-like peptide-1 receptor agonists: a cost-effectiveness study. *Ann Intern Med*. 2022;175(10):1392-1400. doi:10.7326/m21-2941
14. American Diabetes Association. Standards of Medical Care in Diabetes—2021. 2021;44(Supplement 1).
15. Engler C, Leo M, Pfeifer B, et al. Long-term trends in the prescription of antidiabetic drugs: real-world evidence from the Diabetes Registry Tyrol 2012-2018. *BMJ Open Diabetes Research & Care.* 2020;8:e001279. doi:10.1136/ bmjdrc-2020-001279.
16. Tanner C, Wang G, Liu N, et al. Metformin: time to review its role and safety in chronic kidney disease. *Medical Journal of Australia.* 2019;211(1):37-42.
17. Imam TH. Changes in metformin use in chronic kidney disease. *Clinical Kidney Journal.* 2017;10(3):301-304.
18. Razavian N, Blecker S, Schmidt AM, et al. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*. 2015;3(4):277-87. doi:10.1089/big.2015.0020
19. Ji CX, Alaa AM, Sontag D. Large-scale study of temporal shift in health insurance claims. In *Conference on Health, Inference, and Learning*. PMLR. 2023;243-278.
20. Kodialam R, Boiarsky R, Lim J, et al. Deep contextual clinical prediction with reverse distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;35(1):249-258.
21. Goldberg T, Kroehl ME, Suddarth KH, et al. Variations in metformin prescribing for type 2 diabetes. *The Journal of the American Board of Family Medicine*. 2015;28(6):777-784. doi:10.3122/jabfm.2015.06.150064
22. Pilla SJ, Segal JB, Alexander GC, et al. Differences in national diabetes treatment patterns and trends between older and younger adults. *J Am Geriatr Soc*. 2019;67(5):1066-1073. doi:10.1111/jgs.15790
23. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. *Springer*. 2015.
24. Neyman J, Pearson ES. On the problem of most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Seria A, Containing Papers of a Mathematical or Physical Character.* 1933;231(694-706):289-337.
25. Vonesh EF, Chinchilli VM, Pu K. Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics*. 1996;572-587.
26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (methodological)*. 1995;57(1):289-300.
27. Ohlssen DI, Sharples LD, Spiegelhalter DJ. A hierarchical modelling framework for identifying unusual performance in health care providers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2007;170(4):865-890. doi:https://doi.org/10.1111/j.1467-985X.2007.00487.x
28. Marshall EC, Spiegelhalter DJ. Comparing institutional performance using Markov chain Monte Carlo methods. *Statistical analysis of medical data: new developments.* 1998;229-249.

29. Normand SLT, Glickman ME, Gatsonis, CA. Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association.* 1997;92(439):803-814.
30. Spiegelhalter DJ, Aylin P, Best NG, et al. Commissed analysis of surgical performance using routine data: lessons from the Bristol inquiry. *Journal of the Royal Statistical Society Series A: Statistics in Society.* 2002;165(2):191-221.
31. Austin PC. A comparison of Bayesian methods for profiling hospital performance. *Medical Decision Making.* 2002;22(2):163-172.
32. Ohlssen DI, Sharples LD, Spiegelhalter DJ. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statisticals in medicine*. 2007;26(9):2088-2112.
33. Guthrie B, Donnan PT, Murphy DJ, et al. Bad apples or spoiled barrels? Multilevel modelling analysis of variation in high-risk prescribing in Scotland between general practitioners and between the practices they work in. *BMJ Open*. 2015;5(11):e008270. doi:10.1136/bmjopen-2015-008270
34. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J Roy Stat Soc*. 1996;159:385-443.
35. Landon BE, Zaslavsky AM, Souza J, et al. Trends in diabetes treatment and monitoring among medicare beneficiaries. *J Gen Intern Med*. 2018;33(4):471-480. doi:10.1007/s11606-018-4310-4
36. Berkowitz SA, Krumme AA, Avorn J, et al. Initial choice of oral glucose-lowering medication for diabetes mellitus: a patient-centered comparative effectiveness study. *JAMA Intern Med*. 2014;174(12):1955-62. doi:10.1001/jamainternmed.2014.5294
37. Desai NR, Shrank WH, Fischer MA, et al. Patterns of medication initiation in newly diagnosed diabetes mellitus: quality and cost implications. *Am J Med*. 2012;125(3):302.e1-7. doi:10.1016/j.amjmed.2011.07.033
38. Montvida O, Shaw J, Atherton JJ, et al. Long-term trends in antidiabetes drug usage in the U.S.: Real-world evidence in patients newly diagnosed with type 2 diabetes. *Diabetes Care*. 2018;41(1):69-78. doi:10.2337/dc17-1414
39. Krumpal I. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*. 2013;47(4):2025-2047.
40. Trinkley KE, Malone DC, Nelson JA, et al. Prescribing attitudes, behaviors and opinions regarding metformin for patients with diabetes: a focus group study. *Ther Adv Chronic Dis*. 2016;7(5):220-8. doi:10.1177/2040622316657328
41. Oude Wesselink SF, Lingsma HF, Robben PB, et al. Guideline adherence and health outcomes in diabetes mellitus type 2 patients: a cross-sectional study. *BMC health services research*. 2015;15: 1-8.
42. Ackermann RT, Thompson TJ, Selby JV, et al. Is the number of documented diabetes process-of-care indicators associated with cardiometabolic risk factor levels, patient satisfaction, or self-rated quality of diabetes care? The Translating Research in Action for Diabetes (TRIAD) study. *Diabetes Care.* 2006;29(9):2108-2113.
43. Kaplan SH, Griffith JL, Price LL, et al. Improving the reliability of physician performance assessment: identifying the "physician effect" on quality and creating composite measures. *Med Care*. 2009;47:378-387.

