Causal Deep Learning on Real-world Data Reveals the Comparative Effectiveness of Anti-hyperglycemic Treatments

Chinmay Belthangady
Anthem, Inc
Stefanos Giampanis
Anthem, Inc
Will Stedden
Anthem, Inc
Paula Alves
Anthem, Inc
Stephanie Chong
Anthem, Inc
Charlotte Knott
Anthem, Inc
Ivana Jankovic
Anthem, Inc
Beau Norgeot (beaunorgeot@gmail.com)
Anthem, Inc

Research Article

Keywords: Type 2 Diabetes Mellitus, Anti-hyperglycemic, Comparative Effectiveness, Machine Learning, Deep Learning, Causal Modeling, Causal Deep Learning, Real-World Data

Posted Date: June 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-648262/v1

Abstract

Type 2 Diabetes is associated with severe health outcomes, the effects of which are responsible for approximately 1/4 of total U.S. healthcare spending. Current treatment guidelines endorse a massive number of potential anti-hyperglycemic treatment options in various permutations and combinations. Personalized strategies for optimizing treatment selection are lacking. We analyzed real-world data from a nationwide population of over one million diabetics to evaluate the comparative effectiveness of more than 80 different treatment strategies ranging from monotherapy up to combinations of five concomitant classes of drugs across each of 10 clinical subgroups defined by age, insulin dependence, and number of other chronic conditions. Our causal deep learning approach developed on such data allows for more personalized recommendations of treatment selection. We observe significant differences in blood sugar reduction between patients receiving high vs low ranked treatment options and that less than 2% of the population is on a highly ranked treatment. This method can be extended to explore treatment optimization of other chronic conditions.

Introduction

Recent data from the Center for Disease Control and Prevention (CDC) estimates that approximately 13 percent of the adult population of the United States, or about 34 million people, have been diagnosed with diabetes mellitus[1]. When insufficiently managed, diabetes leads to complications including cardiovascular disease, kidney disease, neuropathy, and blindness, any of which can dramatically impair an individual's quality of life. The high incidence of diabetes and concomitant complications put a major burden on the US health care system in terms of care utilization and costs, with one recent report estimating that one of every four healthcare spending dollars in the United States can be directly attributed to diabetes[2].

Diabetes is typically managed by a combination of lifestyle interventions and pharmacological treatments. For the latter, current guidelines stipulate that unless otherwise contra-indicated, initial therapy for type II diabetes mellitus (T2DM) should be metformin[3]. If this first-line therapy is insufficient, combination therapy with anti-hyperglycemic drugs from two or more classes is suggested. There are multiple second-line choices with various risks and benefits, and a clinician may therefore need to attempt multiple treatment combinations before finding one that works for their patient. There have been efforts to determine sequential treatment of diabetes, both with data-driven informatics methods[4] and with expert-curated guidelines[5], however, both of these approaches take into account few patient-specific characteristics and can be ambiguous in suggesting the next best option for an individual patient. Even when glycemic control is achieved, there is currently no simple way to know whether a different combination might be more optimal for a given patient, either by providing greater glycemic control, by simultaneously managing comorbidities, or by providing equivalent control at a lower cost or with fewer total drugs or with less side effects. Given the complexity of potential treatments for T2DM, patients can often benefit from subspecialist management by an endocrinologist.[6] However, the current shortage of

endocrinologists is only projected to grow, thus leaving the majority of T2DM treatment to primary care practitioners (PCPs). Nonetheless, a PCP is often managing several other issues in addition to T2DM and may not be as familiar with diabetes guidelines[7] or the newest treatment modalities[6]. The enormous heterogeneity of treatment decisions observed in daily clinical practice is indicative that optimal treatment regimens have not been identified[8]. Therefore, an understanding of the real-world comparative effectiveness of pharmacological diabetes treatment strategies would be highly desirable to help guide T2DM management.

Research on comparative efficacy and effectiveness of anti-hyperglycemic drugs has been expanding. The ADOPT trial examined the relative efficacies of three monotherapies using a randomized, double-blind study examining the time to monotherapy failure over multiple years on 4360 relatively healthy patients between the ages of 30 and 75[9]. Causal analysis methods, such as the frameworks developed by Rosenbaum and Rubin[10] for observational data are now robust and widespread in clinical research. Meta-analysis[11] and Network Meta-Analysis[12] have made it possible to combine results from multiple cohorts to respectively gain effect insights from a combined pool of patients and leverage both direct and indirect comparisons between treatment arms to reduce measurement uncertainty. These approaches have been applied to a growing body of literature on the effects of T2DM pharmacological interventions. For example, Zhu *et al.* compared Glimepiride against Metformin via a meta-analysis of published results from randomized control trials (RCT) [13]. Similarly, Mearns et al. performed a network meta-analysis on 62 RCTs, with an average duration of four months, including 32,000 total participants with an average baseline HbA1C of 8%, comparing dual therapies from five drug classes added to Metformin[14]. Looking at real-world data, Ryan et al. compared Sodium-Glucose Transport Protein 2 (SGLT2) inhibitors versus non-SGLT2 inhibitor mono-therapies in a meta-analysis on observational data from four large real-world databases[15] and Vashisht et al. evaluated the effectiveness of three second-line monotherapies after metformin treatment by performing meta-analyses on observational data[4]. Although each of these studies has contributed substantially to clinical knowledge, a comprehensive understanding that reflects the realities of daily practice including diverse patients who may be on more than two classes of antihyperglycemics is still missing.

In recent years, there has been a rapid trend towards digitization in the healthcare industry. Patient medical histories are increasingly recorded in electronic format and claims adjudication systems have become streamlined and more automated. This digitization has led to an explosion in the amount of medical data available to learn from. Concurrently, there have been major advances in the fields of artificial intelligence and machine learning[16], allowing algorithms to extract complex signals from increasingly larger amounts of data. In medicine, artificial intelligence models have demonstrated human-level performance in interpreting dermatology[17] and ophthalmology[18] images. Deep neural networks trained on electronic health records (EHR) have been used to estimate the risk of disease onset[19], the risk of hospital readmissions[20], and to forecast the future health state of individuals with complex

diseases[21]. The combination of these factors has made it possible to use artificial intelligence to extract causal insights from large-scale observational studies, which was previously infeasible.

Here, we demonstrate how an approach that combines deep learning, causal inference, and network meta-analysis (NMA) can be used to estimate the real-word comparative effectiveness of combination therapies for T2DM in clinically stratified sub-populations. Using the change in levels of glycated hemoglobin (HbA1c) as the primary outcome of interest, we measure effectiveness by estimating confounder-adjusted average treatment effects (ATE) of each treatment strategy observed using a nation-wide cohort with more than 1 million patients with T2DM and at least one HbA1c measurement. Our work departs from previous research in several important ways: (i) we go beyond single or dual therapies and compare all treatment regimens observed in the data without imposing restrictions on the number of drug classes; results on combinations of up to five drug classes are reported here; (ii) we perform this analysis on 10 sub-groups stratified based on clinical variables to make the rankings more personalized; (iii) we use a recently developed deep-learning-based propensity-score model for causal analysis that scales well to large multi-arm observational studies; and (iv) we perform sensitivity analysis on held-out data in order to assess the extent to which the comparative effectiveness rankings were meaningful and broadly generalizable. These rankings for combination therapies can complement/enhance guideline-based practice and help clinicians make personalized data-driven decisions when formulating treatments for their patients.

