Information, Incentives, and Institutions: Essays on Housing, Media, and Health Care Markets

Inauguraldissertation

zur

Erlangung des Doktorgrades

der

Wirtschafts- und Sozialwissenschaftlichen Fakultät

der

Universität zu Köln

2026

vorgelegt von

M.Sc. Felix Mindl

aus Tübingen

Referent*in:	Prof. Dr. Johannes Münster
Korreferent*in:	Prof. Michael Krause, Ph.D.
Tag der Promotion:	24.09.2026

Contents

Introduction **1**

1 Short-Term Rental Platforms and the Small-Apartment Segment **6**

1.1 Introduction . . . 6
1.2 Institutional Background . . . 9
1.3 Data and Descriptive Analysis . . . 10
1.3.1 Data . . . 10
1.3.2 Professionalism . . . 12
1.3.3 Short-term rental exposure: geography, size, and growth . . . 14
1.3.4 Summary statistics . . . 16
1.4 Empirical Strategy . . . 17
1.4.1 Hedonic model and estimand . . . 17
1.4.2 Identifying assumption . . . 19
1.4.3 Specification, fixed effects, and inference . . . 19
1.4.4 Parallel trends and balance . . . 20
1.5 Results . . . 20
1.5.1 Main estimates . . . 20
1.5.2 Event study . . . 24
1.5.3 Dose-response . . . 24
1.5.4 Magnitude . . . 25
1.6 Robustness . . . 26
1.6.1 Pre-trend sensitivity . . . 26
1.6.2 Randomization inference and placebo tests . . . 27
1.6.3 Few clusters and spatial dependence . . . 28
1.6.4 Rent brake, refugee inflow, and apartment composition . . . 29
1.6.5 Sensitivity to operational choices . . . 31
1.6.6 Flexible entry . . . 32
1.7 Discussion . . . 34
1.7.1 Interpretation and Mechanism . . . 34
1.7.2 Policy Implications . . . 35
1.7.3 Limitations . . . 36
1.8 Conclusion . . . 38

References . . . 39

2 Debunking "Fake News" on Social Media 41

3 Second Opinions, Health Care Cost, and Outcomes in Elective Surgery 42

3.1 Introduction . . . 42
3.2 Related Literature . . . 46
3.2.1 Credence Goods in Economics . . . 46
3.2.2 Empirical Evidence on Second-Opinion Programs in Health Care . . . 49
3.2.3 Healthcare Operations and Provider Decision-Making . . . 50
3.3 Institutional Design and Hypotheses . . . 51
3.3.1 Second-Opinion Programs and the German Health Care Market . . . 51
3.3.2 Mapping the Literature to the Institutional Design . . . 53
3.3.3 Hypotheses Development . . . 54
3.4 Data and Descriptive Statistics . . . 55
3.4.1 Data and Analysis Sample . . . 55
3.4.2 Variables . . . 57
3.4.3 Descriptive Statistics . . . 60
3.5 Empirical Strategy . . . 64
3.5.1 Identification Strategy . . . 64
3.5.2 Constructing a Balanced Sample Using Matching . . . 65
3.5.3 Matched Event Study . . . 66
3.5.4 Discussion of Identifying Assumptions . . . 67
3.6 Results . . . 68
3.6.1 Construction of the Matched Sample . . . 69
3.6.2 Main Effects of Compliance . . . 69
3.6.3 Heterogeneous Effects by Second-Opinion Recommendation . . . 72
3.6.4 Therapy Pathways . . . 77
3.7 Robustness . . . 79
3.7.1 Labor Market Participation Effects . . . 80
3.7.2 Alternative Observation Windows . . . 80
3.7.3 Instrumental Variable Approach . . . 83
3.8 Discussion . . . 86
3.9 Concluding Remarks . . . 89
References . . . 91

A Appendix to Chapter 1 96

A.1 Omitted Tables . . . 96
A.2 Omitted Figures . . . 104
A.3 Dose-response under a continuous-treatment estimator . . . 109

B Appendix to Chapter 2 **111**

C Appendix to Chapter 3 **112**

C.1 Data Construction and Variable Definitions 112

C.1.1 Identifying Second-Opinion Indications 112

C.1.2 Surgical Procedures and Compliance 112

C.1.3 Medication Classification 113

C.2 Sample Construction and Propensity Score Matching 114

C.2.1 Pre-Match Covariate Balance 114

C.2.2 Propensity Score Estimation 114

C.2.3 Matching Diagnostics 117

C.3 Treatment Pathways and Mechanisms 121

C.4 Labor Market Outcomes 123

C.5 Alternative Observation Windows 125

C.6 Instrument Variable Approach 129

Eidesstattliche Erklärung **130**

Declaration on the Use of Generative AI **131**

List of Figures

1.1 Airbnb density terciles across Cologne neighborhoods 14
1.2 Mean Airbnb density by neighborhood tercile, 2010 Q1–2018 Q4 15
1.3 Room-size distribution: Airbnb listings vs. rental advertisements 16
1.4 Within-tercile size gap in log rent, 2010 Q1–2018 Q4 22
1.5 Event study: triple DiD coefficients by quarter relative to treatment 25
1.6 Dose-response: rent effect by Airbnb exposure intensity 26
1.7 Pre-trend sensitivity under the smoothness restriction 28
1.8 Sensitivity of the triple-difference estimate to operational choices 32
1.9 Flexible-entry event study, imputation estimator 33

