Applied Stochastic Models in Business and Industry

WILEY

RESEARCH ARTICLE OPEN ACCESS

Causal Forests for Discovering Diagnostic Language in Electronic Health Records

Alessandro Albano | Chiara Di Maria | Mariangela Sciandra | Antonella Plaia

Department of Economics, Business, and Statistics, University of Palermo, Palermo, Italy

Correspondence: Chiara Di Maria (chiara.dimaria@unipa.it)

Received: 23 December 2024 | **Revised:** 16 July 2025 | **Accepted:** 7 August 2025

Funding: This work was supported by the European Union – Next Generation EU. PRIN 2022 PNRR Project "A unified Italian oral medicine and orthodontic language system: a prototype of Natural language processing application in healthcare" n. P202299ZNW CUP B53D23026050001.

Keywords: adrenal glands | causal forest | causal inference | diabetes | hypothyroidism | MIMIC-III | text analysis

ABSTRACT

Textual analysis has gained significant interest in medical research, particularly for automated patient diagnosis based on clinical narratives. While traditional approaches often focus on associational methods, this paper explores the application of causal forests to analyze textual data from electronic health records (EHRs), aiming to identify causal relationships between specific words and the likelihood of receiving certain medical diagnoses. Utilizing the MIMIC-III dataset, we assess how linguistic factors influence diagnosis probabilities for three conditions: diabetes, hypothyroidism, and adrenal gland disorders. Our findings reveal significant causal links between certain clinical terms and diagnosis probabilities, emphasizing the potential of causal inference techniques to improve the analysis of language in clinical narratives. Additionally, we uncover heterogeneity in treatment effects, demonstrating that specific words can identify high-risk patient subgroups. This study highlights the importance of integrating causal inference in natural language processing within healthcare settings.

1 | Introduction

Textual analysis has recently become one of the most popular streams of research in data science. The possibility of extracting information from massive amounts of text can be of interest in several fields: in engineering, to support decision-making in large-scale projects [1], in business, to analyse the tone and readability of corporate annual disclosures [2, 3], in social and behavioural sciences, to extract latent topics [4] or gain insights into the knowledge representation of individuals [5], and in medicine, to examine clinical notes [6].

The digitalisation of texts produced by healthcare workers is taking place in plenty of hospitals worldwide, with the aim of reducing costs and storage issues and improving patients' outcomes, and several studies have addressed the implications of such a process [7–9]. In particular, many scholars investigated the benefits and the limitations connected to the use of electronic health records (EHR) in the healthcare industry. Silow et al. [10] analysed the performances of nine leading hospitals in the US that adopted EHR, showing an improvement in terms of the efficiency of managerial processes and patients' perceived quality; similar results were obtained by Lin et al. [11], who found that the largest benefit were registered in rural hospitals.

Abbreviations: ATE, average treatment effect; AUTOC, area under the TOC; CATE, conditional average treatment effect; EHR, electronic health record; ICD, International Classification of Diseases; MIMIC III, Medical Information Mart for Intensive Care III; RATE, ranked-weighted verage treatment effect; TOC, targeting operator curve.

These authors contributed equally to this work.

In addition to the advantages of EHR, Menachemi and Collum [12] pointed out also some possible limitations, such as the high maintenance costs and the patients' perceived threat to their privacy, see Uslu and Stausberg [13] for a literature review.

One of the primary uses of EHR's texts is the support of clinical activities, as they can be analysed to achieve different tasks, like information extraction (concepts [14] or relationships [15]), representation learning [16, 17], and outcome prediction [18, 19].

A variety of approaches have been proposed to address these issues, mainly based on machine and deep learning, see Shickel et al. [20] for a review. These approaches provide a representation of the text, highlighting relationships between words, concepts or topics, which are in nature associational. It means that the relationships identified by these models are typically based on statistical correlations within the text, rather than explicit causal links. As a result, while such methods can reveal important patterns and associations, they may not fully capture the underlying causal mechanisms that drive clinical outcomes. For example, natural language processing methods model the relations between words in a sentence and between sentences, while different types of neural networks, like graphical or convolutional ones, represent the texts by exploiting the recurrence of some symptoms and illnesses to obtain disease predictions.

Recently, some scholars started analysing EHR to extract causal relationships. Shen et al. [21] developed a causal discovery approach tailored on EHR to find causes and symptoms of type-2 diabetes; similarly, Chen et al. [22] and Zhang et al. [23] used machine learning to discover potential risk factors of acute kidney injury. Some studies focus on the estimation of causal treatment effects on certain diseases, like Rao et al. [24], who assessed the effect of classes of antihypertensive on incident cancer risk, and Zhang et al. [25], who created a machine learning algorithm to account for unobserved confounders and unbiasedly estimate causal effects. Johnson et al. [26] addressed a related problem, that is, identifying the best treatment to lower systolic blood pressure.

Our work follows this stream of literature, as it aims to assess if and the extent to which the presence of some words in the patients' EHR texts causally affects the probability of receiving a certain diagnosis. Indeed, differently from the aforementioned approaches, we do not explore causal relationships among symptoms or reconstruct causal-effect patterns from sentences; instead, we focus on words and their role to determine (cause) the clinician to diagnose a patient with a disease. We applied our approach to the MIMIC-III dataset [27], considered as a benchmark in the analysis of clinical texts. Specifically, we selected three diseases with different degrees of prevalence among patients (diabetes, hypothyroidism, and adrenal gland dysfunctions) and analysed the causal effect of the frequency of some specific words on the probability of being diagnosed with that disease.

This approach offers concrete practical applications: it can help prioritize patients for further testing based on textual indicators with strong causal effects, provide complementary evidence in cases with ambiguous clinical markers, and identify subtle diagnostic signals that might be overlooked in standard evaluations.

In order to do so, we used causal forests [28, 29], a machine learning method that extends random forests to estimate conditional average treatment effects by recursively partitioning the covariate space. Causal forests adapt to complex, nonlinear relationships between words and diagnosis outcomes without requiring pre-specification of functional forms, that are generally used by full parametric methods. Moreover, it incorporates honest estimation [30], which helps mitigate overfitting and provides valid inference even in the presence of many potential confounders, a common challenge in textual analysis where the confounder space is high-dimensional. Compared to other causal inference approaches, as Bayesian Additive Regression Trees [31], causal forest is computationally efficient for large datasets with many predictors, making it suitable for analyzing the extensive vocabulary in EHRs. While being powerful, the algorithm still maintains a level of interpretability that is crucial in the medical domain, where understanding the driving factors behind predictions is essential.

Another strength of causal forests consists in explicitly modeling and estimating heterogeneous treatment effects, which allows us to identify patient subgroups for whom specific words have stronger or weaker associations with diagnoses. In contrast, other recent machine learning techniques such as Double Machine Learning approaches [32] primarily focus on average treatment effects without exploring effect heterogeneity, thus making the strong assumption that the treatment effect is constant across all units. Finally, causal forests naturally accommodate continuous treatment variables, in contrast to many traditional causal inference methods that are restricted to binary treatments or reduce the problem of continuous treatment to a binary problem by selecting two treatment levels to compare. For example, a traditional propensity score matching or weighting approach [33] typically assumes binary treatments, and while extensions of propensity scores exist for continuous treatments [34], they often struggle with high-dimensional confounder spaces like those present in textual data, where thousands of potential confounding words exist. Other approaches in the literature for causal inference are generally not suitable for textual data. For example, Instrumental Variables methods require identifying valid instruments that affect the treatment (word frequency) but not the outcome (diagnosis) except through the treatment [35]. In textual analysis of EHRs, finding such instruments is particularly challenging, as most linguistic features are interrelated in complex ways.