Results

Inclusion Criteria

Temporal snapshots of patient trajectories beginning at the start of a patient's health history and ending with each subsequent HbA1c lab measurement were generated for each person from their five-year history of medications, diagnoses, procedures, and relevant laboratory values to assess the treatment strategies and resulting causal effect calculations (Figure 1).

Characteristics of the Study Population

Patient variable values (Table 1) show that the study cohort had a mean age of 55, with a baseline HbA1c, EGFR and Creatinine of 10.4%, 95 mL/min/1.72m2, and 0.9 mg/dL respectively. Neighborhood incomes ranged significantly with a median of fifty-five thousand dollars annually. White, Black, and Asian populations were well represented with Whites being in the majority. As expected within a population of patients with significantly elevated blood sugar, a wide range of comorbid conditions were present, with obesity, heart disease, COPD, and renal disease being most prevalent. The test cohort statistically matched with the study cohort on all variables.

Table 1. Summary Statistics of Patient Cohort. (a) Baseline characteristics of the training and test set for confounding variables. Top two tables display mean and standard deviation of continuous variables. Bottom two tables display proportion present for binary variables. Left panels display statistics at the level of sample snapshot, whereas right panels display counts at level of patient. (b) Summary Statistics per Stratum. Training dataset was split in 10 segments based on age, prior use of insulin and comorbidity index values. The number of available treatments varies for each group and is typically higher for larger cohort sizes. ZCTA: Zip Code Tabulation Areas from 2017 American Community Survey; EGFR: Estimated Glomerular Filtration Rate; ASCVD: Atherosclerotic cardiovascular disease.

Feature	Train: Mean ± Std Dev	Test: Mean ± Std Dev
Baseline HbA1c	10.5 ± 1.4	10.5 ± 1.4
ZCTA % Median Income	58651 ± 22589	58651 ± 22504
ZCTA % White	61.8 ± 23.3	61.8 ± 23.2
ZCTA % Native Am.	0.5 ± 0.9	0.5 ± 0.9
ZCTA % Black	16.5 ± 20.2	16.5 ± 20.1
ZCTA % Asian	7.6 ± 10.7	7.6 ± 10.7
Age	55.2 ± 10.6	55.2 ± 10.5
Creatinine Lab	0.9 ± 0.3	0.9 ± 0.3
EGFR Lab	94.6 ± 23.1	94.4 ± 23.0

Feature	Train: % Present	Test: % Present
Cancer	4.6	4.7
Metastatic Carcinoma	0.6	0.6
Connective Tissue Disease	2.2	2.1
Dementia	0.8	0.8
Peptic Ulcer Disease	1.3	1.3
Gastroparesis	0.5	0.5
Paraplegia and Hemiplegia	0.9	0.9
Cerebrovascular Disease	7.6	7.4
Chronic Pulmonary Disease	17.1	17.3
Peripheral Vascular Disease	7.7	7.6
ASCVD	14.5	14.4
Heart Failure	11.1	11.3
Renal Disease	11.3	10.9
End-Stage Renal Disease	0.6	0.6
Chronic Kidney Disease	12.9	12.4
Dialysis	0.3	0.3
Obesity	39.7	40.3
Hypoglycemia	2.3	2.3
Diabetes w Complications	40.5	40.6
Diabetes w/o Complications	95.4	95.5
Mild Liver Disease	11.2	11.1
Severe Liver Disease	0.5	0.5
Fructosamine Test	1.0	0.8

Characteristics of Treatment Pathways

To account for potential differences in patient needs and treatment goals, patient snapshots were assigned to one of ten clinical subgroups on the basis of each patients' age, insulin dependency status, and number of additional chronic health conditions at the time of each index event. The number of snapshots present as well as the number of treatment strategies that subgroups were exposed to tended to decrease with age and disease burden. (Table 2). Though these trends did not decrease monotonically. Subgroup A had the largest number of snapshots (n = 53,532) and treatments (n = 69) and Subgroup H had the fewest snapshots (n = 2,931) and number of treatments (n = 16). There were considerably more insulin naïve snapshots (n=99,116) than insulin dependent (n=24,184). Across insulin status, subgroups <65 years of age had an average of 81,338 snapshots and 162 treatments while those >65 had 24,674 snapshots and 121 treatments. The number of distinct treatment strategies observed in a subpopulation was correlated to the number of patient snapshots present, the larger the population the more unique treatment strategies were observed.

Table 2: Cohort Definitions and Characteristics of Treatment Pathways of T2DM Patient Subgroups. CCI: (unweighted) Charleston Comorbidity Index; Snapshots represent a patient's history prior to an HbA1c laboratory index event as defined in Figure 1. Number of Treatments refers to the number of unique treatment strategies observed in the subgroup as detailed in Figure 1.

Subgroup	Insulin Status	Age	CCI	Number of Snapshots	Number of Treatments
A	Naïve	<65 years	<=2	53532	69
B			>2 & <5	10734	43
C			>=5	17072	50
D		>=65 years	<5	10447	43
E			>=5	7331	37
F	Dependent	<65 years	<=2	7180	35
G			>2 & <5	7177	30
H			>=5	2931	16
I		>=65 years	<5	3471	18
J			>=5	3425	23

Causal Modeling of Optimized Treatment Ranking by Subpopulation

Significant differences existed in the underlying covariate distributions between treatment and control arms in the observational data but were successfully balanced through the BCAUS methodology (Supplemental Figure S1). The confounder-adjusted causal effect of each treatment strategy on each

clinical subpopulation was calculated independently for each subgroup (Figure 2). League plots were produced (Supplemental Figures S12-S21), and treatment strategies were ranked based on causal reduction in HbA1c.

The top 10 most effective treatment strategies for each subgroup (Table 3), revealed that the highest ranked treatment strategy was unique to each subgroup. Indeed, while trends were certainly present, no two clinical subpopulations had a single treatment strategy share the same rank. GLP-1's and metformin, both known to be highly efficacious for blood glucose control[22], are the only classes to appear as monotherapies in any group's top ten ranked treatments, though they only appear for half of the subgroups and never higher than position five. While a monotherapy is not the top-ranked treatment for any subgroup, the rankings clearly show that simply adding more drug classes to a patient's regime is not uniformly best for blood sugar reduction. Hypoglycemic classes known to provide secondary cardioprotective benefits, such as SGLT2's and GLP-1's, feature prominently in the top ten choices for each subgroup, though trends in which is most effective varies by subgroup. Nevertheless, this finding may indicate that there is no need for a trade-off between glucose control and cardio protection and that both can be optimized for simultaneously. Interestingly, in the insulin-naïve groups, insulin-containing regiments tend to rank poorly, suggesting poor real-world effectiveness of a medication known to have high efficacy in Randomized Controlled Trials[22]. A complete listing of the rankings for all treatment strategies across all subpopulations can be found in the Supplement (S22) as well as the measured causal effects, confidence intervals, and observed population sizes for each strategy in each population (S2-S11).