3.1 Timeline of the Structured Second-Opinion Program 52
3.2 Construction of the Analysis Sample 57
3.3 Recommendations of First and Second Opinion 61
3.4 Health Outcomes by Compliance with Second Opinion. 63

A.1 Event study over the full sample window 104
A.2 Pre-trend sensitivity under the relative-magnitudes restriction 105
A.3 Admissible linear differential trends under $\Delta^{SD}(M = 0)$ 105
A.4 Geographic randomization inference: placebo distribution of the triple-difference estimate 106
A.5 In-time placebo: triple-difference estimate at fictitious treatment dates . . 106
A.6 Within-district randomization inference: placebo distribution under spatially constrained relabeling 107
A.7 Conley standard error of the triple-difference estimate by distance cutoff . 108
A.8 Continuous-treatment dose-response on small apartments 110

C.1 Distribution of Time to First Surgery after the Second-opinion Request, by Recommendation; dashed line marks the 180-day compliance window. . 113
C.2 Propensity Score Distributions for Compliers and Non-compliers under the Elastic-net $\lambda_{\min}$ and λ_{1se} Specifications. 114
C.3 Absolute mean standardized differences between compliers and non-compliers under the Elastic-net $\lambda_{\min}$ and λ_{1se} specifications with full matching. . . . 118
C.4 Health Outcomes by Compliance in Matched Sample. 120

List of Tables

1.1 Host concentration among active Airbnb listings (Cologne, 2018 Q4) . . . 12
1.2 Airbnb market structure and break-even with long-term rental income (2018 Q4) . . . 13
1.3 Density distribution across T3 neighborhoods (2018 Q4) . . . 16
1.4 Summary statistics, estimation sample . . . 18
1.5 Pre-treatment balance: normalized differences . . . 21
1.6 Effect of short-term rentals on small-apartment rents . . . 23

3.1 Patients' Treatment Choices by Physician Recommendations . . . 61
3.2 Health Outcomes by Compliance . . . 71
3.3 Heterogeneity in Health Outcomes – Second Opinion: Surgery Discouraged 73
3.4 Heterogeneity in Health Outcomes – Second Opinion: Surgery Recommended 75
3.5 Therapy Pathways by Compliance – Second Opinion: Surgery Discouraged 78
3.6 Labor Market Outcomes by Compliance – Second opinion: Surgery Discouraged . . . 81
3.7 Short-run Health Outcomes by Compliance . . . 82
3.8 Long-run Health Outcomes by Compliance . . . 84
3.9 IV Robustness Check: Effect of Surgical Avoidance on Health Outcomes . 86

A.1 Differential post-2015 demographic shock in T3 . . . 96
A.2 Sensitivity to time-varying neighborhood demographics . . . 97
A.3 Composition of advertised rental stock: pre vs post . . . 98
A.4 Treatment date sensitivity: ±1 quarter . . . 99
A.5 Room-size cutoff sensitivity . . . 100
A.6 Density measure: end-of-sample vs. time-averaged . . . 101
A.7 Group definition: equal terciles vs. 60/20/20 split . . . 102
A.8 Professional threshold sensitivity: 90, 110, and 140 nights . . . 103

C.1 Baseline Characteristics by Compliance . . . 115
C.2 Elastic-net Propensity Score Model: Non-zero Coefficients . . . 116
C.3 Compliance by Propensity Score Quintile . . . 117
C.4 Balance Summary for Full Matching by Elastic-net Propensity Score Specification. . . . 117

C.5 Insurance status in the preferred matched sample (full matching, elastic-net $\lambda_{\min}$). 119
C.6 Treatment Outcomes by Compliance – Second Opinion: Surgery Recommended 122
C.7 Labor Market Outcomes by Compliance – Second Opinion: Surgery Recommended 124
C.8 Short-run Health Outcomes – Second Opinion: Surgery Discouraged 125
C.9 Short-run Health Outcomes – Second Opinion: Surgery Recommended . . 126
C.10 Long-run in Health Outcomes – Second Opinion: Surgery Discouraged . . 127
C.11 Long-run in Health Outcomes – Second Opinion: Surgery Recommended . 128
C.12 Placebo Test: Effect of Instrument on Pre-Period Outcome changes 129

Introduction

This dissertation consists of three self-contained empirical essays on markets that households deal with directly: urban rental housing, social media, and elective health care. Each essay asks a version of the same practical question. Where does the market fall short of serving consumers, how large is the damage, and which remedy works? Each chapter quantifies a concrete distortion with a research design built for causal inference and states its results in euros, percentage points, and avoided operations. All three speak directly to live policy debates on the regulation of short-term rental platforms, the fight against online misinformation, and the design of second-opinion programs in health care.