44. Brown EC, Robicsek A, Billings LK, et al. Evaluating primary care physician performance in diabetes glucose control. *Am J Med Qual*. 2016;31(5):392-9. doi:10.1177/1062860615585138
45. Scott IA, Phelps G, Brand C. Assessing individual clinical performance: a primer for physicians. *Intern Med J*. 2011;41:144-155.
46. Loeb JM. The current state of performance measurement in health care. *Int J Qual Health Care*. 2004;16(suppl 1):i5-i9.
47. Eijkenaar F. Key issues in the design of pay for performance programs. *Eur J Health Econ*. 2013;14:117-131.
48. Eijkenaar F. Pay for performance in health care: an international overview of initiatives. *Med Care Res Rev*. 2012;69:251-276.
49. de Bruin SR, Baan CA, Struijs JN. Pay-for-performance in disease management: a systematic review of the literature. *BMC Health Serv Res*. 2011;11:272.
50. Huang J, Yin S, Lin Y, et al. Impact of pay-for-performance on management of diabetes: a systematic review. *J Evid Based Med*. 2013;6:173-184.
51. Hofer TP, Hayward RA, Greenfield S, et al. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA*. 1999;281(22):2098-105. doi:10.1001/jama.281.22.2098
52. Fung V, Schmittdiel JA, Fireman B, et al. Meaningful variation in performance: a systematic literature review. *Med Care*. 2010;48:140-148.
53. Normand SL, Wolf RE, Ayanian JZ, et al. Assessing the accuracy of hospital clinical performance measures. *Med Decis Making*. 2007;27:9-20.
54. Sequist TD, Schneider EC, Li A, et al. Reliability of medical group and physician performance measurement in the primary care setting. *Med Care*. 2011;49:126-131.
55. Krein SL, Hofer TP, Kerr EA, et al. Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. *Health Serv Res*. 2002;37:1159-1180.

Table 1. Number of patients in each eGFR category who are prescribed metformin or DPP-4i / sulfonylurea as their initial diabetes medication.

eGFR level	Metformin	DPP-4i / Sulfonylurea	Total
eGFR < 30	9	55	64
30 ≤ eGFR < 45	92	204	296
45 ≤ eGFR < 60	709	356	1,065
60 ≤ eGFR < 90	3,851	662	4,513
90 ≤ eGFR	4,178	527	4,705
Total	8,839	1,804	10,643

Figure 1. Individual provider choice of initial treatment for diabetes, by most recent eGFR on x-axis. Each row is one provider. Providers ordered by metformin prescription rates. A subset of the 173 providers with at least 10 patients are shown: 10 with smallest rates, 10 with largest rates, and every third in between. A blue dot indicates the provider prescribed metformin to a patient with that eGFR value. An orange X is the corresponding decision to prescribe DPP-4i / sulfonylurea.