Table 3. Causal Modeling Derived Treatment Strategy Rankings by Clinical Subgroups Treatments are ranked according to their SUCRA scores for each clinical subgroup. Left panel shows Clinical Subgroups where prior treatment did not contain Insulin. Right panel shows Clinal Subgroups where prior treatment contained Insulin. Abbreviations used: INS= Insulin, GLP-1 = Glucagon-Like Peptide-1 Receptor Agonist; SULF: Sulfonylureas; METF = Metformin; MEGL: Meglitinide; DPP-4 = Dipeptidyl Peptidase 4 Inhibitor; TZD = Thiazolidinedione; SGLT2 = Sodium-Glucose Transport Protein 2 Inhibitor; AGI = Alpha-Glucosidase Inhibitor.

	Insulin Naïve Clinical Subgroups					Insulin Dependent Clinical Subgroups				
Treatment Rank	A	B	C	D	E	F	G	H	I	J
1	METF + SGLT2 + TZD	DPP4 + METF + SGLT2 + SULF + TZD	METF + SGLT2	BASAL INS + DPP4 + METF	METF + SGLT2 + SULF	BASAL INS + GLP1 + METF + SGLT2	BASAL INS + GLP1 + SGLT2	BASAL INS + GLP1 + METF + SGLT2	BASAL INS + GLP1 + METF	BASAL INS + GLP1 + METF
2	BASAL INS + DPP4 + METF + SGLT2	GLP1 + SGLT2	GLP1 + METF	DPP4 + METF + TZD	DPP4 + METF + SULF	BASAL INS + GLP1 + SGLT2	BASAL INS + DPP4 + METF + SGLT2	BASAL INS + METF + SGLT2	METF + MIXED INS	BASAL INS + DPP4 + METF + SULF
3	GLP1 + METF + SGLT2	GLP1 + METF + SGLT2	METF + TZD	GLP1 + METF	METF + SULF + TZD	BASAL INS + BOLUS INS + GLP1 + METF	BASAL INS + GLP1 + METF + SGLT2	BASAL INS + GLP1 + METF	BASAL INS + DPP4 + METF	BASAL INS + BOLUS INS + METF + SULF
4	DPP4 + SGLT2	DPP4 + METF + SGLT2	GLP1 + METF + SGLT2 + SULF	BASAL INS + DPP4 + METF + SULF	METF + SULF	DPP4 + METF	BASAL INS + GLP1 + SULF	BASAL INS + BOLUS INS + GLP1 + METF	BASAL INS + GLP1	BASAL INS + BOLUS INS + DPP4
5	DPP4 + METF + SGLT2 + SULF + TZD	GLP1 + METF + SGLT2 + SULF	GLP1 + METF + SGLT2	DPP4 + METF + SGLT2	DPP4 + METF	BASAL INS + DPP4 + METF + SGLT2	BASAL INS + GLP1 + METF	BASAL INS + GLP1 + METF + SULF	BASAL INS + METF + SULF	METF
6	METF + SGLT2	BASAL INS + GLP1	GLP1 + SULF	SULF + TZD	METF + TZD	BASAL INS + METF + SGLT2 + SULF	BASAL INS + BOLUS INS + SGLT2	BASAL INS + BOLUS INS + GLP1	BASAL INS + DPP4 + METF + SULF	BASAL INS + BOLUS INS + GLP1
7	BASAL INS + GLP1 + METF	GLP1 + METF	METF + SGLT2 + SULF	DPP4 + METF	SULF + TZD	BASAL INS + BOLUS INS + GLP1	BASAL INS + DPP4 + GLP1 + METF	BASAL INS + BOLUS INS + METF + SGLT2	BASAL INS + BOLUS INS + METF	BASAL INS + METF
8	DPP4 + METF + SGLT2	DPP4 + METF + SGLT2 + SULF	BASAL INS + GLP1 + METF	GLP1	BASAL INS + METF + SULF	GLP1 + METF	BASAL INS + METF + SGLT2 + SULF	BASAL INS + DPP4 + METF + SGLT2	BASAL INS + METF	BASAL INS + DPP4 + METF
9	GLP1 + METF	METF + SGLT2	GLP1 + METF + SULF	METF + SGLT2 + SULF	BASAL INS + DPP4 + SULF	BASAL INS + GLP1 + METF	BASAL INS + GLP1 + METF + SULF	BASAL INS + DPP4 + METF + SULF	METF	BASAL INS + BOLUS INS
10	GLP1 + SGLT2 + SULF	GLP1	DPP4 + METF + SULF + TZD	GLP1 + METF + SULF	METF	BASAL INS + GLP1 + METF + SULF	BASAL INS + GLP1	BASAL INS + GLP1	BASAL INS + BOLUS INS	BASAL INS + GLP1

Causal Effect of Treatment Ranking Group on Blood Sugar Reduction

When treatments were divided into three groups based on rank (high:1-3, middle: 4-10, low: below 10) significant differences in patient outcomes between highly ranked treatments and all other choices were observed for nine out of ten subgroups (Figure 3). The differences were significant clinically as well as statistically, persisted even after controlling for differences between patients that received highly ranked choices versus others, and generalized extraordinarily well to the test subgroups. A sensitivity analysis revealed a consistent dose response relationship between highly ranked, middle ranked, and low ranked treatment strategies (Supplement Figure S23).

Ranking Group Prescription Patterns in Real-World Observational Data

After determining that treatment strategies with higher ranks caused larger reductions in blood sugar, we examined the study population to measure the distribution of high, medium, and low ranked treatment strategies provided to patients in each clinical subgroup (Table 4). We found that a less than two percent of patients were provided a highly ranked treatment choice and the vast majority were on low ranked treatment options. Across all subgroups, the average treatment rank per snapshot was 28. We observed the lowest rates of concordance among the younger or relatively healthier subgroups. The likelihood of switching treatments between snapshots was 47 percent. When patients switched treatments, 49 percent of those switches lead to a new treatment with a better rank (with an average improvement of ten positions of rank), while fifty-one percent of switches lead to treatments with a worse rank for the patient (with an average decrease in rank of eleven). Overall, changes did not lead to improvements, the mean change in treatment rank across all changes was decrease of 1.

Table 4: Real-world Observations of subgroup Treatment Ranking Concordance

Concordant percent values indicate the percent of each group that received a drug recommended for their subgroup as being ranked high (rank 1-3), medium (rank 4-10) or low (rank 11-n). One percent of patients in Subgroup A where on a highly ranked treatment, seven percent where on a middle-ranked treatment, ninety-two where on a low ranked treatment.

