Overlabeling:
Causal Evidence from a Top Medication

Manuel Hoffmann[a*] Zoey Chopra[b*] Hsu-Hang Yeh[c] Elizabeth A. McAninch[d] Pascal Geldsetzer[e]

Abstract (236)

Early disease detection is commonly assumed to be helpful for treatment initiation, disease management, and prevention of complications. For several chronic conditions, the medical community is labeling predisposing disease states at lower than diagnostic thresholds, such as prediabetes and prehypertension, to improve monitoring of disease progression. While labeling may improve awareness of disease progression, treatment at these lower thresholds may prompt concerns for overdiagnosis and overtreatment. We investigate one of the most globally prevalent diseases, hypothyroidism, which is treated with one of the most prescribed medications worldwide, levothyroxine. Using commercial claims data from 2003 to 2021, we assess two common thresholds in thyroid-stimulating hormone (TSH) levels, a lower monitoring threshold of subclinical hypothyroidism (label) where prescription is controversial, and a higher prescription threshold of overt hypothyroidism (diagnosis) where prescription is recommended. We estimate local average treatment effects via a regression discontinuity design. We find no jump in prescriptions at the higher diagnostic threshold. However, we observe a substantial increase in prescriptions and associated costs at the lower labeling threshold without significant improvements across health outcomes. A back-of-the-envelope calculation reveals that moving the labeling threshold closer to the diagnostic threshold can generate cost savings exceeding $1.5 billion annually among the U.S. adult population.

Keywords: overlabeling, natural experiment, regression discontinuity design, overdiagnosis, overprescription, overtreatment, hypothyroidism, levothyroxine

[a] University of California, Irvine, Paul Merage School of Business, 4293 Pereira Dr, Irvine, CA 92697, United States, E-Mail: manuel.hoffmann@uci.edu.
[b] Zoey Chopra, Department of Economics, University of Michigan, 611 Tappan Ave, Ann Arbor, MI 48109, USA, E-Mail: zchopra@umich.edu.
[c] Hsu-Hang Yeh, Biomedical Data Science, School of Medicine, Stanford University Stanford, California, USA, E-Mail: ericyeh@stanford.edu.
[d] Elizabeth A. McAninch, Division of Endocrinology, Department of Medicine, Stanford University, 300 Pasteur Drive, Stanford, CA, USA, E-Mail: lizzymac@stanford.edu. ORCiD: 0000-0003-3993-4663
[e] Pascal Geldsetzer, Division of Primary Care and Population Health, Department of Medicine, Stanford University, 1265 Welch Road, Stanford, CA, USA, E-Mail: pgeldsetzer@stanford.edu.
[*] Denotes co-authorship.

Introduction

Early disease detection has promise to reduce disease burden, prevent disease progression and associated costs, and improve patient outcomes by addressing health concerns before complications arise. One way to achieve earlier detection is to expand diagnostic thresholds (i.e., increase test sensitivity) to include pre-disease labeling thresholds. Examples of this practice include diabetes with prediabetes, hypertension with prehypertension and hypothyroidism with subclinical hypothyroidism. Recently, the medical community has called for a measure of preclinical obesity as well (Iacobucci 2025). In specific instances, such as in the case of cervical cancer, pre-disease labels have proven effective (Viera 2011). However, expanding diagnostic thresholds may inadvertently result in overlabeling, a phenomenon whereby heightened focus on "pre-disease" states yields excessive categorization of at-risk or diseased individuals, even when the actual likelihood of a significant health decline is minimal (Viera 2011).

There are numerous potential downstream consequences that may arise from overlabeling, including overdiagnosis and overtreatment.[1] Patients who might never experience adverse outcomes from their labeled diagnoses (i.e., labels) may be exposed to greater psychological strain or unnecessary treatment with associated side effects and financial burden. At the population level, excessive healthcare utilization may ultimately raise health insurance premia overall (Levine & Mulligan 2015). There is particular concern about overprescription, a type of overtreatment prevalent in the U.S. healthcare system (Ooi 2020). Overprescription occurs when medications are prescribed without strict adherence to medical guidelines, despite the availability of expert recommendations (Lyu 2017). This practice contributes to polypharmacy, financial burden, dissatisfaction with treatment (Peterson 2018), and drug-related morbidity and mortality (Hepler

[1] Overlabeling does not need to accompany overdiagnosis, but overdiagnosis usually accompanies overlabeling.

2001, Patel & Zed 2002). It is essential to investigate the costs and benefits from expanding diagnostic criteria to include pre-disease labels and investigate overlabeling concerns seriously.

To better understand labeling and any associated downstream consequences, we study hypothyroidism, which affects over 11.7% of U.S. Americans (Wyne et al. 2023) and an estimated 10% globally (Chiovato et al. 2019).[2] Hypothyroidism, or underactive thyroid, is a disease of the endocrine system that reflects insufficient production of thyroid hormone (thyroxine, or T4) from the thyroid gland. Common symptoms include fatigue, cold sensitivity, muscle ache, weight gain, and emotional and cognitive difficulty.

Traditional diagnosis is made with laboratory measurements, including thyroid stimulating hormone (TSH, or thyrotropin), which stimulates the thyroid gland to produce and secrete thyroid hormone.[3] Higher than normal levels of TSH suggest lower than normal thyroid hormone production, because the endocrine system is attempting to stimulate increased thyroid hormone production. In the United States, patients with normal free T4 levels and mildly elevated TSH levels (>4.5mIU/L and <10mIU/L) receive the pre-disease label of subclinical hypothyroidism. Patients with significantly elevated TSH levels (>10mIU/L) receive the diagnosis of overt hypothyroidism. A substantial share of patients with subclinical hypothyroidism eventually progress to overt hypothyroidism (Fatourechi 2009).

The standard of care for hypothyroidism management is levothyroxine, which acts as a thyroid hormone replacement (Jonklaas et al. 2014). Medical guidelines generally recommend that patients with subclinical hypothyroidism labels be monitored but not treated with levothyroxine

[2] The statistics here are provided based on the two main types of hypothyroidism identified via blood panels and referred to below as subclinical and overt hypothyroidism.
[3] Serum free T4 (FT4) is also commonly recommended in medical guidelines as an additional measure but seldom used (Sheehan 2016).

(Jonklaas et al. 2014, Calissendorff and Falhammar 2020, Biondi and Cappola 2022), especially in elderly patients (Du Puy 2021). For patients with overt hypothyroidism diagnoses, levothyroxine treatment is indicated and has been documented to improve outcomes (Jonklaas et al. 2014, Biondi and Cappola 2022).

Levothyroxine consistently ranks within the top five medications prescribed globally (Iversen et al. 2022, ClinCalc 2023, HealthPrep 2023). Perhaps unsurprisingly, then, overprescription of levothyroxine has been described for decades as a potential concern (Friedman 1992), both in the United States and worldwide (Ross et al. 1990, De Whalley 1995, Canaris 2000). About 30% of patients prescribed levothyroxine may not meet diagnostic criteria for overt hypothyroidism (Brito et al. 2021). In addition to the emotional weight and financial burden of an additional prescription, levothyroxine's side effect profile may increase patient risk for adverse events, unnecessarily.

One channel for overlabeling is information delivery, namely how patient laboratory (i.e., lab) values related to diagnostic thresholds are conveyed to clinicians. Rather than provide separate flags for patients with TSH lab values meeting labeling criteria for subclinical hypothyroidism and patients with TSH lab values meeting diagnostic criteria for overt hypothyroidism, electronic health record systems coarsely flag all TSH lab values outside of normal limits in the same way (see Figure S1 for illustrative examples for laboratory reports). From the perspective of clinicians, subclinical and overt hypothyroidism patients receive the same flag. If clinicians use a heuristic that associates this TSH flag with levothyroxine prescription, then clinicians may overprescribe subclinical hypothyroidism patients.

Establishing and quantifying the causal effects of labeling is critically important for policymakers and healthcare providers to introduce evidence-based solutions. Because it would be

unethical to assign patients to likely overlabeled conditions using a randomized controlled trial, the causal relationship between labeling and its sequelae is difficult to measure. Instead, we employ a regression discontinuity design (RDD) to study two thresholds of hypothyroidism salient for clinician decision-making, one for overt hypothyroidism – the prescription threshold where prescription is recommended – and one for subclinical hypothyroidism – the monitoring threshold where prescription is not guideline-directed. In our setting, prescription at the monitoring threshold may provide evidence of overlabeling. The RDD is a statistical method for obtaining quasi-random causal evidence from retrospective real-world observational data when certain verifiable statistical assumptions are met. The approach has been successfully used in a wide variety of research studies across a plethora of disciplines (e.g., see Almond 2010, Jensen 2015, Geneletti 2015, Venkataramani 2016, Lemp et al. 2023, Eyting et al. 2025) with strong evidence for internal and external validity (Chaplin et al. 2018). The RDD allows investigation of the causal effect of medical guidelines that introduce differential recommendations for clinical actions at a given threshold. For example, if a drug is prescribed with a higher likelihood above a certain labeled threshold for which there is no explicit recommendation to treat, then the treatment effect can be interpreted as evidence of overlabeling. The identifying assumption for causal interpretation of the RDD treatment effect is that patients slightly above or below the threshold are considered equal in all attributes other than their minimally different laboratory measurements, thus forming natural treatment and control groups.

This study is among the first to provide causal evidence of overlabeling and associated downstream effects on health and financial outcomes. We also contribute methodologically to the literature on overdiagnosis and overtreatment, where it is challenging to employ causal identification strategies due to ethical limitations on randomization. Using commercial claims data

for millions of patients over nearly two decades, we observe a significant jump in the diagnosis rate of hypothyroidism and prescription rate of levothyroxine at the TSH monitoring threshold of 4.5 mIU/L, at which guidelines recommend pre-disease labels of subclinical hypothyroidism but do not recommend prescription (Jonklaas et al. 2014). If we had observed improvements in patient health accompanying these diagnoses and prescriptions, then we would consider early treatment at the TSH monitoring threshold to be an effective policy. However, we find that health outcomes do not change in a clinically meaningful way at the monitoring threshold. Furthermore, associated healthcare costs increase substantially. We interpret these increases in costs without accompanying increases in benefits to be consistent with overlabeling. A back-of-the-envelope calculation reveals substantial potential to reduce costs for the US adult population. Cost saving estimates range from $1.64 billion within one year up to $2.77 billion within five years when moving the monitoring threshold closer to the prescription threshold at which treatment is guideline-directed.

Methods

Data source and sample selection

We use health insurance data from Optum's de-identified Clinformatics® Data Mart (CDM) database, which include claims for diagnoses, prescriptions, and laboratory test results for 90,857,976 patients across 10,236,502 providers from 2003 to 2021 (Detailed definitions of each outcome are provided in Table S1 and TSH laboratory, Levothyroxine prescription, and hypothyroidism diagnoses codes are available in Tables S2-S3). We focus on 15,683,706 unique patients with TSH laboratory measurements, since TSH is considered the best and (often) only test needed for evaluation of hypothyroidism (Sheehan 2016).[4] In clinical decision-making, upper

[4] In further checks in the supplements, we show tests with serum free T4 (FT4) measures.

limits of TSH measurements (i.e., maximum thresholds of the normal reference range) play a critical role, because they are prominently highlighted in medical records viewed by clinicians (Albert 2024). Values exceeding the upper limit may prompt further diagnostic actions, including monitoring, and/or medical treatment. The majority (85%) of TSH measurements have a laboratory-specific upper limit of 4.5 mIU/L or 5.5 mIU/L, and the modal threshold (56.1%) is 4.5 mIU/L. This upper limit of 4.5 mIU/L is particularly relevant, since it is used as the threshold for subclinical hypothyroidism labels (Wilson and Curry, 2005). Therefore, for estimation of our main treatment effects, we focus on patients who receive TSH labs with an associated upper limit of 4.5 mIU/L. Patients with prior levothyroxine prescriptions, prior hypothyroidism diagnoses, prior hyperthyroidism treatments, history of thyroid cancer, history of thyroidectomy, first levothyroxine prescription after 1 year from initial TSH measurement, pregnancy at time of TSH test, missing demographic information (e.g., race, age), and those younger than 18 years of age at time of first TSH test are excluded. The primary cohort for estimates around the 4.5 mIU/L threshold includes 772,715 patients (see Figure S2). Patients with alternative TSH values and upper limits are included in supplementary analyses.[5]

--- Table 1 ---

Table 1 provides descriptive statistics of patients identified by different bandwidths. The large bandwidth contains information ranging from TSH values between 0.5 mIU/L to 15 mIU/L and includes both the upper limit of 4.5 mIU/L (monitoring threshold associated with subclinical hypothyroidism label) and 10 mIU/L (prescription threshold associated with overt hypothyroidism). In the large bandwidth, 50.50% of patients are between the ages of 30 to 60 years, 54.50% are female, and 64.40% are white. Dates of TSH measurements range from January 1, 2003,

[5] We have a robustness check for all results with the 5.5 mIU/L upper limit which is available upon request.

through November 19, 2021. 49.2% of patients are continuously enrolled in CDM for at least 1 year after initial TSH measurement. Baseline hypertension and diabetes mellitus were found in 35% and 14% of patients, respectively. To estimate a global quadratic polynomial around the upper lab limit (and monitoring) threshold of 4.5 mIU/L, we construct a medium bandwidth of 2.0 to 7.0 mIU/L. However, neither the large nor medium bandwidths are sufficiently narrow to obtain a local average treatment effect via RDD. Hence, we construct a narrow bandwidth from 3.25 to 5.75 mIU/L. This bandwidth has a slightly lower share of adult patients (42.30% aged 30 to 60) and higher shares of female (56%), and white (67.70%) patients with baseline hypertension (39%) and diabetes mellitus (16%) measures but is otherwise similar to the broader patient pool. For each outcome of interest, we also construct optimal bandwidths following Imbens and Kalyanaraman (2012).

Satisfying the relevant statistical assumptions for RDD

The RDD allows us to measure the impact of crossing guideline thresholds on prescriptions, health outcomes, and financial outcomes due to quasi-random assignments above and below the TSH upper limit threshold of 4.5 mIU/L. With RDD assumptions satisfied, we can make causal intent-to-treat statements for patients that are sufficiently near the threshold.

The RDD requires that there is no clinician manipulation of laboratory values, in which case certain patients may be more likely above the threshold and selected to receive treatment relative to others. In our setting, laboratory tests are decentralized. Blood is drawn in a medical clinic or external laboratory; external laboratories outside clinician purview conduct assays and upload results to electronic medical record systems for clinicians to view. In this process, laboratory technicians lack incentives to manipulate results. Statistically, we do not observe evidence of

manipulation and find no change in the frequency of TSH values or covariates across the threshold (see Figures S3-S7 and Table S4).

Further, one may worry about other interventions occurring at the same threshold, in which case we cannot attribute our treatment effects to subclinical hypothyroidism labeling, alone. To our knowledge, no other relevant clinical decision-making interventions occur at the TSH threshold of 4.5 mIU/L. Therefore, we attribute any changes in outcomes to the salient subclinical hypothyroidism label.

Results

Prescription, diagnoses, and costs increase at monitoring threshold but not prescription threshold

Guidelines distributed by the American Thyroid Association (ATA) and the Association of American Family Physicians (AAFP) recommend levothyroxine prescription for management of overt hypothyroidism, which diagnosis occurs at TSH levels above 10 mIU/L (i.e., prescription threshold). In contrast, guidelines recommend routine monitoring but not prescription at TSH levels above 4.5 mIU/L (i.e., monitoring threshold) but below 10 mIU/L (more details under Methods/Description of hypothyroidism guidelines). Therefore, one may expect a large jump in prescriptions at the TSH threshold of 10 mIU/L but not at the TSH threshold of 4.5 mIU/L. Any jump observed at the monitoring threshold without accompanying health benefits is thus evidence for overlabeling. We attribute jumps at the monitoring threshold to threshold salience, since it is at this threshold that TSH values are first flagged as above normal limits.

--- Figure 1 ---

Figure 1 depicts prescription, diagnoses, and thyroid-related cost outcomes one year after initial TSH measurement.[6] In Figure 1 Panel A we do not observe any significant discontinuity in the share of patients with prescriptions above versus below the prescription threshold of 10 mIU/L.[7] However, we find a visually detectable jump in the share of patients with prescriptions at the monitoring threshold of 4.5 mIU/L, which coincides with the threshold for subclinical hypothyroidism labels. The jump at 4.5 mIU/L is a sizeable 15 percentage points (pp.) from a baseline level of 10%, which corresponds to a 150% relative increase in levothyroxine prescriptions. These results are robust to other prescription windows (see Figure S8).

Given the jump in the prescription propensity at the monitoring threshold, one may expect an associated increase in hypothyroidism diagnosis at this threshold. Indeed, Figure 1 Panel B shows an increase of around 20 pp. in the likelihood of diagnosis at the 4.5 mIU/L threshold, which corresponds to a 200% relative increase. When evaluating patients that receive both prescription and diagnosis (Figure S9 Panel A), we find a jump consistent in magnitude to that observed for prescriptions only. This is reassuring, since we would expect prescriptions to be accompanied by a formal diagnosis (e.g., for insurance purposes). From a financial perspective, we may expect cost increases to accompany observed increases in the share of prescriptions and diagnoses at the monitoring threshold. Figure 1 Panel C shows an increase in thyroid-related costs by $22 (19% above baseline).[8] These costs appear to be driven by levothyroxine prescription and refill costs (Figure S9 Panel B and Figure S10) and TSH laboratory test costs (Figure S9 Panel C).

Stark increases in prescriptions and corresponding costs over the monitoring threshold suggest that clinicians do not strictly adhere to medical guidelines for hypothyroidism management.

[6] Table 1 displays the treatment effects for these laboratory outcomes.
[7] The coefficient is insignificant at the five percent level.
[8] Thyroid related costs are defined based on thyroid-related prescriptions, laboratory measures, and procedures.

The results are robust to alternative specifications and bandwidths (see Figures S11-13). However, to claim that observed labeling practices are sub-optimal additionally requires that increases in prescriptions and diagnoses at the subclinical hypothyroidism threshold are not accompanied by associated improvements in health outcomes. Therefore, we study patient outcomes across two dimensions of health: laboratory-based measures and diagnosis-based measures.

No meaningful effects based on laboratory or diagnosis-based measures

To better understand whether prescription and diagnosis behavior are associated with improved health for patients with subclinical hypothyroidism labels, we study (1) whether there are any beneficial changes in thyroid-relevant laboratory outcomes, and (2) whether diagnosis-based health outcomes improve, for patients just above versus just below the monitoring threshold.

--- Figure 2 ---

Figure 2 depicts changes in laboratory outcomes one year after initial TSH measurement. Treatment effects are estimated linearly using the narrow TSH bandwidth of 3.25 to 5.75 mIU/L.[9] If levothyroxine prescriptions are beneficial, a substantial drop in TSH values is expected. However, Figure 2 Panel A shows a statistically significant but clinically insignificant decrease in TSH values of -0.16 mIU/L (3.5% relative to the baseline). While we do not observe a substantial change in average TSH values, one might still expect changes at the extremes of TSH values. Overly aggressive hypothyroidism treatment may result in iatrogenic thyrotoxicosis, a medically induced condition defined by too much thyroid hormone, rather than too little, from excessive levothyroxine use. Indeed, Figure 2 Panel B shows a significant increase in the fraction of TSH values below the lowest normal TSH value (0.4 mIU/L), which corresponds to the threshold of

[9] Table 2 Panel B displays the treatment effects for these laboratory outcomes.

concern for thyrotoxicosis. While concern for thyrotoxicosis is rare and affects only 2% of our sample below the monitoring threshold, this share increases by 0.4 pp (20% relative to the baseline) above the monitoring threshold, warranting attention, since thyrotoxicosis left untreated may result in serious medical complications (Ross et al. 2016).

A substantial literature discusses the benefits of levothyroxine in reducing low-density lipoprotein (LDL) cholesterol for individuals with hypothyroidism (Kotwal et al. 2020). If levothyroxine were beneficial for patients at the threshold, then a reduction in LDL would be expected. However, Figure 2 Panel C shows no significant reduction in LDL at the monitoring threshold. If at all, we observe an insignificant increase in LDL. Similarly, studies have documented that levothyroxine may improve kidney function (e.g., see Hennessey et al., 2021), which is measured using the estimated Glomerular Filtration Rate (eGFR). However, Figure 2 Panel D shows no clinically meaningful increase in eGFR at the monitoring threshold.[10]

--- Figure 3 ---

Figure 3 illustrates changes in diagnosis-based health outcomes one year after initial TSH measurement.[11] Thyroid function is biologically linked to both skeletal and cardiovascular physiology, and several studies have found associations between thyroid pathology and fracture risk (Tuchendler and Bolanowski 2014) as well as thyroid pathology and cardiovascular risk (Razvi et al. 2018). Were hypothyroidism effectively treated at the monitoring threshold, we might expect decreases in the share of patients with fractures and cardiovascular complications above versus below the threshold. Figure 2 Panel A and Panel B show the share of patients with fracture

[10] Importantly, statistically significant estimates, as shown in Table 2 Panel B, are not robust to alternative functional form specifications either. A robustness check using a quadratic shows an insignificant treatment effect on eGFR (available upon request).

[11] Table 2 Panel C displays the treatment effects for these diagnosis-based health outcomes.

diagnoses and CVD hospitalizations, respectively. We find no significant difference in shares of affected patients at the monitoring threshold. We also study aggregate measures of health, including all-cause hospitalization and all-cause mortality, which provide a comprehensive view into health and capture other aspects of health that may come from early disease management. To justify the current treatment regime for patients with subclinical hypothyroidism labels but not overt hypothyroidism diagnoses, we might expect reductions in the likelihood of hospitalization and mortality at the monitoring threshold. However, both Panel C and Panel D, which depict all-cause mortality and all-cause hospitalization, respectively, do not indicate any change when crossing this threshold.

All results from Figures 1-3 are robust to longer time horizons of 2, 3, or 5 years (Table S5-7), inclusion of covariates, different weights (Table S8), an instrumental variables strategy (Table S9), and pooling of the two most prevalent upper limits of the reference range, i.e. 4.5 mIU/L and 5.5 mIU/L (Table S10). The results are also robust to alternative specifications and different bandwidths (see Figures S14-21). As such, we find no evidence of clinically meaningful changes in laboratory-based or diagnosis-based proxies for patient health across the TSH monitoring threshold of 4.5 mIU/L.

Moving monitoring threshold towards prescription threshold could improve health

Treatment for patients with subclinical hypothyroidism labels at the TSH monitoring threshold of 4.5 mIU/L appears to increase costs without improving—and potentially harming— health outcomes. We interpret these findings about pre-disease labeling practices to be evidence of overlabeling via channels of overdiagnosis and overtreatment (in this context, overprescription).

One feasible solution to mitigate overlabeling concerns could be to move the monitoring threshold used for subclinical hypothyroidism labeling closer to the prescription threshold used for overt hypothyroidism diagnosis. Following Dong and Lewbel (2015), we investigate the implications of this policy by simulating treatment effects for counterfactual labeling thresholds (see Methods/Statistical analysis: extrapolation away from the threshold).

--- Table 3 ---

Table 3 shows extrapolated treatment effects from the status quo monitoring threshold of 4.5 mIU/L to new thresholds, including 5.0 mIU/L, 6.0 mIU/L, and 7.0 mIU/L, the last of which has been recommended by the literature (Ross 2022). Table 3 Panel A shows that jumps in levothyroxine prescriptions, hypothyroidism diagnoses, and thyroid-related costs increase when moving the monitoring threshold towards the prescription threshold. These increases are expected—as TSH values approach diagnostic criteria for overt hypothyroidism, and patient symptoms likely increase, clinicians should increasingly manage concerns for hypothyroidism. Table 3 Panel B shows simulated treatment effects for laboratory outcomes. As the monitoring threshold moves further to the right, we observe stronger predicted reductions in future TSH values and increases in eGFR values, suggestive improvements to lab-based proxies for health. However, we also predict increases in LDL values and in the share of TSH values in the range of thyrotoxicosis, which may warrant concern. While we predict some changes in laboratory outcomes when moving the threshold, Table 3 Panel C shows that diagnosis-based health outcomes only marginally change.

Back-of-the-envelope cost calculation when moving the threshold upwards

While thyroid-related costs increase when extrapolating the monitoring threshold upwards, these costs alone are not sufficient for a holistic understanding of the impact of changes to the monitoring threshold. For example, this measure does not contain visits to the doctor, which monitoring can comprise a large portion of healthcare expenditure and be expected to decrease as the share of monitored patients decreases (i.e., as the monitoring threshold is shifted upwards). Further, hypothyroidism may be associated with additional comorbidity burden or adverse effects from treatment (Taylor et al. 2024) that may not be captured without other cost measures. Additionally, behavioral effects from early hypothyroidism labels could lead patients to request healthcare services that they otherwise would not seek. To account for these possibilities, we investigate more comprehensive cost measures.