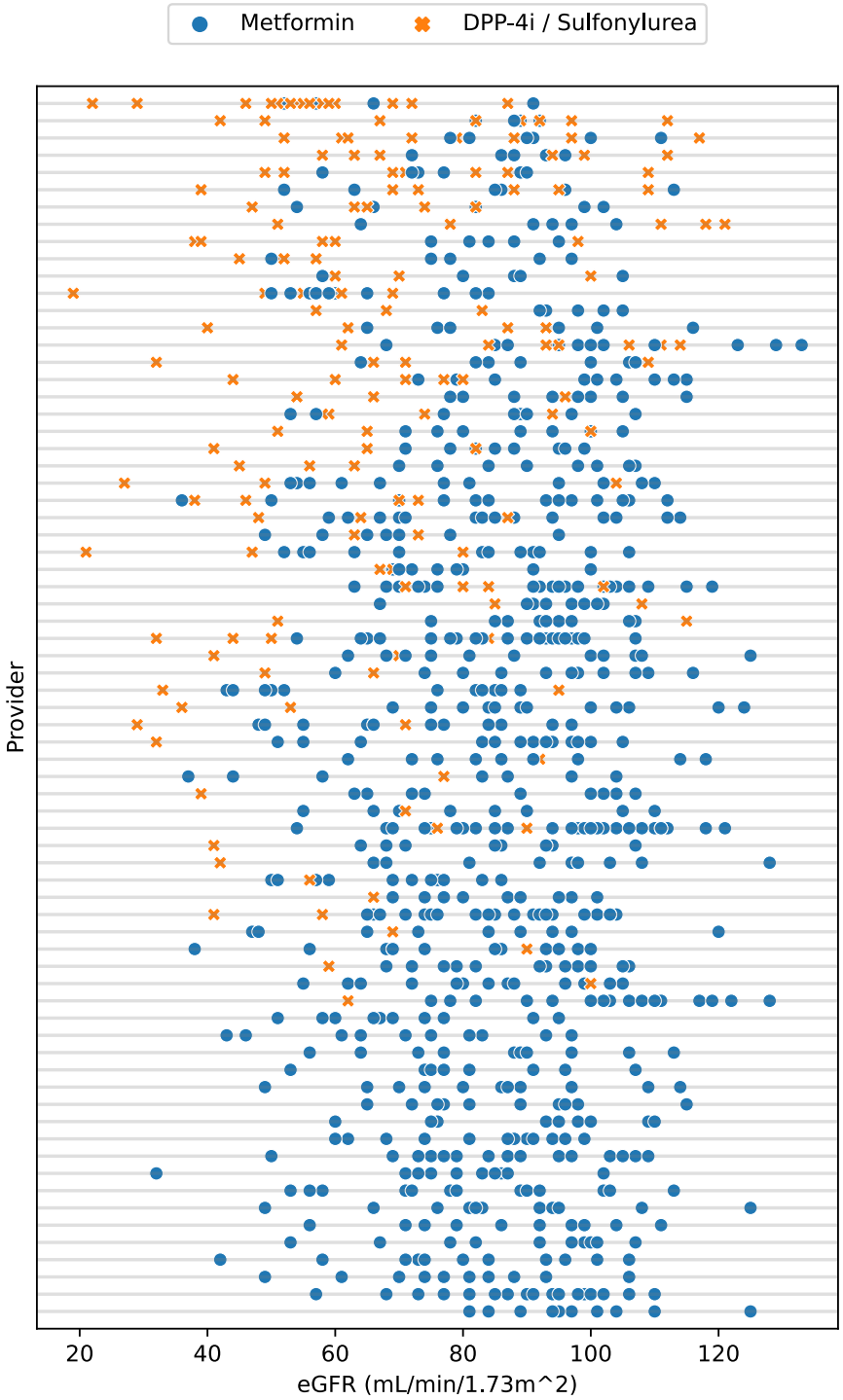
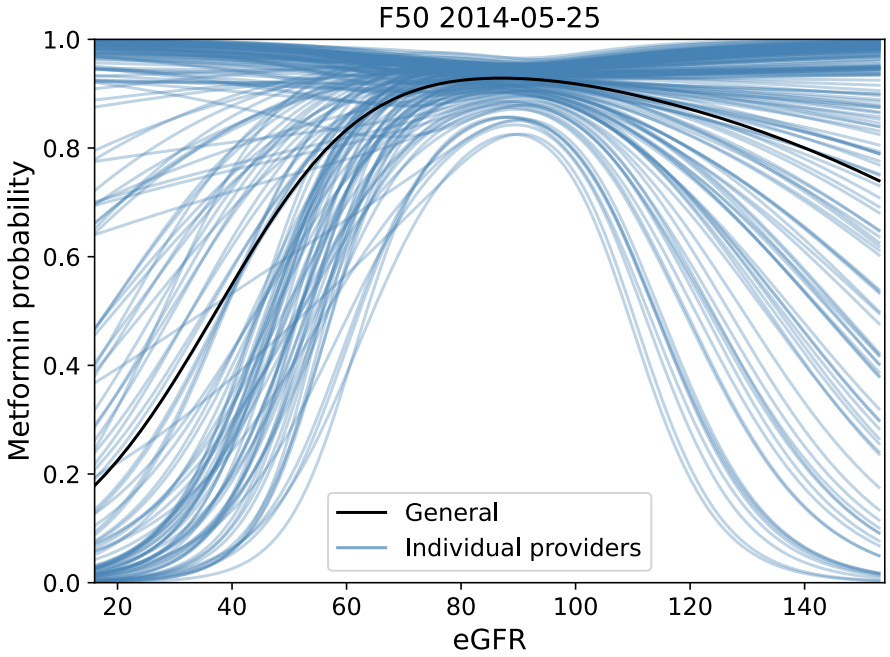