Subgroup	Number of Snapshots	% Concordant Top 3	% Concordant Top 4-10	% Concordant (>10 or none)
A	53532	1.0%	6.8%	92.2%
B	10734	2.3%	6.6%	91.1%
C	17072	1.5%	9.1%	89.4%
D	10447	1.5%	6.8%	91.7%
E	7331	3.6%	7.4%	89.0%
F	7180	3.8%	9.0%	87.2%
G	7177	2.5%	11.0%	86.5%
H	2931	8.5%	36.0%	55.5%
I	3471	9.8%	37.5%	52.7%
J	3425	3.7%	25.7%	70.6%

Discussion

In this study, we examined the anti-hyperglycemic treatment strategies over a five-year period in a nationwide cohort of patients with type II diabetes whose index HbA1C was 9% or higher. We found over 80 different strategies of drug class combination, ranging from monotherapy to combinations of five distinct drug classes. This enormous heterogeneity persisted even after accounting for age, number of comorbidities, and status of insulin dependence. We then performed a network meta-analysis using deep causal models on the cohort's observational data to rank treatment strategies for the ten clinical subgroups based on efficacy in lowering HbA1C. We found that the rankings differed between each of the subgroups, were sensitive even for infrequently observed treatments, and that they generalized well to patients in our test set. Highly ranked treatments were clinically and statistically significantly better than other choices. There were considerable differences between which treatments where best for each of the clinical subgroups and cardioprotective drug classes featured prominently, though the specific class and combination was subgroup dependent. We observed that fewer than 2% of patients were on an optimal treatment and that treatment switches, when they occur, move patients into worse strategies as often as they result in better strategies.

That treatment strategies for diabetes mellitus are massively heterogenous is well known. While Hripcsak *et al*.[8] observed that 10% of diabetic patients in their international study had treatment pathways that were unique specifically to that individual, the authors hazard that the variability was not a sign of personalization but rather that "it may point to a failure of the field to converge on an effective treatment". To our knowledge, no previous studies have examined comparative efficacy between all observed treatment strategies in multiple clinically relevant real-world sub-groups. The mono-therapy[9] and dual-therapy[14] results found in this work are reasonably consistent with prior published results, that were limited to those two options. However, differences in cohort sizes and inclusion criteria make direct comparisons difficult. For example, the ADOPT trial contained fewer than 5,000 participants and excluded patients with more advanced disease that would not be eligible for monotherapy. Mearns *et al.* combined

all patients from dual-therapy trials, regardless of age, disease severity, or comorbid conditions, making it impossible to directly compare results to our clinically stratified subgroups.

We found no treatment strategies that were in violation of current standard of care guidelines provided by UpToDate. While it is reassuring that guidelines are followed, it also reinforces the concern that guidelines may be insufficient at guiding treatment choices for blood sugar reduction. Instead of contradicting current best practices, our findings provide clarity on which strategies may be best when the guidelines provide many to choose from. It is perhaps not surprising that so few patients are on a treatment that may be optimal for them. Patients with highly elevated HbA1C are, by definition, those who have not yet found a treatment strategy sufficient to control their blood sugar. Additionally, since the current guidelines unilaterally suggest a progressive approach from mono to dual therapy, followed by experimentation within dual-therapy before adding more drug classes, there is a diffusion effect that necessitates a long time until sufficient experimentation has occurred to identify a good strategy for many patients.

Although we believe that this work has potentially significant clinical value, and may even provide the signal necessary for the field to identify effective treatments that Hripcsak *et al*. have called for[8], there are two important limitations. First, we have defined effectiveness exclusively on the grounds of HbA1C reduction. This choice is reasonable given that HbA1C as a surrogate endpoint is the most used outcome for clinical trials and that our study population is comprised of patients with highly elevated blood sugar, however, there are additional clinical endpoints, particularly those related to cardiovascular outcomes, that are relevant for patients with diabetes. While many of these clinical endpoints are strongly correlated with HbA1c, conclusions about how treatment protocols that utilized the rankings derived here would impact these endpoints cannot be drawn from this work. That said, the strong presence of treatments with secondary protective effects in the top ranked choices may indicate that while the effect on cardiovascular and renal outcomes is not captured here, never-the-less the treatment choices that are most effective for glycemic control are top choices for secondary protective effects as well. A second limitation is that we calculated impact on HbA1C only at the follow-up measurement after a treatment was assigned (a 6-month median window). This time period is sufficient to see the effects of medication changes considering the half-life of hemoglobin, including the slower-acting thiazolidinediones[22], but may not be perfectly indicative of long-term trends and tolerability. Given the length of time it takes for diabetes-related complications to develop, causal attribution to specific treatment strategies is clouded by the many patient-related factors that can change over such a length of time, such as the course of treatment. However, a longer study period that tracks clinical endpoints as well as laboratory endpoints is desirable and could be feasible as datasets such as this grow over time. Future studies could leverage our methodology to define effectiveness by distance from a specified blood sugar value for each clinical sub-population, instead of absolute HbA1c reduction. Alternatively, maximal risk reduction for microvascular or macrovascular outcomes could be used as the endpoint instead of HbA1c. However, we

believe such investigations would be most robustly served by a prospective study tracking the impact on multiple clinical endpoints from prescribing high-ranked treatment strategies to achieve subgroup-specific targets.

Although the treatment rankings presented in this work are fixed, they can supplement guidelines to support many different approaches to decision making. For example, based on patient needs and clinician preferences, some may choose not to prescribe the highest ranked treatment but instead the highest-ranked option that involves the smallest change from the current regimen. Alternatively, patients for whom compliance may be a concern, selecting a treatment that optimizes rank with the fewest number of total drugs would be an option. For every highly ranked strategy in our study that contained many different drug classes, there was usually a simpler combination with nearby rank. The enormous variety of ways in which these comparative efficacy rankings could be utilized may be best leveraged by software with a performant, intuitive user interface to return the optimized results for a given patient target. Such software could also provide additional metrics captured in this study, like the number of patients observed on each treatment strategy and the clinical and demographic parameters associated with each person, which are not possible to display in the context of individual patients within a manuscript like this. Additional convenience functions, such as the removal of contraindicated treatments from the rankings list for each individual patient may be desired.

Taken together, these findings have important implications for personalizing type 2 diabetes mellitus recommendations without any tradeoff for cardiovascular protection. In real-world terms, we may be to provide more effective, more personalized treatment strategies to 98% of established diabetic patients immediately, although future work will have to confirm these hypotheses. We believe that the approach outlined here represents a concrete step towards a functional learning healthcare system, and that it is immediately extensible to other conditions beyond diabetes mellitus that have complex pharmacological treatment patterns such as hypertension, asthma, chronic obstructive pulmonary disease, depression, and congestive heart failure. By forestalling adverse events that arise from unmanaged chronic diseases, such learning systems could greatly reduce patient suffering and lead to significant reductions in healthcare costs.