The three essays are connected by overlapping lines rather than a single framework. Chapters 1 and 3 are applied microeconomics on rich field data, roughly 90,000 rental advertisements matched to property-level short-term rental records in one case and comprehensive administrative health insurance claims in the other. In both, the central challenge is to separate causal effects from the selection that pervades observational data. Platform operators locate where rents would have risen anyway, and patients who differ in their health histories choose different treatments. Chapters 1 and 2 study markets reshaped by digitization. Airbnb and Facebook scaled from niche services to market-defining platforms within a few years, far faster than regulation adapted, and both chapters measure what this platform growth does to outcomes consumers and policymakers care about, rents in one case and beliefs in the other. Chapters 2 and 3, finally, share an information problem. Consumers must judge the quality of what they are told, be it a health claim in a social media post or a surgeon's recommendation, and both chapters evaluate remedies that protect them against misleading content or interested advice.

The comparison across this last pair is instructive, because the same broad problem calls for remedies of very different depth. A social media post is not an expert good in the strict sense. With modest guidance, users can assess plausibility, check sources, and spot manipulation themselves. Chapter 2 shows that a brief media literacy intervention, ten tips read once, equips them to do exactly that. Elective surgery sits at the opposite end of the spectrum. Medical treatment is a classic credence good. Patients cannot self-diagnose at any reasonable cost, so no brief training will do. Consumer protection requires a second, independent expert whose advice is decoupled from the treatment margin, as Chapter 3 documents. In both markets, institutional solutions can carry the remedy to scale. Platforms can be required to display accuracy tips (the tips we test were developed

by Facebook itself, though Facebook never promoted them at scale), and media literacy training in school curricula can be strengthened. In health care, German statutory health insurers introduced structured second-opinion programs decentrally, well before the legislature created a statutory right to a second opinion in 2015. Housing completes the picture from the regulatory side. Chapter 1 delivers the benchmark magnitude, measured in a setting where existing regulation was unenforced or legally void, against which platform-targeted regulation can be judged.

Read in sequence, the chapters thus move from diagnosis to remedy. Chapter 1 measures the harm from unregulated platform growth. Chapter 2 compares two remedies head-to-head in a randomized experiment. Chapter 3 evaluates a remedy that has already been institutionalized.

Short-Term Rental Platforms and the Small-Apartment Segment The first chapter estimates the effect of Airbnb growth on long-term rents in Cologne, one of Germany's tightest rental markets in a country where more than half of all households rent. The empirical challenge is that short-term rental activity concentrates in neighborhoods that are attractive to tenants and tourists alike, so simple comparisons confound the platform's effect with neighborhood demand. The chapter addresses this with a triple difference-in-differences design built on roughly 90,000 individual rental advertisements and property-level short-term rental data. Identification exploits three margins that jointly characterize professional Airbnb activity: its spatial concentration in central neighborhoods, its concentration in apartments of three rooms or fewer, and its sharp growth after mid-2015. Large apartments in the same neighborhoods absorb neighborhood-wide demand and amenity shocks. Small apartments elsewhere absorb citywide trends in the segment. What remains is the effect on the housing stock Airbnb directly competes for.

Airbnb growth raised rents on small apartments in high-density neighborhoods by 5.9 percent, about 36 euros per month. Rents in this segment rose by 24 percent from the pre- to the post-treatment period, so the Airbnb effect accounts for roughly one quarter of that growth. The effect rises monotonically with exposure, appears as a sharp break in the event study, and survives a design-based inference battery: randomization inference in both geography and time, wild cluster bootstrap, spatial standard errors, and pre-trend sensitivity bounds. The mechanism is professional operating intensity, that is, hosts who withdraw units from long-term supply or outbid ordinary households for them. This is precisely the lever on which platform-targeted instruments such as night caps and registration requirements act. Because Cologne's housing protection ordinance was unenforced and its rent brake legally void during the sample period, the estimate is a clean benchmark for what platform growth does to rents when policy is passive, and a yardstick against which the enforcement regime introduced in Cologne in 2021–2022 can be evaluated.