Figure 2. Predicted probability of prescribing metformin versus eGFR for a 50-year-old female without heart failure whose first-line treatment is initiated on 2014-05-25. Metformin probabilities estimated from models without random effects (black) and with provider-specific random effects (blue).



Appendix A: Cohort Definition

Diabetes drugs were defined as drugs that are descendants of the "drugs used in diabetes" concept in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) and contain at least one of the following ingredients: metformin, sitagliptin, vildagliptin, saxagliptin, linagliptin, gemigliptin, anagliptin, teneligliptin, acetohexamide, carbutamide, chlorpropamide, glycyclamide, tolcyclamide, metahexamide, tolazamide, tolbutamide, glibenclamide, glyburide, glibornuride, gliclazide, glipizide, gliquidone, glisoxepide, glyclopyramide, or glimepiride.

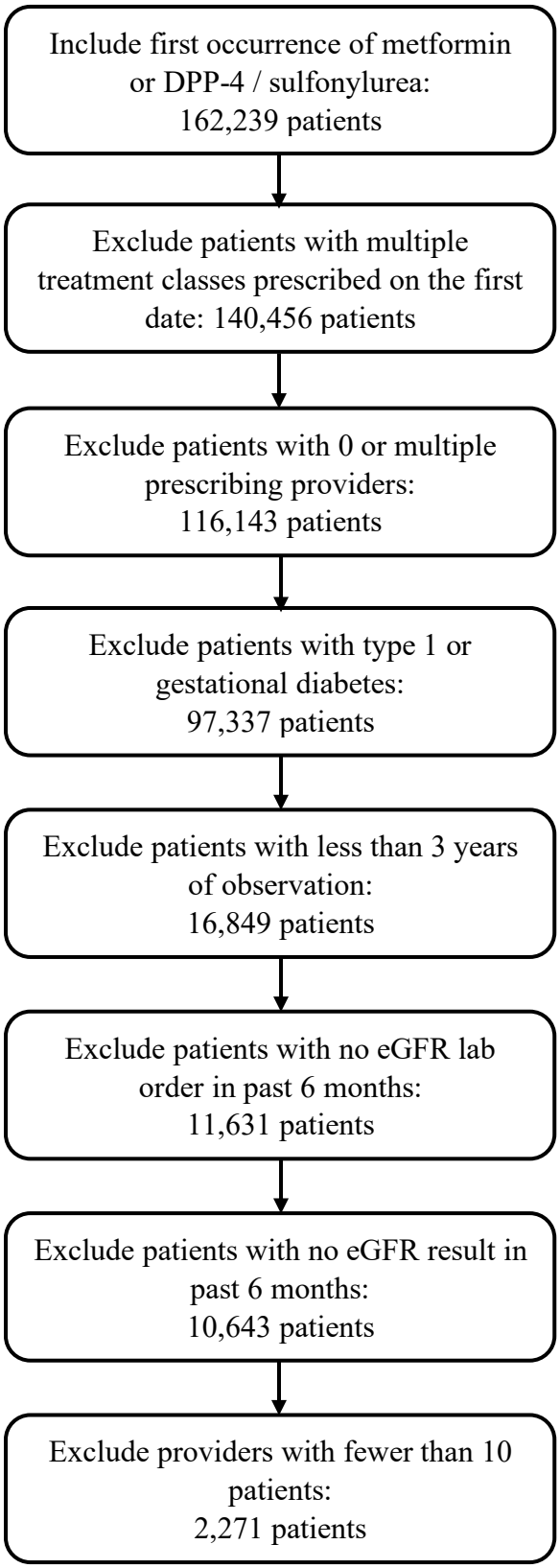
Estimated glomerular filtration rate (eGFR) measurements were defined by all lab concepts that started with "Glomerular filtration rate/1.73 sq M". This included both CKD-EPI that was commonly used through 2017 and MDRD that became more predominant starting in 2018. Because some eGFR measurements are race-adjusted, for black or African American patients, we prioritized measurement concepts that specified "among blacks" over concepts that did not specify race and only used concepts that specified "among non-blacks" if no other measurements were available. For patients of other races or without race specified, the priorities of concepts that specified "among blacks" and "among non-blacks" were flipped. If multiple measurement values at the same priority level exist on the most recent measurement date, the maximum value was taken. Heart failure was defined by concepts that are descendants of the SNOMED concept "Heart failure".