Table S11 depicts treatment effects on our monitoring threshold and extrapolated thresholds as well as a back-of-the-envelope calculation based on 1-year estimates for a plethora of cost outcomes. Table S11 Panel A mirrors Table 3 with a focus on a range of broader cost categories.[12] We find increases in costs from visits to general physicians (GP), the emergency room (ER), and inpatient hospitalizations when crossing the monitoring threshold. Indeed, one-year total standardized costs increase by around $256 on average per person, which is driven by inpatient costs.[13] These findings appear to be consistent with the idea that patients with additional comorbidities (i.e., from an additional diagnosis of hypothyroidism) receive more unnecessary costly attention from physicians, particularly in the inpatient setting. Indeed, we observe increases in general physician visits, emergency room visits, and new and established patient visits (Figure

[12] We have estimates for more granular information on costs for all the subcategories, including costs on comorbidities. All measures of costs are run with the main sample of 772,715 patients where patients that did not incur costs of a particular type receive a cost of zero.

[13] We also find significant increases in total standardized costs after year two and onwards (available upon request).

298 S22 Panels A-D), all of which are consistent with increased monitoring. Importantly, in
299 counterfactual simulations, total standardized costs drop substantially by $424 when moving from
300 4.5 mIU/L to 7.0 mIU/L. These results suggest substantial financial benefits from prescribing and
301 diagnosing at higher thresholds, which likely stem from more tailored management of patients with
302 concern for hypothyroidism.[14]

303 We use treatment effects from total standardized costs for a back of the envelope calculation
304 in Table S11 Panel B to better understand the counterfactual world in which labeling would occur
305 at a TSH value of 7 mIU/L rather than 4.5 mIU/L, as has been recommended by the medical
306 literature (Ross 2022). First, we multiply the prescription jump of 13.16 pp with the sample size of
307 patients in small and medium bandwidths (772,715 and 6,996,054 patients, respectively). Next, we
308 multiply this product with the extrapolated treatment effect at the 7 mIU/L threshold for one year
309 (-$424.85) and five years (-$717.85).[15] Finally, we interpret these results as estimated cost savings
310 of about $43 to $72 million for the small sample and $391 to $660 million for the medium sample
311 after one and five years, respectively. If instead we use the population of affected patients (11% of
312 266 million adults) across the United States, we arrive at total annual costs savings ranging from
313 $1.6 to $2.7 billion after one and five years, respectively.[16]

314 **Discussion**

315 This study investigates costs and benefits of pre-disease labeling for one of the most globally
316 prevalent diseases, hypothyroidism. Utilizing a regression discontinuity design, we find evidence

[14] The histogram of TSH values in Figure S3 provides additional supporting evidence for this notion.
[15] An equivalent version of the back-of-the envelope calculation could have used the jump in diagnosis of 19.67 pp. However, the 13.16 pp. allows for a more conservative estimate and as such, we use it here.
[16] We estimated the subclinical hypothyroidism prevalence to be 11% by scaling the 11.7% total hypothyroidism prevalence from Wyne et al. (2023) by 94%—the proportion of subclinical cases (TSH ≥4.5 and <10 mIU/L) within all hypothyroidism cases (TSH ≥4.5 mIU/L) found in our claims data. Overall, we consider the estimates to be conservative with a separate upper bound for costs savings based on charges exhibiting a higher savings potential (available upon request).

of overprescribing near the upper limit of normal TSH measures, which coincides with the labeling threshold for subclinical hypothyroidism, at which monitoring is guideline-recommended but not pharmaceutical treatment. We also observe associated increases in thyroid-related and total costs without evidence of improvements in health outcomes. Counterfactual simulations suggest that moving the threshold used to label subclinical hypothyroidism may yield potential cost savings in the United States from $1.6 to $2.7 billion. These results provide a basis for policymakers and practitioners alike to reconsider pre-disease labeling practices and address overlabeling concerns in hypothyroidism.

The "real world" results in this paper are, in a nuanced manner, comparable to those of a randomized clinical trial (RCT). While RCTs are valuable, their statistical power is often constrained by sample size. By investigating threshold treatment behavior on diagnoses, prescriptions, and monitoring practices in general, we add to the literature from previously published RCTs. It is reassuring that our results are consistent with findings from several RCTs, notably on eGFR (Tang 2018) and fractures (Mooijaart 2019).[17] On the other hand, mortality benefits for patients younger than 65 years compared to older patients were not observed (Peng 2021), nor did we find significant decreases in LDL (Li 2017), for patients more likely to receive levothyroxine (i.e., those above the monitoring threshold).

Applying rigorous quasi-experimental methods to advance our understanding of overlabeling

This study demonstrates overlabeling near the upper normal limit of TSH for patients with subclinical hypothyroidism labels. In this context, the consequences of overlabeling are realized through overdiagnosis and overtreatment. Overlabeling is not well studied. To our knowledge, only

[17] Similar to Mooijaart (2019), we also find null effects on the outcomes of atrial fibrillation and other cardiovascular events (available upon request).

two papers explore related issues but in a separate context of prediabetes and with a focus on the diagnostic threshold (Iizuka et al. 2021, Alalouf et al. 2024). Concerns about overdiagnosis and overtreatment have been documented for decades, especially in the most recent opioid crisis (Welch et al. 2011, Brownlee et al. 2017, Makary et al. 2017). However, this phenomenon has seldom been investigated rigorously within a causal framework for such a large population nor at such low TSH levels, which is essential for making prescription decisions (Feller 2018). Though levothyroxine prescription is guideline-recommended at TSH levels of 10 mIU/L for overt hypothyroidism, we find diagnosis and prescription increases at TSH levels of 4.5 mIU/L for subclinical hypothyroidism labels (the upper limit of the reference range for many laboratories), at which threshold guidelines recommend monitoring but not prescribing. We also find associated increases in healthcare expenditure. Importantly, we do not find associated improvements in thyroid-related or aggregate measures of health outcomes.[18] Therefore, this study finds evidence for overlabeling driven by overdiagnosis and overprescription.

Costs from overlabeling are especially important to consider. Spending is a major challenge in the U.S. healthcare system. U.S. health expenditure accounted for 17.6% of GDP in 2023, which amounts to $ 4.9 trillion (CMS 2023), and is projected to grow to 20% of GDP by 2033 (Keehan 2025). Relative to other industrialized nations, after excluding R&D expenditure, the U.S. has the highest prescription drug costs per capita (Kesselheim et al. 2016). In fact, prescriptions account for 41.9% of Medicare spending (CMS 2023). Saving on unnecessary medications could lead to large reductions in overall costs, especially for commonly prescribed drugs like levothyroxine, without reducing care quality. Similarly, healthcare inpatient and outpatient visit costs account for

[18] This result is consistent with findings in other causal studies of health interventions, where increased engagement or utilization did not lead to measurable health benefits (Hoffmann et al. 2025).

around 61% of overall U.S. healthcare expenditure (HealthSystem Tracker 2024), implying a large potential for reduction in costs from optimizing visits towards efficiently labeled patients who would benefit most from monitoring.

We also note significant cost increases from TSH laboratory tests. These costs are not necessarily sub-optimal in the status quo, because TSH tests are used for guideline-directed monitoring of subclinical hypothyroidism, however if the monitoring threshold is too low, then these costs are unnecessarily high. While we find evidence of monitoring activity, large jumps in prescriptions and accompanying costs suggest that clinicians do not strictly adhere to medical guidelines for hypothyroidism management. Perhaps this could be rationalized by acknowledging clinical pressures. Short visit times common in primary care may lead to overprescription (Neprash et al. 2023). Behavioral interventions seem unlikely to drive downstream outcomes, since standard of care does not include recommendations beyond thyroid-specific monitoring or prescribing. Therefore, it is plausible that any change in medical and financial outcomes when crossing the 4.5 mIU/L monitoring threshold is driven by prescriptions, tests, and diagnosis-related visits.

The regression discontinuity analysis jointly with an extrapolation exercise offers an analytical way to estimate potential savings in healthcare expenditure achieved by counterfactual pre-disease labeling thresholds. These results suggest that moving the monitoring threshold from TSH values of 4.5 mIU/L to 7 mIU/L could result in overall healthcare savings of around $391 million to $660 million for one and five years, respectively, within the sample. Extrapolating to the United States adult population, we obtain a cost savings estimate near $1.6 to 2.7 billion after one year or five years, respectively. These savings arise without clinically meaningful changes to patient health outcomes. As such, the potential economic value from increasing the labeling threshold of subclinical hypothyroidism from 4.5 mIU/L to 7 mIU/L is immense.

This large observational study has several limitations. The data used has limited national coverage and only includes patients of UnitedHealthcare, which comprised about 12% of the insured U.S. population in 2022 (Statista, 2022). Commercial claims data have limited coverage of publicly insured individuals, namely Traditional Medicare and Medicaid beneficiaries (Gunaseelan et al. 2019). A limitation of the regression discontinuity design is that the treatment effects estimated at 4.5 mIU/L are local average treatment effects that are only valid close to the threshold. Results from the extrapolation exercise must similarly be interpreted, and the reader must draw conclusions about whether 7 mIU/L is already too far from the threshold. We argue that our robustness checks (e.g., global polynomial functional form, alternative bandwidths) allow for such an exercise. Further, while this is one of the largest samples on the topic of overlabeling, and certainly in the context of hypothyroidism, the data is gathered in the context of the U.S. healthcare system. Additional studies in other contexts are required to validate the external validity of these causal findings.[19] Moreover, commercial claims do not contain information on patient-provider interactions, and hence, the conversations guiding clinical decision-making for patient encounters are unknown. While we have shown evidence consistent with clinician prescribing behavior, we cannot rule out the role played by patient preferences. However, were patient preferences to account for a significant share of prescriptions, one may expect these preferences to be amplified for even higher serum TSH values, but we observe no large visible jump in prescriptions at serum TSH values of 10 mIU/L.[20]

[19] It is encouraging that the authors find similar jumps in prescribing behavior of levothyroxine in the United States using samples from the Merative Marketscan (previously: IBM Marketscan) data which include patients from large employers, managed care organizations, hospitals, EMR providers, Medicare and Medicaid. The authors similarly find a similar jump in prescribing behavior in the United Kingdom using the CPRD data.

[20] See also Table S12.

Another limitation of this study is that we can neither observe whether patients are symptomatic nor measure changes in quality of life. While guidelines from both the United States Preventive Task Force and the Endocrine society do not recommend treatment at TSH values below 10 mIU/mL (Helfand 2004, Surks 2004, Rugge 2015, Bekkering 2019), joint ATA and American Association of Clinical Endocrinologists (Garber 2012) state that the benefits for patients between TSH levels of 4.5 mIU/L and 10 mIU/L are less certain: prescriptions can be considered when patients are symptomatic. However, due to the non-specific nature of hypothyroidism-induced symptoms (e.g., fatigue), defining "symptomatic" consistently is challenging. Even if patients do not receive TSH tests because of symptoms, they may still *ex post* attribute non-specific symptoms to hypothyroidism after receiving an abnormal TSH value.

However, there are multiple reasons why improvements in symptoms or quality of life are unlikely to drive the jump in prescriptions at the monitoring threshold. First, there is no reason to expect more symptoms for patients with TSH values marginally above versus below the monitoring threshold of 4.5 mIU/L. Second, we do not find a smaller jump for the subgroup of patients undergoing regular health examinations, which is consistent with the idea that prescriptions are not driven by hypothyroidism-related symptoms (for further analyses with subgroups that are more likely to be symptomatic – e.g., older adults or adults with FT4 labs drawn –, see *statistical analysis: heterogeneous treatment effects*). These findings suggest that other factors may play a role in observed overlabeling patterns, including but not limited to malpractice concerns, patient-driven demand, difficulties in accessing previous medical records, stress, and exhaustion (Lyu 2017). Therefore, we might expect any vague indication for subclinical hypothyroidism to contribute substantially to overprescription of levothyroxine. One might also wonder whether treatment at the monitoring threshold might benefit quality of life. However, a substantial body of evidence reports

no improvements to quality of life (Feller et al. 2018, Chaker et al. 2022, Hegedüs et al. 2022) after the prescription of levothyroxine at the monitoring threshold for subclinical hypothyroidism.[21] To conclude, we find it unlikely that quality of life or symptoms are driving our results.

Overall, the findings from this study are relevant to policymakers and healthcare organizations responsible for clinical guidelines and to clinicians responsible for treatment decisions. Pre-disease labels like subclinical hypothyroidism are not limited to hypothyroidism but exist for an expanding group of common chronic conditions, including diabetes, hypertension, and obesity. While early labeling may be an effective tool for disease prevention and management, this work illustrates key tradeoffs. Establishing labels too early may result in unnecessary medical intervention and substantial financial burdens without associated clinically meaningful benefits. In the context of hypothyroidism, we suggest increasing the current labeling threshold as well as establishing continuous policy evaluations for calibration of optimal labeling guidelines. An alternative but consistent approach would be to improve personalized medicine through artificial intelligence such as machine learning methods. As such, our findings highlight the need for a more nuanced approach to diagnostic and treatment thresholds common in clinical settings.

[21] Even for older adults, which group may be expected to be more likely to experience positive effects, the literature does not find any evidence on improvements in quality of life (Zhao et al. 2022).

Main references

Alalouf, M., Miller, S., & Wherry, L. R. (2024). What difference does a diagnosis make? Evidence from marginal patients. *American Journal of Health Economics*, 10(1), 97-131.

Albert, Charna (2024). Digging into the interpretation of TSH results. College of American Pathologists. URL: https://www.captodayonline.com/digging-into-the-interpretation-of-tsh-results.

Almond, D., Doyle Jr, J. J., Kowalski, A. E., & Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, 125(2), 591-634.

Bekkering, G. E., Agoritsas, T., Lytvyn, L., Heen, A. F., Feller, M., Moutzouri, E., Abdulazeem, H., Aertgeerts, B., Beecher, D., Brito, J.P., Farhoumand, P.D., Ospina, N.S., Rodondi, N., van Driel, M., Wallace, E., Snel, M., Okwen, P.M., Siemieniuk, R., Vandvik, P.O., Kuijpers, T., & Vermandere, M. (2019). Thyroid hormones treatment for subclinical hypothyroidism: a clinical practice guideline. *BMJ*, 365.

Biondi, B., & Cappola, A. R. (2022). Subclinical hypothyroidism in older individuals. *The Lancet Diabetes & Endocrinology*, 10(2), 129-141.

Brito, J. P., Ross, J. S., El Kawkgi, O. M., Maraka, S., Deng, Y., Shah, N. D., & Lipska, K. J. (2021). Levothyroxine use in the United States, 2008-2018. *JAMA Internal Medicine*, 181(10), 1402-1405.

Brito, J. P., Deng, Y., Ross, J. S., Choi, N. H., Graham, D. J., Qiang, Y., Rantou, E., Wang, Z., Zhao, L., Shah, N.D. & Lipska, K.J. (2022). Association between generic-to-generic

levothyroxine switching and thyrotropin levels among US adults. *JAMA Internal Medicine*, 182(4), 418-425.

Brownlee, S., Chalkidou, K., Doust, J., Elshaug, A. G., Glasziou, P., Heath, I., Nagpal, S., Saini, V., Srivastava, D., Chalmers, K., & Korenstein, D. (2017). Evidence for overuse of medical services around the world. *The Lancet*, 390(10090), 156-168.

Calissendorff, J., & Falhammar, H. (2020). To treat or not to treat subclinical hypothyroidism, what is the evidence?. *Medicina*, 56(1), 40.

Chaker, L., Razvi, S., Bensenor, I. M., & Azizi, F. Elizabeth N. Pearce and Robin P. Peeters (2022). Hypothyroidism. *Nature Reviews Disease Primers*, 8(30).

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37(2), 403-429.

Chen, S., Kuhn, M., Prettner, K., Bloom, D. E., & Wang, C. (2021). Macro-level efficiency of health expenditure: Estimates for 15 major economies. *Social Science & Medicine*, 287, 114270.

CMS (2023). NHE Fact Sheet. URL: https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.

ClinCalc (2023). The Top 300 of 2020, Provided by the ClinCalc DrugStats Database. URL: https://clincalc.com/DrugStats/Top300Drugs.aspx.

De Whalley, P. H. I. L. I. P. (1995). Do abnormal thyroid stimulating hormone level values result in treatment changes? A study of patients on thyroxine in one general practice. *British Journal of General Practice*, 45(391), 93-95.

Drake, T. (2021). Levothyroxine Is Overprescribed More Than You Realize. *Clinical Thyroidology*, 33(8), 339-341.

Du Puy, R. S., Poortvliet, R. K., Mooijaart, S. P., Den Elzen, W. P., Jagger, C., Pearce, S. H., Arai, Y., Hirose, N., Teh, R., Menzies, O., Rolleston, A., Kerse, N., & Gussekloo, J. (2021). Outcomes of thyroid dysfunction in people aged eighty years and older: an individual patient data meta-analysis of four prospective studies (Towards Understanding Longitudinal International Older People Studies Consortium). *Thyroid*, 31(4), 552-562.

Eyting, M., Xie, M., Michalik, F., Heß, S., Chung, S., & Geldsetzer, P. (2025). A natural experiment on the effect of herpes zoster vaccination on dementia. *Nature*, 1-9.

Fatourechi, V. (2009, January). Subclinical hypothyroidism: an update for primary care physicians. In *Mayo Clinic Proceedings*, Vol. 84, No. 1, pp. 65-71.

Feller, M., Snel, M., Moutzouri, E., Bauer, D. C., De Montmollin, M., Aujesky, D., Ford, I., Gussekloo, J., Kearney, P. M., Mooijaart, S., Quinn, T., Stott, D., Westendorp, R., Rodondi, N. & Dekkers, O. M. (2018). Association of thyroid hormone therapy with quality of life and thyroid-related symptoms in patients with subclinical hypothyroidism: a systematic review and meta-analysis. Jama, 320(13), 1349-1359.

Friedman, D., Reed, R. L., & Mooradian, A. D. (1992). The prevalence of overmedication with levothyroxine in ambulatory elderly patients. *Age*, 15, 9-13.

Garber, J. R., Cobin, R. H., Gharib, H., Hennessey, J. V., Klein, I., Mechanick, J. I., Pessah-Pollack, R., Singer, P. A., & Woeber for the American Association of Clinical Endocrinologists and American Thyroid Association Taskforce on Hypothyroidism in Adults, K. A. (2012). Clinical practice guidelines for hypothyroidism in adults: cosponsored by the American Association of Clinical Endocrinologists and the American Thyroid Association. *Thyroid*, 22(12), 1200-1235.

Geneletti, S., O'Keeffe, A. G., Sharples, L. D., Richardson, S., & Baio, G. (2015). Bayesian regression discontinuity designs: incorporating clinical knowledge in the causal analysis of primary care data. *Statistics in Medicine*, 34(15), 2334-2352.

Gunaseelan, V., Kenney, B., Lee, J. S. J., & Hu, H. M. (2019). Databases for surgical health services research: Clinformatics Data Mart. *Surgery*, 165(4), 669-671.

HealthPrep (2023). The 10 Most Prescribed Drugs In The World. URL: https://healthprep.com/articles/medications/the-10-most-prescribed-drugs-in-the-world/

HealthSystem Tracker (2024). What drives health spending in the U.S. compared to other countries? URL: https://www.healthsystemtracker.org/brief/what-drives-health-spending-in-the-u-s-compared-to-other-countries.

Hegedüs, L., Bianco, A. C., Jonklaas, J., Pearce, S. H., Weetman, A. P., & Perros, P. (2022). Primary hypothyroidism and quality of life. *Nature Reviews Endocrinology*, 18(4), 230-242.

Helfand, M. (2004). Screening for subclinical thyroid dysfunction in nonpregnant adults: a summary of the evidence for the US Preventive Services Task Force. *Annals of Internal Medicine*, 140(2), 128-141.

Hennessey, J. V., Weir, M. R., Soni-Brahmbhatt, S., Duan, Y., & Gossain, V. V. (2021). Effect of levothyroxine on kidney function in chronic kidney disease with subclinical hypothyroidism in US veterans: a retrospective observational cohort study. *Advances in Therapy*, 38, 1185-1201.

Hepler, C. D. (2001). Regulating for outcomes as a systems response to the problem of drug-related morbidity. *Journal of the American Pharmaceutical Association*, (1996), 41(1), 108-115.

Hoffmann, M., Mosquera, R., & Chadi, A. (2025). Vaccines at Work: Experimental Evidence from a Health Campaign. *Management Science*.

Iacobucci, G. (2025). Define obesity as clinical or pre-clinical for more accurate diagnosis, says global commission. *BMJ* 388-(r78).

Jensen, V. M., & Wüst, M. (2015). Can Caesarean section improve child and maternal health? The case of breech babies. *Journal of Health Economics*, 39, 289-302.

Jonklaas, J., Bianco, A. C., Bauer, A. J., Burman, K. D., Cappola, A. R., Celi, F. S., Cooper, D. S., Kim, B.W., Peeters, R.P., Rosenthal, M.S. & Sawka, A. M. (2014). Guidelines for the treatment of hypothyroidism: prepared by the american thyroid association task force on thyroid hormone replacement. *Thyroid*, 24(12), 1670-1751.

Iizuka, T., Nishiyama, K., Chen, B., & Eggleston, K. (2021). False alarm? Estimating the marginal value of health signals. *Journal of Public Economics*, 195, 104368.

Iversen, D. B., Andersen, N. E., Dalgård Dunvald, A. C., Pottegård, A., & Stage, T. B. (2022). Drug metabolism and drug transport of the 100 most prescribed oral drugs. *Basic & Clinical Pharmacology & Toxicology*, 131(5), 311-324.

Keehan, S. P., Madison, A. J., Poisal, J. A., Cuckler, G. A., Smith, S. D., Sisko, A. M., Fiore, J. A., & Rennie, K. E. (2025). National Health Expenditure Projections, 2024–33: Despite Insurance Coverage Declines, Health To Grow As Share Of GDP: Article features National Health Expenditure Projections, 2024-33. *Health Affairs*, 10-1377.

Kesselheim, A. S., Avorn, J., & Sarpatwari, A. (2016). The high cost of prescription drugs in the United States: origins and prospects for reform. *Jama*, 316(8), 858-871.

Kotwal, A., Cortes, T., Genere, N., Hamidi, O., Jasim, S., Newman, C. B., Prokop, L. J., Murad, M.H., & Alahdab, F. (2020). Treatment of thyroid dysfunction and serum lipids: a systematic review and meta-analysis. *The Journal of Clinical Endocrinology & Metabolism*, 105(12), 3683-3694.

Lemp, J. M., Bommer, C., Xie, M., Michalik, F., Jani, A., Davies, J. I., Bärnighausen, T., Vollmer, S., & Geldsetzer, P. (2023). Quasi-experimental evaluation of a nationwide diabetes prevention programme. *Nature*, 624(7990), 138-144.

Levine, D., & Mulligan, J. (2015). Overutilization, overutilized. *Journal of Health Politics, Policy and Law*, 40(2), 421-437.

Li, X., Wang, Y., Guan, Q., Zhao, J., & Gao, L. (2017). The lipid-lowering effect of levothyroxine in patients with subclinical hypothyroidism: A systematic review and meta-analysis of randomized controlled trials. *Clinical Endocrinology*, 87(1), 1-9.

Lillevang-Johansen, M., Abrahamsen, B., Jørgensen, H. L., Brix, T. H., & Hegedüs, L. (2018). Over-and under-treatment of hypothyroidism is associated with excess mortality: a register-based cohort study. *Thyroid*, 28(5), 566-574.

Lyu, H., Xu, T., Brotman, D., Mayer-Blackwell, B., Cooper, M., Daniel, M., Wick, E.C., Saini, V., Brownlee, S., & Makary, M.A. (2017). Overtreatment in the United States. *PloS One*, 12(9), e0181970.

Neprash, H. T., Mulcahy, J. F., Cross, D. A., Gaugler, J. E., Golberstein, E., & Ganguli, I. (2023). Association of primary care visit length with potentially inappropriate prescribing. In *JAMA Health Forum* (Vol. 4, No. 3, pp. e230052-e230052). American Medical Association.

Makary, M. A., Overton, H. N., & Wang, P. (2017). Overprescribing is major contributor to opioid crisis. *BMJ*, 359.

Mooijaart, S. P., Du Puy, R. S., Stott, D. J., Kearney, P. M., Rodondi, N., Westendorp, R. G., den Elzen, W.P.J., Postmus, I., Poortvliet, R.K.E., van Heemst, D., van Munster, B.C., Peeters, R.P., Ford, I., Kean, S., Messow, C.-M., Blum, M.R., Collet, T.-H., Watt, T., Dekkers, O.M., Jukema, J.W., Smit , J.W.A., Langhorne, P., & Gussekloo, J. (2019). Association between levothyroxine treatment and thyroid-related symptoms among adults aged 80 years and older with subclinical hypothyroidism. *JAMA*, 322(20), 1977-1986.

Ooi, K. (2020). The pitfalls of overtreatment: Why more care is not necessarily beneficial. *Asian Bioethics Review*, 12(4), 399-417.