Debunking Fake News: Fact-Checking versus Media Literacy The second chapter, joint with Lara Marie Berger, Anna Kerkhof, and Johannes Münster, turns from prices to beliefs. Policymakers fund two remedies against online misinformation, professional fact-checking and media literacy campaigns, yet no study had compared their effects on consumers' beliefs within one design. We conduct a pre-registered randomized survey experiment with roughly 3,000 German social media users, exposing participants to real Facebook posts containing true and false claims about Covid-19 vaccines and nutrition, and re-surveying them about two weeks later. One treatment group sees professional fact-checks. Another receives a brief media literacy intervention, Facebook's ten "tips to spot false news."

The two remedies work differently. Fact-checking reduces the perceived credibility of exactly the false posts it corrects (by 7–8 percentage points) but leaves untreated misinformation essentially untouched. The media literacy intervention reduces the credibility of false posts across topics, including topics no fact-check addressed, by 6–10 percentage points, improves incentivized factual knowledge, raises stated willingness to vaccinate, and its effects persist two weeks after a one-time exposure. Neither intervention makes participants distrust true information. Both improve discernment rather than breed skepticism. A pre-registered follow-up experiment with 4,000 new participants and environmental misinformation confirms the gap, with media literacy outperforming fact-checking by 10–14 percentage points. Mechanism evidence points to capability. Media literacy makes participants slow down, search for information, and identify untrustworthy elements in postings. Participants who receive fact-checks instead rely on the verdict and spend less time with the posts themselves. Both interventions work best for supporters of the far-right AfD, the group most likely to find the false posts credible, who also report less certainty about their prior knowledge. This suggests that outreach to seemingly entrenched audiences pays. Since only a small fraction of false content can ever be fact-checked, the results make a case for cheap, scalable media literacy interventions as the front line against misinformation.

Second Opinions in Elective Surgery The third chapter, joint with Yero Samuel Ndiaye and Daniel Wiesen, studies the starkest information problem of the three: the physician as expert in a credence goods market. German institutions long rewarded surgical volume, and comparable patients underwent disc surgery up to thirteen times more often across municipalities. We analyze a structured second-opinion program for elective surgery run by a large statutory health insurer, using administrative claims data from 2013 to 2019. The program's defining feature is institutional. The second physician is an experienced surgeon in the relevant field, but she does not perform the surgery even if she advises it. This separation of diagnosis from treatment decouples her diagnostic payoff from the surgical margin and aligns her incentives toward diagnostic accuracy rather than

surgical volume.

In orthopedics and neurosurgery, first and second physicians disagree in 56 percent of cases, and the disagreement is asymmetric in the direction physician-induced demand predicts: second physicians contradict 63 percent of recommendations *for* surgery but only 14 percent of recommendations *against* it. Seventy percent of patients comply with the second opinion. Using propensity score matching combined with patient fixed-effects event studies, and an instrumental variable strategy as a check, we estimate that compliance reduces monthly health expenditures by roughly 430–480 euros over eighteen months, about the size of baseline monthly spending, with fewer outpatient visits and hospital days and no adverse effects on sick leave. The savings concentrate almost entirely among patients advised *against* surgery. For them, avoided surgeries remain avoided. Symptoms are managed with conservative therapy, painkiller use does not rise, and no catch-up surgery or delayed deterioration appears within the observation window. The program thus converts a theoretical prediction, that separating diagnosis from treatment curbs overtreatment, into realized health and spending outcomes. With take-up at fewer than three requests per 100,000 insured per year, substantial welfare gains from scaling the program remain unrealized.

Common Threads Three lessons cut across the chapters without collapsing them into one framework. First, the harm is concentrated, not diffuse. Airbnb's rent effect loads onto small apartments in high-exposure neighborhoods rather than spreading evenly across the city. Misinformation's damage loads onto the claims no fact-check reaches and onto the audiences most likely to believe them. Unnecessary treatment loads onto recommendations for surgery, not against it. Averages computed over whole markets understate what the affected group experiences, and policy that targets the concentration buys the most welfare per euro, whether through night caps on professional operating intensity, interventions aimed at susceptible audiences, or second-opinion programs for high-disagreement specialties.

Second, the effective remedy is the one matched to the depth of the underlying problem. Where consumers can learn to judge for themselves, enabling them is enough. A brief media literacy intervention builds a skill that generalizes to claims nobody has checked yet, while a fact-check corrects one false claim and leaves the next one untouched. Where consumers cannot judge for themselves, as in elective surgery, protection must come from the supply side, and there the design details carry the weight. A second-opinion program that merely routed patients to another treating surgeon would reproduce the original incentive problem, so removing the second physician's financial stake in the treatment is what makes the correction credible. In housing, likewise, the instrument that acts on the conversion incentive dominates instruments that merely ban unit types and invite substitution.