Methods

Study Cohort Definition and Data Preparation

We analyzed electronic health records from 56.4 million members in Anthem Inc's subscriber population between December 1st, 2014 and January 1st, 2020. The records included approximately five billion insurance claims (for diagnoses, procedures, and drug prescriptions or refills) as well as lab test results for the associated patients. Clinical filters were designed to distinguish between major sub-types of

diabetes, and patients with Type I, anyone under 18 years of age, or Gestational Diabetes were excluded from the study (Figure 1a). We also excluded individuals with histories of diabetes ketoacidosis, cystic fibrosis, or solid-organ transplants as a safety precaution because they are highly complex patients who would clearly benefit from subspecialist care and the rankings developed herein are targeted towards PCPs managing typical patients with Type II Diabetes Mellitus (T2DM). We removed patients with recent HbA1c's below 9% to focus on those who were clearly eligible for treatment strategies beyond first line. This resulted in a study population of 1.2 million individuals with T2DM.

The health status of any individual evolves with time. Since the study period in our work spanned several years, to properly account for this evolution, we split each individual's health history into a series of temporal snapshots as shown in Fig. 1(b). Each snapshot terminated in a pair of HbA1c lab measurements. The time period between the two labs was considered as the observation period. The age of the individual in a particular snapshot and any clinical covariates that were treated as confounders were measured as of the date of the first lab of the pair. Individuals with only a single HbA1c lab reported were excluded. Only snapshots where the observation period was between 90 and 365 days were retained, and the rest were excluded, resulting in a final study population of 141,625 patient snapshots. As shown in Fig 1(c), an individual was considered as treated by a particular anti-hyperglycemic drug at the time of a particular HbA1c lab event if it was prescribed prior to that lab and if the number of days of supply plus a grace period of 30 days (for non-adherence) extended past the lab date. When multiple such drugs existed, the individual was considered as treated by the combination of these drugs. Diabetes drugs were identified only by their class names (e.g. SGLT2 inhibitors, sulfonylureas, etc.) and non-diabetes drugs were excluded. Prior treatment was the regimen used to treat the individual in the period prior to the observation window between the two labs. All further analysis was performed on this pseudo-population of patient snapshots.

Many clinical and social factors are known to be associated with diabetic treatment selection and HbA1c outcomes. For example, kidney function as well as the presence of various comorbid conditions may result in contraindications for certain antihyperglycemic drug classes and may also influence the HbA1c value that the prescribing clinician targets for an individual. Additionally, Social Determinants of Health (SDoH) such as patient race, income, and location are known to influence both treatment selection and health outcomes. In order to control for these confounding factors so that an accurate estimate of the causal effect of treatment strategies could be obtained, we included all comorbidities present in the history of each patient using diagnostic definitions defined by the Charslon Comorbidity Index (CCI)[23], as well as the most recent EGFR and Creatinine values at the time of each snapshot. Race is known to be reported at very low levels both within EHRs and claims data. Accordingly, we chose to use census-derived data on the racial and economic profiles of each patient's neighborhood using zip codes. We believe that these are weak surrogates for true SDoH markers, but that to use them is still significantly

better to ignore SDoH completely from large scale clinical studies. Table 1 provides the summary statistics of all covariates that were treated as confounders for causal inference.

Snapshots were stratified into ten clinical sub-populations that were defined based on age, number of additional chronic diseases, using disease definitions from the unweighted Charslon Comorbidity Index (CCI)[23], and prior insulin use as evidenced by the presence or absence of insulin treatment as of the first HbA1c lab of the observation period. Summary statistics for the 10 segments are shown in Table 1(b). For comparison all summary statistics in Table 1 are reported at the individual as well as patient snapshot levels.

Causal Inference Modeling

We considered several methods for the causal inference analysis presented here. Because there are multiple possible combinations of treatments, the number of head-to-head comparisons that need to be performed is extremely large. Propensity score matching[10] or weighting[24] methods are widely used for observational studies but are considered "do-it-yourself,"[25] in that the propensity score model must be checked for correct specification after it is trained, and, when incorrectly specified, it has to be retrained by modifying model parametrization or feature engineering. Automated methods like Bayesian Additive Regression Trees[26] have yielded good performance on benchmark datasets[25], but rely on Monte Carlo sampling and are therefore prohibitively slow for the number of comparisons necessary in this study. We recently introduced[27] a technique called BCAUS (Balancing Covariates Automatically Using Supervision) that scales well to massive multi-arm studies. BCAUS consists of a neural-network propensity model that is trained using a joint loss given by

$$\mathcal{L}_{TOTAL} = \mathcal{L}_{BCE} + \nu\mu\mathcal{L}_{BIAS}.$$

The first term, $\mathcal{L}_{BCE}$ is a binary cross-entropy loss which penalizes incorrect treatment assignment, while the second, $\mathcal{L}_{BIAS}$ is a loss term which explicitly tries to minimize imbalance between inverse probability weighted covariates. Details of the training process are described in Supplementary Materials and a comparison with other state-of-the-art neural-network-based methods on benchmark datasets has been described elsewhere[27]. For each pairwise comparison between diabetes treatments, a separate BCAUS model was trained. The propensity score outputs of trained models were used to estimate average treatment effects using Inverse Probability of Treatment Weighting (IPTW). A bootstrapping procedure was used to compute standard errors and confidence intervals. The input data for NMA consisted of the estimated ATEs and standard errors.

Network Meta-Analysis

An average treatment effect value measured via a *direct* causal comparison between two treatments has to be consistent with values that are *indirectly* estimated (under the transitivity assumption) by comparing each treatment of the pair with intermediary treatments and then computing differences. Separate network graphs were constructed for the 10 clinical subgroups where every treatment node was connected with every other treatment node. Edges representing observational studies where all confounding covariates were not balanced were trimmed and Bayesian NMA was performed over the resultant graph. Heterogeneity in the treated populations was accounted for by using a random-effects, (hierarchical) model, uninformative priors were set, and a Markov Chain Monte Carlo (MCMC) sampling procedure was used to construct posterior distributions of average treatment effect values for all treatment pairs. To determine relative ranks, samples were drawn from the posterior predictive distributions of average treatment effects of all treatments compared against metformin, which was set as the baseline treatment. For each draw, treatments were ranked in ascending order of average treatment effect values (i.e. higher ranks for more negative values), and a mean rank was computed for each treatment across all draws. This mean rank was normalized to compute the SUCRA score. Treatments were ranked in descending order of SUCRA scores such that the treatment that reduced HbA1c by the largest amount relative to metformin had the highest rank. This ranked list of treatments was returned to all members of the subgroup. Further details of the training procedure are available in Supplementary Materials.

Ranking Algorithm Development

A schematic of the workflow used to generate treatment rankings is shown in Fig. 2. Patient snapshots were split randomly within each sub-group into training (80%) and hold-out (20%) sets such that in each set the relative sizes of the 10 clinical subgroups was approximately the same and that no patient was present in both. In each subgroup of the training set, all treatments combinations with a cohort size greater than 35 were chosen and head-to-head average treatment effect estimation were run comparing each treatment with every other treatment within that subgroup. Average treatment effect estimation was performed using a neural-network-based propensity-score model described in Methods. For the 10 subgroups considered here, a total of 15,198 neural networks were trained, one for each observational study. The propensity-score models were used to estimate pairwise average treatment effects and associated 95% confidence intervals. These values were used to construct a densely connected network graph, where each node represented a treatment and edges connecting two nodes represented the case-control study between the respective treatments. Bayesian network meta-analysis was performed (see Methods) to compute network-synthesized meta-analytic average treatment effects and 95% credible intervals compared against a baseline treatment. The baseline treatment was set to metformin since it is the first-line therapy for T2DM and all ATEs were measured relative to metformin. Samples were drawn from the posterior predictive distributions of the trained model to compute Surface Under the Cumulative

RAnking curve (SUCRA) scores for all treatments (see Supplement). Treatments were sorted in decreasing order of this score to generate comparative effectiveness rankings.