Patel, P., & Zed, P. J. (2002). Drug-related visits to the emergency department: how big is the problem? *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 22(7), 915-923.

Peng, C. C. H., Huang, H. K., Wu, B. B. C., Chang, R. H. E., Tu, Y. K., & Munir, K. M. (2021). Association of thyroid hormone therapy with mortality in subclinical hypothyroidism: a

systematic review and meta-analysis. *The Journal of Clinical Endocrinology & Metabolism*, 106(1), 292-303.

Peterson, S. J., Cappola, A. R., Castro, M. R., Dayan, C. M., Farwell, A. P., Hennessey, J. V., Kopp, P.A., Ross, D.S., Samuels, M.H., Sawka, A.M., Taylor, P.N., Jonklaas, J., & Bianco, A.C. (2018). An online survey of hypothyroid patients demonstrates prominent dissatisfaction. *Thyroid*, 28(6), 707-721.

Ross D.S., Daniels G.H., & Gouveia D. (1990). The use and limitations of a chemiluminescent thyrotropin assay as a single thyroid function test in an out-patient endocrine clinic. *The Journal of Clinical Endocrinology & Metabolism*, 71(3), 764-769.

Ross, D. S., Burch, H. B., Cooper, D. S., Greenlee, M. C., Laurberg, P., Maia, A. L., Rivkees, S.A., Samuels, M., Sosa, J.A., Stan, M.N. & Walter, M. A. (2016). 2016 American Thyroid Association guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis. *Thyroid*, 26(10), 1343-1421.

Rugge, J. B., Bougatsos, C., & Chou, R. (2015). Screening and treatment of thyroid dysfunction: an evidence review for the US Preventive Services Task Force. *Annals of Internal Medicine*, 162(1), 35-45.

Sacarny, Adam, David Yokum, and Shantanu Agrawal. 2017. "Government-Academic Partnerships in Randomized Evaluations: The Case of Inappropriate Prescribing." *American Economic Review*, 107 (5): 466-70.

Schneider, C., Feller, M., Bauer, D. C., Collet, T. H., da Costa, B. R., Auer, R., Peeters, R. P., Brown, S. J., Bremmer, A. P., O'Leary, P. C., Feddema, P, Leedman, P. J., Aujesky, D., Walsh,

J. P. & Rodondi, N. (2018). Initial evaluation of thyroid dysfunction-Are simultaneous TSH and fT4 tests necessary?. *PloS One*, 13(4), e0196631.

Sheehan, M. T. (2016). Biochemical testing of the thyroid: TSH is the best and, oftentimes, only test needed–a review for primary care. *Clinical Medicine & Research*, 14(2), 83-92.

Silverman, H. M. (2011). The Pill Book: The Illustrated Guide to the Most-prescribed Drugs in the United States. Bantam.

Silverstein, W. K., & Grady, D. (2021). Overuse of Levothyroxine in Patients With Subclinical Hypothyroidism: Time to "Leve"-Out-Thyroxine. *JAMA Internal Medicine*, 181(10), 1286-1287.

Stanford Center for Population Health Sciences (2021). Optum DOD (v5.0). Redivis. (Dataset) https://redivis.com/datasets/31mg-exzk3fja0?v=5.0.

Statista (2022). Market share of leading health insurance companies in the United States as of 2022. Source: https://www.statista.com/statistics/761446/leading-us-health-insurers-in-the-us-covered-lives.

Surks, M. I., Ortiz, E., Daniels, G. H., Sawin, C. T., Col, N. F., Cobin, R. H., Franklyn, J.A., Hershman, J.M., Burman, K.D., Denke, M.A., Gorman, C., Cooper, R.S., & Weissman, N. J. (2004). Subclinical thyroid disease: scientific review and guidelines for diagnosis and management. *JAMA*, 291(2), 228-238.

Taylor, P.N., Medici, M.M., Hubalewska-Dydejczyk, A., & Boelaert, K. (2024). Hypothyroidism, The Lancet, 404(10460), 1347-1364.

Tang, W., Chen, Q., Chen, L., Chen, S., Shao, X., & Wang, X. (2018). Favorable effect of levothyroxine on nutritional status of patients with stage 3-4 chronic kidney disease. *Acta Endocrinologica* (Bucharest), 14(3), 338.

Tuchendler, D., & Bolanowski, M. (2014). The influence of thyroid dysfunction on bone metabolism. *Thyroid Research*, 7(1), 12.

Venkataramani, A. S., Bor, J., & Jena, A. B. (2016). Regression discontinuity designs in healthcare research. *BMJ*, 352.

Viera, A. J. (2011). Predisease: when does it make sense?. *Epidemiologic Reviews*, 33(1), 122-134.

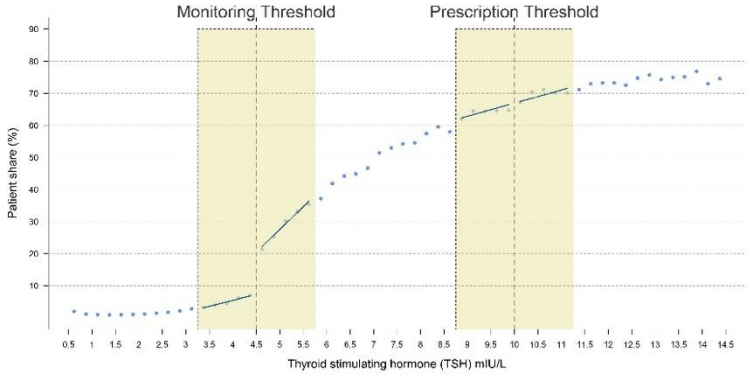
Welch, H. G., Schwartz, L. M., & Woloshin, S. (2011) - Overdiagnosed: Making People Sick in the Pursuit of Health.

Wyne, K. L., Nair, L., Schneiderman, C. P., Pinsky, B., Antunez Flores, O., Guo, D., Barger B. & Tessnow, A. H. (2023). Hypothyroidism prevalence in the United States: a retrospective study combining national health and nutrition examination survey and claims data, 2009–2019. *Journal of the Endocrine Society*, 7(1), bvac172.

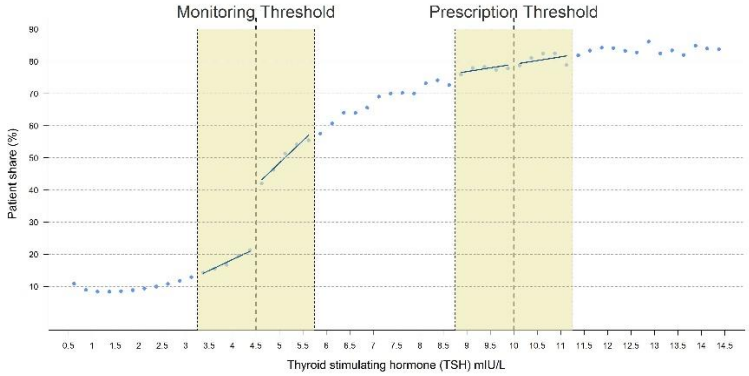
Zhao, C., Wang, Y., Xiao, L., & Li, L. (2022). Effect of levothyroxine on older patients with subclinical hypothyroidism: a systematic review and meta-analysis. *Frontiers in Endocrinology*, 13, 913749.

Figure 1. Prescription uptake and costs across thyroid stimulating hormone

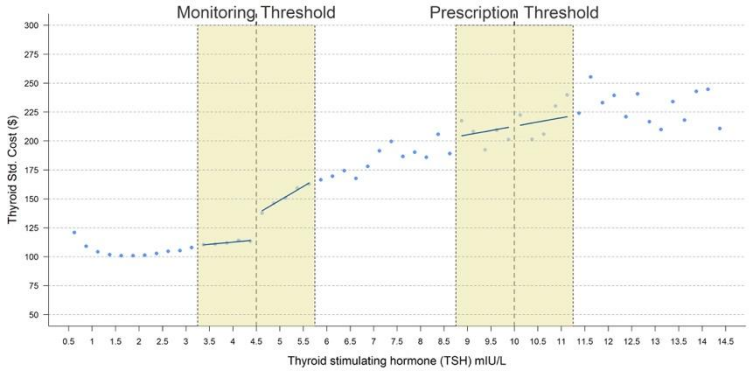
Panel A: Levothyroxine prescriptions



Panel B: Hypothyroidism diagnosis

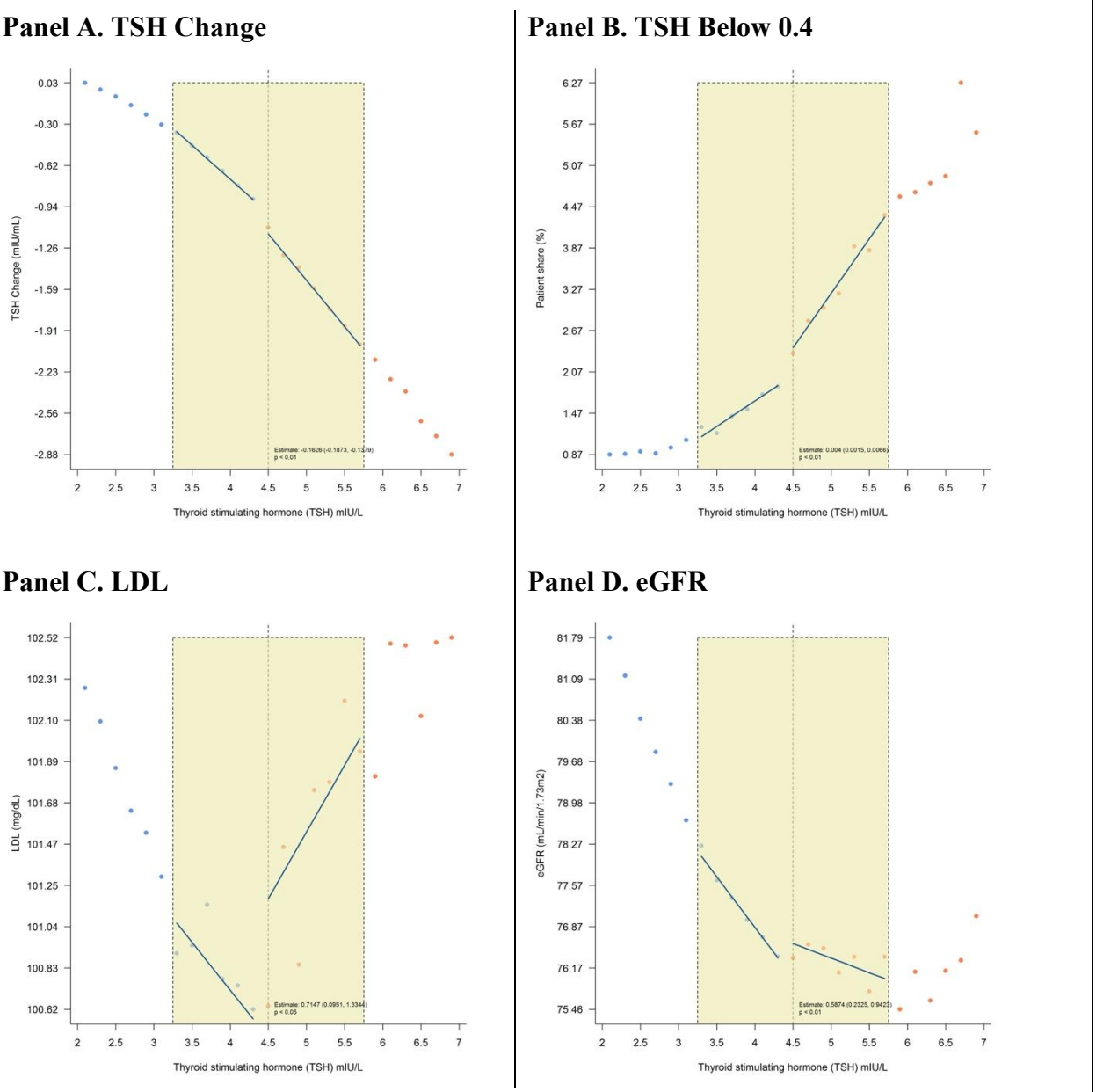


Panel C: Thyroid std. costs



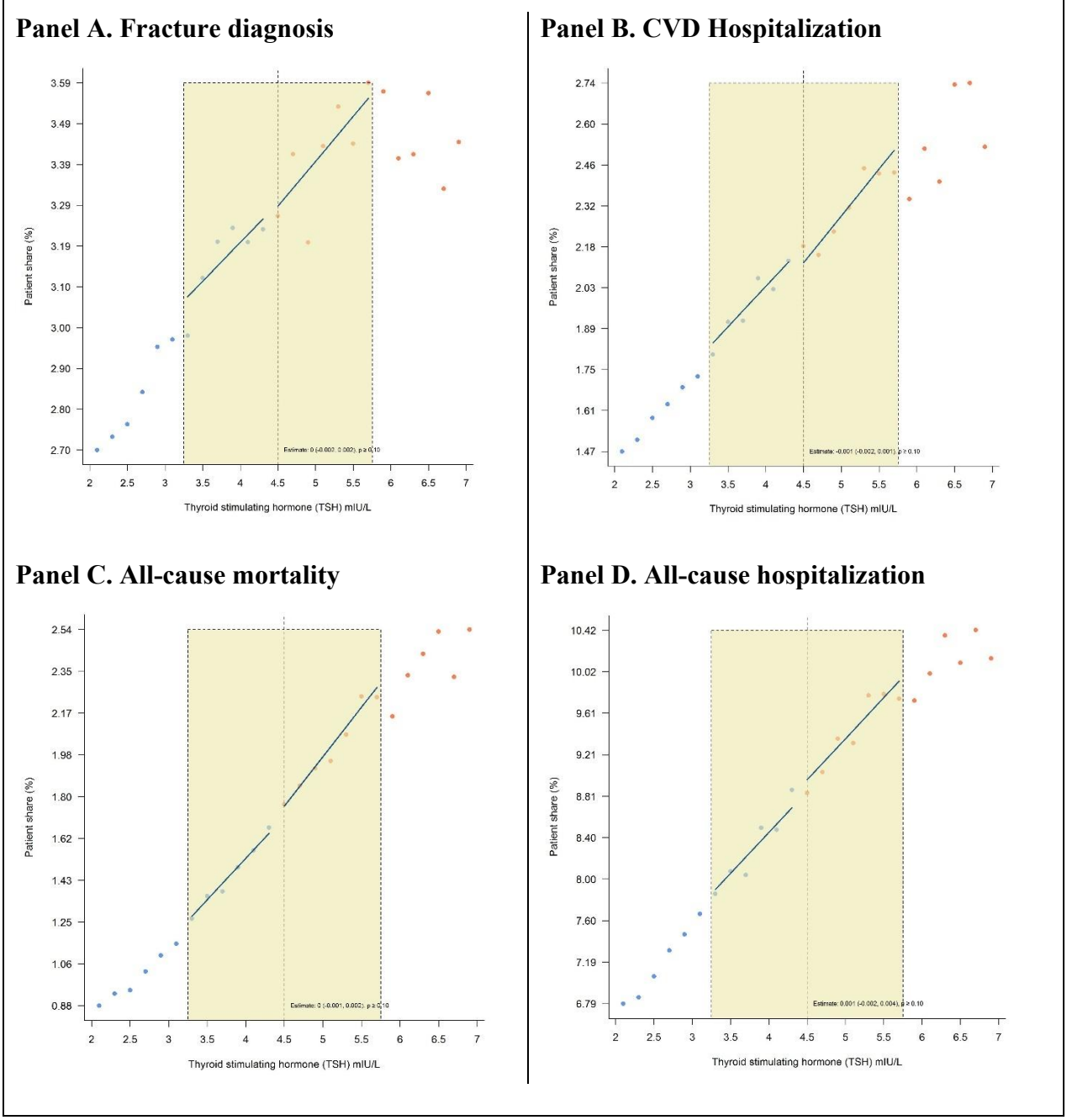
Note: The figure in Panel A displays Levothyroxine prescription 12 months after the first TSH laboratory measurement (mIU/L) across the largest sample bandwidth from 0.5 to 15 mIU/L. We allow for up to 180 days to pass within which Levothyroxine prescriptions may occur. Panel B and Panel C display hypothyroidism diagnosis and thyroid-related standardized costs, respectively. The bandwidth of 3.25 to 5.75 mIU/L and 8.75 to 11.25 is used for estimation of a polynomial of degree one.

Figure 2. Laboratory measurements over thyroid stimulating hormone



Note: The figure displays laboratory outcome measurements across the laboratory threshold for subclinical hypothyroidism using a medium bandwidth of 2 to 7 mIU/L for display purposes and 3.25 to 5.75 mIU/L for estimation of a polynomial of degree one. Outcome variables are displayed within 12 months after the TSH laboratory measurement using the last available measure for outcome variables in Panel A and B and the average for Panel C and D. The confidence intervals are based on heteroskedasticity robust standard errors.

Figure 3. Diagnosis-based health outcomes over thyroid stimulating hormone



Note: This figure shows the treatment effects on diagnosis-based health outcomes within one year after the TSH laboratory measurement when crossing the threshold of 4.5 mIU/L from left to right where subclinical hypothyroidism is indicated on the right side of the threshold and monitoring is recommended. The figures displays the share of patients within the medium bandwidth of 2.0 to 7.0 mIU/L for display purposes and 3.25 to 5.75 mIU/L for estimation using a polynomial of degree one. Outcome measures in each panel are displayed within 12 months after the TSH laboratory measurement using the average for each outcome variable. The confidence intervals are based on heteroskedasticity robust standard errors.

Table 1. Descriptive Statistics: sample selection from 2003-2021

	Bandwidth					
	0.5-15		**2.0-7.0**		**3.25-5.75**	
	Mean	SD	Mean	SD	Mean	SD
Running Variable						
TSH (mIU/L)	2.08	1.25	3.00	0.95	4.03	0.64
Laboratory Characteristics	Mean	SD	Mean	SD	Mean	SD
LDL	102.30	34.06	101.62	34.23	101.01	34.48
TG	140.39	103.93	143.68	102.98	145.80	101.21
eGFR	82.17	22.10	79.37	22.11	77.05	22.38
Comorbidities	Mean	SD	Mean	SD	Mean	SD
Diabetes Mellitus	0.14	0.35	0.15	0.36	0.16	0.37
Hypertension	0.35	0.48	0.37	0.48	0.39	0.49
Demographic Characteristics	%	N	%	N	%	N
Age (in Years)*						
18-20	1.90	132,284	1.70	50,092	1.70	12,979
20-30	11.80	827,983	10.10	289,483	8.80	68,509
30-40	15.90	1,111,798	13.60	388,628	11.60	89,763
40-50	17.30	1,210,363	15.80	453,991	14.20	109,861
50-60	17.30	1,211,660	17.10	489,740	16.50	128,212
60-70	20.00	1,404,760	22.00	630,654	23.60	182,698
70-80	11.40	800,888	13.70	392,767	15.80	122,466
>80	4.40	308,053	5.90	169,734	7.80	60,315
Male	45.50	3,191,188	45.30	1,297,632	44.00	340,633
Female	54.50	3,816,077	54.70	1,567,249	56.00	434,117
White	64.40	4,511,578	67.20	1,926,168	67.70	524,192
Hispanic	17.10	1,198,353	17.00	488,100	17.20	133,153
Black	12.70	888,129	9.70	277,823	8.70	67,354
Asian	5.80	409,729	6.00	172,998	6.50	50,104
Sample Size	6,996,054		2,865,089		772,715	

Note: The table displays descriptive statistics such as the arithmetic mean (Mean), standard deviation (SD) for patients from 2003-2021 who had a first laboratory test of the thyroid stimulating hormone (TSH). The first two columns contain statistics for the full sample, the next two columns from the medium sample and the last two columns are for our main sample around 4.5 mIU/L. The running variable is the laboratory test of the TSH measured in milliunits per liter (mIU/L). All measures are assessed at baseline, i.e., prior to TSH measurement. *Open on left, closed on right.

Table 2. 1-year treatment effects at early hypothyroidism threshold

	Baseline	Treatment effects	Sample size
Panel A: Prescription, diagnosis, and costs			
Prescription	0.07	0.1316	772,715
		(0.1278, 0.1354)	
Diagnosis	0.15	0.2054	772,715
		(0.2009, 0.2100)	
Thyroid std. costs	114.52	22.2583	772,715
		(18.5002, 26.0165)	
Panel B: Laboratory outcomes			
TSH change	-0.99	-0.1626	228,390
		(-0.1864, -0.1388)	
TSH below 0.4	0.02	0.0040	228,390
		(0.0016, 0.0065)	
LDL	100.44	0.7147	244,146
		(0.0958, 1.3337)	
eGFR	75.99	0.5874	317,727
		(0.2359, 0.9389)	
Panel C: Diagnosis-based health outcomes			
Fracture	0.03	-0.0001	772,715
		(-0.0019, 0.0018)	
CVD hospitalization	0.02	-0.0006	772,715
		(-0.0021, 0.0009)	
All-cause hospitalization	0.09	0.0011	772,715
		(-0.0018, 0.0041)	
All-cause mortality	0.02	0.0005	772,715
		(-0.0009, 0.0018)	

Note: This table shows the treatment effects on outcomes within one year after the TSH laboratory measurement when crossing the threshold of 4.5 mIU/L from left to right where subclinical hypothyroidism is indicated on the right side of the threshold and monitoring is recommended. The baseline is the estimate of being directly below the cutoff of 4.5 mIU/L. All statistics are shown within the smallest bandwidth within 3.25 mIU/L to 5.75 mIU/L. Confidence intervals are shown using heteroskedasticity robust standard errors.

Table 3. Extrapolated treatment effects from early hypothyroidism threshold

Threshold:	4.5 mIU/L	5.0 mIU/L	6.0 mIU/L	7.0 mIU/L
Panel A: Prescription, diagnosis, and costs				
Prescription	0.1316	0.1834	0.2871	0.3908
	(0.1278, 0.1354)	(0.1804, 0.1865)	(0.2796, 0.2946)	(0.3774, 0.4042)
Diagnosis	0.2054	0.2481	0.3334	0.4187
	(0.2009, 0.2100)	(0.2442, 0.2520)	(0.3245, 0.3422)	(0.4032, 0.4341)
Thyroid std. costs	22.2583	32.3348	52.4876	72.6404
	(18.5002, 26.0165)	(28.5001, 36.1695)	(43.7641, 61.2110)	(57.8823, 87.3984)
Panel B: Laboratory outcomes				
TSH change	-0.1626	-0.2594	-0.4530	-0.6465
	(-0.1873, -0.1379)	(-0.2873, -0.2315)	(-0.5088, -0.3971)	(-0.7368, -0.5563)
TSH below 0.4	0.0040	0.0082	0.0166	0.0251
	(0.0015, 0.0066)	(0.0055, 0.0110)	(0.0110, 0.0223)	(0.0157, 0.0344)
LDL	0.7147	1.3033	2.4804	3.6575
	(0.0951, 1.3344)	(0.6590, 1.9476)	(1.1658,3.7950)	(1.4875, 5.8276)
eGFR	0.5874	1.2020	2.4313	3.6605
	(0.2325, 0.9423)	(0.8343,1.5670)	(1.6783, 3.1843)	(2.4156, 4.9054)
Panel C: Diagnosis-based health outcomes				
Fracture	-0.0001	0.0001	0.0004	0.0007
	(-0.0019, 0.0018)	(-0.0018, 0.0019)	(-0.0035, 0.0042)	(-0.0058, 0.0071)
CVD hospitalization	-0.0006	-0.0004	0.0001	0.0006
	(-0.0021, 0.0009)	(-0.0019, 0.0012)	(-0.0031, 0.0033)	(-0.0047, 0.0059)
Hospitalization	0.0011	0.0012	0.0012	0.0013
	(-0.0018, 0.0041)	(-0.0018, 0.0041)	(-0.0050, 0.0074)	(-0.0090, 0.0116)
All-cause mortality	0.0005	0.0008	0.0015	0.0022
	(-0.0009, 0.0018)	(-0.0005, 0.0022)	(-0.0013, 0.0044)	(-0.0026, 0.0071)

Note: This table shows the treatment effects when we move the threshold of subclinical hypothyroidism of 4.5 mIU/L, where monitoring is recommended, closer to the threshold of overt hypothyroidism of 10 mIU/L, where prescription is recommended. Column 1 shows the impact of being just below vs above the threshold of subclinical hypothyroidism while the following columns show the impact of being just below and above the new extrapolated thresholds of 5 mIU/L, 6 mIU/L, and 7 mIU/L. The confidence intervals are in parenthesis and are based on heteroskedasticity robust standard errors.