Third, the chapters share a methodological commitment. The research design has to match the question, and identification has to be transparent. Where policy variation is absent, Chapter 1 builds a quasi-experimental design whose assumptions are probed with randomization inference and sensitivity bounds rather than asserted. Chapter 2 randomizes because beliefs are the outcome and selection into exposure is fatal. Where randomization is infeasible but administrative data are rich, Chapter 3 combines matching on treatment histories with within-patient event studies and an instrument. Across designs, the chapters state their results in concrete units, such as euros per month, percentage points of credibility, or avoided operations, so that they can be weighed directly in the policy decisions they inform.

Structure of the Dissertation

Each of the following chapters is a self-contained research paper with its own introduction, conclusion, and list of references. Each chapter reproduces the text of the underlying research paper verbatim and in full. Each chapter is preceded by the abstract of the underlying paper. All appendix material, including additional tables, figures, robustness checks, and experimental instructions, is gathered in a joint appendix at the end of the dissertation, organized by chapter.

Chapter 1 is single-authored. Chapter 2, "Debunking 'fake news' on social media: Immediate and short-term effects of fact-checking and media literacy interventions," is joint work with Lara Marie Berger, Anna Kerkhof, and Johannes Münster. All authors contributed equally to the development of the research idea and the design of the experiments. Lara and I implemented the experiments and conducted most of the data analysis. Anna prepared most of the initial draft, while Lara and I were primarily responsible for preparing the revised and published version. Chapter 3 is joint work with Yero Samuel Ndiaye and Daniel Wiesen. Daniel initiated the project, and he and I prepared the initial application to obtain the administrative claims data. All authors contributed equally to the development of the research idea. Yero and I prepared and analyzed the data, developed the empirical strategy, and wrote most of the initial draft. Daniel embedded the paper in the healthcare operations literature and edited and revised the manuscript. All co-authors have consented to the use of the joint work in this dissertation.

1 Short-Term Rental Platforms and the Small-Apartment Segment: Rental-Price Effects in Cologne

Abstract

I estimate that Airbnb growth raised rents on small apartments in high-density neighborhoods of Cologne by 5.9% (€36 per month, €430 per year), roughly one quarter of the rise in the segment's average rent between the pre- and post-treatment periods. The triple difference-in-differences design combines approximately 90,000 individual rental advertisements with property-level short-term rental data and exploits three sources of variation that jointly characterize professional Airbnb activity: spatial concentration in central neighborhoods, size concentration in apartments of three rooms or fewer, and sharp growth after mid-2015. The design isolates the segment-specific effect from neighborhood-wide demand and amenity channels. The effect rises monotonically with exposure intensity, emerges as a sharp break in the event study, and lies beyond its randomization-inference distribution in both geography and time; its significance is robust to few treated clusters and spatial dependence. The effect operates through professional short-term rental activity competing with long-term tenancy for the same housing stock, the lever on which platform-targeted regulation directly acts.

1.1 Introduction

Housing affordability has become a defining concern of urban policy across Europe and North America. Rents in major cities have outpaced incomes over the past decade, and the question of what drives this divergence has moved to the center of public debate. One contested factor is the rise of digital short-term rental platforms. Airbnb, the dominant platform globally, grew from roughly 50,000 listings in 2011 to over six million worldwide by 2019. In Germany alone, approximately 130,000 units were offered on the platform in January 2020. Where professional hosts concentrate their listings, the concern is that

This chapter is single-authored. Thank you Lisa, Frida and Jule for your support and inspiration. I am grateful for valuable comments and suggestions from Anna Kerkhof, Johannes Münster, Yero Ndiaye, and Tom Zimmermann.

housing units are withdrawn from the long-term rental market and redirected to higher-yielding short-term use, with consequences for the rents paid by long-term tenants.

The economic rationale for such an effect is straightforward. When the expected revenue from short-term letting exceeds the net return from a long-term contract, profit-maximizing hosts shift units into professional short-term use. Whether this translates into measurable rent increases for long-term tenants depends on the local supply elasticity of housing, the ease with which units can be shifted between short-term and long-term use, and the intensity of platform penetration. Answering this question empirically has proven difficult. Short-term rental activity concentrates in neighborhoods that are already attractive to tenants, landlords, and tourists alike, confounding a direct comparison of rents in high-exposure and low-exposure areas.

An empirical literature on the United States and Southern Europe has made progress on this identification problem using a range of strategies. Barron et al. (2020) estimate a national elasticity of 0.018% per 1% increase in Airbnb listings across U.S. ZIP codes using a shift-share instrument. Horn and Merante (2017) document a 0.4% average rent increase per standard deviation of Airbnb density in Boston, rising to 3.1% in popular tourist destinations. Garcia-López et al. (2020) find that Airbnb raised rents by 1.9% for the average Barcelona neighborhood and by up to 7% in high-activity areas. Franco and Santos (2021) document comparable effects in Lisbon and Porto. Koster et al. (2021) exploit staggered Home Sharing Ordinances in Los Angeles County to estimate rent effects of approximately 2%. Evidence for Germany is more limited. Duso et al. (2024) exploit two sequential regulatory interventions in Berlin to estimate that each additional commercial Airbnb listing increases rents by 1.3–2.4%. These studies establish that short-term rental growth raises rents in high-exposure markets. They estimate the effect averaged over neighborhoods or in cross-section, without isolating the housing segment Airbnb most directly competes with, namely small apartments where Airbnb activity concentrates.