As shown in Figure 4, we started by including the first prescription of one of the specified diabetes drugs. Then, we excluded any patients who are prescribed both metformin and a drug in the sulfonylurea or DPP-4 inhibitor class. Because we are studying variation across providers, the treatment decision must be clearly attributed to a single prescribing doctor. The prescribing providers were identified as the providers listed on the insurance claim for the medication fill. We excluded patients who have 0 or multiple providers associated with the first-line treatment. These drugs were typically only prescribed to patients with some form of diabetes. Because not all patients had a type 2 diabetes diagnosis code associated with the prescription, we did not require a type 2 diabetes code. Instead, to focus on type 2 diabetes, we excluded patients who had a type 1 diabetes, gestational diabetes, neonatal diabetes, or pregnancy-related code at any point in their history to remove other common use cases for these medications. Then, to ensure that the prescription is the first diabetes drug a patient receives, the patient must be observed for at least 95% of the 3 years preceding the treatment date. Although this requirement significantly reduced the cohort size, it is important for guaranteeing we are indeed studying first-line treatment decisions.

The next set of exclusion criteria was based on the eGFR measurement. Because only recent eGFR measurements were taken into consideration when making treatment decisions, we excluded patients who did not have an eGFR lab in the 6 months prior to the initial medication fill. Because the most recent measurement was used for decision-making, if the result was not available for the most recent measurement, the patient was also excluded from the cohort. After applying all these criteria, the cohort included 10,643 patients. We used this cohort to obtain descriptive statistics regarding prescribing practices and the relation to eGFR.

Finally, to characterize the policy for a specific provider, the provider must have sufficient patients. For the analysis on provider-driven variation, we limited our population to patients who were seen by providers with at least 10 patients who met the inclusion and exclusion criteria. This limited the final cohort to a total of 2,271 patients seen by 173 providers.

Figure 4. Inclusion-exclusion criteria for cohort.



Appendix B: Details on Statistical Models and Tests for Variation

We assessed whether the initial treatment class is significantly associated with eGFR level using a chi-squared test. The null hypothesis is metformin prescription and eGFR level are independent, that is, metformin prescription rates are the same across all eGFR levels. The chi-squared test statistic is 280,356.3. With 4 degrees of freedom, p-value < 0.0001. The test rejects the null, and we can conclude metformin prescription rates vary significantly across different eGFR levels.

Then, we performed a generalized likelihood ratio test (GLRT) to assess whether the initial treatment decision is significantly different across providers even after accounting for patient characteristics. This GLRT compared the data likelihood under two generalized linear models that predict treatment from patient characteristics—one with provider-specific random effects and another without provider-specific random effects. There is significant provider variation if the model with provider-specific random effects is a much better fit.