Methods

Description of hypothyroidism guidelines

The American Thyroid Association (ATA) deems it as merely acceptable to *consider* starting levothyroxine prescription for patients with subclinical hypothyroidism. They do not have a recommendation in place to start levothyroxine for patients below TSH values of 10 mIU/L. In contrast, they state levothyroxine "is the treatment of choice for patients with" overt hypothyroidism, i.e., for patients with TSH values above 10 mIU/L (Jonklaas 2014). The Association of American Family Physicians (AAFP) has guidelines for treatment with Levothyroxine that are a bit easier to read than the ones from the ATA. In 2001 – prior to the first available diagnosis in the 2003-2021 CDM data –, subclinical hypothyroidism started at a threshold value for TSH at 6 mIU/L, with a recommendation to treat when the TSH laboratory measurement was above 10 mIU/L (Hueston 2001, Wilson et al. 2021) of what we call the prescription threshold. At the time, personalized care with Levothyroxine was recommended to physicians at the prescription threshold while 4.5 mIU/L. A jump at 4.5 mIU/L or at any laboratory threshold below 10, however, is evidence against personalized care and rather evidence that a large group of individuals are treated similarly. Two decades later, the AAFP updated the guidelines. Wilson and Curry (2005) mention a threshold of 4.5 mIU/L explicitly for subclinical hypothyroidism at the beginning of our study period. The general guidelines for subclinical hypothyroidism did not drastically change with the subclinical threshold value remaining at 4.5 mIU/L (Wilson et al. 2021). The authors recommend again a treatment with Levothyroxine only for patients with laboratory values above 10. Ross (2022) further states that Levothyroxine treatment is not necessary for patients unless the TSH value exceeds 7 mIU/L.

Data source and study population

Optum's de-identified Clinformatics® Data Mart Database (CDM or Clinformatics®) is derived from a database of administrative health claims for members of large commercial and Medicare Advantage health plans. Clinformatics® utilizes medical and pharmacy claims to derive patient-level enrollment information,

healthcare costs, and resource utilization information. The population is geographically diverse, spanning all 50 states and is statistically de-identified under the Expert Determination method consistent with HIPAA and managed according to Optum® customer data use agreements1,2. CDM administrative claims submitted for payment by providers and pharmacies are verified, adjudicated and de-identified prior to inclusion. We leverage information from 2003 to 2021 through the Stanford Center for Population Health Sciences (2021) and the University of Michigan. Our study population excludes patients with a prior hypothyroidism diagnosis, children, and pregnant women. Further details on the study population are included in the patient selection diagram in the supplementary materials.

Outcome and treatment variables

Our treatment variable is an indicator function determining crossing a threshold while the outcome variables are measures of health expenditures and health outcomes where the latter are subdivided into laboratory measures and diagnosis-based health measures. In the main text we show one-year outcome variables while the supplementary material includes additional results for 2-years, 3-years, and 5-years after the first laboratory measurement.

Statistical analysis: main analysis

To better understand the concept of overlabeling, we employ a regression discontinuity design to estimate the treatment effects of crossing the subclinical threshold of 4.5 mIU/L for our outcomes of interests. We leverage exogenous quasi-random variation from laboratory thyroid stimulating hormone (TSH) blood test thresholds to test whether they induce a jump in our outcomes of interest through the following econometric model:

$$Y_i = \alpha_0 + \alpha_1 r_i + \alpha_2 \mathrm{I}(r_i > 0) + \alpha_3 \mathrm{I}(r_i > 0) r + \varepsilon_i \quad (1)$$

where $r_i = TSH - t_i$, i.e., the normalized running variable, and $\mathrm{I}(r > 0)$ is an indicator on the normalized running variable which is one if the normalized running variable is greater than zero. The outcome variable

Y_i is a binary measuring whether an individual received a prescription of levothyroxine within twelve months after the first TSH laboratory test measurement or a diagnosis of hypothyroidism as well as laboratory outcomes, diagnosis-based health outcomes, and cost outcomes. The threshold t_i is the upper normal limit of TSH in the laboratory where the sample for individual *i* is measured. In our main estimation, $t_i = 4.5$ which overlaps with the monitoring threshold of 4.5 mIU/L which defines subclinical hypothyroidism in the United States. In the above model, we are interested in the parameter α_2 which informs us for Levothyroxine prescription whether there is a statistically significant and large jump in prescriptions due to the laboratory threshold while allowing us to control for different linear slopes on the left- vs the right-hand side of the threshold. Similarly, we are interested in the intent-to-treat effect on the laboratory outcomes, diagnosis-based health outcomes, and cost outcomes of being just above the threshold vs just below the threshold.

Statistical analysis: sensitivity analysis

In our main text, we use the bandwidth of 3.25-5.75 mIU/L for local polynomials and 2-7 mIU/L for global polynomials. We vary the bandwidth, as shown in the supplementary material (Figure S.12 to S.26) and show the robustness of the results per outcome based on the local and quadratic polynomial for the two bandwidths. We further use the optimal bandwidth based on Imbens and Kalyanaraman (2012) as a robustness check in the supplementary material. We find that all our results are robust to different bandwidths and functional forms.

Statistical analysis: heterogeneous treatment effects

We ran three heterogeneity analyses to contextualize our evidence on overlabeling. Since we do not have direct information on the symptoms of patients, we check sub-group for which we hypothesize symptomatic prescriptions are more or less likely. As a measure of overlabeling, we use levothyroxine prescriptions and analyze heterogeneity by age, general health-checkup visits, and free T4. We know that hypothyroidism is

a disease that affects the elderly more strongly. Hence, we expect stronger prescription behavior for the elderly since they are the ones that are more likely to have symptoms. As such, Figure S23 splits prescriptions across TSH laboratory tests by the young, i.e. below 65 (Panel A) and the elderly, i.e. above 65 (Panel B). We do not find a medically meaningfully larger jump for the elderly relative to the young. If at all, the jump in prescribing behavior is larger for the young (10 pp.) relative to the elderly (13 pp.).

An alternative way to proxy for symptoms is by inspecting TSH tests that are occurring during a regular health check-up. Tests that occur during regular check-ups are less likely to be symptom driven, so if prescriptions are less likely to be symptom-based, then we would not expect large differences for the regular and non-regular health check-up groups. Figure S24 shows that the jump in prescribing behavior is very similar during regular health check-ups (Panel A) relative to outside regular health check-ups (Panel B), with a 12.30 pp. vs 12.54 pp. jump, respectively.

We run a last check on FT4, as shown in Figure S25. Since the actual diagnosis is theoretically dependent on whether free T_4 is low, i.e., below a level of 0.8 ng/dL (below the lower normal value), we separately condition on the subset of patients for which FT4 measures are available (32% or around 2.8 million patients within the medium bandwidth). Panel A shows the majority of individuals that never receive FT4 measures but who are nevertheless prescribed levothyroxine and are diagnosed as hypothyroid. It is very unlikely that those patients have all symptoms, which implies that they are not treated correctly. Despite the missing measures, we see a strong jump in prescribing behavior. Similarly, Panel B shows a jump for patients with regular or high FT4 values that should not be diagnosed as hypothyroid. Given the jump that is independent of FT4, FT4 does not seem to be used as a guiding second measure to prescribe levothyroxine, even though it is recommended to make a diagnosis. Finally, there are few patients which are actually tested for FT4 and exhibit low FT4. Panel C shows that patients with low FT4 exhibit the same jump as in our main Figure 1 but at a higher level.

To conclude, the subgroup evidence implies that the groups that are more likely to have symptoms are not more likely to drive our jump in prescribing behavior which adds to the idea that prescription

behavior at the threshold is not symptom-related and indeed, that overprescription as a mechanism of overlabeling is occurring at the 4.5 mIU/L subclinical hypothyroidism threshold.

Statistical analysis: extrapolation away from the threshold

To extrapolate treatment effects away from the threshold, we rely on the method from Dong and Lewbel (2015) who introduce the concept of the treatment effect derivative (TED). They define the extrapolated treatment effect as follows:

$$\tau(t_{new}) \approx \tau(t_{old}) + (t_{new} - t_{old}) \times TED \quad (4)$$

where t_{new} is the new threshold for extrapolation, t_{old} is the old real threshold and τ is the treatment effect evaluated at a particular threshold. The TED in the reduced form is the change in the slope when crossing the threshold, i.e. it is the "coefficient of the interaction term between the treatment T and $X - [t]$ in a (local) linear regression of Y on a constant, T, $X - [t]$, and $(X - [t])\,T$" (Dong and Lewbel 2015). The new extrapolated treatment effect $\tau(t_{new})$ is approximately identical to the treatment effect τ at the regular threshold t_{old} while moving along the running variable by $t_{new} - t_{old}$ steps adjusting for the TED. Thew new extrapolated treatment standard errors are constructed using the Delta method.

Ethics

This study has been approved by the Stanford University Center for Population Health Sciences IRB under the study protocol IRB-40974 and by the University of Michigan IRB under the protocol HUM00247279. It was determined to be not subject to approval.

Methods References

Angrist, J. D., & Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512), 1331-1344.

Bertanha, M., & Imbens, G. W. (2020). External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, 38(3), 593-612.

Chiovato, L., Magri, F., & Carlé, A. (2019). Hypothyroidism in context: where we've been and where we're going. *Advances in Therapy*, 36(Suppl 2), 47-58.

Dong, Y., & Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, 97(5), 1081-1092.

Hueston, W. J. (2001). Treatment of hypothyroidism. American family physician, 64(10), 1717.

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3), 933-959.

Jonklaas, J., Bianco, A. C., Bauer, A. J., Burman, K. D., Cappola, A. R., Celi, F. S., Cooper, D.S., Kim, B.W., Peeters, R.P., Rosenthal, M.S. & Sawka, A.M. (2014). Guidelines for the treatment of hypothyroidism: prepared by the American Thyroid Association task force on thyroid hormone replacement. *Thyroid*, 24(12), 1670-1751.

Razvi, S., Jabbar, A., Pingitore, A., Danzi, S., Biondi, B., Klein, I., Peeters, R., Zaman, A. & Iervasi, G. (2018). Thyroid hormones and cardiovascular function and diseases. *Journal of the American College of Cardiology*, 71(16), 1781-1796.

Ross, D.S. (2022). Treating hypothyroidism is not always easy: When to treat subclinical hypothyroidism, TSH goals in the elderly, and alternatives to levothyroxine monotherapy. *Journal of Internal Medicine*, 291(2), 128-140.

Wilson, G. R., & Curry Jr, W. R. (2005). Subclinical thyroid disease. *American Family Physician*, 72(8), 1517-1524.

Wilson, S. A., Stem, L. A., & Bruehlman, R. D. (2021). Hypothyroidism: diagnosis and treatment. *American Family Physician*, 103(10), 605-613.

Significance Statement: Overlabeling can be a substantial driver of unnecessary health expenditures without benefit in health outcomes, but it is often difficult to quantify these costs. In the case of hypothyroidism diagnosis and levothyroxine prescription at a subclinical hypothyroidism threshold, a regression discontinuity design identified an optimized cost-to-benefit threshold that would lead to substantial cost reductions of millions of dollars per year for patients in the United States of America.

Acknowledgement: This study would not be possible without using Optum's de-identified Clinformatics® Data Mart Database (2003-2021) also commonly known as CDM. We thank the Center for Population Health Sciences at Stanford as well as the University of Michigan and the CDM team for data access. We further thank Dan Ly, Raf Van Gestel, Manasvini Singh, Mayra Pineda Torres, Maike Hohberg, and seminar participants at the 11th IRDES-LIRAES Workshop on Applied Health Economics and Policy Evaluation, the 14th Annual Conference of the American Society of Health Economists, and the Department of Medical Statistics at the University of Goettingen for their valuable feedback.

Funding: This study received no external funding. ZC's effort was supported in part by an NIA training grant to the Population Studies Center at the University of Michigan (T32AG000221). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contribution: MH, ZC, and HY designed the study and interpreted the data. MH collected initial data, and ZC and HY collected the main data and generated outputs for plots and tables. ZC, HY, and MH analyzed the data. MH and ZC wrote the manuscript, with input from HY, EM, and PG.

Competing interests: The authors declare no competing interests.

Additional information: Supplementary Material is available for this paper. Correspondence and requests for material should be addressed to Manuel Hoffmann, Email: manuel.hoffmann@uci.edu.

Supplementary Information for

Overlabeling:
Causal Evidence from a Top-Medication

Manuel Hoffmann, Zoey Chopra, Hsu-Hang Yeh, Elizabeth A. McAninch, Pascal Geldsetzer

Correspondence to: mhoffmann@hbs.edu

This PDF file includes:
Figs. S1 to S25
Tables S1 to S27

Fig S1. Examples of laboratory reports

Panel A: TSH Below Reference Range

Test Name	Result	Flag	Reference Range	Lab
TSH	0.03	LOW	0.40-4.50 mIU/L	RGA

Panel B: TSH Within Reference Range

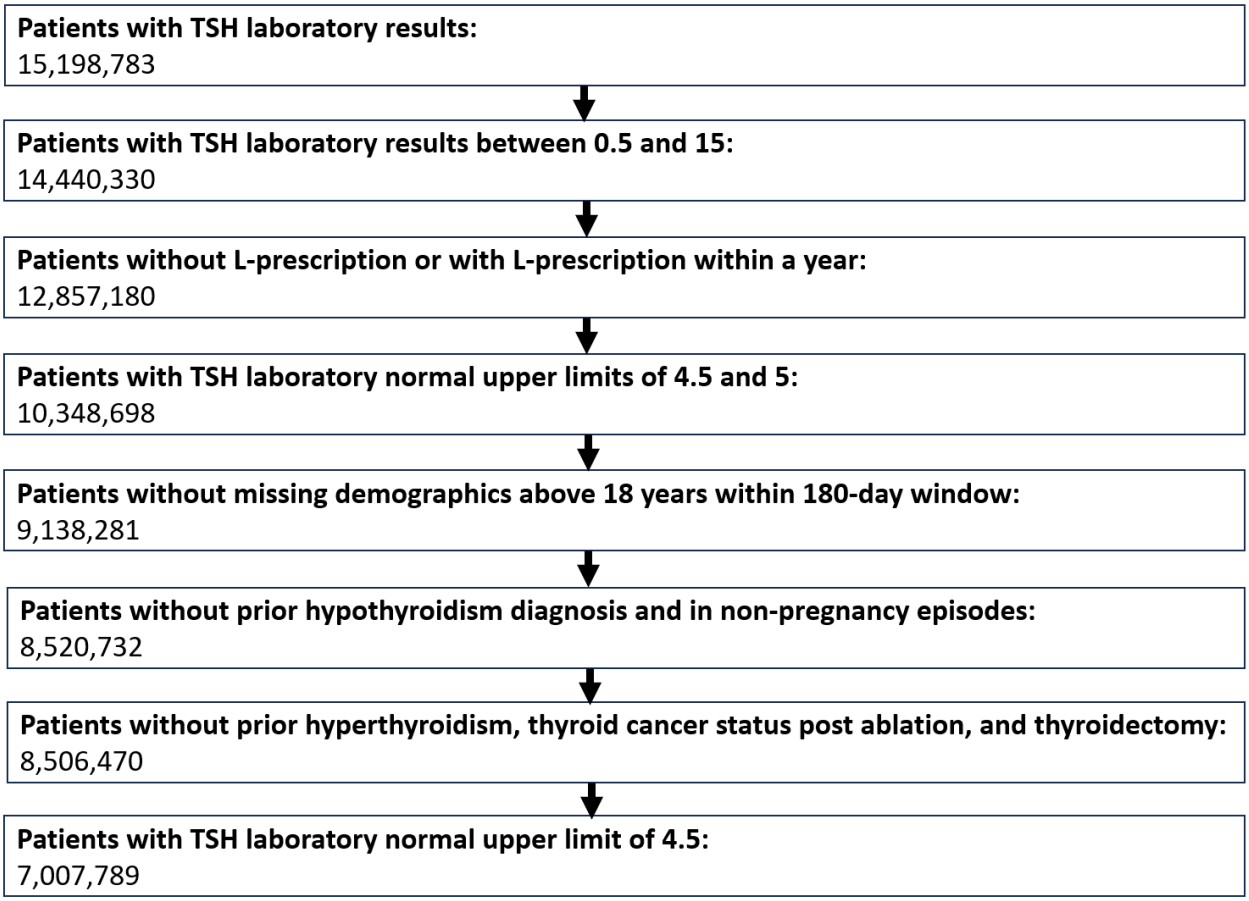
Test Name	Result	Flag	Reference Range	Lab
TSH	3.57	NORMAL	0.40-4.50 mIU/L	RGA

Panel C: TSH Above Reference Range

Test Name	Result	Flag	Reference Range	Lab
TSH	6.45	HIGH	0.40-4.50 mIU/L	RGA

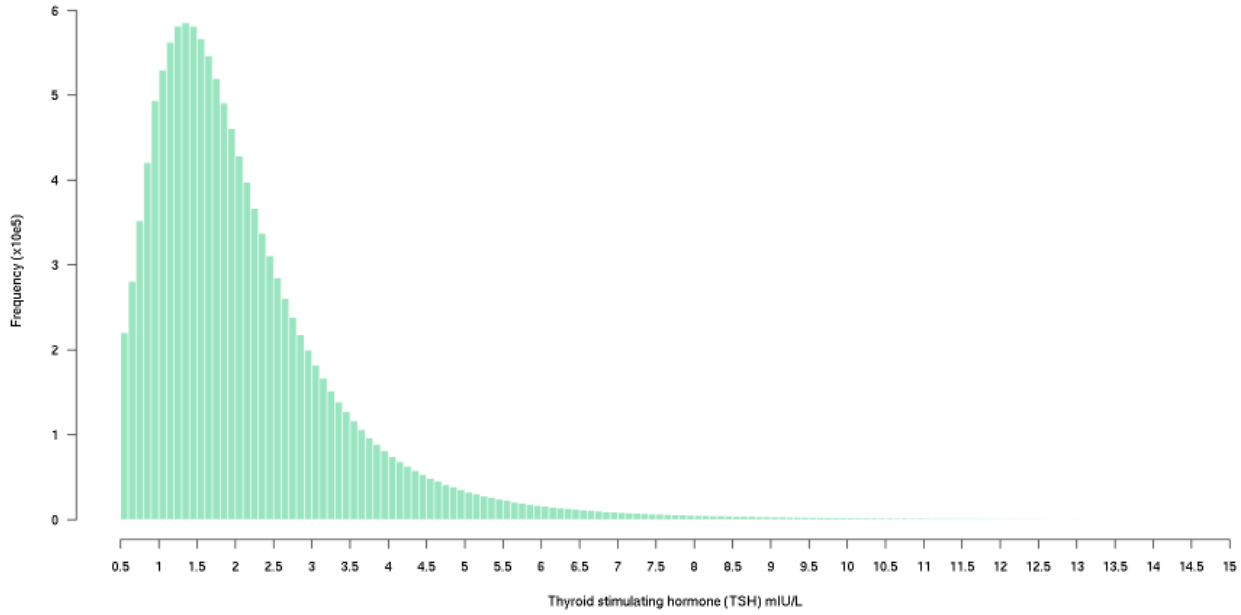
Note: The figure provides illustrative examples of three laboratory reports. Panel A shows a TSH value below the reference range and a blue flag for hyperthyroidism. Panel B shows a TSH value within the reference range and it has no colored flag. Panel C shows a TSH value above the reference range and a red flag for subclinical hyperthyroidism. We exclude the header that usually contains patient information and fully focus on clinical information. RGA stands for Reference Group Analysis which implies that the results are contextualized against a reference population.

Fig S2. Patient selection diagram



Note: The patients with Thyroid stimulating hormone (TSH) results are selected based on the LOINC codes 11579-0, 3016-3, 11580-8. We excluded 180,813 patients under the age of 18 years. The 180 day window excludes 43,274 patients. There are 492,845 patients with prior hypothyroidism diagnosis and 124,704 patients with pregnancy episodes that were excluded. We further excluded 11,414 patients with a prior hyperthyroidism diagnosis, 70 patients with prior thyroid cancer status post ablation, and 2,778 patients with a prior thyroidectomy.

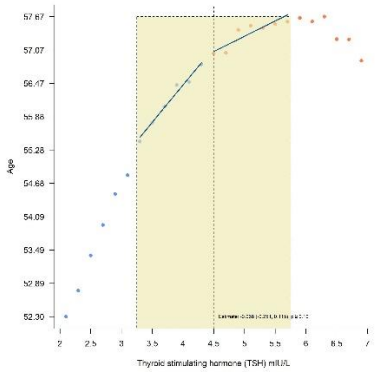
Fig S3. Histogram of baseline thyroid stimulating hormone (TSH)



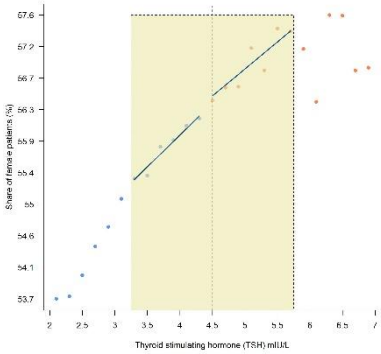
Note: The figure displays the absolute frequency of TSH laboratory measurement across the TSH values.

Fig S4. Demographic characteristics across TSH – polynomial of degree 1

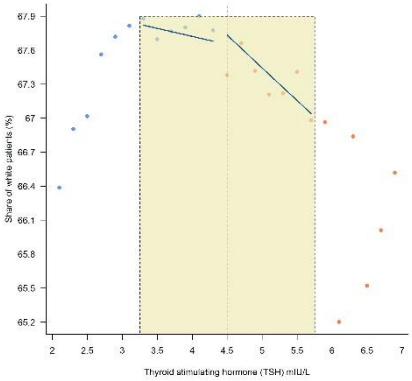
A. Age



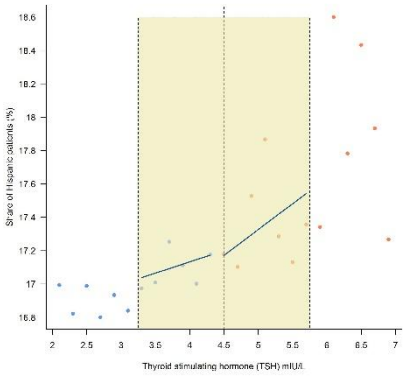
B. Female



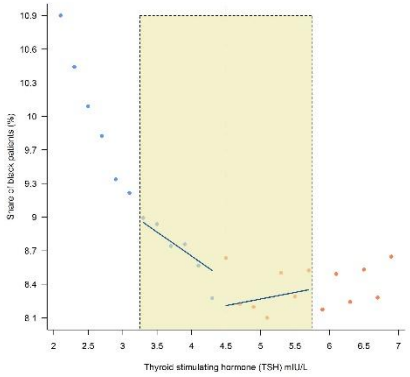
C. White



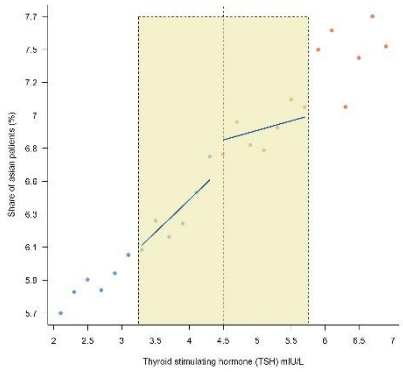
D. Hispanic



E. Black



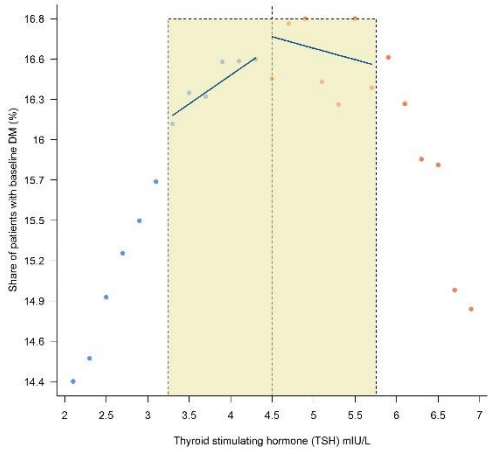
F. Asian



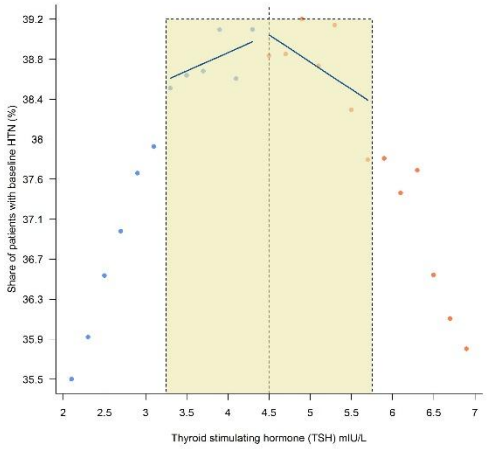
Note: The figure displays the averages of demographic characteristics across thyroid stimulating hormone (TSH) laboratory measurement. The blue lines depict the local linear estimation on the left- and right hand side of the cutoff of 4.5 mIU/L. Panel A contains the mean age whereas panel B through F contains the demographics of being female, White, Hispanic, Black and Asian as a fraction over all patients. All measures are assessed at baseline prior to TSH measurement.