Standardized difference plots for confounder adjustment analysis are shown in Supplementary Fig. S1. Forest plots for network meta-analysis and league tables for each subgroup are shown in Supplementary Figs. S2-S21.

Ranking Validation Procedure

To investigate the degree to which the rankings generalized to new patients while generating an estimate of the improvement to HbA1c over existing practices if rankings were used to guide treatment decisions, we retrospectively compared outcomes between patients whose physicians happened to have prescribed a top-3 ranked treatment choice for them versus selecting any other treatment option. Snapshots in each clinical subgroup were divided into concordant cohorts (where the actual prescribed treatment matched one of the top-3 recommendations) and non-concordant cohorts (where a patient was provided any treatment ranked 4 or lower). Differences in the mean change in HbA1c between and the non-concordant cohorts were calculated for all subgroups for both training and test populations. If the difference in means was found to be statistically significant, an additional confounder-adjusted observational study was performed between the cohorts to measure whether the differences in means was directly attributable to the differences in treatment strategy ranks.

To further investigate if the rankings demonstrate an internally consistent effect, we performed a sensitivity analysis by splitting patient snapshots of each subgroup in the training dataset into three concordance cohorts: i) the “top” cohort is concordant with treatments ranked 1-3; ii) the “middle” cohorts is concordant with treatments ranked 4-10; and iii) the “bottom” cohort is concordant with treatments ranked 11 and below. We estimated confounder-adjusted average treatment effect values, comparing the top versus bottom groups and the middle versus bottom groups. If the ranks are internally consistent, we would expect a titration effect where-in the top outperforms the middle and the middle outperforms the bottom.

Declarations

This manuscript reports findings that were obtained as a part of Healthcare Operations quality improvement.

Funding and Conflicts of Interest

All authors were employed by Anthem, Inc.

References

1 Center for Disease Control and Prevention. National Diabetes Statistics Report. (Atlanta, GA, 2020).

2 American Diaetes Association. Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care* **41**, 917–928 (2018).

3 Davies, M. J. *et al.* Management of Hyperglycemia in Type 2 Diabetes, 2018. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care* **41**, 2669, doi:10.2337/dci18-0033 (2018).

4 Vashisht, R. *et al.* Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin: Analysis From the Observational Health Data Sciences and Informatics Initiative. *JAMA Network Open* **1**, e181755-e181755, doi:10.1001/jamanetworkopen.2018.1755 (2018).

5 Association, A. D. Pharmacologic Approaches to Glycemic Treatment:Standards of Medical Care in Diabetes—2019. *Diabetes Care* **42**, S90, doi:10.2337/dc19-S009 (2019).

6 Setji, T. L., Page, C. y., Pagidipati, N. & Goldstein, B. A. Differences in Achieving Hba1C Goals Among Patients Seen by Endocrinologists and Primary Care Providers. *Endocrine Practice* **25**, 461-469, doi:https://doi.org/10.4158/EP-2018-0405 (2019).

7 Beaser, R. S. *et al.* Coordinated Primary and Specialty Care for Type 2 Diabetes Mellitus, Guidelines, and Systems: An Educational Needs Assessment. *Endocrine Practice* **17**, 880-890, doi:https://doi.org/10.4158/EP10398.OR (2011).

8 Hripcsak, G. *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proceedings of the National Academy of Sciences* **113**, 7329-7336, doi:10.1073/pnas.1510502113 (2016).

9 Kahn, S. E. *et al.* Glycemic Durability of Rosiglitazone, Metformin, or Glyburide Monotherapy. *New England Journal of Medicine* **355**, 2427-2443, doi:10.1056/NEJMoa066224 (2006).

10 Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41-55 (1983).

11 Chalmers, T. C., Matta, R. J., Smith, H. & Kunzler, A.-M. Evidence Favoring the Use of Anticoagulants in the Hospital Phase of Acute Myocardial Infarction. *New England Journal of Medicine* **297**, 1091-1096, doi:10.1056/nejm197711172972004 (1977).

12 Lumley, T. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* **21**, 2313-2324, doi:https://doi.org/10.1002/sim.1201 (2002).

13 Zhu, H. *et al.* Comparative efficacy of glimepiride and metformin in monotherapy of type 2 diabetes mellitus: meta-analysis of randomized controlled trials. *Diabetology & Metabolic Syndrome* **5**, 70, doi:10.1186/1758-5996-5-70 (2013).

14 Mearns, E. S. *et al.* Comparative efficacy and safety of antidiabetic drug regimens added to metformin monotherapy in patients with type 2 diabetes: a network meta-analysis. *PloS one* **10**, e0125879 (2015).

15 Ryan, P. B. *et al.* Comparative effectiveness of canagliflozin, SGLT2 inhibitors and non-SGLT2 inhibitors on the risk of hospitalization for heart failure and amputation in patients with type 2 diabetes mellitus: a real-world meta-analysis of 4 observational databases (OBSERVE-4D). *Diabetes, Obesity and Metabolism* **20**, 2585-2597 (2018).

16 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).

17 Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118, doi:10.1038/nature21056 (2017).

18 Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402-2410, doi:10.1001/jama.2016.17216 (2016).

19 Li, L. *et al.* Disease risk factors identified through shared genetic architecture and electronic medical records. *Sci Transl Med* **6**, 234ra257-234ra257, doi:10.1126/scitranslmed.3007191 (2014).

20 Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* **1**, 18, doi:10.1038/s41746-018-0029-1 (2018).

21 Norgeot, B. *et al.* Assessment of a Deep Learning Model Based on Electronic Health Record Data to Forecast Clinical Outcomes in Patients With Rheumatoid Arthritis. *JAMA Network Open* **2**, e190606-e190606, doi:10.1001/jamanetworkopen.2019.0606 (2019).

22 Noble, J., Baerlocher, M. O. & Silverberg, J. Management of type 2 diabetes mellitus. Role of thiazolidinediones. *Can Fam Physician* **51**, 683-687 (2005).

23 Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases* **40**, 373-383 (1987).

24 Rosenbaum, P. R. Model-based direct adjustment. *Journal of the American Statistical Association* **82**, 387-394 (1987).

25 Dorie, V., Hill, J., Shalit, U., Scott, M. & Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science* **34**, 43-68 (2019).

26 Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217-240 (2011).

27 Belthangady, C. S., Will; Norgeot, Beau. Minimizing Bias in Massive Multi-Arm Observational Studies with BCAUS: Balancing Covariates Automatically Using Supervision. *Submitted* (2021).