While there is growing interest in applying causal machine learning methods to text analysis [36], to the best of our knowledge, this is the first paper applying causal forest to analyse texts[1] , and in the context of EHR. This is consistent with recent literature, including the methodological review by Rehill [38]. Given the widely use of textual data in different contexts and its usefulness in clinical practice, we believe that this work can contribute to automatizing the diagnostic processes, helping clinical workers in choosing the best therapy for patients. Indeed, since EHRs are at the heart of hospital operations (a classic business process), our technique can be used as a decision-support module. This directly speaks to the "business" of healthcare as it allows to optimize patient screening to reduce costs, shorten time-to-diagnosis, and enhance overall patient management. Therefore, the healthcare industry can largely benefit from such tools through improved efficiency and better allocation of clinical resources.

The paper is structured as follows: in Section 2, we describe causal forests; in the third section, we present the data analysis and the results. A discussion will follow.

2 | Methods

Let W represent the treatment variable of interest and Y denote the outcome. We assume the existence of potential confounders X, which may influence both the treatment and the outcome, leading to biased estimates of the causal effect.

To measure the impact of W on Y, we utilize the *Average Treatment Effect* (ATE), a fundamental concept in causal inference, often framed in terms of *counterfactuals* [39]. Counterfactuals describe hypothetical scenarios, such as "If I had woken up on time, I would not have been late to work." Let $Y(w)$ denote the potential outcome that Y would assume if W were set (possibly contrary to fact) to a specific value w. If W is binary, the ATE of W on Y is the average of the difference $Y(1) - Y(0)$, that is, the extent to which the outcome would change, on average, moving from a scenario where all individuals did not receive the treatment to a scenario where all of them are treated.

Mathematically, the ATE is expressed as:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

which, under the assumption of no unmeasured confounders (e.g., in randomized experiments), simplifies to the difference in conditional means:

$$\text{ATE} = \mathbb{E}[Y|W = 1] - \mathbb{E}[Y|W = 0]$$

This formulation holds when the assignment of W is independent of the potential outcomes given X, a condition known as *ignorability* or *unconfoundedness*.

In observational settings, confounders of the $W-Y$ relationship are ubiquitous, and it is plausible that causal effects would differ among subgroups, or, in other words, the causal effect would assume different values *conditional* on a set of covariates X. This causal effect takes the name of conditional average treatment effect (CATE) and it is denoted by $\tau(X)$:

$$\tau(x) = \mathbb{E}[(Y(1) - Y(0))|X = x] \quad (1)$$

Following Athey et al. [29, 30], for each subject $i = 1, \ldots, n$, it is possible to write the relationship between Y, W and X as

$$Y_i = \tau(X_i)W_i + f(X_i) + \varepsilon_i \quad (2)$$

where f is an unknown function describing the dependence of the outcome on the covariates, and ε is an error term having zero mean conditional on W and X.

To estimate $\tau(X_i)$, we introduce two key functions: the *propensity score* $e(x) = \mathbb{E}[W|X_i = x]$, which measures the probability of receiving treatment given X, and the *conditional mean* $m(x) = \mathbb{E}[Y|X_i = x] = f(x) + \tau e(x)$, the expected outcome given the covariates. Subtracting $m(x)$ from both sides of Equation (2) leads to the following *residual-on-residual regression*:

$$Y_i - m(x) = \tau(X_i)(W_i - e(x)) + \varepsilon_i \quad (3)$$

Equation (3) shows that the CATE can be estimated using a residual-on-residual regression.

The difficulty is linked to the fact that $m(x)$ and $e(x)$ are unknown. However, since we are just interested in "good" estimates of them, we do not need to specify parametric models like logistic regression, but we can use machine learning tools like random forests. If it is possible to identify a neighbourhood $\mathcal{N}(x)$ where τ is approximately constant, the CATE can be estimated through a weighted residual-on-residual regression for X_i in the neighbourhood. The weights are estimated by combining the adaptive neighbourhood weights proposed by Breiman [40] with the approach in Equation (3) proposed by Robinson [41]. Indeed, Breiman [40] constructs forests that split covariates in order to maximize the squared difference of outcome means in the subgroups, and the resulting weights are used in the regression. The algorithm proposed by Wager and Athey [42] instead, puts splits in the forest to maximize the squared difference of $\hat{\tau}$ in the subgroups, where $\hat{\tau}$ is estimated by running Robinson's residual-on-residual regression for each possible split point.

In the case of continuous or discrete treatments, Athey et al. [30] extend the CATE framework by proposing a consistent estimator based on the *covariance* between W and Y, conditional on X. This estimator takes the form:

$$\text{CATE} = \frac{\text{Cov}(W, Y|X)}{\text{Var}(W|X)} \quad (4)$$

This ratio captures the linear relationship between W and Y within subgroups defined by X, allowing for estimation of treatment effects even in non-binary settings.

If the estimates of $\tau(X_i)$ differ among subgroups, this *treatment effect heterogeneity*, meaning the treatment varies depending on individual characteristics. Causal effect heterogeneity can be further investigated using the Ranked-weighted Average Treatment Effect (RATE). The RATE is a metric that assesses whether there are any benefits in treating just a fraction of units chosen according to a prioritizing rule $S(X)$, that is, a rule to assign treatment to specific subjects having certain characteristics. It can be considered as an area under the curve, called Targeting Operator Curve (TOC), defined as

$$TOC(q) = \mathbb{E}[Y(1) - Y(0) \mid \hat{\tau} \geq F^{-1}(1 - q)] - \mathbb{E}[Y(1) - Y(0)] \quad (5)$$

where F is the distribution function of the prioritizing rule $S(X)$, and q represents the fraction of the population being treated. The *area under the TOC (AUTOC)* provides a summary measure of the benefit of selective treatment based on the estimated CATE. If the scoring rule $S(X)$ effectively identifies individuals with significantly different treatment benefits, we would expect the RATE metric to be positive. On the other hand, if it performs poorly or if there are minimal benefits to stratifying treatment, we would expect the metric to be negative or close to zero, respectively.

3 | Data Analysis

We analysed data from MIMIC-III (Medical Information Mart for Intensive Care III) [27], a freely accessible database including

TABLE 1 | Empirical Bayes-smoothed log-odds ratios for selected terms across three diagnoses. Horizontal dashed lines separate positive from negative values.

Diabetes			Hypothyroidism			Disorders of adrenal glands		
Term	***n***	**Log-odds**	**Term**	***n***	**Log-odds**	**Term**	***n***	**Log-odds**
Insulin	31,395	68.6	tsh	2585	15.0	Insufficiency	905	7.24
Units	27,063	46.9	chf	3595	6.34	Steroid (s)	1048	6.37
Sugar (s)	7015	30.7	uti	1820	5.40	Prednisone	1160	5.56
Subcutaneous	8805	24.6	afib	2318	5.34	Hypotension	1 121	4.01
Renal	23,006	17.6	Female	3030	3.58	Stress	374	2.45
Obese/obesity	6830	17.6	Woman	1899	3.24	Vancomycin	776	1.90
Dialysis	5685	17.3	Man	751	−4.54	micu	623	1.59
Cad	11,580	17.1	Head	5159	−4.94	Aortic	730	−1.38
Foot	5101	16.9	Cardiovascular	589	−5.88	Postoperative	126	−1.41
Hepatitis	2105	−13.2	Male	1631	−6.58	Valve	704	−1.46
Breast	1839	−14.1	Hepatitis	543	−7.28	Coronary	310	−1.59
Screening	268	−14.7	Baby	85	−8.02	Artery	664	−1.62
Bilirubin	880	−14.8						
Head	12,301	−15.7						
Baby	197	−19.7						

several pieces of information regarding patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, from 2001 to 2012. The dataset contains a total of 61,532 hospital admissions for 38,597 unique patients.

For each patient, there are details about their clinical history in the form of free texts, such as discharge summaries, progress notes, and nursing notes. Patients' diagnoses are recorded through the ICD-9 coding system, which has a hierarchical structure, that is, 18 chapters representing macro-categories of diseases and subchapters classifying them more accurately. Patients can receive more than one diagnosis, that is, more than a single ICD-9 code, and they are ordered according to the severity of the disease, so that the first ICD-9 codes correspond to the conditions that need main cures.