Fig S5. Comorbidities and laboratory characteristics across TSH – polynomial of degree 1

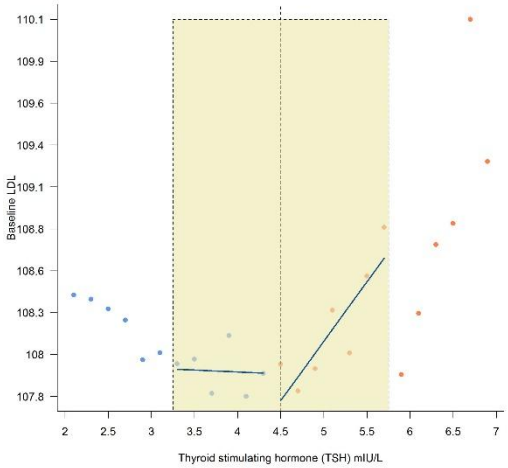
A. Diabetes Mellitus



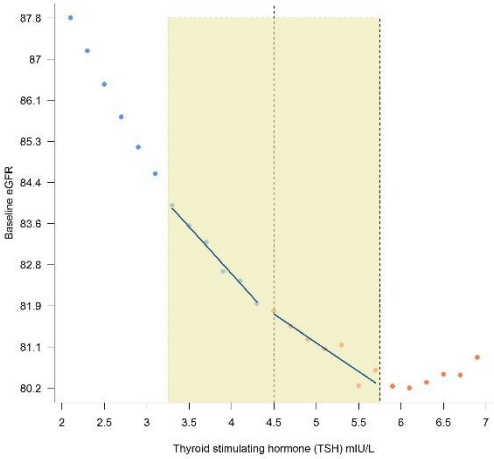
B. Hypertension



C. LDL



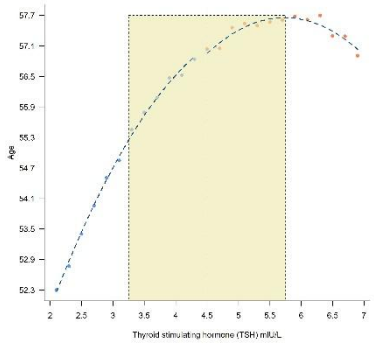
D. eGFR



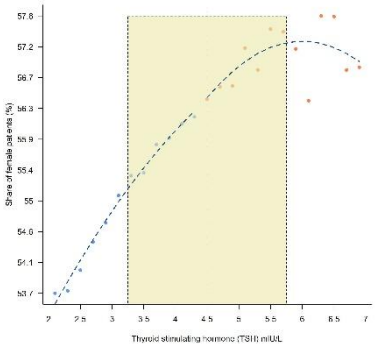
Note: The figure displays the averages of comorbidities and laboratory characteristics across thyroid stimulating hormone laboratory (TSH) measurement. The blue lines depict the local linear estimation on the left- and right hand side of the cutoff of 4.5 mIU/L. Panel A and B contain the fraction of patients with diabetes mellitus and hypertension whereas panel C and D contain levels of low-density lipoprotein (LDL) and the estimated Glomerular Filtration Rate (eGFR). All measures are assessed at baseline prior to TSH measurement.

Fig S6. Demographic characteristics across TSH – polynomial of degree 2

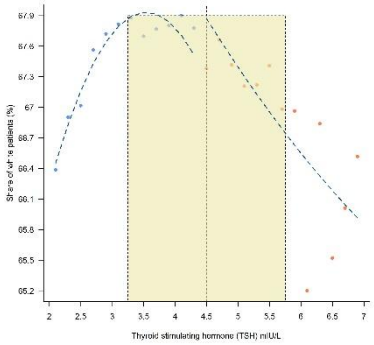
A. Age



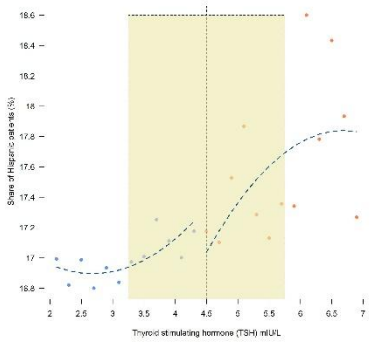
B. Female



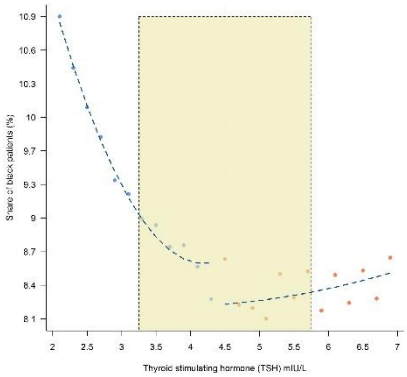
C. White



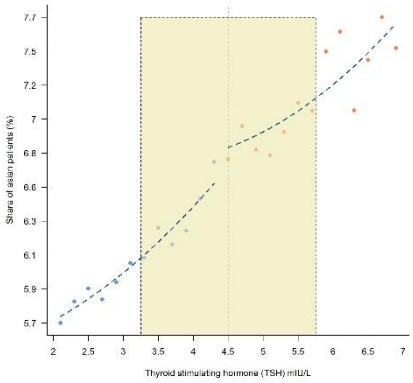
D. Hispanic



E. Black



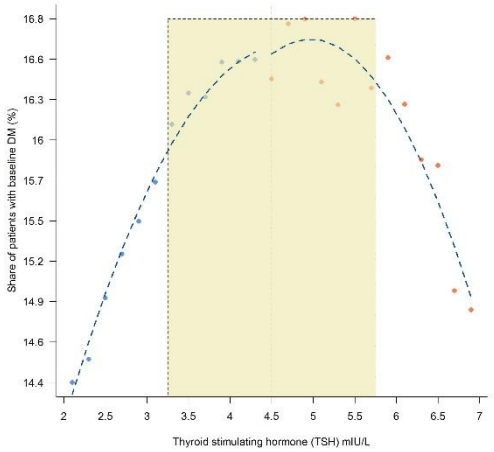
F. Asian



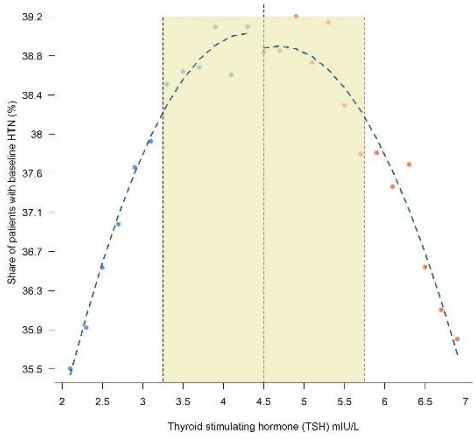
Note: The figure displays the averages of demographic characteristics across thyroid stimulating hormone (TSH) laboratory measurement. The blue lines depict the global polynomial estimation of degree two on the left- and right hand side of the cutoff of 4.5 mIU/L. Panel A contains the mean age whereas panel B through F contains the demographics of being female, White, Hispanic, Black and Asian as a fraction over all patients. All measures are assessed at baseline prior to TSH measurement.

Fig S7. Comorbidities and laboratory characteristics across TSH – polynomial of degree 2

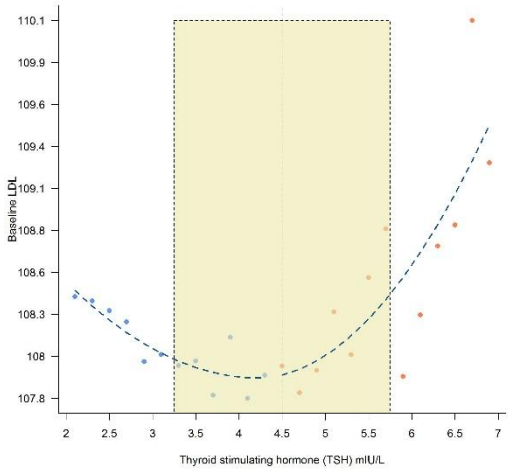
A. Diabetes Mellitus



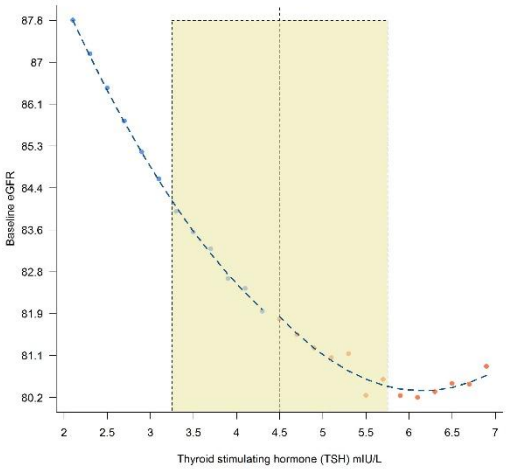
B. Hypertension



C. LDL

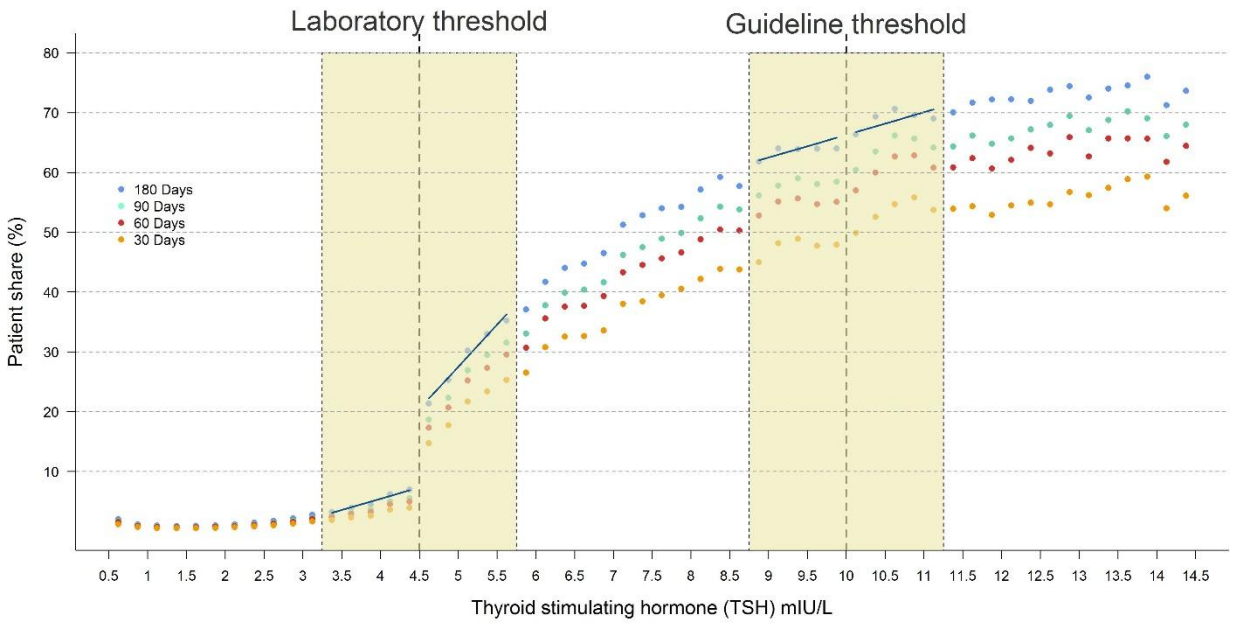


D. eGFR



Note: The figure displays the averages of comorbidities and laboratory characteristics across thyroid stimulating hormone laboratory (TSH) measurement. The blue lines depict the global polynomial estimation of degree two on the left- and right hand side of the cutoff of 4.5 mIU/L. Panel A and B contain the fraction of patients with diabetes mellitus and hypertension whereas panel C and D contain levels of low-density lipoprotein (LDL) and the estimated Glomerular Filtration Rate (eGFR). All measures are assessed at baseline prior to TSH measurement.

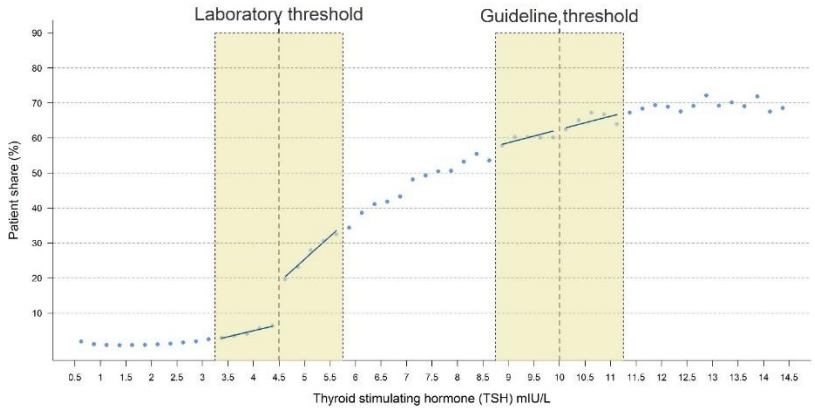
Fig S8. L-prescriptions within 30 to 180 Days across thyroid stimulating hormone



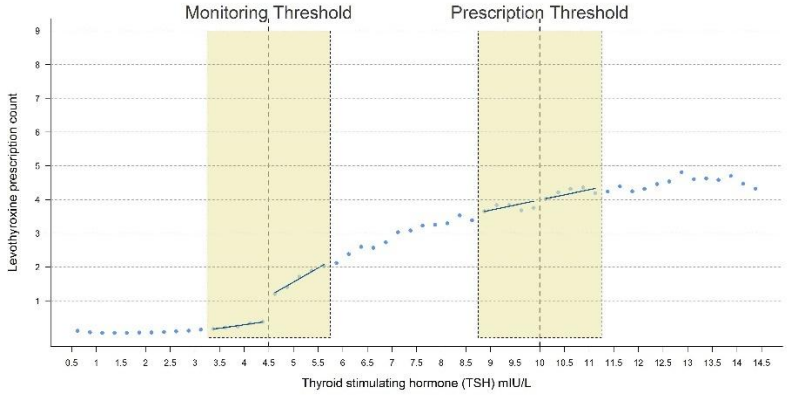
Note: The figure displays the share of patients receiving Levothyroxine (L)-prescriptions for 30, 60, 90, and 180 days after the laboratory measurement across the thyroid stimulating hormone (TSH) laboratory measurements. For our estimation we allow for a window of 180 days.

Fig S9. Prescriptions and hypothyroidism diagnosis across thyroid stimulating hormone

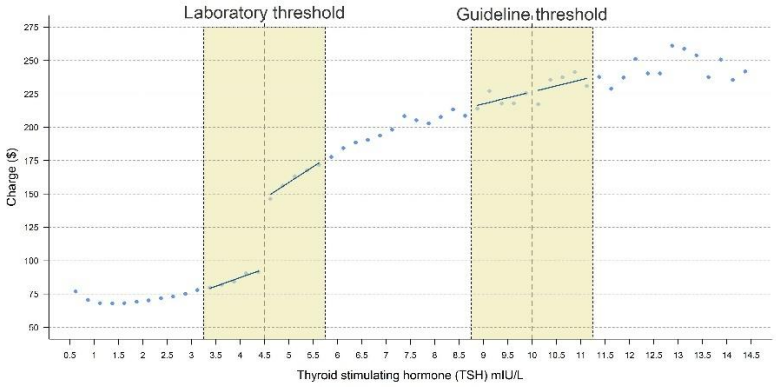
A. Prescription and hypothyroidism diagnosis jointly



B. Continuous prescription



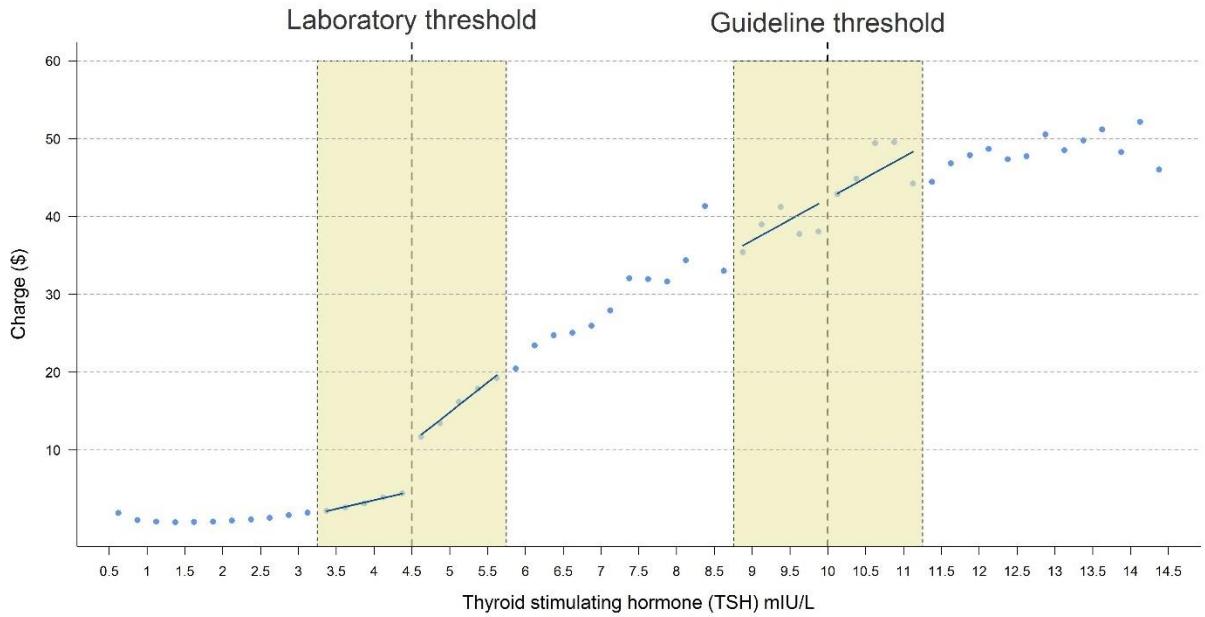
C. TSH test costs



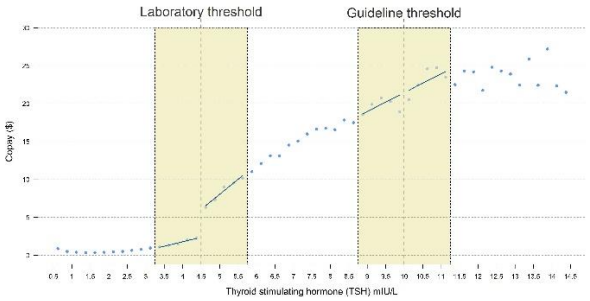
Note: The figure displays the primary exposure to a Levothyroxine prescription and hypothyroidism diagnosis as well as TSH test costs 12 months after the first thyroid stimulating hormone (TSH) laboratory measurement (mIU/L). Panel A shows average prescription and hypothyroidism diagnosis jointly, Panel B shows the number of prescriptions, and Panel C shows TSH test costs.

Fig S10. Prescription costs across thyroid stimulating hormone

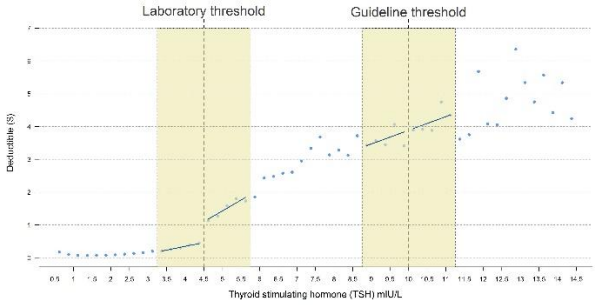
A. Charge



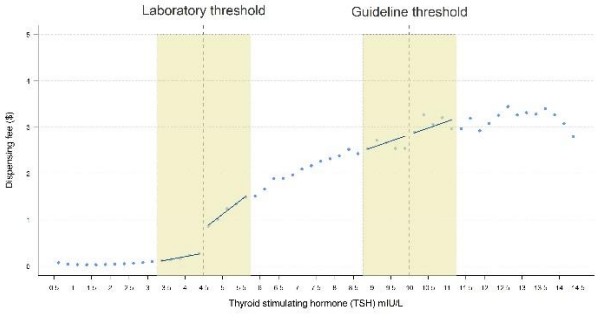
B. Copay Amount



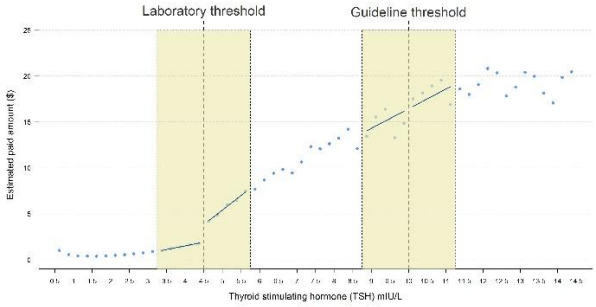
C. Deductible Amount



D. Dispensing Fee

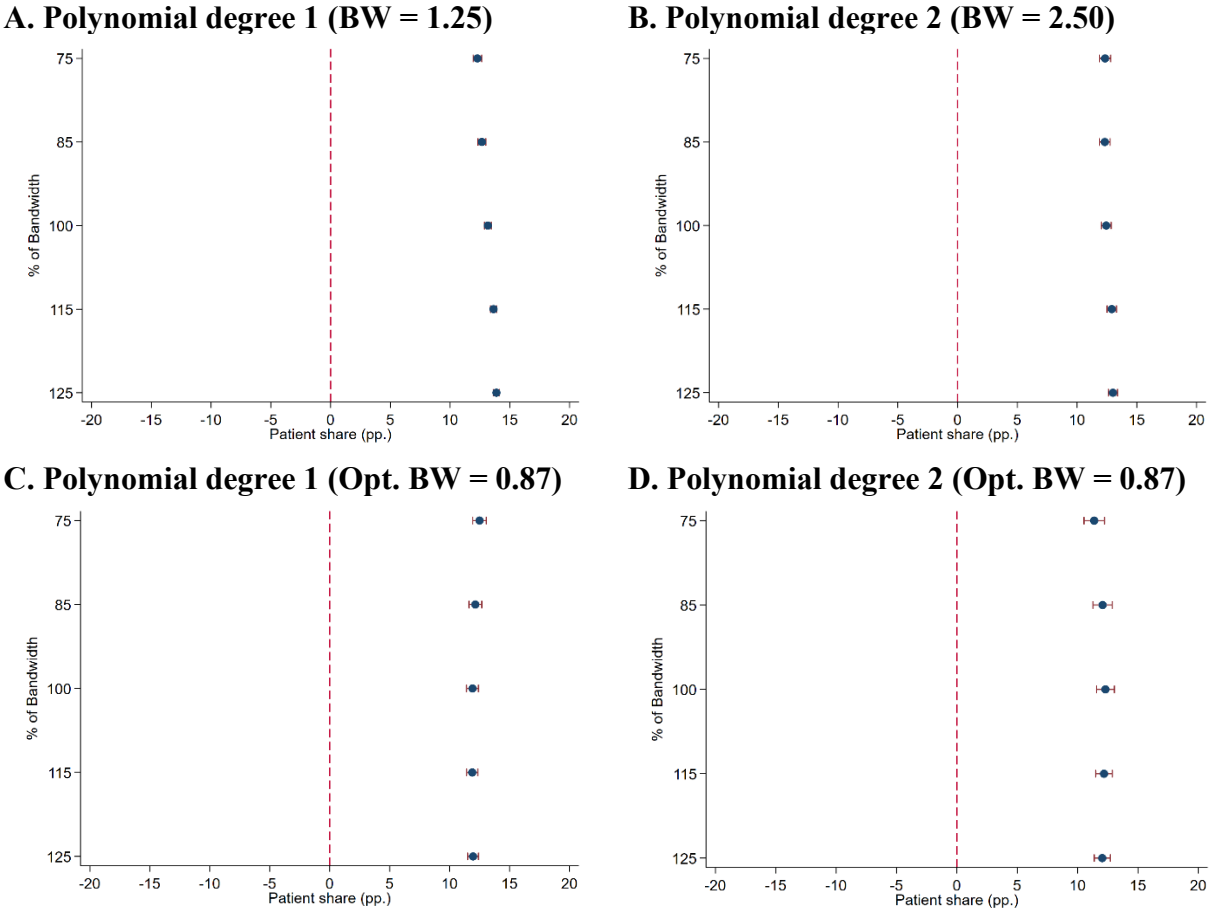


E. Estimated Paid Amount



Note: The figure displays Levothyroxine prescription costs 12 months after the first thyroid stimulating hormone (TSH) laboratory measurement (mIU/L). All costs are displayed in US-Dollars. Panel A shows the average charge. Panel B through E provides the average copay amount, deductible, dispensing fee and estimated paid amount. The estimated paid amount is calculated by subtracting the copay amount, and deductible amount from the charged amount.