For this test, we first fitted a generalized linear model with binary indicators for heart failure and sex and restricted cubic spline features for eGFR, age, and treatment date. We chose restricted cubic splines to allow for non-linearities and avoid extrapolation issues. A restricted cubic spline is defined by knots, which can be viewed as the turning points. The curve is smooth and continuous. Between each pair of knots, the function is cubic. Before the first knot and after the last knot, the function is linear. For k knots placed at $t_1 < t_2 < \cdots < t_k$, this is achieved by adding the following features constructed from the original feature x for $i = 1, \ldots, k-2$:

$$x_i = (x - t_i)_+^3 - (x - t_{k-1})_+^3 \frac{t_k - t_i}{t_k - t_{k-1}} + (x - t_k)_+^3 \frac{t_{k-1} - t_i}{t_k - t_{k-1}}$$

Note that $a_+ = a$ if $a > 0$ and $a_+ = 0$ if $a \leq 0$. The continuous features x are normalized to have mean 0 and standard deviation 1 prior to constructing the restricted cubic spline features. Then, the new features are again placed on the same scale by dividing by $(t_k - t_1)^2$.

A standard reference for restricted cubic splines suggests tuning the number of knots but not the positions of the knots as the latter does not make much of a difference.[52] For eGFR, age, and treatment date, we tried linear functions (no restricted cubic spline features) and 3 or 4 knots set at their suggested quantiles.

Table 2. Knot positions for each feature.

eGFR	Knot 1	Knot 2	Knot 3	Knot 4
3 knots	56	85	106	
4 knots	49	77	93	112
Age	Knot 1	Knot 2	Knot 3	Knot 4
3 knots	47	62	80	
4 knots	42	57	66	83
Treatment date	Knot 1	Knot 2	Knot 3	Knot 4
3 knots	2013-07-17	2016-09-14	2019-09-06	
4 knots	2013-02-18	2015-08-29	2017-09-26	2020-10-21

We used the glm method in R to fit generalized linear models without random effects and glmer to fit generalized linear models with provider-specific random intercepts or correlated random intercepts and slopes for the eGFR features. glmer estimates random effects from the frequentist perspective. Let $X_{eGFR} \in \mathbb{R}^{n \times d_{eGFR}}$ include both the normalized eGFR level and any restricted cubic spline features created for eGFR for all n patients. X_{age} and X_{date} are defined analogously for age and date. X_{HF} and X_{sex} are binary indicators for heart failure and sex. The fixed slopes are denoted by β, and the random slope for eGFR is denoted by b_{eGFR}^{j} for provider j. The fixed intercept is denoted by γ, and the random intercept is denoted by g^{j} for provider j. p is the probability of prescribing metformin. The random effects model can be specified as

$$ln\left(\frac{p}{1-p}\right) = X_{eGFR} \times \left(\beta_{eGFR} + b_{eGFR}^{j}\right) + X_{age} \times \beta_{age} + X_{date} \times \beta_{date} + X_{HF} \times \beta_{HF} + X_{sex} \times \beta_{sex} + \left(\gamma + g^{j}\right)$$

The random effects follow the structure $\left[b_{eGFR}^{j},\ g^{j}\right] \sim \mathcal{N}(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{(d_{eGFR}+1)\times(d_{eGFR}+1)}$ is a parameter that is estimated alongside the fixed effects via maximum likelihood estimation. Σ is a symmetric covariance matrix, so the number of additional parameters introduced is $\frac{d_{eGFR}(d_{eGFR}+1)}{2}$. Once the parameters are estimated, b_{eGFR}^{j} and g^{j} can be predicted deterministically for each provider j and used to predict the probability of prescribing metformin to a particular patient. b_{eGFR}^{j} is omitted in a model with only random intercepts.

A generalized likelihood ratio test for goodness-of-fit of random effect models assesses the following hypotheses:[25]

H_0: The data distribution can be fit by a model in family $\mathcal{M}_0$ without provider-specific random effects. $\mathcal{M}_0$ is defined as the class of generalized linear models predicting the probability of prescribing metformin from eGFR, age, treatment date, sex, and heart failure with 0, 3, or 4 knots in eGFR, age, and treatment date and no provider-specific random effects. In other words, $\left[b_{eGFR}^{j},\ g^{j}\right] = 0$. $\mathcal{M}_0$ is defined by 12 parameters.