Since our aim is investigating which words have a significant causal effect on the probability of being diagnosed with a certain condition, we discarded too common illnesses like cardiovascular or respiratory diseases, focusing on three diseases with different prevalence among patients. Specifically, we selected diabetes (30%), hypothyroidism (10%) and adrenal gland dysfunctions (1%). We chose these diseases because they vary in prevalence and impact, providing a diverse set of conditions for examining how different clinical terms and notes might influence diagnosis. Diabetes, for instance, is a highly prevalent chronic disease that typically features in a range of clinical narratives and may exhibit strong associations with particular vocabulary, whereas hypothyroidism, being less common, might show more complex patterns of linguistic association. Adrenal gland dysfunctions, being rare, offer a unique case to explore whether certain clinical terms are disproportionately predictive of such infrequent conditions.

The outcome variable we are interested in is whether a specific disease is present or absent in a patient's ICD-9 codes, regardless of the order they appear. For a given disease, we aim to identify relevant treatment variables W. To do this, we compute an empirical Bayes-smoothed log-odds ratio [43] comparing each term's prevalence in patients with the diagnosis versus those without. Conceptually, this metric quantifies how much more (or less) likely a word is to appear in the diagnosed group, while adjusting for its overall frequency in the corpus. The resulting score is a continuous measure of the strength of association, useful for ranking the words most likely associated with the disease under exam. The selected words along with their log-odds ratio are reported in Table 1.

After selecting the list of relevant words that could serve as treatments, we analyze each word individually by running a separate causal forest for each one. Specifically, the number of occurrences of each word was treated as a discrete treatment (Equation 4), while the set of other words except those used as treatment acted as potential confounders. This allowed us to isolate their impact on outcomes, adjusting for the other words in the text.

First, we analyzed the average treatment effect of single words, that is, the average change in the probability of receiving the diagnosis of a specific disease occurring when increasing by one unit the frequency of the word selected as a treatment.

In a second step, we investigated the effect heterogeneity by transforming the exposure into a binary variable, that is, the presence or absence of a specific word in the discharge note, using the CATE as a prioritizing rule. In other words, we want to evaluate if treating only the individuals with the largest estimated CATEs (i.e., assigning the selected word to their discharge note) brings some benefit compared to the overall ATE.

Before performing the statistical analysis, we carried out all the standard data cleaning procedures, including tokenization and removal of stop words, while stemming was not applied. These steps were implemented in `R` using the `tidytext` package [44]. To perform causal inference, we employed the `grf` package [28] in `R`, which allows us to run causal forests as described in [29, 30, 45]. The `R` code used for the analyses is publicly available in the accompanying GitHub repository.

3.1 | Diabetes Data

First, we analyzed the most common disease among the selected ones, that is diabetes. The ATE and RATE estimates are shown in Figure 1. Figure 2 shows the TOC curves of four words, belonging to the list of select words, displaying interesting yet different behaviours. Indeed, as we will discuss later, some of them correspond to words better capturing heterogeneity among patients, while some are not effective for identifying patient subgroups benefiting more from the treatment.

Starting with the ATE estimates (Figure 1, left plot) for diabetes, we see that certain terms are strongly causally linked to the likelihood of a diabetes diagnosis. *Insulin* has the highest positive ATE estimate, around 0.07. This result is expected, as insulin is a central treatment in diabetes management, particularly in Type 1 diabetes or advanced Type 2 diabetes. Other terms like *sugar (s)*, *cad* (coronary artery disease), *obese/obesity*, *dialysis*, and *foot* also display positive causal effects, ranging between 0.02 and 0.04. These are all conditions commonly associated with diabetes: sugars are directly linked to blood glucose levels, obesity is a well-known risk factor for diabetes, and diabetes is often complicated by coronary artery disease and renal complications, sometimes leading to dialysis and foot issues (like diabetic foot ulcers). Terms such as *subcutaneous*, *units*, and *renal* show smaller but still positive effects, signifying their relevance to diabetes, although to a lesser degree than insulin or sugars. These terms relate to insulin administration (subcutaneous injections, insulin units) and kidney-related complications (renal issues are common in diabetes).

On the negative side, terms like *bilirubin*, *screening*, *hepatitis*, and *breast* show significant negative effects, indicating that these conditions are associated with a lower likelihood of diabetes diagnosis. Indeed, *bilirubin* and *hepatitis* are more related to liver diseases. In contrast, terms such as *baby* and *head* exhibit confidence intervals that include zero, suggesting that their potential protective effect is not statistically significant.

The right plot of Figure 1 displays the RATE estimates, which assess how well these terms capture heterogeneity and serve as targeting rules for identifying different subgroups in diabetes diagnoses. *Insulin* once again stands out with the highest RATE estimate, around 0.08, also confirmed by the TOC curve in Figure 2a, which shows a clear significant linear negative trend. This indicates that the term is not only causally linked to diabetes but is also highly effective at identifying subgroups of patients with different probabilities of diagnosis. This makes sense given the direct role of insulin in diabetes management. Terms like *subcutaneous*, *units*, and *sugars* also display strong RATE estimates. These terms are critical in the treatment and monitoring of diabetes (insulin administration and glucose monitoring), making them effective in identifying individuals with the condition. In particular, we can note that the TOC of the word *subcutaneous* has very narrow confidence intervals, denoting low uncertainty about the role of this word.

On the other hand, the term *baby*, which had a non-significant ATE, shows a relatively large RATE estimate (even if not significant). This suggests some heterogeneity in how this term might identify different subpopulations with varying diabetes risks, possibly related to gestational diabetes or pediatric diabetes, though the relationship is not directly causal. Terms such as *foot*, *renal*, and *cad* have intermediate RATE estimates (around 0.03-0.05), indicating that while they are less effective than insulin at capturing subgroup heterogeneity, they still play a role in identifying high-risk patients, such as those with diabetes complications (e.g., foot ulcers, kidney disease, and heart disease). Terms like *breast*, *head*, *obesity*, *dialysis* and *hepatitis* display low RATE estimates, mirroring their weaker causal relationship seen in the ATE plot. These terms are not effective at differentiating subgroups for diabetes diagnosis as shown in Figures 2c and 2d.

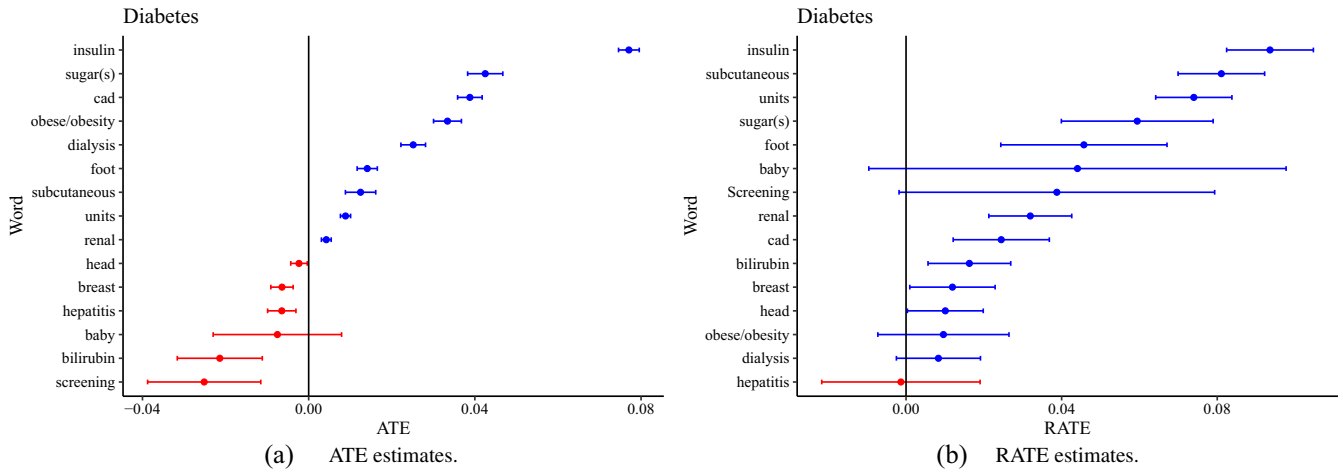


FIGURE 1 | Estimates for the Average Treatment Effect (ATE) on the left and rank-weighted Average Treatment Effect (RATE) on the right for diabetes-related terms, along with their corresponding 95% confidence intervals.