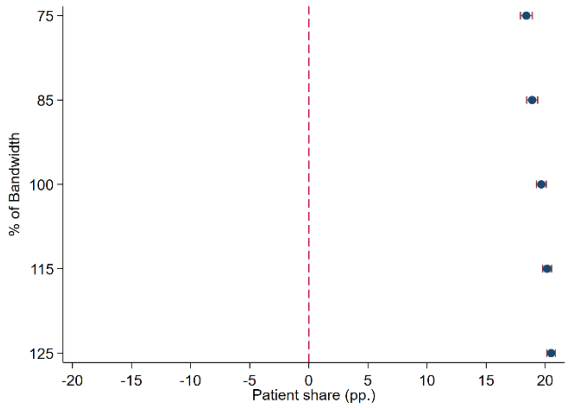
Fig S11. Varying bandwidths of threshold impact on Levothyroxine prescription



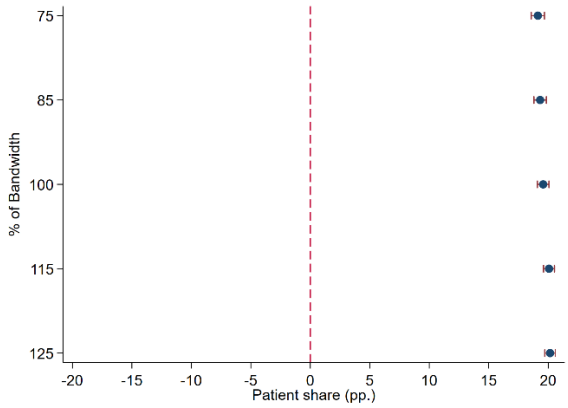
Note: The figure displays the average treatment effects on TSH test costs at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S12. Varying bandwidths of threshold impact on hypothyroidism diagnosis

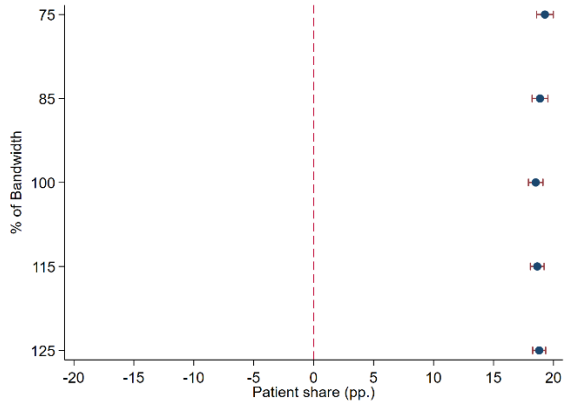
A. Polynomial degree 1 (BW = 1.25)



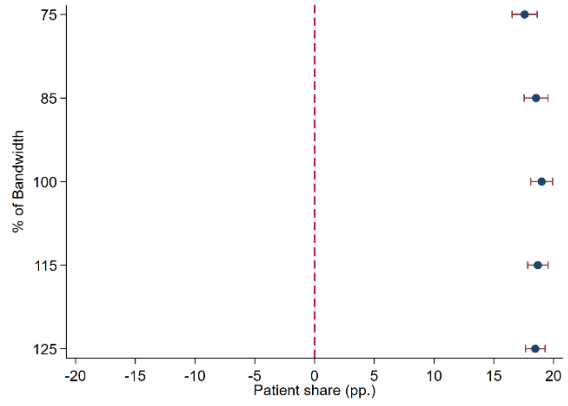
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 0.87)



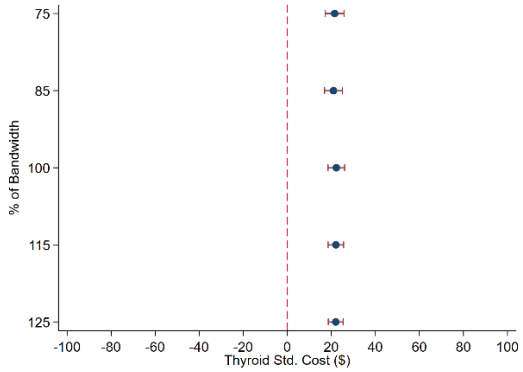
D. Polynomial degree 2 (Opt. BW = 0.87)



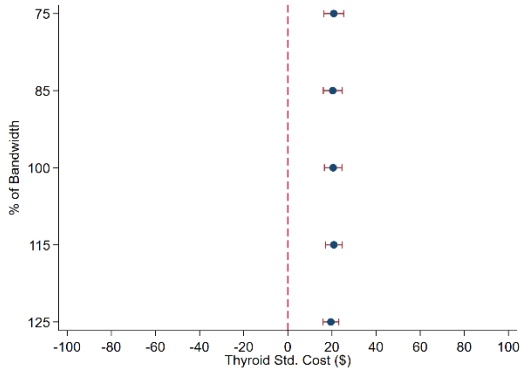
Note: The figure displays the average treatment effects on TSH test costs at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S13. Varying bandwidths of threshold impact on thyroid std. costs

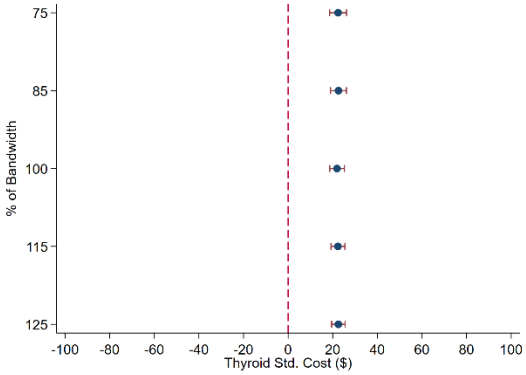
A. Polynomial degree 1 (BW = 1.25)



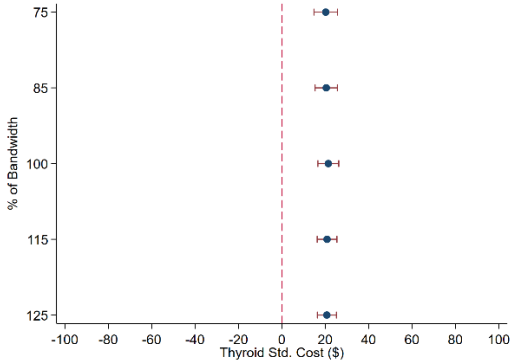
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 0.87)



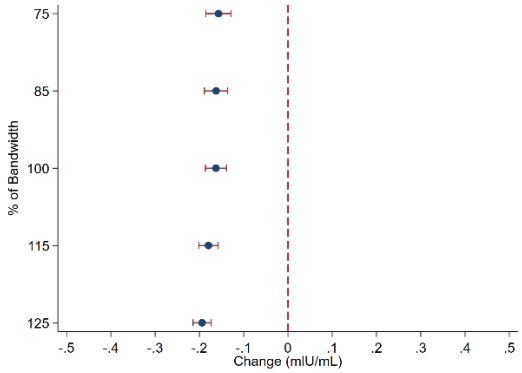
D. Polynomial degree 2 (Opt. BW = 0.87)



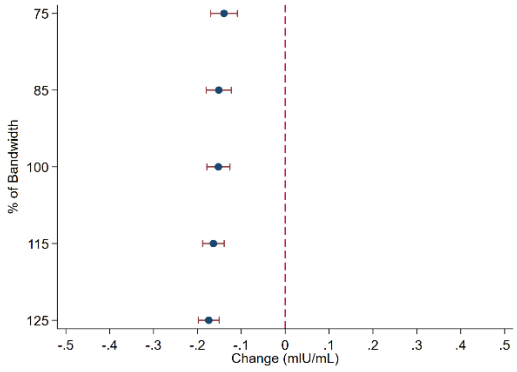
Note: The figure displays the average treatment effects on TSH test costs at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows the effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S14. Varying bandwidths of threshold impact on TSH change

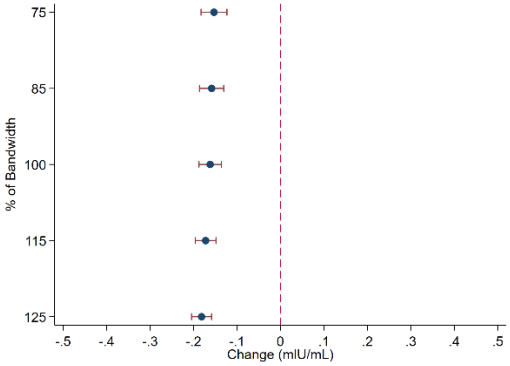
A. Polynomial degree 1 (BW = 1.25)



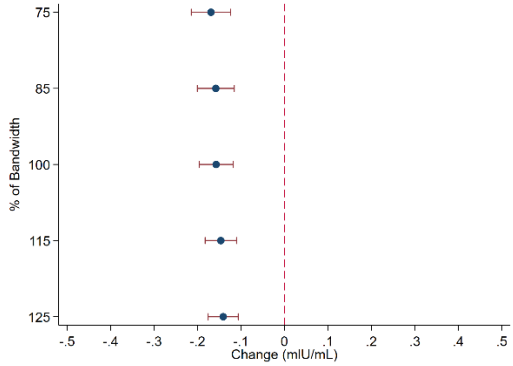
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 0.96)



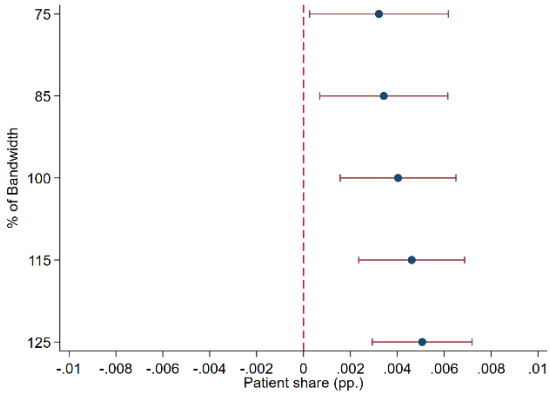
D. Polynomial degree 2 (Opt. BW = 0.96)



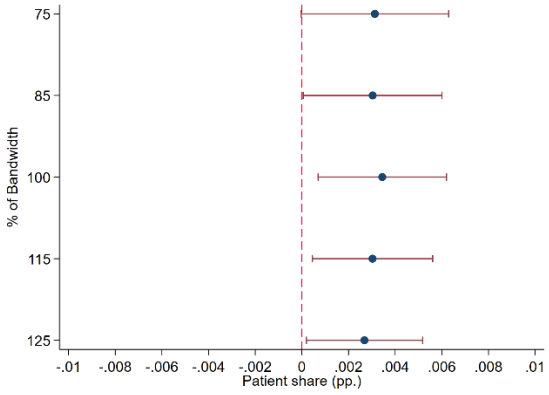
Note: The figure displays the average treatment effects on TSH change at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S15. Varying bandwidths of threshold impact on TSH below 0.4

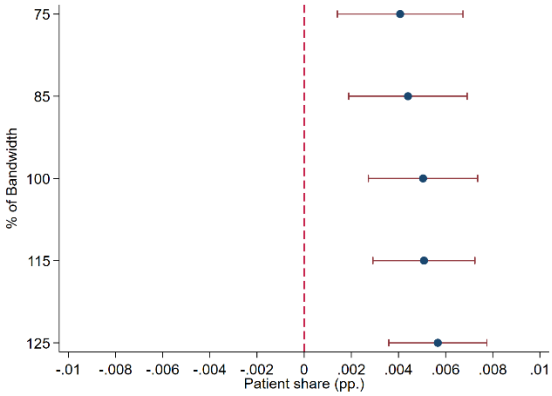
A. Polynomial degree 1 (BW = 1.25)



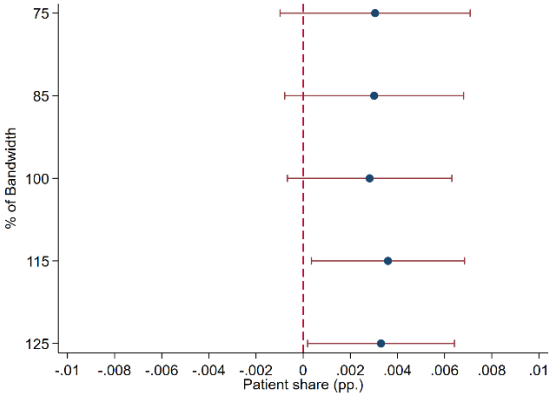
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 1.94)



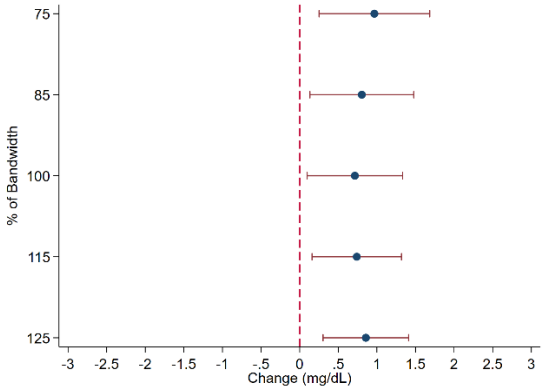
D. Polynomial degree 2 (Opt. BW = 1.83)



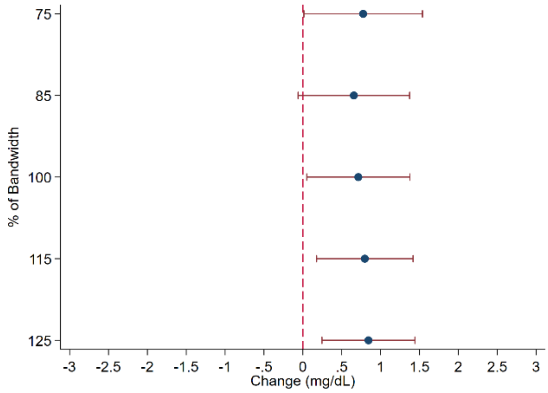
Note: The figure displays the average treatment effects on TSH below 0.4 at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S16. Varying bandwidths of threshold impact on LDL

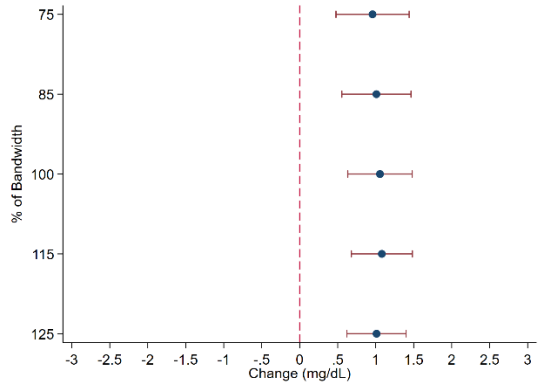
A. Polynomial degree 1 (BW = 1.25)



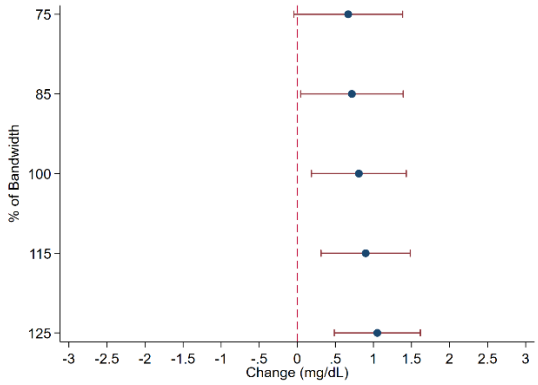
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 2.39)



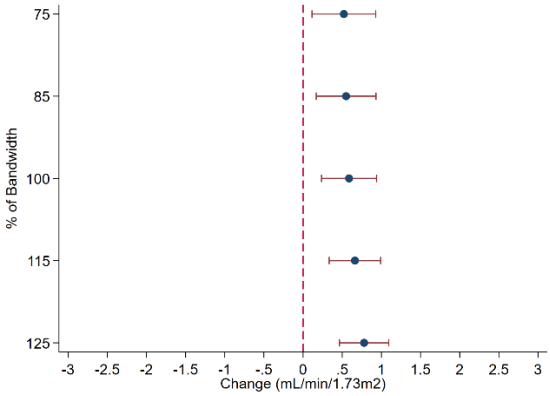
D. Polynomial degree 2 (Opt. BW = 2.39)



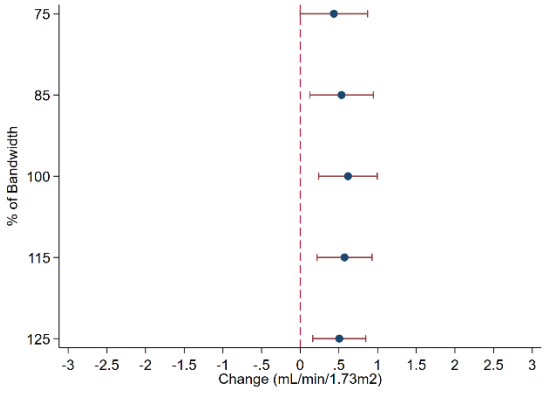
Note: The figure displays the average treatment effects on LDL at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S17. Varying bandwidths of threshold impact on eGFR

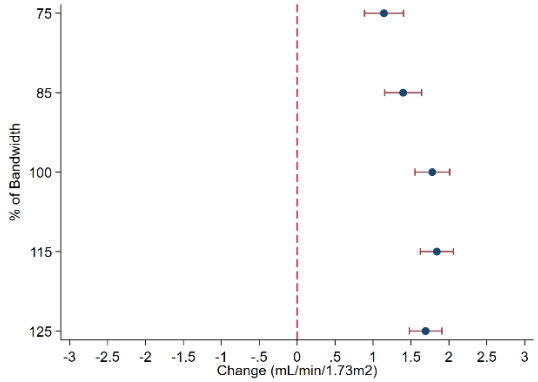
A. Polynomial degree 1 (BW = 1.25)



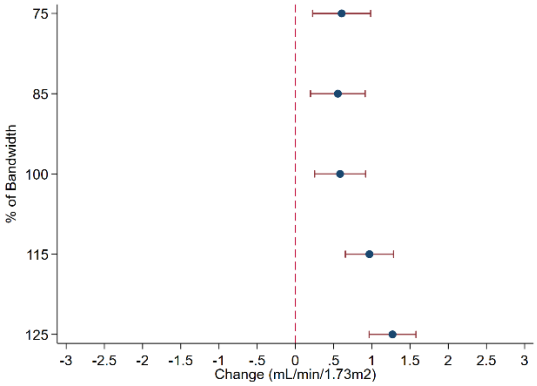
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 2.66)



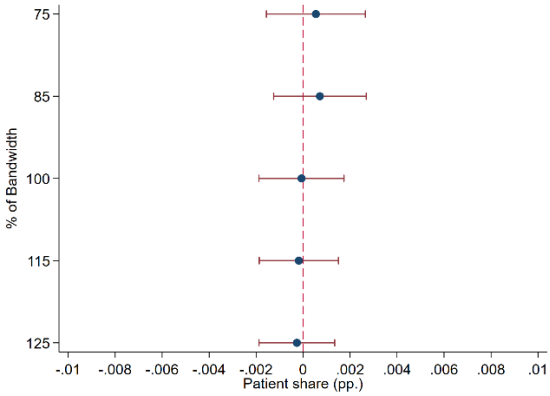
D. Polynomial degree 2 (Opt. BW = 2.66)



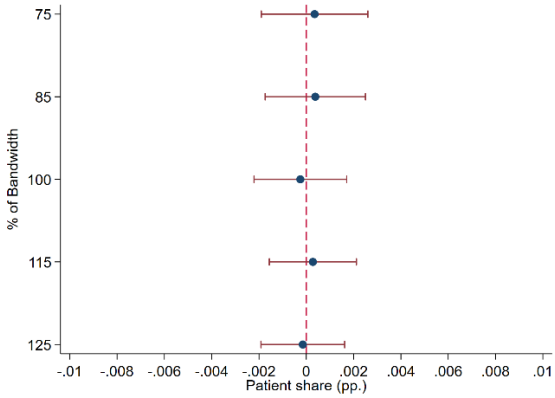
Note: The figure displays the average treatment effects on eGFR at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S18. Varying bandwidths of threshold impact on Fracture Diagnosis

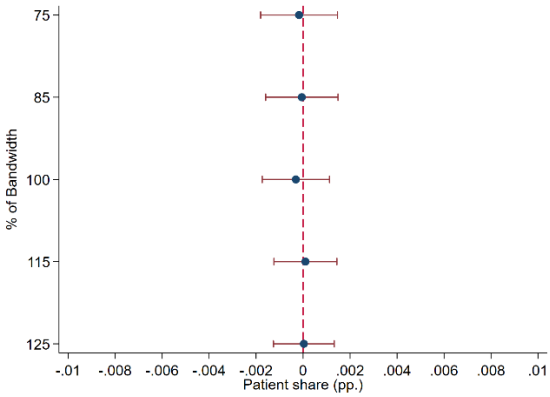
A. Polynomial degree 1 (BW = 1.25)



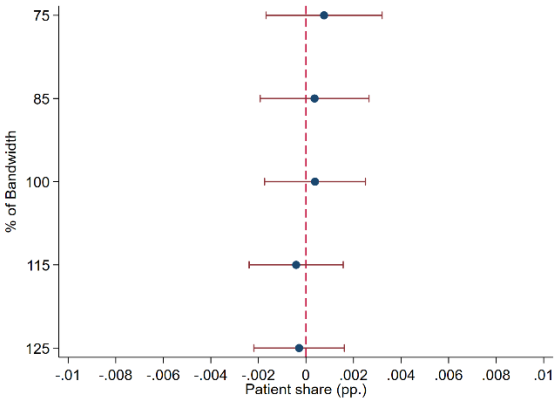
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 2.01)



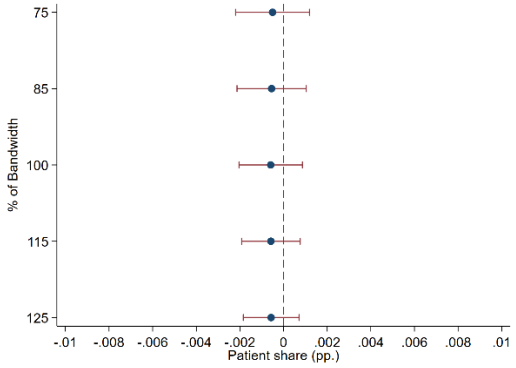
D. Polynomial degree 2 (Opt. BW = 2.01)



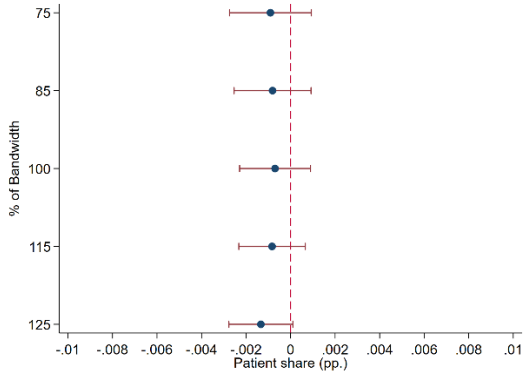
Note: The figure displays the average treatment effects on Fracture Diagnosis at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S19. Varying bandwidths of threshold impact on Cardiovascular Hospitalization

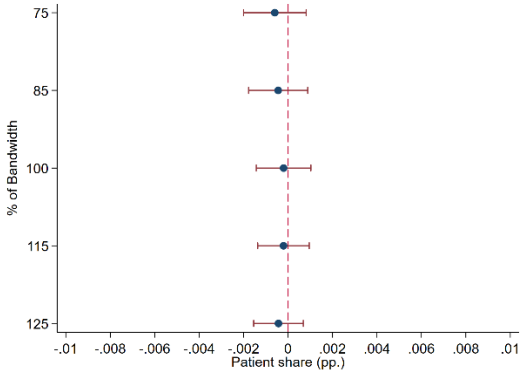
A. Polynomial degree 1 (BW = 1.25)



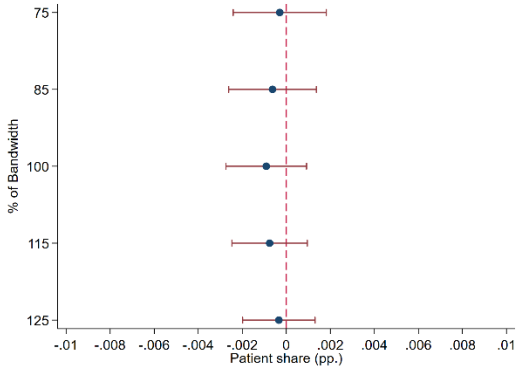
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 2.05)



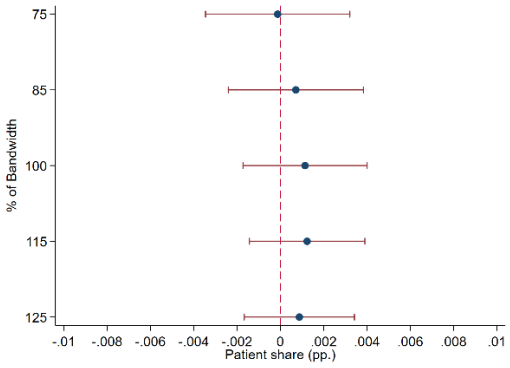
D. Polynomial degree 2 (Opt. BW = 2.05)