This paper provides causal evidence on the rental-price effect of Airbnb growth using a triple difference-in-differences design for the city of Cologne over 2010 Q1–2018 Q4. Professional short-term rental activity in Cologne exhibits three sources of variation that the design exploits jointly: it concentrates in specific central neighborhoods (geography), targets small apartments of three rooms or fewer (size), and grew sharply after mid-2015 when the platform crossed meaningful scale (time). The triple DiD identifies the segment-specific rent change for small apartments in high-density neighborhoods, netting out both differential geographic trends that would affect all apartment sizes uniformly and differential size trends that would affect small apartments across all neighborhoods. The identifying assumption is weaker than the standard parallel-trends requirement: it fails if an unobserved shock differentially affected small apartments specifically in high-density neighborhoods at the same time as Airbnb growth.

The Cologne setting is particularly well-suited to this design. Platform-based short-term rental activity crossed meaningful scale only in mid-2015, producing a sharp treatment onset. Airbnb dominates the Cologne short-term rental market with a 95% share, making the platform a reliable proxy for the sector as a whole. The housing protection

ordinance in place since 2014 was not enforceable against short-term rental activity during the sample period, and the 2015 rent brake, later ruled legally void from its enactment, carried no enforceable sanction on advertised new-contract rents. The data are micro-level rather than aggregate: approximately 90,000 individual rental advertisements with a rich set of apartment characteristics, combined with property-level short-term rental listings and a neighborhood-level panel of housing stock and demographics.

I find that Airbnb growth increased rents on small apartments in high-density neighborhoods by 5.9%. The effect corresponds to €36 per month or €430 per year at the mean rent of affected units, equivalent to one quarter of the rise in the subgroup's average rent between the pre- and post-treatment periods.

The result is stable across specification choices, lies beyond its randomization-inference distribution in both time and geography, and is unaffected by controls for time-varying neighborhood demographics. The effect rises monotonically with exposure intensity: splitting the treated tercile by Airbnb density yields a gradient from 0.045 to 0.067, a pattern difficult to reconcile with a confounder. An event study shows a clear post-treatment break, and a pre-period trend test fails to reject flatness. Because such a non-rejection is weak evidence on its own, the causal case does not rest on it: it rests on the randomization and dose-response evidence, with sensitivity bounds making explicit how large a pre-trend violation would be required to overturn the result (Rambachan and Roth, 2023). Significance is robust to few treated clusters and spatial dependence. The pattern isolates a segment-specific Airbnb effect, distinct from neighborhood-wide demand and amenity channels.

The magnitude sits in the upper range of existing high-exposure estimates in the international literature, consistent with the tight focus on the directly competing market segment. This paper makes three contributions:

1. It provides the first estimate of the segment-specific rental-price effect of short-term rental platforms for a German city. The triple difference-in-differences design uses platform-growth variation rather than policy-induced contraction, and identifies the average treatment effect on small apartments in high-density neighborhoods over a 14-quarter post-treatment window. The causal claim rests on a design-based credibility suite: randomization inference on both assignment margins, inference robust to few treated clusters and spatial dependence, sensitivity bounds that quantify how large a pre-trend violation would have to be to overturn the result, and robustness to operational choices.

2. It documents that the rent effect is concentrated in a specific segment of the housing stock, small apartments in high-density neighborhoods, distinct from neighborhood-wide demand and amenity channels absorbed by the design. The distinction is policy-relevant: registration requirements and night caps act directly on professional short-term operating intensity, the lever through which the segment-specific effect operates, while neighborhood-wide channels lie outside the reach of any platform-targeted instrument.

3. It exploits a setting that isolates the Airbnb effect from the regulatory confounders prevalent in comparable studies. The Cologne housing protection ordinance existed throughout the sample but was not enforced against short-term rental activity, and no rent regulation bound meaningfully on advertised new-contract rents. The estimated effect is the counterfactual against which regulatory interventions on short-term rental platforms can be evaluated.

Section 1.2 describes the German rental market and the regulatory environment for short-term rental platforms. Section 1.3 presents the data and the spatial and size concentration of Airbnb activity that motivates the design. Section 1.4 develops the triple DiD specification and its identifying assumption. Section 1.5 presents the main estimate, the event study, the dose-response, and the magnitude. Section 1.6 tests the identifying assumption and the inference: pre-trend sensitivity bounds, randomization inference, few-cluster and spatial corrections, contemporaneous confounders, and specification robustness. Section 1.7 discusses mechanism, policy, and limitations. Section 1.8 concludes.