Figures

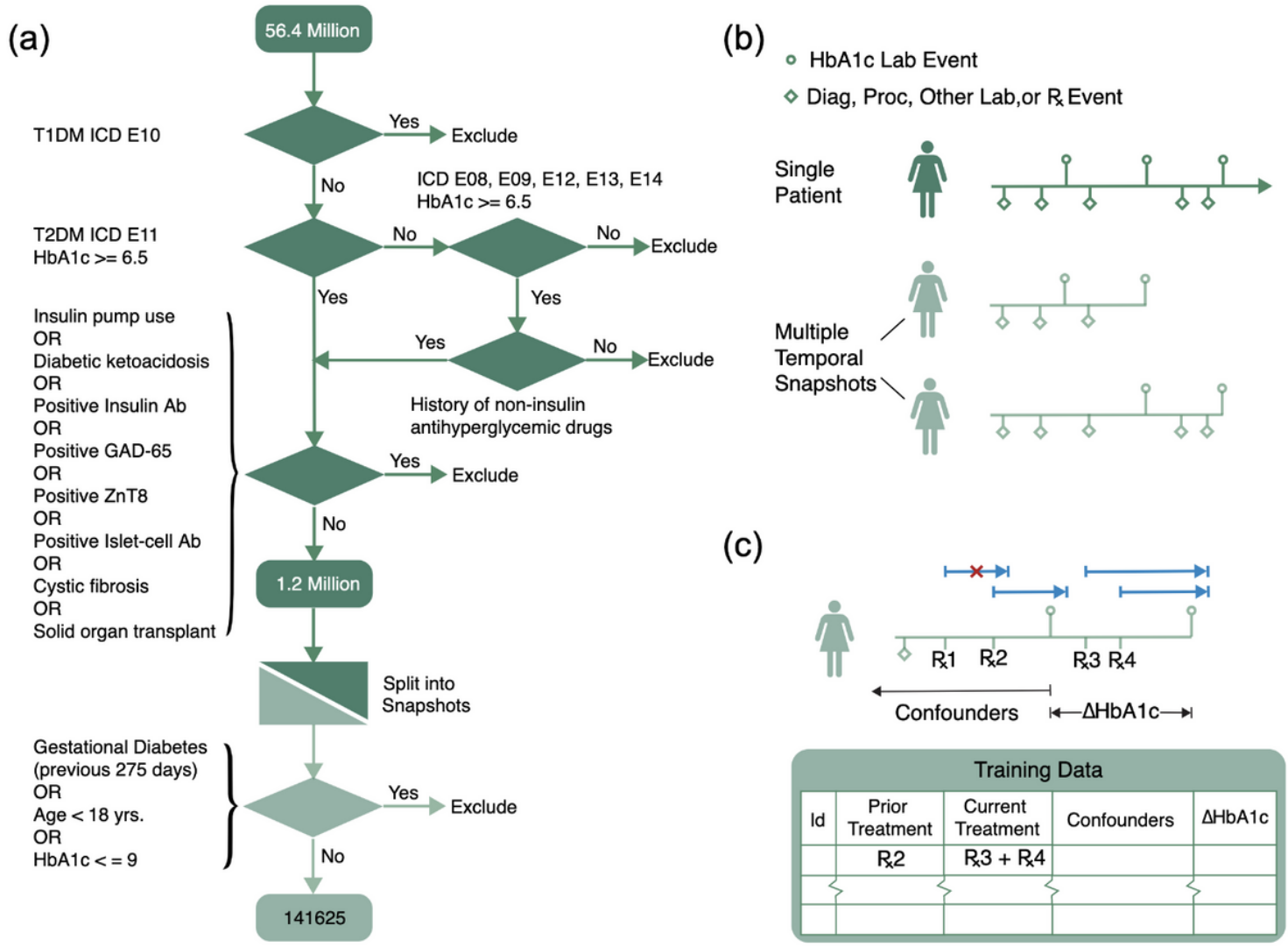

Figure 1

Study Cohort Definition and Data Preparation (a) Figure 1. Study Cohort Definition and Data Preparation (a) Clinical filters were designed to identify patients with Type 2 Diabetes (1.2 Million individuals) and retain only those with well-established disease. (b) Each patient's health history was split into a series of temporal snapshots, beginning at the start of a patient's health history and ending with each subsequent

HbA1c lab measurement. Only snapshots where the duration between the lab pairs was between 90 to 365 days were retained and the rest were excluded, resulting in a final study population of 141,625 patient snapshots. All further analyses were conducted on the snapshots. (c) A patient was considered to have been treated by a particular anti-hyperglycemic drug at the time of a given HbA1c lab event if it was prescribed prior to the lab and if the number of days supply (blue arrows) extended past the lab date. When multiple such drugs existed, the individual was considered treated by the combination of these drugs. Prior treatment was the regimen used to treat the individual in the period prior to the observation window between the two labs.