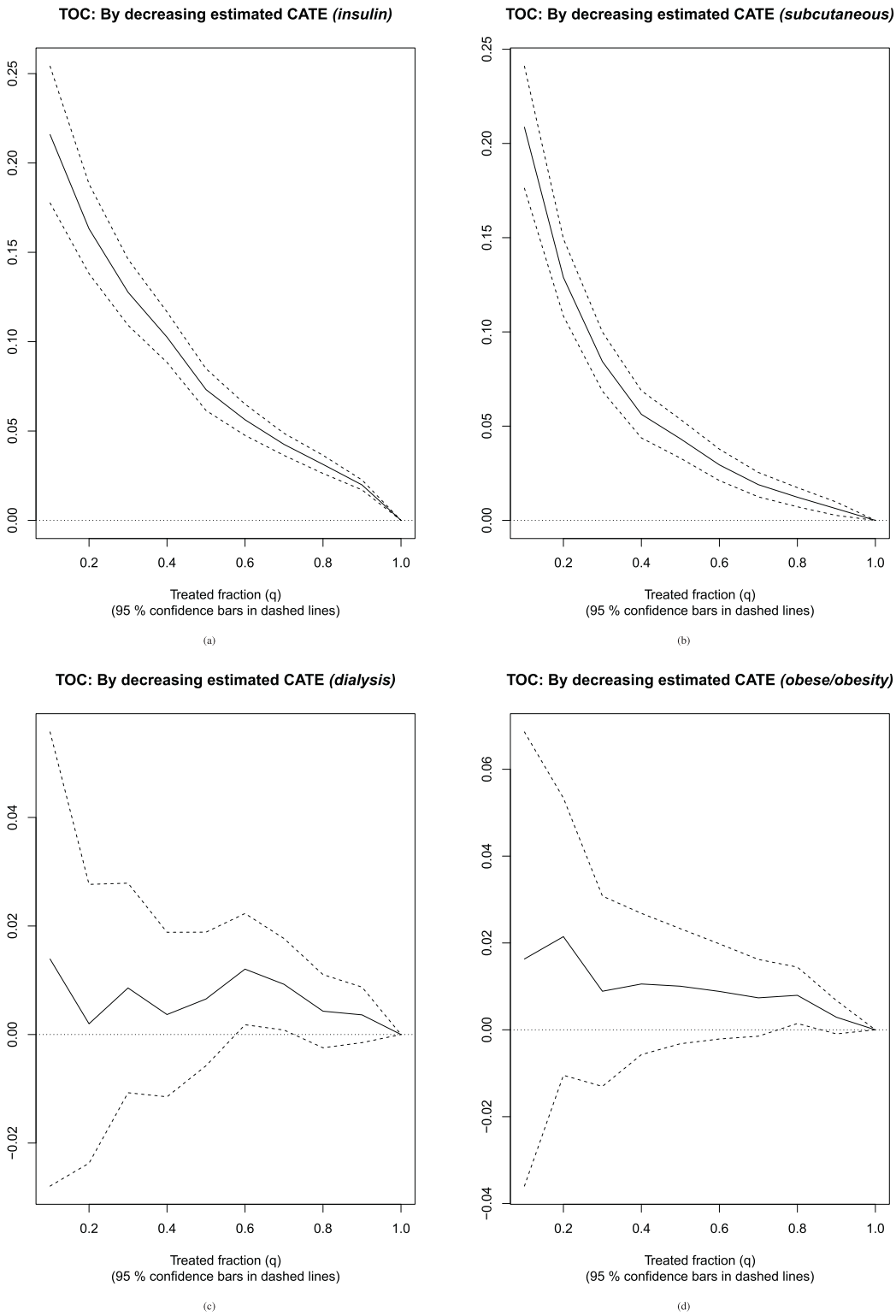
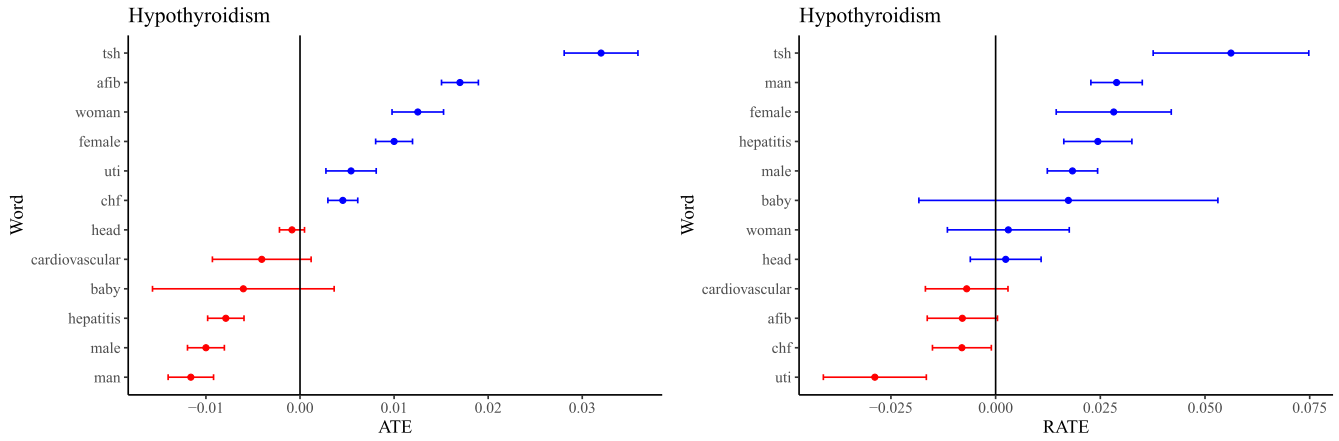


FIGURE 2 | Estimate TOCs for diabetes-related terms, along with their corresponding 95% confidence intervals.

3.2 | Hypothyroidism Data

The second selected disease is hypothyroidism. In Figure 3 the ATE and RATE estimates of the list of words selected are shown. While, in Figure 4 we display the TOCs of four words chosen following the same line of reasoning used for diabetes.

Among the words selected through log odds, there is *levothyroxine* which we decided not to display since it is the active principle used to treat this disease, so it is not particularly informative (the estimated ATE is 0.263). This means that if all the texts were "treated" so that they contain the one more occurence of the word *levothyroxine*, the probability that



(a) ATE estimates.

(b) RATE estimates.

FIGURE 3 | Estimates for the Average Treatment Effect (ATE) on the left and rank-weighted Average Treatment Effect (RATE) on the right for hypothyroidism-related terms, along with their corresponding 95% confidence intervals. (a) ATE estimates, (b) RATE estimates.

the patients would be diagnosed with hypothyroidism would increase by 0.26.

The word with the highest ATE is *tsh*, which is the thyroid-stimulating hormone, a hormone produced by the pituitary gland that regulates the production of thyroid hormones from the thyroid gland. As regards biological sex, the observed effects for *female* and *woman* (0.010 and 0.012, respectively) signify that the diagnosis of hypothyroidism tends to be more likely among females in contrast to males, whose effect sizes for *male* and *man* are negative (−0.010 and −0.011). Notably, all these coefficients demonstrate statistical significance, being distinctly different from zero. This result is consistent with empirical evidence, since hypothyroidism is more common in women than men.

Several terms in the data frame are related to medical conditions. For instance, *afib* (atrial fibrillation), *hepatitis*, *chf* (congestive heart failure), and *uti* (urinary tract infection) represent specific health conditions. Except for hepatitis, all of them show a positive causal effect, and this may be because all these conditions are often associated with hypothyroidism, so they may concur with its diagnosis.