1.2 Institutional Background

Germany is a nation of renters. Approximately 53% of households rent their primary residence, the second-highest tenant share in the OECD after Switzerland. In large cities the share is higher still, roughly 85% in Berlin and over 75% in Cologne.[0] Urban tenant shares in major cities outside Germany are comparable: 76% in Manhattan, 62% in San Francisco, and 72% in the London borough of Westminster.[1] Any shock that diverts housing units from the long-term rental market to short-term tourist use therefore reaches a large share of urban residents in any of these settings.

Existing tenancies are regulated through the *Mietspiegel* (reference-rent index). Municipalities compile the index from rents in new contracts and adjustments over the preceding four to six years, stratified by apartment size, construction period, location, and amenities. Landlords may raise rents for sitting tenants only up to the reference rent, and by at most 15–20% over any three-year period.[2] New-contract rents, by contrast, were unregulated until the rent brake of June 2015, which caps them at 10% above the local reference rent. The Cologne implementation of the brake, however, was later ruled invalid from its enactment, and recovering excess rent required a qualified objection by the tenant, so the cap carried no effective sanction during the sample period.[3]

[0] Eurostat EU-SILC 2022.

[1] U.S. Census Bureau ACS 2020; ONS Census 2021 (tenure by local authority).

[2] §558 BGB. The cap is 20% by default. State governments may lower it to 15% in designated tight markets.

[3] §556d BGB, introduced by the *Mietrechtsnovellierungsgesetz* of April 2015. Cologne was included in the NRW *Mietpreisbegrenzungsverordnung* effective July 1, 2015. The Amtsgericht Köln (2019) held, and the Landgericht Köln (2020) confirmed, that the Cologne designation lacked the statutorily required justification of tight-market status and was therefore invalid from its enactment (*ex tunc*). The defect could not be cured retroactively, and the NRW government re-enacted a corrected ordinance only in July 2020. Newly built units, comprehensively modernized units, and grandfathered prior rents were

Neither the reference-rent index nor the rent brake covers short-term rentals. A unit converted from long-term to short-term use exits the regulated market entirely. The rental data used here, advertised asking rents on new contracts, are observed in the least regulated segment, where supply-side shocks from platform-based short-term rentals pass through to prices most directly.

Airbnb launched in Germany around 2011 and grew rapidly over the following decade. In Cologne, the first listings appeared in 2011, but the platform reached meaningful scale only in mid-2015, when cumulative active listings exceeded 500. Growth accelerated sharply thereafter: by end-2019 nearly 6,200 units were listed (Mindl and Arentz, 2020). Airbnb dominates the Cologne short-term rental market with a 95% platform share. HomeAway (3%) and Booking.com (2%) are marginal. Section 1.3 documents the spatial, host, and size composition of this activity in detail.

Cologne's regulatory environment during the sample period was permissive. On July 1, 2014 the city adopted a *Wohnraumschutzsatzung* (housing protection ordinance) under the North Rhine-Westphalia *Wohnungsaufsichtsgesetz* (WAG NRW). The ordinance was a general instrument to protect residential stock in a tightening market: it placed all residential-to-nonresidential conversion (extended vacancy, commercial or professional use, demolition, *and* short-term tourist letting) under a blanket permit requirement.[4] Enforcement against short-term operators was negligible: the WAG NRW required no platform registration and no disclosure of host identities, and anonymous profiles made systematic detection nearly impossible. The city assigned three staff positions to enforce the entire ordinance across all categories of misuse.[5]

The housing protection ordinance neither targeted nor meaningfully constrained short-term rental activity, and the rent brake applied to new contracts only from July 2015 with weak enforcement and void legal status. Cologne thus offers a rental market in which platform-based short-term rental growth can be observed without substantial regulatory confounding.

1.3 Data and Descriptive Analysis

1.3.1 Data

I combine three data sources: individual rental advertisements, short-term rental platform data, and administrative records on housing stock and demographics.

in any case exempt. Studies of asking-rent responses to the rent brake document limited compliance and modest, transitory effects (Mense et al., 2023; Breidenbach et al., 2022). Section 1.6.4 examines the rent brake as a potential confounder of the treatment effect.

[4] §4 Wohnraumschutzsatzung der Stadt Köln, July 4, 2014. Short-term platforms were not yet a significant policy concern in Cologne at enactment.