H_1: The data distribution is better fit by a model in family $\mathcal{M}_1$ with provider-specific random effects. $\mathcal{M}_1$ is defined as the class of generalized linear models with the same possible covariate sets as $\mathcal{M}_0$ and either only provider-specific intercepts or both provider-specific intercepts and provider-specific slopes for the eGFR features. In other words, $b_{eGFR}^{j} = 0$ and $g^{j} \sim \mathcal{N}(0, \sigma^2)$ or $\left[b_{eGFR}^{j},\ g^{j}\right] \sim \mathcal{N}(0, \Sigma)$. $\mathcal{M}_1$ is defined by 18 parameters.

The G-statistic is

$$G = -2 \log \frac{\max_{\theta_0 \in \mathcal{M}_0} \widehat{\mathbb{P}}\left(Y \mid X;\ \theta_0\right)}{\max_{\theta_1 \in \mathcal{M}_0 \cup \mathcal{M}_1} \widehat{\mathbb{P}}\left(Y \mid X;\ \theta_1\right)}$$

The best number of knots for eGFR, age, and treatment date are selected in each model family based on maximum likelihood. This G-statistic follows a χ^2 distribution with 6 degrees of freedom under the null hypothesis.[24] Predictions were obtained from the models in R, and the GLRT was run in Python.

Appendix C: Connections to Related Work

Past studies related to variation in clinical practice for type 2 diabetes focused on process-of-care indicators. These procedures included annual assessments of labs such as estimated glomerular filtration rate (eGFR) and hemoglobin A1c (HbA1c), annual examinations for diabetic retinopathy or neuropathy, and self-management training from nurses.[41-43] These studies found that poor adherence to these process-of-care indicators was associated with worse patient-reported outcomes. Another study looked at how well different providers were able to maintain blood glucose levels as measured by HbA1c.[44]

As mentioned in Section 2.2, our method to test for variation across providers can be viewed in the context of other approaches for identifying outlying providers.[27-34] We focused on testing whether variation exists across all providers rather than testing whether a single provider is an outlier. While the detection of outlying providers would imply that variation exists across all providers, testing for variation across all providers is feasible in more contexts. With limited data, there may be insufficient power to identify individual providers as outliers, but there may be sufficient power to identify variation exists across all providers. This difference in power is due to having few samples per provider, whereas the total number of samples across all providers is much larger. To make an analogy to a more common statistical procedure, when testing whether the means of individual groups are significantly non-zero, a t-test may not have the power to detect a non-zero mean in any single group if all groups have few samples. However, when testing whether the mean of any group in a large collection of groups is non-zero, a F-test has more power to detect the existence of a non-zero mean using the samples from all groups.

The generalized linear models fitted in our method can be used to test whether an individual provider is an outlier by running a GLRT using only samples from that provider. Our method also differs from existing approaches in this regard. To place our approach in the classification of methods reviewed by Ohlssen et al,[27] our method falls under the category of estimating and testing with an explicit alternative. The approach Ohlssen et al[27] propose also falls in this category. They fitted Bayesian models with provider-specific random effects. Then, they simulated expected outcome rates for a provider under the null hypothesis that the patients were seen by a provider who is somewhat extreme but would still be considered normal. Next, they assessed if the observed rate for a potential outlying provider was more extreme than this simulated rate. Similar to our method, this simulation also accounts for the distribution of patient features seen by a particular provider. In contrast though, our method evaluates the likelihood of the outcome for each patient, while their method evaluates the overall prescription rate for each provider. Our method is also more conservative for small per-provider sample size, which is important when labeling a provider as an outlier leads to severe consequences. To illustrate why, suppose a provider only has 10 patients and never prescribes metformin. As we found, the GLRT does not produce small p-values with only 10 samples. In contrast, their approach would run a

large number of simulations. These simulations are likely to yield non-zero metformin prescription rates because the Bayesian approach regularizes the provider-specific effects to be smaller than observed, especially when sample size is small. As a result, the observed rate of 0 metformin prescriptions will almost always be more extreme than the simulation, so the p-value will be extremely small from a large number of simulations. This exaggerates the significance that could be attained from such a small sample size.