The term *baby* shows a negative but non-significant effect because of the low number of occurrences. The negative sign is, however, meaningful since hypothyroidism is a disease that typically affects adults. Finally, terms such as *head* and *cardiovascular* relate to anatomical structures or physiological systems, but their effects are not significant.

In the right plot (RATE estimates), we observe how different terms perform in identifying heterogeneity in the relationship between specific words and hypothyroidism diagnoses. Once again, the word with the highest RATE estimate is *tsh*, with an even more pronounced effect than in the ATE plot, approaching 0.06. Specifically, from Figure 4a the estimated $TOC(q)$, that is, the difference between the causal effect obtained treating just $q \times 100\%$ of subjects and the overall ATE, is significant up to $q = 0.4$, starting around 0.15 and decreasing sharply as more patients are treated. This suggests that the term *tsh* is highly effective as a targeting rule, highlighting its key role in identifying patients with hypothyroidism.

In terms of gender, the terms *female* and *male* show a positive RATE, slightly above 0.025, which indicates that these terms not only have a causal relationship with hypothyroidism (as seen in the ATE plot) but also serve as strong indicators to identify heterogeneity between patients who are likely to be diagnosed. As regards Figure 4b, the TOC values are always significant, indicating that treating with the word female only a fraction a patients (according to CATE) is always more beneficial that treating everyone.

This confirms the robustness of the diagnosis patterns, where females are more frequently diagnosed with hypothyroidism compared to males. The terms *man* and *woman* display smaller effects in the RATE plot, suggesting that these terms show less heterogeneity in how they relate to hypothyroidism diagnosis.

Interestingly, *hepatitis*, which had a non-significant ATE, now shows a slightly positive RATE estimate. This suggests that while the direct causal link between hepatitis and hypothyroidism may be weak, there may be specific subgroups or contexts where the presence of hepatitis serves as an indicator for hypothyroidism.

Medical conditions such as *afib*, *chf*, and *uti*, which had positive ATE estimates, continue to show positive RATE estimates, showing their effectiveness in targeting subgroups of patients who are more likely to receive the diagnosis.

The term *baby*, which had a non-significant negative ATE, also has a non-significant positive RATE estimate. As a matter of fact, the corresponding TOC curve confidence intervals include 0 (see Figure 4c). Similarly, anatomical and physiological terms like *head* and *cardiovascular* exhibit non-significant RATE estimates, confirming they can not be used as prioritization rules.

Finally, the term *uti* presents an interesting dynamic in this analysis. While it shows a positive significant ATE, the derived negative significant RATE suggests that *uti* may not effectively identify at-risk subgroups (as shown in Figure 4d). This contrast indicates

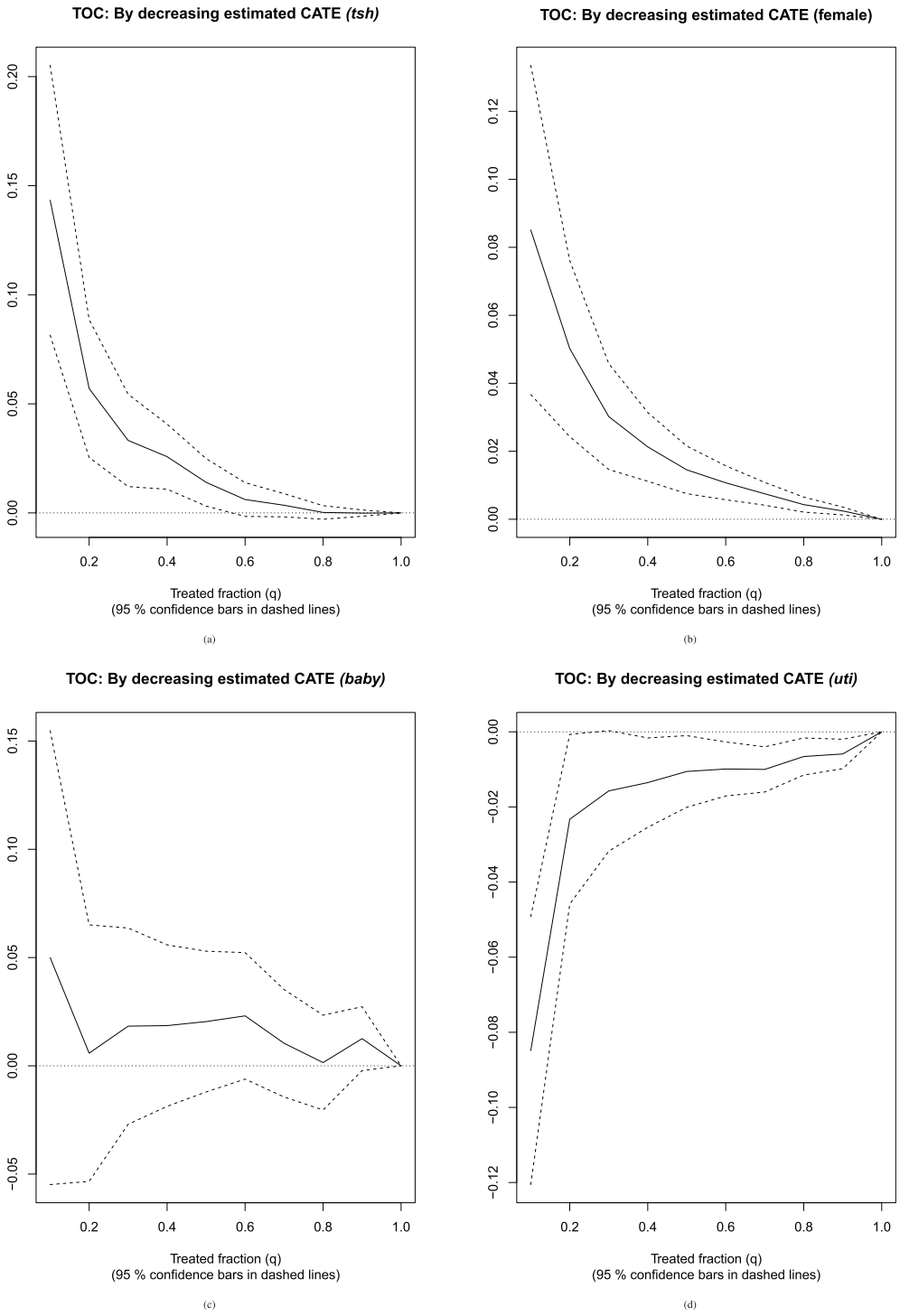


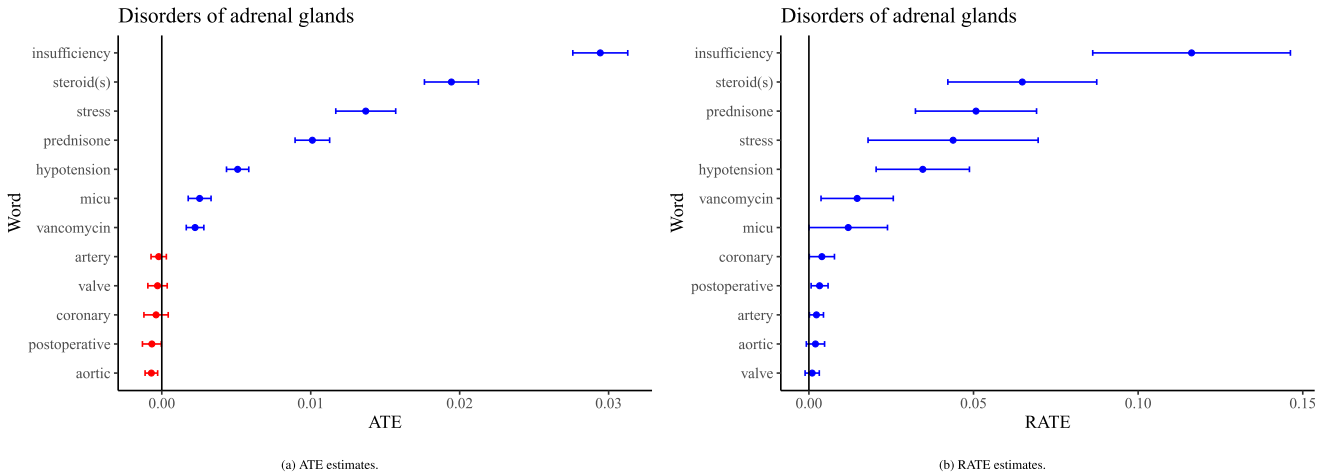
FIGURE 4 | Estimate TOCs for hypothyroidism-related terms, along with their corresponding 95% confidence intervals.

that, although there is an average positive effect, using *uti* as a sole indicator for targeting treatment could be less effective, as it does not fully capture the variability in patient risk profiles.