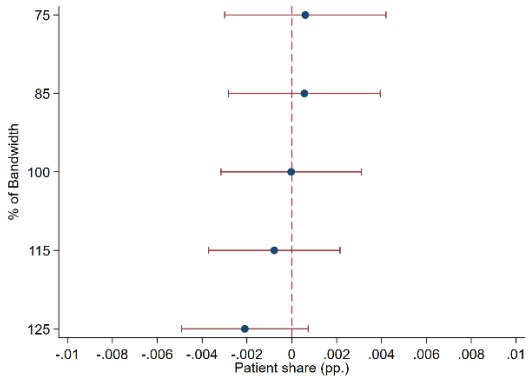
Note: The figure displays the average treatment effects on CVD Hospitalization at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S20. Varying bandwidths of threshold impact on All-cause Hospitalization

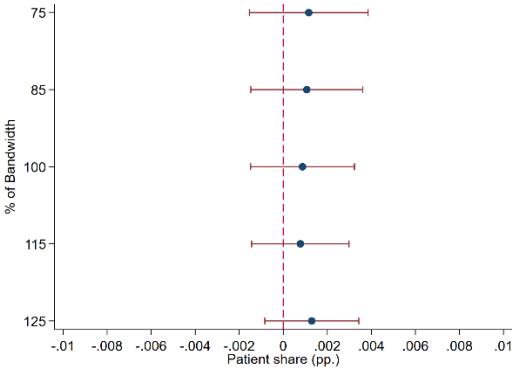
A. Polynomial degree 1 (BW = 1.25)



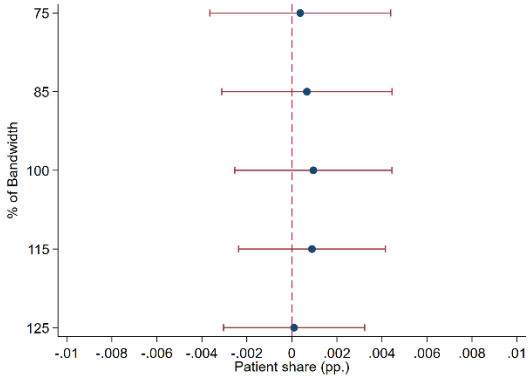
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 1.74)



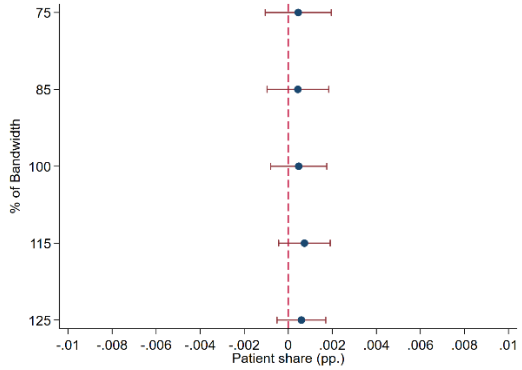
D. Polynomial degree 2 (Opt. BW = 1.74)



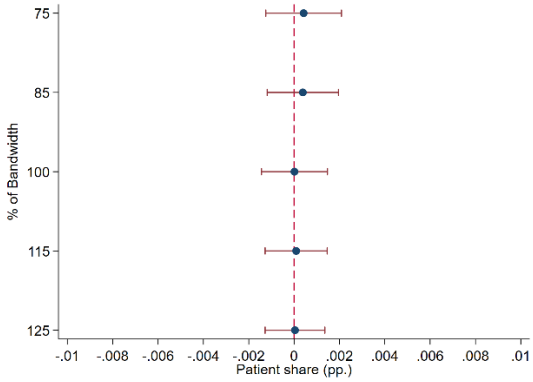
Note: The figure displays the average treatment effects on All-cause Hospitalization at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S21. Varying bandwidths of threshold impact on All-cause Mortality

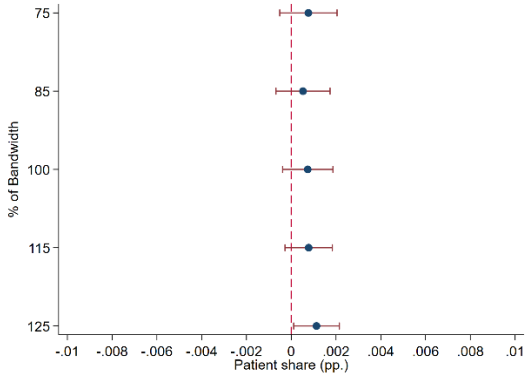
A. Polynomial degree 1 (BW = 1.25)



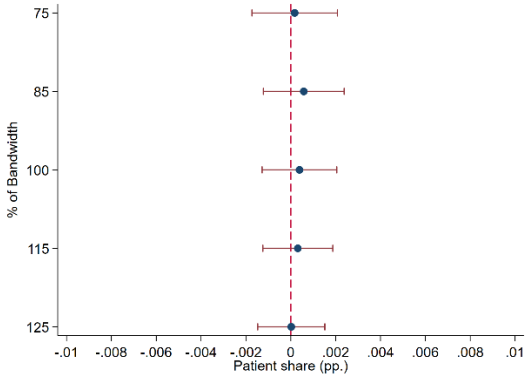
B. Polynomial degree 2 (BW = 2.50)



C. Polynomial degree 1 (Opt. BW = 1.57)



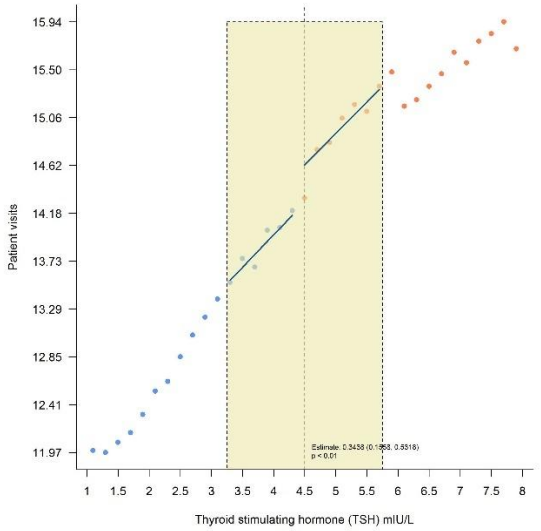
D. Polynomial degree 2 (Opt. BW = 1.57)



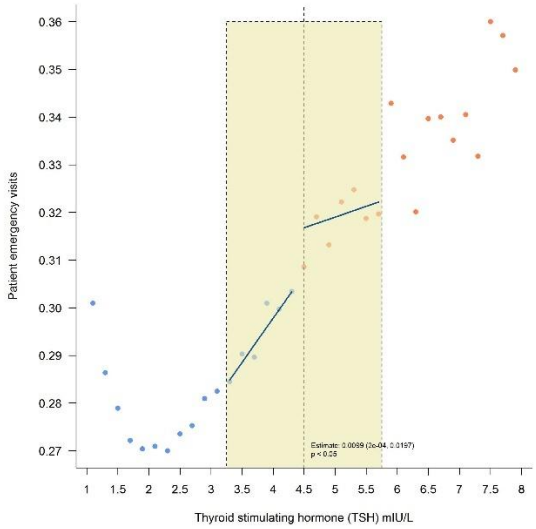
Note: The figure displays the average treatment effects on All-Cause Mortality at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Each figure varies the bandwidth from 75% to 125%. Panel A (B) shows effects using a local linear (quadratic) specification with a bandwidth of 1.25 (2.50). Panel C (D) uses the optimal bandwidth instead and shows again the local linear (quadratic) specification.

Fig S22. Types of visits across thyroid stimulating hormone

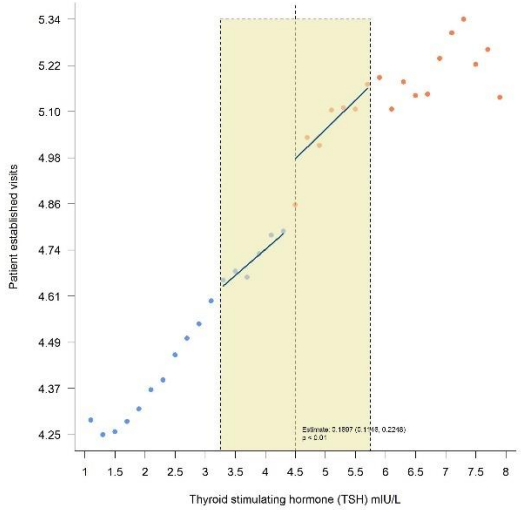
A. General Physician Visits



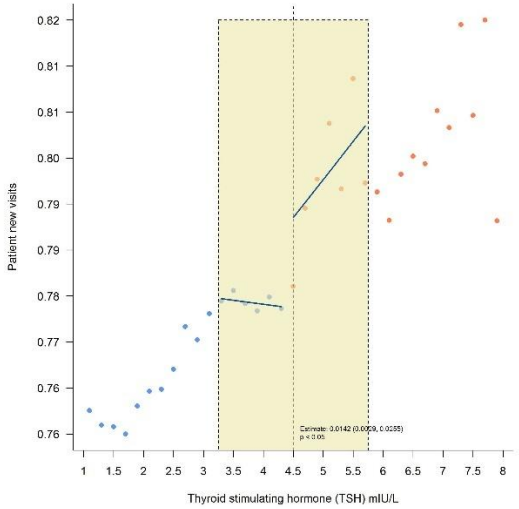
B. Emergency Room Visits



C. Established Visits



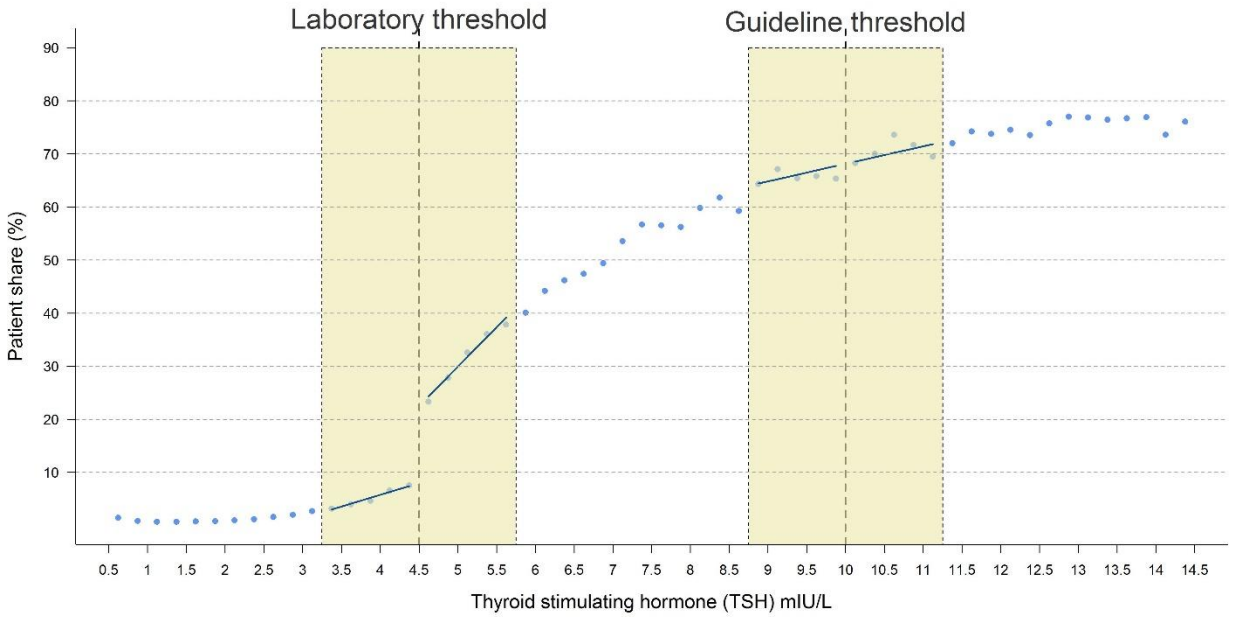
D. New Visits



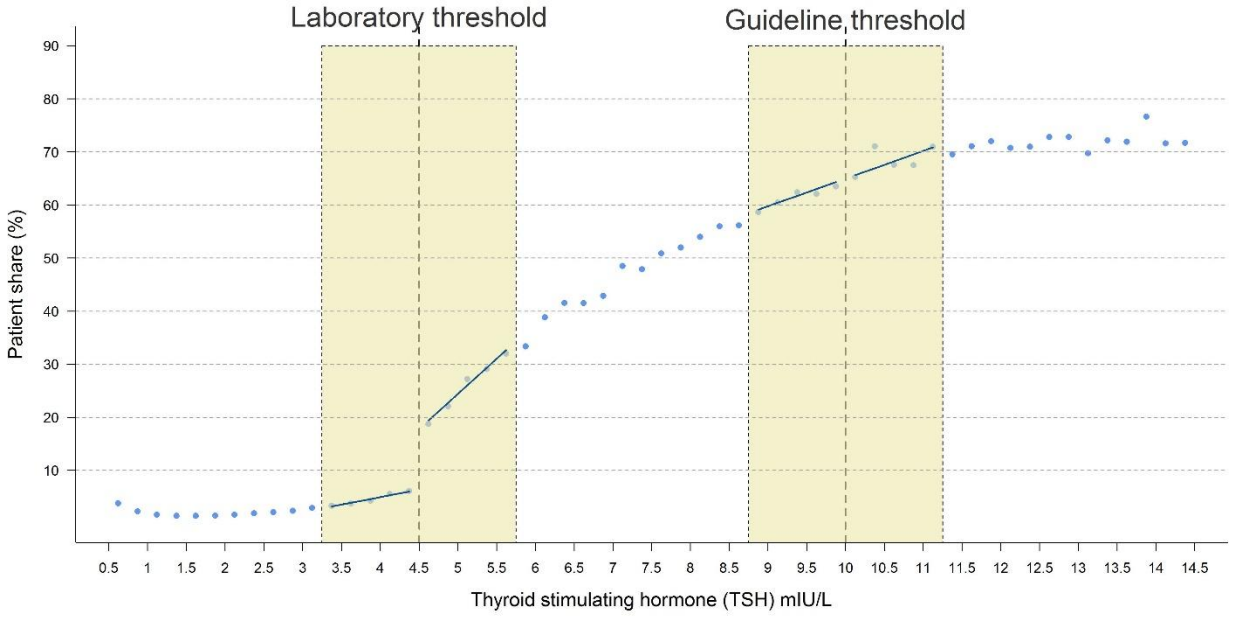
Note: The figure displays the average treatment effects on visits at the threshold of 4.5 mIU/L using the thyroid stimulating hormone (TSH) laboratory measurement as a running variable. Panel A through D show treatment effects using a local linear specification on general physician visits, emergency room visits, established visits, and new visits with a bandwidth of 1.25. The confidence intervals are based on heteroskedasticity robust standard errors.

Fig S23 L-prescriptions across thyroid stimulating hormone based on Age

A. L-prescription for the young



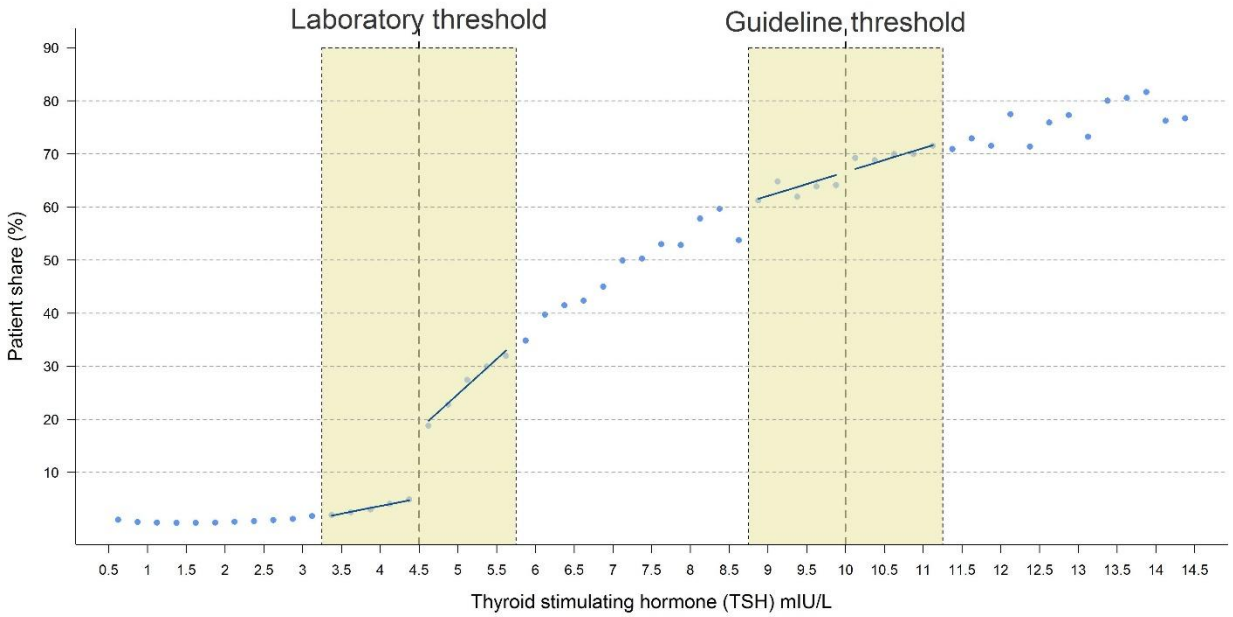
B. L-prescription for the elderly



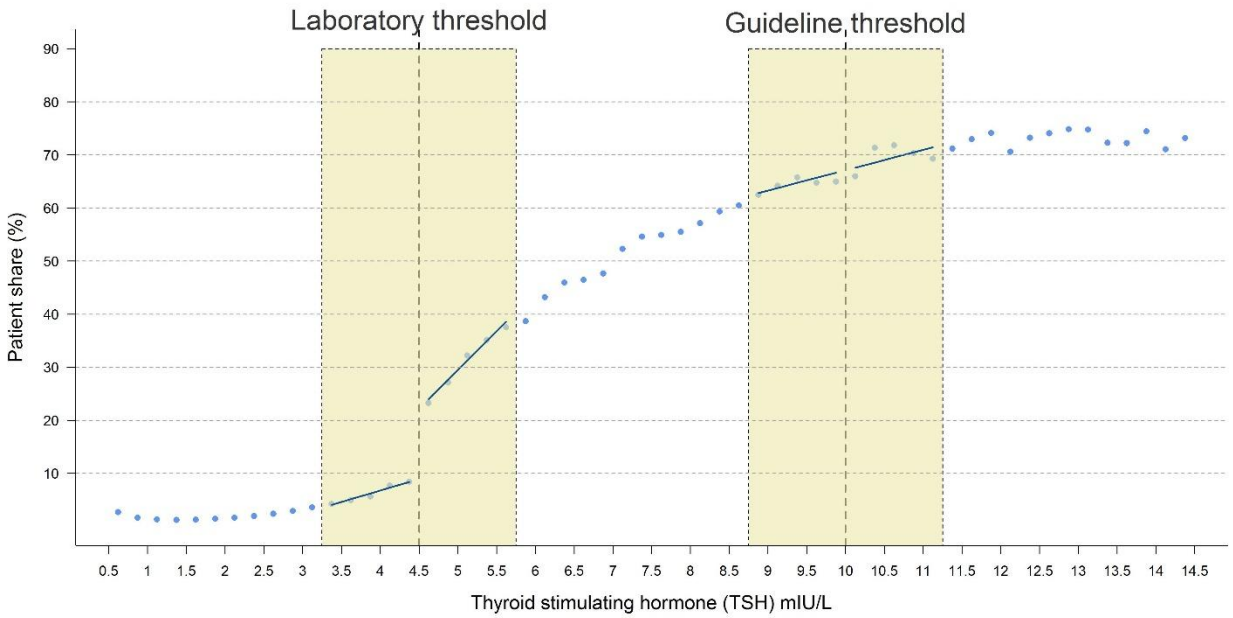
Note: The figure displays the primary exposure to a Levothyroxine prescription 12 months after the first thyroid stimulating hormone (TSH) laboratory measurement (mIU/L) across different age groups. We define young as individuals below 65 years of age and old as including and above 65 years of age. Panel A (B) shows average prescriptions for the young (elderly).

Fig S24. L-prescriptions across thyroid stimulating hormone based on health checkups

A. L-prescription during regular health check-ups



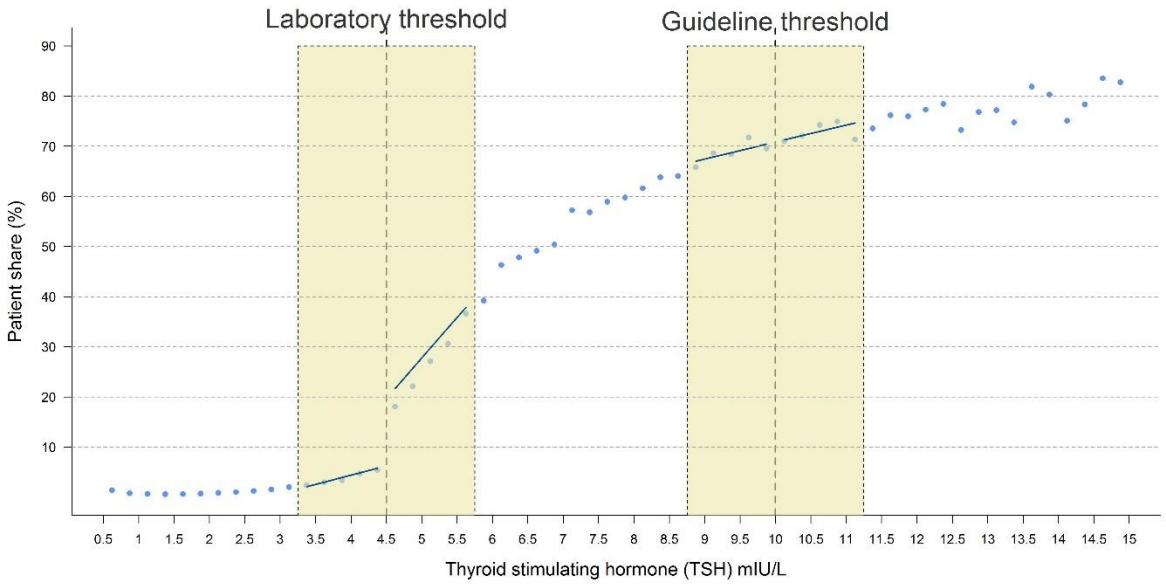
B. L-prescription outside of regular health check-ups



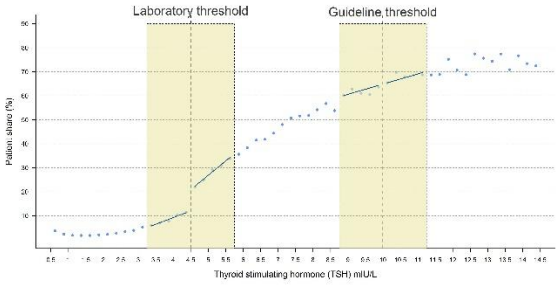
Note: The figure displays the primary exposure to a Levothyroxine prescription 12 months after the first thyroid stimulating hormone (TSH) laboratory measurement (mIU/L) for differential health check-up visits. Panel A (B) shows average prescriptions for patients during (outside of) regular health check-ups.

Fig S25. Prescriptions across thyroid stimulating hormone based on FT4

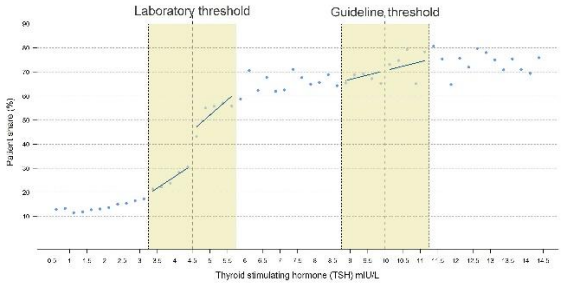
A. Prescriptions for missing FT4



B. Prescriptions for Regular or High FT4



C. Prescriptions for Low FT4



Note: The figure displays the primary exposure to a Levothyroxine prescription 12 months after the first thyroid stimulating hormone (TSH) laboratory measurement (mIU/L). Panel A through C shows Levothyroxine prescriptions for the 4.5 mIU/L upper limit of subclinical hypothyroidism for patients that did not receive any free thyroxine (FT4) measure, who received a regular or high FT4 measure, and for those that received a low FT4 measure, respectively.