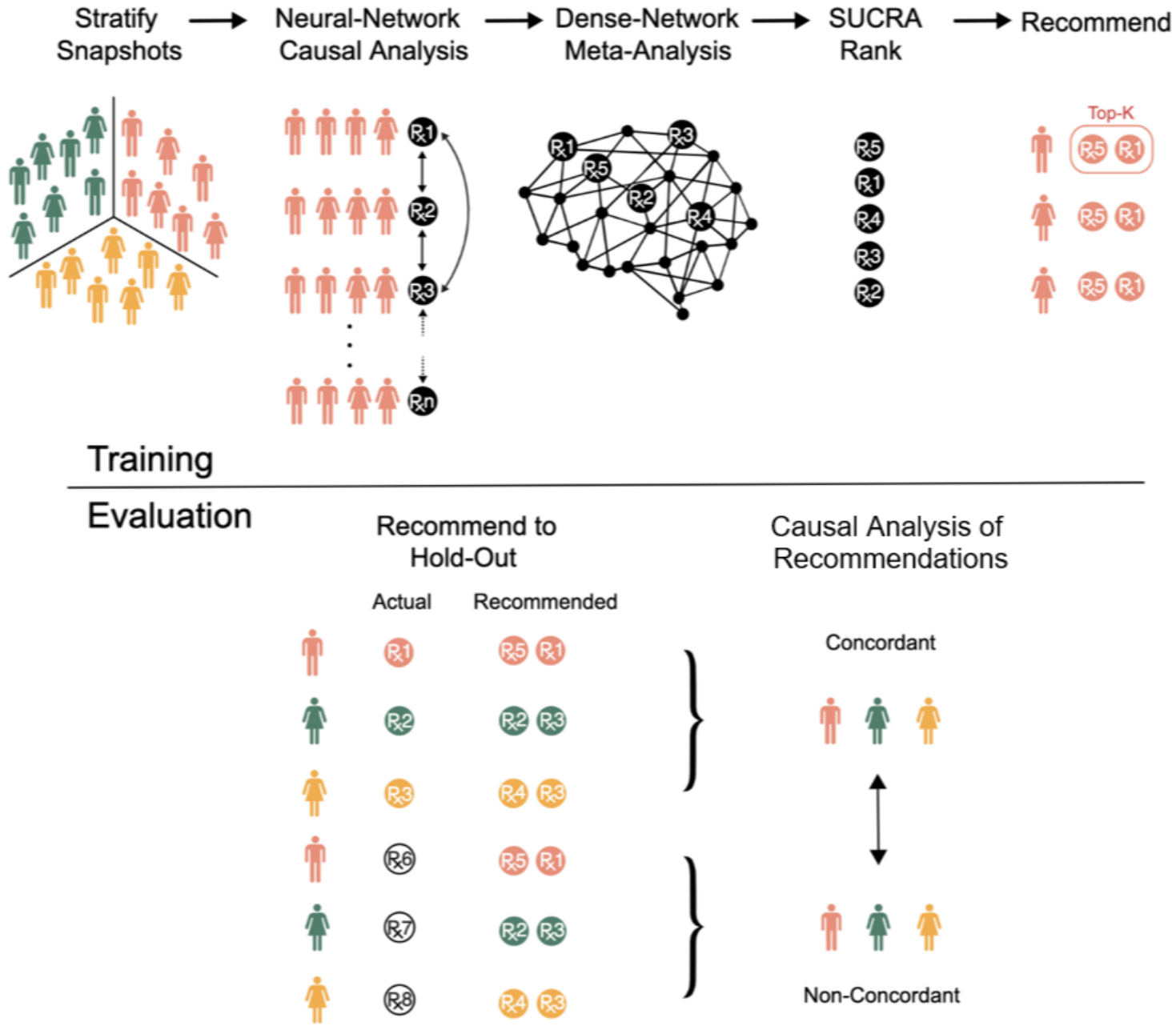

Figure 2

Schematic of Ranking Generation and Analysis Snapshots were split into Training (80%) and Hold-Out (20%) datasets. Patients were stratified into 10 clinical subgroups based on age, number of

comorbidities, and prior insulin use (Table 2). For each clinical subgroup, all treatments with cohort size > 35 were selected and case-control observational studies were performed comparing every treatment with every other treatment using a neural-network-based propensity-score model for causal inference. A densely connected network graph was constructed with treatments as nodes and edges connecting treatments via measured Average Treatment Effect (ATE) values. Bayesian Network Meta-Analysis (NMA) was performed to compute network-synthesized ATEs compared against a baseline treatment which was set to Metformin (the first-line therapy for T2DM). Treatments were ranked by their Surface Under the Cumulative RAnking curve (SUCRA) scores and the Top-K were returned to each member of the subgroup. To gauge the efficacy of the causal recommender engine, recommendations were generated for each patient in the hold-out dataset and the observed change in HbA1c was recorded. This change in HbA1c between the concordant cohort (where the prescribed treatment was one of the top-three ranked treatments) and the non-concordant cohort was used to estimate the confounder-adjusted ATE of the recommendations based on rank.

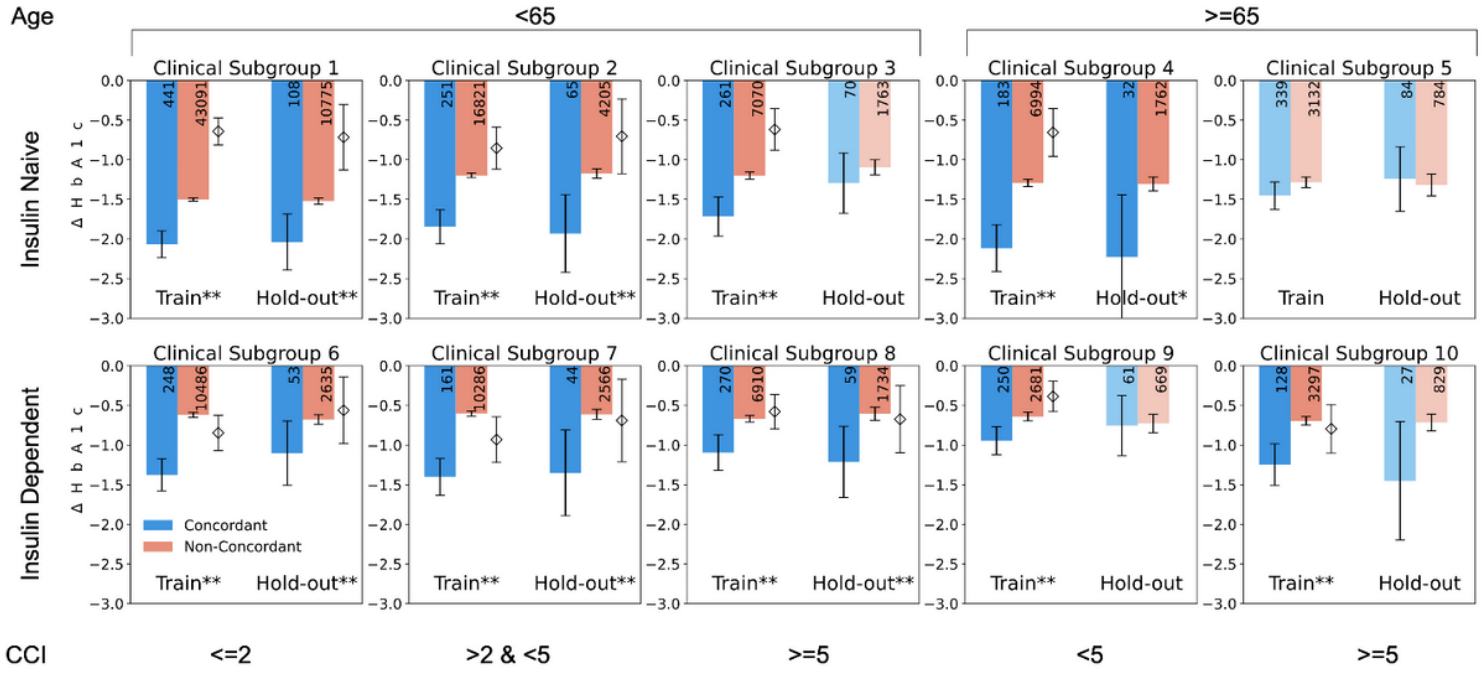

Figure 3

Causal Effect of Treatment Ranking Groups on Blood Sugar Reduction. Evaluations for concordant (blue) and non-concordant (red) cohorts for all clinical subgroups. An individual is considered concordant if their current treatment matches one of the top-K=3 recommendations for their clinical subgroup and non-concordant otherwise. Training and hold-out set results are shown. * denotes difference of means between concordant and non-concordant cohorts is statistically significant ($p < 0.05$). ** denotes that the confounder-adjusted Average Treatment Effect (ATE) of the causal recommender is also statistically significant ($p < 0.05$). Diamonds show ATE values of the causal recommender for cases with two asterisks. Numbers on bars denote the number of individuals in each cohort. CCI = (unweighted) Charlson Comorbidity Index.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CausalRecSupplementaryMaterialsBN3.docx
- Table3.png