3.3 | Disorders of Adrenal Glands

Moving to Disorders of Adrenal Glands, the ATE and RATE estimates are shown in Figure 5. Once again, Figure 6 represents the TOCs of four words selected as done in diabetes and hypothyroidism analysis.

The word *insufficiency* shows the highest ATE estimate, close to 0.03. This is meaningful as adrenal insufficiency (a condition where the adrenal glands do not produce enough hormones) is a direct and major disorder associated with the adrenal glands. Following *insufficiency*, the terms *steroid (s)* and *prednisone* show significant positive effects (between 0.01 and 0.02), reflecting the well-documented association between adrenal disorders and steroid therapy. The adrenal glands produce steroids, and conditions like adrenal insufficiency are often treated with synthetic steroids, such as prednisone. The term *stress* also shows a positive



(a) ATE estimates.

(b) RATE estimates.

FIGURE 5 | Estimates for the Average Treatment Effect (ATE) on the left and rank-weighted Average Treatment Effect (RATE) on the right for Disorders of Adrenal Glands related terms, along with their corresponding 95% confidence intervals. (a) ATE estimates, (b) RATE estimates.

ATE estimate, likely reflecting the fact that adrenal glands produce cortisol, a hormone related to stress, and adrenal disorders can involve abnormal responses to stress. *Hypotension* appears with a smaller but significant positive effect. Indeed, hypotension (low blood pressure) is a common symptom of adrenal insufficiency, which explains its positive causal effect on the adrenal gland disorders diagnosis. Terms such as *micu* (medical intensive care unit) and *vancomycin* (an antibiotic) exhibit positive effects, although their impact is relatively modest compared to other factors. Terms related to cardiovascular health (e.g., *artery*, *valve*, *aortic*, *coronary*) are associated with a reduced probability of adrenal gland disorders, indicating that these terms may be more closely linked to other unrelated conditions. However, given the lack of statistical significance, these results should be interpreted with caution.

Looking at the RATE estimates (right plot), *insufficiency* shows the highest RATE estimate, close to 0.13, and the corresponding TOC clearly shows treatment heterogeneity. This indicates that the term is highly effective as a targeting rule for distinguishing between different patient subgroups. Indeed, adrenal insufficiency is a hallmark condition of adrenal dysfunction. The terms *steroid (s)* and *prednisone* also have large RATE estimates, reinforcing their role as key indicators for identifying different subgroups of patients with adrenal disorders. Notably, the TOC curve of the word *steroid (s)* is shown in Figure 6b, acknowledging the parallel findings for *prednisone*.

The term *stress* has a moderately high RATE estimate, indicating that it plays a role in identifying heterogeneity in adrenal disorders, reflecting the adrenal glands' role in managing stress via hormone production (e.g., cortisol). *Hypotension*, while showing a smaller ATE, still has a moderately high RATE estimate, suggesting that while its causal link to adrenal disorders is weaker, it is effective at targeting certain subgroups of patients. Other terms like *vancomycin, micu*, and cardiovascular terms (*artery, valve, aortic*) show smaller RATE estimates. These terms are not so effective at distinguishing subgroups with adrenal disorders, as shown in Figures 6d and 6c (just the words *valve* and *aortic*), which aligns with their lower causal relationship in the ATE plot.

4 | Conclusions

This study examines the causal relationships between specific words in electronic health records (EHRs) and the probability of receiving certain medical diagnoses. By using causal forests as our analytical approach, we show that this method allows us to identify disease-related words and quantify how a one-unit increase in their frequency causally affects, on average, the probability of receiving the diagnosis of a specific disease. In addition, we also investigated their effect heterogeneity, assessing whether treating only the patients with the largest estimated CATEs leads to a different outcome compared to the ATE.

Our analysis focused on three diseases with varying prevalence: diabetes, hypothyroidism, and adrenal gland disorders. We found that certain terms have strong causal effects on the probability of diagnosis. For diabetes, the term insulin was the most effective treatment, with an estimated average treatment effect (ATE) of around 0.07. This aligns with established knowledge about the role of insulin in managing diabetes, particularly in patients with Type 1 and advanced Type 2 diabetes. Other terms, such as sugar (s), obesity, and dialysis, also showed positive associations, further confirming that the language in clinical notes reflects important clinical realities.

In the case of hypothyroidism, the term tsh (thyroid-stimulating hormone) had the highest ATE, indicating a strong connection to hypothyroidism diagnoses. Additionally, the positive effects associated with female-related terms (female and woman) highlight the gender differences seen in thyroid disorders, which is consistent with existing research that shows a higher prevalence of hypothyroidism in women. This emphasizes the need for gender-specific approaches in both diagnosis and treatment.

For adrenal gland disorders, the term insufficiency had the highest ATE, followed by steroid (s) and prednisone. These findings are relevant because they confirm established links between these terms and adrenal dysfunction.