Table S1. Description of outcome variables

Outcome Variable	Definition
Panel A: Prescription, diagnosis, and costs	
Prescription	Binary, indicating if a patient filled a Levothyroxine prescription within 1 year, 2 years, 3 years, or 5 years of TSH measurement
Diagnosis	Binary, indicating if a patient was diagnosed with hypothyroidism within 1 year, 2 years, 3 years, or 5 years of TSH measurement
Thyroid (std.) costs	Continuous, sum of thyroid costs charged to a patient's insurer in 2021 US dollars (USD) within 1 year, 2 years, 3 years, or 5 years of TSH measurement. Thyroid costs include TSH lab costs, Levothyroxine prescription costs, thyroid ablation costs, and thyroidectomy costs. Optum constructs standard pricing algorithms that account for variation in allowed payments across plans and providers separately for facility inpatient, facility outpatient, professional, ancillary, and pharmacy claims. These standardizations are provided from 2004-2021 and indexed to 2021 USD. Standardized costs for 2003 data are assumed to be $0 rather than missing.
Panel B: Laboratory outcomes	
Change in TSH value	Continuous, difference between first TSH value and the latest TSH value within 1 year, 2 years, 3 years, or 5 years of TSH measurement.
TSH below 0.4	Binary, indicating if a patient has any TSH value below the lower limit of normal range within 1 year, 2 years, 3 years, or 5 years of TSH measurement.

LDL	Continuous, mean serum LDL with units of mg/dL within 1 year, 2 years, 3 years, or 5 years of TSH measurement. Values below 0 or above 1,000 were excluded.
eGFR	Continuous, mean estimated glomerular filtration rate (eGFR) by CKD-EPI equation in 2009 using the race, gender, and creatinine values measured within 1 year, 2 years, 3 years, or 5 years of TSH measurement (Levey 2009). Creatinine values below 0 or above 1000 were excluded.
Panel C: Diagnosis-based health outcomes	
Fracture	Binary, indicating if a patient has any record of fracture diagnoses (as defined by ICD-9 and ICD-10 codes) within 1 year, 2 years, 3 years, or 5 years of TSH measurement.
CVD Hospitalization	Binary, indicating if a patient has any record of hospitalization associated with cardiovascular (CVD) diagnoses (as defined by ICD-9 and ICD-10 codes) on admission within 1 year, 2 years, 3 years, or 5 years of TSH measurement.
All-Cause Mortality	Binary, indicating if a patient has a record of death within 1 year, 2 years, 3 years, or 5 years of TSH measurement. Because the format of date of death only included year and month, the first day of each month was assumed to be the date of death.
All-cause Hospitalization	Binary, indicating if a patient has any record of hospitalization within 1 year, 2 years, 3 years, or 5 years of TSH measurement.
Other Measures	
TSH lab (std.) costs	Continuous, sum of TSH lab costs charged to a patient's insurer in 2021 US dollars (USD) within 1 year, 2 years, 3 years, or 5 years of TSH measurement. Optum constructs standard pricing algorithms that account for variation in allowed payments across plans and providers separately for facility inpatient,

	facility outpatient, professional, ancillary, and pharmacy claims. These standardizations are provided from 2004-2021 and indexed to 2021 USD. Standardized costs for 2003 data are assumed to be $0 rather than missing.
Charge	Continuous, sum of charges billed by a patient's insurer inflation-adjusted to 2021 US dollars (USD) using the Medical Care component of the Consumer Price Index for All Urban Consumers (CPI-U) within 1 year, 2 years, 3 years, or 5 years of TSH measurement (see Dunn et al. 2018 and BLS 2025).
Out-of-pocket costs	Continuous, sum of out-of-pocket costs (deductible, copay, coinsurance) paid by a patient inflation-adjusted to 2021 US dollars (USD) using the Health care component of Personal Consumption Expenditure (PCE) price within 1 year, 2 years, 3 years, or 5 years of TSH measurement. (see Dunn et al. 2018 and Census 2025)
General Physician Visits	Continuous, recorded number of health maintenance visits for a patient (as defined by ICD-9 and ICD-10 and CPT codes) within 1 year, 2 years, 3 years, or 5 years of TSH measurement.
Emergency Room Visits	Continuous, recorded number of visits for a patient associated with an emergency room (as defined by CPT codes) within 1 year, 2 years, 3 years, or 5 years of TSH measurement.
Established Visits	Continuous, recorded number of established (return) visits for a patient (as defined by CPT codes) within 1 year, 2 years, 3 years, or 5 years of TSH measurement.
New Visits	Continuous, recorded number of new visits for a patient (as defined by CPT codes) within 1 year, 2 years, 3 years, or 5 years of TSH measurement.
Covariates	

Age	Continuous, measured by difference in years between year of birth and year at time of TSH measurement. Patients with missing values were excluded.
Female	Binary, provided by data provider. Patients with missing values were excluded.
White race	Binary, provided by data provider. Patients with missing values were excluded.
Hispanic ethnicity	Binary, provided by data provider. Patients with missing values were excluded.
Black race	Binary, provided by data provider. Patients with missing values were excluded.
Asian race	Binary, provided by data provider. Patients with missing values were excluded.
Baseline diabetes mellitus	Binary, indicating if a patient has any record of diabetes diagnoses (as defined by ICD-9 and ICD-10 codes) prior to TSH measurement.
Baseline hypertension	Binary, indicating if a patient has any record of hypertension diagnoses (as defined by ICD-9 and ICD-10 codes) prior to TSH measurement.
Baseline LDL	Continuous, mean serum LDL with units of mg/dL in the 1 year prior to TSH measurement. Values below 0 or above 1,000 were excluded.
Baseline eGFR	Continuous, mean estimated glomerular filtration rate (eGFR) by CKD-EPI equation in 2009 using the race, gender, and creatinine values measured in the 1 year prior to TSH measurement. Values below 0 or above 1,000 were excluded.

Explanatory Variable	**Definition**
TSH	Thyroid stimulating hormone (TSH) laboratory test with upper limits of 4.5 mIU/L (or 5.5 mIU/L in robustness checks)

Table S2. Identification of Levothyroxine Prescriptions

Drug Name	Drug type
Eltroxin	Brand
Euthyrox	Brand
Levothroid	Brand
Levoxyl	Brand
Roxin	Brand
Synox	Brand
Synthroid	Brand
Thyronorm	Brand
Thyrosec	Brand
Thyrowin	Brand
Thyrox	Brand
Tirosint	Brand
Unithroid	Brand
Levothyroxine	Generic
Thyroxine	Generic
Liothyronine	Generic

Note: All drugs here are considered equivalents of levothyroxine and are treated as identical.

Table S3. Codes for laboratory measurements and diagnosis

Name	Code	Code type
Laboratory Measurement: TSH	11579-0	LOINC
Laboratory Measurement: TSH	3016-3	LOINC
Laboratory Measurement: TSH	11580-8	LOINC
Diagnosis: Hypothyroidism	243*	ICD-9
Diagnosis: Hypothyroidism	244*	ICD-9
Diagnosis: Hypothyroidism	E890*	ICD-10
Diagnosis: Hypothyroidism	E00*	ICD-10
Diagnosis: Hypothyroidism	E02*	ICD-10
Diagnosis: Hypothyroidism	E03*	ICD-10

Note: The above-mentioned codes capture 96% of the variation of all TSH codes. The diagnosis of hypothyroidism is not differentiated into subclinical and overt hypothyroidism and, hence, in clinical practice, the catch all hypothyrodisim diagnosis is used.

Table S4. Discontinuity in baseline characteristics at eligibility threshold.

Variable	Discontinuity at eligibility threshold	Sample size
Age	-0.0679 (-0.2516, 0.1158)	772,715
Female	0.0010 (-0.0041, 0.0061)	772,662
White race	0.0008 (-0.0040, 0.0056)	772,715
Hispanic ethnicity	-0.0003 (-0.0042, 0.0036)	772,715
Black race	-0.0023 (-0.0052, 0.0005)	772,715
Asian race	0.0018 (-0.0008, 0.0044)	772,715
Baseline diabetes mellitus	0.0007 (-0.0032, 0.0045)	772,715
Baseline hypertension	-0.0000 (-0.0051, 0.0050)	772,715
Baseline LDL	-0.1685 (-0.5837, 0.2467)	611,411
Baseline eGFR	0.1321 (-0.1082, 0.3724)	704,214
Frequency (McCrary Test)	0.0013 (-0.0015, 0.0041)	

Note. The table shows whether the demographics, baseline comorbidities, baseline laboratory measures, and the frequency of TSH measures are smooth across the 4.5 mIU/L threshold. The variables on the left are used as outcome variable in the linear regression discontinuity specification.

Table S5. 2-year treatment effects at early hypothyroidism threshold

	Baseline	Treatment effects	Sample size
Panel A: Prescription, diagnosis, and costs			
Prescription	0.07	0.1316	772,715
		(0.1278, 0.1354)	
Diagnosis	0.18	0.2054	772,715
		(0.2008, 0.2101)	
Thyroid std. costs	220.16	40.8992	772,715
		(35.1136, 46.6848)	
Panel B: Laboratory outcomes			
TSH change	-0.97	-0.1519	342,660
		(-0.1726, -0.1313)	
TSH below 0.4	0.02	0.0075	342,660
		(0.0052, 0.0098)	
LDL	102.09	0.0918	357,877
		(-0.4093, 0.5928)	
eGFR	77.63	0.4573	424,397
		(0.1514, 0.7631)	
Panel C: Diagnosis-based health outcomes			
Fracture	0.05	0.0009	772,715
		(-0.0014, 0.0032)	
CVD hospitalization	0.03	-0.0010	772,715
		(-0.0029, 0.0008)	
All-cause hospitalization	0.13	0.0020	772,715
		(-0.0015, 0.0055)	
All-cause mortality	0.03	0.0001	772,715
		(-0.0018, 0.0019)	

Note: This table shows the treatment effects on outcomes within two years after the TSH laboratory measurement when crossing the threshold of 4.5 mIU/L from left to right where subclinical hypothyroidism is indicated on the right side of the threshold and monitoring is recommended. The baseline is the estimate of being directly below the cutoff of 4.5 mIU/L. All statistics are shown within the smallest bandwidth within 3.25 mIU/L to 5.75 mIU/L. Confidence intervals are shown using heteroskedasticity robust standard errors.

Table S6. 3-year treatment effects at early hypothyroidism threshold

	Baseline	Treatment effects	Sample size
Panel A: Prescription, diagnosis, and costs			
Prescription	0.07	0.1316	772,715
		(0.1278, 0.1354)	
Diagnosis	0.19	0.2026	772,715
		(0.1979, 0.2073)	
Thyroid std. costs	295.51	49.4821	772,715
		(42.2457, 56.7186)	
Panel B: Laboratory outcomes			
TSH change	-0.97	-0.1358	375,882
		(-0.1559, -0.1157)	
TSH below 0.4	0.02	0.0082	375,882
		(0.0059, 0.0106)	
LDL	102.31	-0.0386	388,418
		(-0.5141, 0.4369)	
eGFR	78.03	0.4687	451,813
		(0.1712, 0.7663)	
Panel C: Diagnosis-based health outcomes			
Fracture	0.06	0.0013	772,715
		(-0.0012, 0.0038)	
CVD hospitalization	0.04	-0.0001	772,715
		(-0.0022, 0.0020)	
All-cause hospitalization	0.16	0.0038	772,715
		(-0.0001, 0.0076)	
All-cause mortality	0.05	-0.0002	772,715
		(-0.0024, 0.0020)	

Note: This table shows the treatment effects on outcomes within three years after the TSH laboratory measurement when crossing the threshold of 4.5 mIU/L from left to right where subclinical hypothyroidism is indicated on the right side of the threshold and monitoring is recommended. The baseline is the estimate of being directly below the cutoff of 4.5 mIU/L. All statistics are shown within the smallest bandwidth within 3.25 mIU/L to 5.75 mIU/L. Confidence intervals are shown using heteroskedasticity robust standard errors.

Table S7. 5-year treatment effects at early hypothyroidism threshold

	Baseline	Treatment effects	Sample size
Panel A: Prescription, diagnosis, and costs			
Prescription	0.07	0.1316	772,715
		(0.1278, 0.1354)	
Diagnosis	0.21	0.1995	772,715
		(0.1948, 0.2042)	
Thyroid std. costs	388.08	65.6059	772,715
		(56.6012, 74.6106)	
Panel B: Laboratory outcomes			
TSH change	-0.98	-0.1295	398,799
		(-0.1494, -0.1096)	
TSH below 0.4	0.02	0.0091	398,799
		(0.0067, 0.0115)	
LDL	102.39	-0.0219	409,895
		(-0.4805, 0.4368)	
eGFR	78.22	0.4845	470,618
		(0.1919, 0.7771)	
Panel C: Diagnosis-based health outcomes			
Fracture	0.08	0.0015	772,715
		(-0.0013, 0.0043)	
CVD hospitalization	0.05	0.0000	772,715
		(-0.0023, 0.0023)	
All-cause hospitalization	0.19	0.0042	772,715
		(0.0002, 0.0083)	
All-cause mortality	0.07	-0.0007	772,715
		(-0.0032, 0.0019)	

Note: This table shows the treatment effects on outcomes within five years after the TSH laboratory measurement when crossing the threshold of 4.5 mIU/L from left to right where subclinical hypothyroidism is indicated on the right side of the threshold and monitoring is recommended. The baseline is the estimate of being directly below the cutoff of 4.5 mIU/L. All statistics are shown within the smallest bandwidth within 3.25 mIU/L to 5.75 mIU/L. Confidence intervals are shown using heteroskedasticity robust standard errors.

Table S8. 1-year treatment effect at early hypothyroidism threshold varying covariates and weights

	(1)	(2)	(3)	(4)
Panel A: Prescription, diagnosis, and costs				
Prescription	0.1316	0.1270	0.1246	0.1204
	(0.1278, 0.1354)	(0.1229, 0.1312)	(0.1205, 0.1286)	(0.1160, 0.1247)
Diagnosis	0.2054	0.2091	0.1945	0.1984
	(0.2009, 0.2100)	(0.2041, 0.2142)	(0.1896, 0.1993)	(0.1929, 0.2038)
Thyroid std. costs	22.2583	20.3938	21.5521	19.6814
	(18.5002, 26.0165)	(16.5974, 24.1901)	(17.6704, 25.4338)	(15.8885, 23.4743)
Panel B: Laboratory outcomes				
TSH change	-0.1626	-0.1480	-0.1603	-0.1470
	(-0.1864, -0.1388)	(-0.1750, -0.1211)	(-0.1873, -0.1334)	(-0.1764, -0.1177)
TSH below 0.4	0.0040	0.0045	0.0038	0.0041
	(0.0016, 0.0065)	(0.0019, 0.0071)	(0.0010, 0.0065)	(0.0013, 0.0070)
LDL	0.7147	0.4125	0.8957	0.5386
	(0.0958, 1.3337)	(-0.0739, 0.8990)	(0.2238, 1.5677)	(0.0126, 1.0646)
eGFR	0.5874	0.0228	0.5297	0.0362
	(0.2359, 0.9389)	(-0.1357, 0.1813)	(0.1453, 0.9142)	(-0.1351, 0.2076)
Panel C: Diagnosis-based health outcomes				
Fracture	-0.0001	0.0001	0.0007	0.0008
	(-0.0019, 0.0018)	(-0.0019, 0.0022)	(-0.0013, 0.0027)	(-0.0014, 0.0030)
CVD hospitalization	-0.0006	-0.0008	-0.0004	-0.0009
	(-0.0021, 0.0009)	(-0.0024, 0.0009)	(-0.0021, 0.0012)	(-0.0027, 0.0009)
All-cause hospitalization	0.0011	0.0003	0.0007	0.0005
	(-0.0018, 0.0041)	(-0.0029, 0.0034)	(-0.0024, 0.0039)	(-0.0030, 0.0039)
All-cause mortality	0.0005	0.0003	0.0005	0.0005
	(-0.0009, 0.0018)	(-0.0011, 0.0018)	(-0.0010, 0.0020)	(-0.0010, 0.0021)
Covariates	No	Yes	No	Yes
Weights	Uniform	Uniform	Triangular	Triangular

Note: This table shows the treatment effects on outcomes within one years after the TSH laboratory measurement when crossing the threshold of 4.5 mIU/L from left to right where subclinical hypothyroidism is indicated on the right side of the threshold and monitoring is recommended. Column (1) and (3) show the estimated effects without covariates varying the weighting scheme from uniform to triangular while column (4) and (5) show the estimated effects with covariates (as shown in Table S4 while similarly varying the weighting scheme. The baseline is the estimate of being directly below the cutoff of 4.5 mIU/L. All statistics are shown within the smallest bandwidth within 3.25 mIU/L to 5.75 mIU/L. Confidence intervals are shown using heteroskedasticity robust standard errors.

Table S9. 1-year complier treatment effects from prescription at early hypothyroidism threshold

	Baseline	Treatment effects	Sample size
Panel A: Prescription, diagnosis, and costs			
Prescription (Endogenous Regressor)			
Diagnosis	0.04	1.5568 (1.5253, 1.5883)	772,715
Thyroid std. costs	102.38	167.0136 (136.8765, 197.1508)	772,715
Panel B: Laboratory outcomes			
TSH change	-0.74	-1.2860 (-1.4777, -1.0944)	228,390
TSH below 0.4	0.01	0.0373 (0.0179, 0.0567)	228,390
LDL	99.95	5.1366 (0.9972, 9.2761)	244,146
eGFR	75.60	4.0785 (1.6337, 6.5233)	317,727
Panel C: Diagnosis-based health outcomes			
Fracture	0.03	-0.0026 (-0.0167, 0.0115)	772,715
CVD hospitalization	0.02	-0.0059 (-0.0172, 0.0054)	772,715
All-cause hospitalization	0.09	0.0077 (-0.0146, 0.0300)	772,715
All-cause mortality	0.02	0.0024 (-0.0075, 0.0123)	772,715

Note: This table shows the 2SLS regression complier treatment effects from prescription of Levothyroxine on outcomes within one year after the TSH laboratory measurement when crossing the threshold of 4.5 mIU/L from left to right where subclinical hypothyroidism is indicated on the right side of the threshold and monitoring is recommended. The baseline is the estimate of being directly below the cutoff of 4.5 mIU/L. All statistics are shown within the smallest bandwidth within 3.25 mIU/L to 5.75 mIU/L. Confidence intervals are shown using heteroskedasticity robust standard errors.

Table S10. 1-year treatment effects at early hypothyroidism threshold 4.5 and 5.5 mIU/L pooled

	Baseline	Treatment effects	Sample size
Panel A: Prescription, diagnosis, and costs			
Prescription	0.08	0.1378	817,734
		(0.1340, 0.1416)	
Diagnosis	0.16	0.2112	817,734
		(0.2067, 0.2156)	
Thyroid std. costs	113.89	22.5198	817,734
		(18.8453, 26.1942)	
Panel B: Laboratory outcomes			
TSH change	-1.04	-0.1724	241,575
		(-0.1971, -0.1477)	
TSH below 0.4	0.02	0.0037	241,575
		(0.0011, 0.0063)	
LDL	100.89	0.7271	254,766
		(0.1206, 1.3336)	
eGFR	76.09	0.6444	330,979
		(0.2975, 0.9912)	
Panel C: Diagnosis-based health outcomes			
Fracture	0.03	-0.0003	817,734
		(-0.0020, 0.0015)	
CVD hospitalization	0.02	-0.0007	817,734
		(-0.0021, 0.0008)	
All-cause hospitalization	0.09	0.0004	817,734
		(-0.0024, 0.0033)	
All-cause mortality	0.02	0.0003	817,734
		(-0.0010, 0.0016)	

Note: This table shows the treatment effects on outcomes within one years after the TSH laboratory measurement when crossing the threshold of 4.5 mIU/L or 5.5 mIU/L from left to right where subclinical hypothyroidism is indicated on the right side of the threshold and monitoring is recommended. The baseline is the estimate of being directly below the cutoff of 4.5 mIU/L or 5.5 mIU/L. All statistics are shown within the smallest normalized bandwidth of -1.25 mIU/L to 1.25 mIU/L. Confidence intervals are shown using heteroskedasticity robust standard errors.

Table S11. Granular cost effects and back-of-the-envelope calculation using 1-year estimates

Panel A: In-depth cost effects

	Threshold:			
	4.5 mIU/L	**5.0 mIU/L**	**6.0 mIU/L**	**7.0 mIU/L**
GP visit std. costs	14.86	17.90	24.00	30.09
	(9.32, 20.40)	(12.22, 23.59)	(12.35, 35.64)	(10.79, 49.39)
ER visit std. costs	7.58	-1.29	-19.05	-36.80
	(-1.55, 16.71)	(-10.22, 7.63)	(-37.30, -0.79)	(-67.38, -6.21)
Inpatient std. costs	122.59	5.42	-228.93	-463.27
	(-82.53, 327.71)	(-198.22, 209.05)	(-687.92, 230.06)	(-1242.43, 315.89)
Total std. costs	256.49	120.22	-152.31	-424.85
	(-34.35, 547.33)	(-161.97, 402.41)	(-739.56, 434.93)	(-1413.25, 563.55)

Panel B: Back-of-the-envelope calculation

Statistics	
Small Bandwidth Sample Size (n_S^{CDM})	772,715
Medium Bandwidth Sample Size (n_M^{CDM})	6,996,054
Prescription Jump (p)	13.16 pp
1-year cost treatment effect at 7 mlU/L (c1)	-$424.85
5-year cost treatment effect at 7 mlU/L (c5)	-$717.85
Scaled Subclinical Hypothyroidism Prevalence of Americans (Wyne et al. 2023) (t)	11%
US Adult Population in 2021 (Census, 2025) (n_{2021}^{US})	266,978,268
Cost-Analysis at extrapolated 7 mlU/L threshold:	
Small Sample USD Savings ($n_S^{CDM} \times p \times c1$)	$43,202,697
Medium Sample USD Savings ($n_M^{CDM} \times p \times c1$)	$391,151,198
US Prevalence Sample USD Savings ($n_{2021}^{US} \times t \times p \times c1$)	$1,641,950,680
Small Sample USD Savings ($n_S^{CDM} \times p \times c5$)	$72,997,660
Medium Sample USD Savings ($n_M^{CDM} \times p \times c5$)	$660,910,645
US Prevalence Sample USD Savings ($n_{2021}^{US} \times t \times p \times c5$)	$2,774,330,460

Note: This table shows the treatment effects when we move the threshold of subclinical hypothyroidism of 4.5 mIU/L, where monitoring is recommended, closer to the threshold of overt hypothyroidism of 10 mIU/L, where prescription is recommended. Column 1 shows the impact of being just below vs above the threshold of subclinical hypothyroidism while the following columns show the impact of being just below and above the new extrapolated thresholds of 5 mIU/L, 6 mIU/L, and 7 mIU/L. The confidence intervals are in parenthesis and based on standard errors that are derived using the Delta method. Results on surgery, ICU, imaging, and procedures costs are available upon request.

Table S12. 1-year treatment effects at prescription threshold 10 mIU/L

	Baseline	Treatment effects	Sample size
Panel A: Prescription, diagnosis, and costs			
Prescription	0.68	0.0185	21,406
		(-0.0071, 0.0441)	
Diagnosis	0.76	0.0127	21,406
		(-0.0109, 0.0364)	
Thyroid std. costs	188.26	10.0236	21,406
		(-16.7149, 36.7622)	
Panel B: Laboratory outcomes			
TSH change	-5.63	-0.0251	13,214
		(-0.2396, 0.1894)	
TSH below 0.4	0.10	0.0146	13,214
		(-0.0063, 0.0356)	
LDL	108.27	-0.1191	6,892
		(-3.6806, 3.4424)	
eGFR	78.31	0.1440	9,150
		(-1.8939, 2.1820)	
Panel C: Diagnosis-based health outcomes			
Fracture	0.03	0.0032	21,406
		(-0.0062, 0.0126)	
CVD hospitalization	0.02	0.0034	21,406
		(-0.0046, 0.0114)	
All-cause hospitalization	0.10	-0.0028	21,406
		(-0.0195, 0.0139)	
All-cause mortality	0.02	0.0021	21,406
		(-0.0064,0.0106)	

Note: This table shows the treatment effects on outcomes within one years after the TSH laboratory measurement when crossing the threshold of 10 mIU/L from left to right where overt hypothyroidism is indicated on the right side of the threshold and prescription is recommended. The baseline is the estimate of being directly below the cutoff of 10 mIU/L. All statistics are shown within the smallest bandwidth of 8.75 mIU/L to 11.25 mIU/L. Confidence intervals are shown using heteroskedasticity robust standard errors.

Supplements References

BLS (2025). United States Bureau of Labor Statistics. BLS Data Viewer. https://data.bls.gov/dataViewer/view/timeseries/CUUR0000SAM

Census (2025). National Population by Characteristics: 2020-2024, https://www.census.gov/data/tables/time-series/demo/popest/2020s-national-detail.html, version from 8/11/2025.

Dunn, A., Grosse, S. D., & Zuvekas, S. H. (2018). Adjusting health expenditures for inflation: a review of measures for health services research in the United States. Health services research, 53(1), 175-196.

Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y., Castro III, A. F., Feldman, H. I., ... & CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) *. (2009). A new equation to estimate glomerular filtration rate. Annals of internal medicine, 150(9), 604-612.