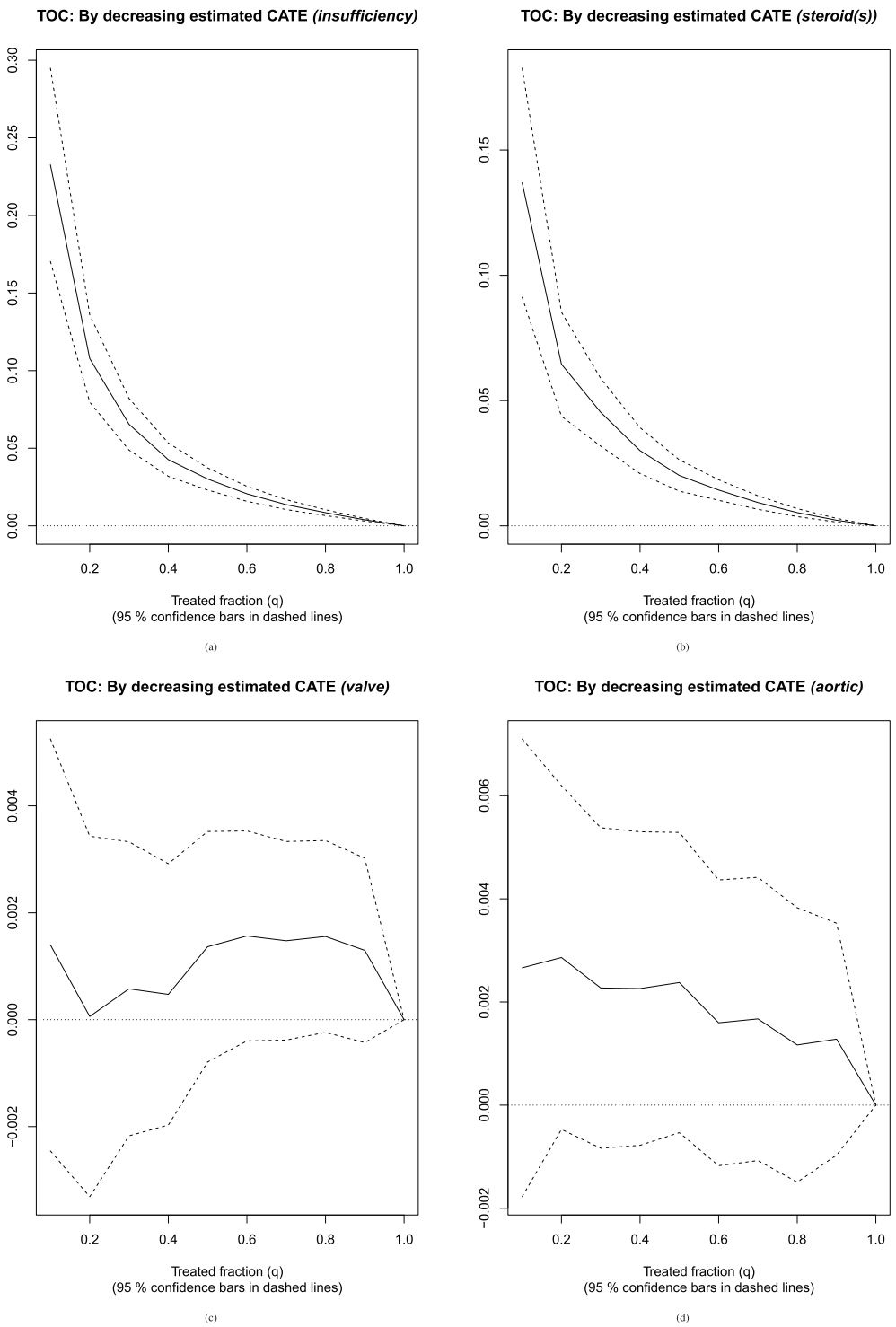


FIGURE 6 | Estimate TOCs for adrenal glands disorders-related terms, along with their corresponding 95% confidence intervals.

It is worth noting that our analysis did not identify only words with a clearly established link with the diseases (e.g., insulin and diabetes), but also words indicating comorbidities, like renal dysfunctions in diabetic patients, or hepatitis for hypothyroidism, and hypotension for adrenal glands' disorders. This applies also in the other direction, since we discovered words which instead reduce the average probability of being diagnosed with one of the chosen diseases. This aspect is particulrly interesting in the case of words related to gender, age or other demographics, and other diseases. For example, the word *male* reduces the probability of being diagnosed with hypothyroidism. Another very interesting example is given by the relationship between the word *bilirubin* and diabetes, since it has a negative ATE on the probability of receiving such a diagnosis. This is in line with medical studies that found the protective role of high levels of bilirubin against diabetes [46–48].

Our analysis of effect heterogeneity using Conditional Average Treatment Effect (CATE) analysis provided valuable insights into how specific terms can help identify patient subgroups at

different risks for diseases. For example, the term *tsh* not only had a significant ATE in hypothyroidism but also effectively differentiated between patients based on their risk levels. This capability to identify high-risk groups through language in EHRs highlights the importance of personalizing clinical assessments. In addition, we tested our approach on diseases showing different prevalence among patients and in all cases we managed to recover relevant words causally affecting the probability of receiving the diagnosis of interest. Thus, this suggests that our method helps to identify a specific "vocabulary" that can guide clinicians towards the correct diagnosis, both for common diseases, such as diabetes, and for rarer conditions like adrenal gland disorders.

Overall, our findings contribute to the ongoing conversation about integrating Natural Language Processing (NLP) techniques in clinical settings. The practical implications of our findings extend beyond academic interest to clinical application. While we acknowledge that common diseases are primarily diagnosed through standardized biochemical criteria (e.g., glycemic thresholds, TSH measurements), our methodology offers complementary benefits to established diagnostic protocols. First, by identifying linguistic patterns that causally affect diagnosis probability, our approach can help prioritize patients for biochemical screening, potentially enabling earlier intervention. Second, in cases with borderline biochemical markers, our language-based causal indicators could provide supplementary evidence to inform clinical decision-making. Third, the identification of comorbidity-related terms can alert clinicians to screen for commonly associated conditions, enhancing comprehensive care. The approach shows particular promise for rare conditions where specialized testing might otherwise be overlooked, and in resource-limited settings where linguistic analysis could help prioritize diagnostic resources.

Our study presents some limitations. First, the analysis depends on the quality and completeness of the EHR data, which may vary. Missing or incomplete clinical notes, inconsistencies in documentation, and variations in how diagnoses are coded could impact the accuracy of the associations identified in the study. Second, we assumed that the words in the clinical notes are a sufficient set of confounders, but demographic information not included in EHRs may have a role in the identified relationships. Third, the interpretation of clinical terms is context-dependent, while, in our analysis, we studied each word as a single unit. This single-word approach introduces certain constraints when interpreting our results, particularly for high-frequency clinical indicators. For example, terms related to patient demographics (such as age and gender) or common clinical measurements appear frequently across patient records. While these terms show measurable causal effects in our analysis, their full clinical significance may be partially captured because their diagnostic relevance often depends on specific values and interactions with other clinical parameters. This context-dependency is particularly evident for words that derive their clinical significance from surrounding qualifiers, specific numerical values, or their placement within complex clinical expressions. For instance, the causal impact of terms like "age" is underestimated in our single-word analysis, despite being well-established risk factors. Despite these limitations, our causal forest methodology successfully identified numerous clinically meaningful relationships between specific words and diagnoses, demonstrating its value even with the constraints of single-word analysis.

For future research, we recommend applying this methodology to a wider range of datasets from different healthcare systems to broaden our findings. Furthermore, investigating the interactions between patient demographics and linguistic factors (also including groups of words and context-dependency) in EHRs could offer deeper insights into how language shapes diagnostic processes. Ultimately, this work aims to connect clinical practice with data science, paving the way for a healthcare environment where analyzing textual data is a standard part of patient care, leading to improved outcomes and more personalized healthcare experiences.

Acknowledgments

This research was funded by European Union–Next Generation EU. PRIN 2022 PNRR Project "A unified Italian oral medicine and orthodontic language system: a prototype of Natural language processing application in healthcare" n. P202299ZNW CUP B53D23026050001.

Disclosure

The authors have nothing to report.

Ethics Statement

The authors have nothing to report.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in PHYSIONET at https://archive.physionet.org/mimic2/.

Endnotes

[1] This version of the paper substantially expands upon a previously published short version [37] by introducing new methodological insights, additional empirical results, and a more detailed discussion of the implications.

References

1. F. Browne, N. Rooney, W. Liu, et al., "Integrating Textual Analysis and Evidential Reasoning for Decision Making in Engineering Design," *Knowledge-Based Systems* 52 (2013): 165–175.

2. T. Loughran and B. McDonald, "Measuring Readability in Financial Disclosures," *Journal of Finance* 69, no. 4 (2014): 1643–1671.

3. Y. Luo and L. Zhou, "Textual Tone in Corporate Financial Disclosures: A Survey of the Literature," *International Journal of Disclosure and Governance* 17, no. 2 (2020): 101–110.

4. D. C. Atkins, T. N. Rubin, M. Steyvers, M. A. Doeden, B. R. Baucom, and A. Christensen, "Topic Models: A Novel Method for Modeling Couple and Family Text Data," *Journal of Family Psychology* 26, no. 5 (2012): 816–827, https://doi.org/10.1037/a0029607.

5. R. Iliev, M. Dehghani, and E. Sagi, "Automated Text Analysis in Psychology: Methods, Applications, and Future Developments," *Language*

and Cognition 7, no. 2 (2015): 265–290, https://doi.org/10.1017/langcog.2014.30.

6. H. Dong, M. Falis, W. Whiteley, et al., "Automated Clinical Coding: What, Why, and Where We Are?," *NPJ Digital Medicine* 5, no. 1 (2022): 159.

7. M. Paul, L. Maglaras, M. A. Ferrag, and I. Almomani, "Digitization of Healthcare Sector: A Study on Privacy and Security Concerns," *ICT Express* 9, no. 4 (2023): 571–588.

8. C. Hufnagl, E. Doctor, L. Behrens, C. Buck, and T. Eymann, "Digitisation Along the Patient Pathway in Hospitals," Research Paper (2019).

9. L. V. Lapão, *The Future of Healthcare: The Impact of Digitalization on Healthcare Services Performance* (Springer, 2019), 433–446.

10. S. Silow-Carroll, J. N. Edwards, and D. Rodin, "Using Electronic Health Records to Improve Quality and Efficiency: The Experiences of Leading Hospitals," *Issue Brief (Commonwealth Fund)* 17 (2012): 1–40.

11. Y. K. Lin, M. Lin, and H. Chen, "Do Electronic Health Records Affect Quality of Care? Evidence From the HITECH Act," *Information Systems Research* 30, no. 1 (2019): 306–318.

12. N. Menachemi and T. H. Collum, "Benefits and Drawbacks of Electronic Health Record Systems," *Risk Management and Healthcare Policy* 4 (2011): 47–55, https://doi.org/10.2147/RMHP.S12985.

13. A. Uslu and J. Stausberg, "Value of the Electronic Medical Record for Hospital Care: Update From the Literature," *Journal of Medical Internet Research* 23, no. 12 (2021): e26323, https://doi.org/10.2196/26323.

14. A. N. Jagannatha and H. Yu, "Bidirectional Recurrent Neural Networks for Medical Event Detection in Electronic Health Records," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2016), 473–482.

15. X. Lv, Y. Guan, J. Yang, and J. Wu, "Clinical Relation Extraction With Deep Learning," *International Journal of Hybrid Information Technology* 9, no. 7 (2016): 237–248.

16. T. Tran, T. D. Nguyen, D. Phung, and S. Venkatesh, "Learning Vector Representation of Medical Objects via EMR-Driven Nonnegative Restricted Boltzmann Machines (eNRBM)," *Journal of Biomedical Informatics* 54 (2015): 96–105.

17. Y. Choi, C. Y. I. Chiu, and D. Sontag, "Learning Low-Dimensional Representations of Medical Concepts," In *AMIA Joint Summits on Translational Science Proceedings* (AMIA Joint Summits on Translational Science, 2016), 41–50.

18. R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients From the Electronic Health Records," *Scientific Reports* 6 (2016): 26094, https://doi.org/10.1038/srep26094.

19. Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to Diagnose With LSTM Recurrent Neural Networks," 2016.

20. B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE Journal of Biomedical and Health Informatics* 22, no. 5 (2018): 1589–1604, https://doi.org/10.1109/JBHI.2017.2767063.

21. X. Shen, S. Ma, P. Vemuri, M. R. Castro, P. J. Caraballo, and G. J. Simon, "A Novel Method for Causal Structure Discovery From EHR Data and Its Application to Type-2 Diabetes Mellitus," *Scientific Reports* 11, no. 1 (2021): 21025, https://doi.org/10.1038/s41598-021-99990-7.

22. W. Chen, Y. Hu, X. Zhang, et al., "Causal Risk Factor Discovery for Severe Acute Kidney Injury Using Electronic Health Records," *BMC Medical Informatics and Decision Making* 18, no. Suppl 1 (2018): 13, https://doi.org/10.1186/s12911-018-0597-7.

23. J. Zhang, E. Kummerfield, G. Hultman, et al., "Application of Causal Discovery Algorithms in Studying the Nephrotoxicity of Remdesivir Using Longitudinal Data From the EHR," *AMIA Annual Symposium Proceedings* 2022 (2023): 1227–1236.

24. S. Rao, M. Mamouei, G. Salimi-Khorshidi, et al., "Targeted-BEHRT: Deep Learning for Observational Causal Inference on Longitudinal Electronic Health Records," *IEEE Transactions on Neural Networks and Learning Systems* 35, no. 4 (2024): 5027–5038, https://doi.org/10.1109/TNNLS.2022.3183864.

25. L. Zhang, Y. Wang, A. Ostropolets, J. J. Mulgrave, D. M. Blei, and G. Hripcsak, "The Medical Deconfounder: Assessing Treatment Effects With Electronic Health Records (EHRs)," *Proceedings of Machine Learning Research* 1 (2019): 22.

26. K. W. Johnson, B. S. Glicksberg, R. A. Hodos, K. Shameer, and J. T. Dudley, "Causal Inference on Electronic Health Records to Assess Blood Pressure Treatment Targets: An Application of the Parametric g Formula," *Pacific Symposium on Biocomputing* 23 (2018): 180–191.

27. A. E. Johnson, T. J. Pollard, L. Shen, et al., "MIMIC-III, a Freely Accessible Critical Care Database," *Scientific Data* 3, no. 1 (2016): 160035, https://doi.org/10.1038/sdata.2016.35.

28. J. Tibshirani, S. Athey, E. Sverdrup, and S. Wager, "GRF: Generalized Random Forests," R Package Version 2.3.1 (2023).

29. S. Athey and S. Wager, "Estimating Treatment Effects With Causal Forests: An Application," *Observational Studies* 5, no. 2 (2019): 37–51.

30. S. Athey, J. Tibshirani, and S. Wager, "Generalized Random Forests," *Annals of Statistics* 47, no. 2 (2019): 1148–1178, https://doi.org/10.1214/18-AOS1709.

31. H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian Additive Regression Trees," *Annals of Applied Statistics* 4, no. 1 (2010): 266–298, https://doi.org/10.1214/09-AOAS285.

32. V. Chernozhukov, D. Chetverikov, M. Demirer, et al., "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal* 21, no. 1 (2018): C1–C68, https://doi.org/10.1111/ectj.12097.

33. P. R. Rosenbaum and D. B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70, no. 1 (1983): 41–55.

34. K. Hirano and G. W. Imbens, *The Propensity Score With Continuous Treatments* (John Wiley & Sons, Ltd., 2004), 73–84.

35. J. D. Angrist, G. W. Imbens, and D. B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91, no. 434 (1996): 444–455, https://doi.org/10.1080/01621459.1996.10476902.

36. F. Villarroel Ordenes, G. Packard, J. Hartmann, and D. Proserpio, "Using Traditional Text Analysis and Large Language Models in Service Failure and Recovery," *Journal of Service Research* 28 (2025): 10946705241307678.

37. A. Albano, C. D. Maria, M. Sciandra, and A. Plaia, "Causal Forests for Electronic Health Records," in *Accepted for Publication in the Conference Proceedings of the Italian Statistical Society Meeting, to Appear in the Italian Statistical Society Series on Advances in Statistics (ISSSAS)* (Springer, 2025).

38. P. Rehill, "How Do Applied Researchers Use the Causal Forest? A Methodological Review," *International Statistical Review* 93 (2025): 288–316.

39. P. W. Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association* 81, no. 396 (1986): 945–960.

40. L. Breiman, "Random Forests," *Machine Learning* 45 (2001): 5–32.

41. P. M. Robinson, "Root-N-Consistent Semiparametric Regression," *Econometrica* 56 (1988): 931–954.

42. S. Wager and S. Athey, "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association* 113, no. 523 (2018): 1228–1242.

43. B. L. Monroe, M. P. Colaresi, and K. M. Quinn, "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict," *Political Analysis* 16, no. 4 (2017): 372–403, https://doi.org/10.1093/pan/mpn018.

44. J. Silge and D. Robinson, "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R," *Joss* 1, no. 3 (2016): 37, https://doi.org/10.21105/joss.00037.

45. S. Yadlowsky, S. Fleming, N. Shah, E. Brunskill, and S. Wager, "Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects," *Journal of the American Statistical Association* 120, no. 549 (2024): 38–51.

46. B. Zhu, X. Wu, Y. Bi, and Y. Yang, "Effect of Bilirubin Concentration on the Risk of Diabetic Complications: A Meta-Analysis of Epidemiologic Studies," *Scientific Reports* 7, no. 1 (2017): 41681.

47. T. Inoguchi, N. Sonoda, and Y. Maeda, "Bilirubin as an Important Physiological Modulator of Oxidative Stress and Chronic Inflammation in Metabolic Syndrome and Diabetes: A New Aspect on Old Molecule," *Diabetology International* 7 (2016): 338–341.

48. T. D. Hull and A. Agarwal, "Bilirubin: A Potential Biomarker and Therapeutic Target for Diabetic Nephropathy," *Diabetes* 63, no. 8 (2014): 2613–2616.