[5] A revised ordinance took effect on July 1, 2019, and the subsequent *Wohnraumstärkungsgesetz* (WohnStG NRW, effective July 1, 2021) introduced a platform registration requirement and raised the maximum fine substantially. The 2021 framework also established a 90-day annual cap on permit-free short-term letting, consistent with Berlin, London, San Francisco, and (since 2025) Paris, a threshold that motivates the 110-night cutoff used to define professional hosting in Section 1.3.2. Both postdate the sample period. See Mindl and Arentz (2020).

Rental listings. I obtain the rental data from the City of Cologne via FUB (Forschungsdatenzentrum für Bildungs-, Wissenschafts- und Kulturstatistik, operated by IGES), which aggregates listing advertisements across major German rental platforms.[6] Each observation is a unique advertisement with the full set of characteristics visible to prospective tenants: cold rent per square meter, living area, number of rooms, year-built class, and amenity indicators (elevator, fitted kitchen, balcony, parking, and others). Advertisements are geocoded to one of Cologne's 86 *Stadtteile* (neighborhoods), the spatial unit throughout.

The raw extract covers 92,918 listings between 2010 Q1 and 2018 Q4. I remove exact duplicates and listings with missing or invalid neighborhood codes, and drop advertisements that remained online for more than four quarters, whose extended visibility signals commercial or misclassified listings rather than genuine offers.[7] Where rooms are missing but floor area is observed, rooms are imputed from the area-to-rooms mapping in the non-missing sample. After these restrictions, 89,878 listings remain across all 86 neighborhoods. The dependent variable throughout is the log of cold rent per square meter at the last observed listing date.

Short-term rental listings. Short-term rental data come from AirDNA, a commercial provider tracking Airbnb and HomeAway (now Vrbo). The property-level cross-section (scraped June 16, 2020) contains 25,372 listings with coordinates, room type, creation date, calendar activity, reviews, host portfolio size, and average daily rate. A companion monthly spell file records occupancy and revenue. Standard cleaning removes listings without an identified host, without signs of activity, or of non-residential type (hotels, hostels, bed-and-breakfasts), and spatially joins each listing to its neighborhood. The cleaned cross-section contains 15,987 listings across 86 neighborhoods, operated by 12,556 hosts. A quarterly spell panel expands each listing's activity window from its creation date to its last calendar update, trimmed to 2010 Q1–2018 Q4 to match the rental data.

Housing stock and demographics. Housing stock data from the Statistical Office of the City of Cologne report residential units per neighborhood and year, 2010–2018. They enter as the denominator of the density measure introduced in Section 1.3.3. Neighborhood-level demographic controls (population, unemployment, and completed new construction) and city-level tourist overnight stays serve as time-varying controls in robustness specifications. The main specification relies on neighborhood and quarter fixed effects rather than these controls.

[6] See `https://fub.iges.com/`. FUB aggregates advertisements from ImmobilienScout24, Immonet, Immowelt, and several regional portals, providing broader coverage than single-platform sources.

[7] The four-quarter threshold is conservative. Relaxing it to seven quarters retains roughly 3,000 additional observations without affecting the results.

1.3.2 Professionalism

Not all Airbnb listings compete with the long-term market. A host renting a spare room a few weekends a year does not meaningfully reduce tenant housing. A unit let to tourists more than half the year is, in practice, off the long-term market. The relevant margin is professional hosting with listings operated at a scale that would otherwise serve as permanent rental housing.

Every professional listing removes one unit of long-term supply. This supply effect is universal. A subset additionally exerts demand-side pressure: operators, here called entrepreneurs, who rent or buy long-term units to sublet on Airbnb. Because expected short-term revenue exceeds the going rent, their bid exceeds an ordinary household's, raising the equilibrium price on comparable units. The entrepreneurial channel layers on top of the supply channel and operates only on the subset who entered the long-term market to acquire housing units for short-term operation.

A listing is professional if it is booked for more than 110 nights per year or if its host operates multiple simultaneous listings. HomeAway listings (slightly more than 3% of the raw sample) lack a host identifier and are excluded. Professional activity is concentrated. Table 1.1 summarizes listings across hosts active in 2018 Q4, when 5,441 listings are run by 4,377 hosts. The largest host operates 132 listings. The top 5% of hosts control 18% of active listings and the top 10% control 26%. Multi-listing hosts account for 29% of the active stock, while 88% of hosts operate a single listing. The multi-listing leg of the definition captures this tail. Operators of dozens of simultaneous listings are not casual home-sharers, and their inventory is committed to professional short-term use.

Table 1.1: Host concentration among active Airbnb listings (Cologne, 2018 Q4)

	Value
Active listings	5,441
Active hosts	4,377
Mean listings per host	1.24
Max listings per host	132
Share of hosts with multiple listings	11.8%
Share of listings held by multi-listing hosts	29.0%
Share of listings held by top 1% of hosts	9.5%
Share of listings held by top 5% of hosts	18.1%
Share of listings held by top 10% of hosts	26.2%

Notes: HomeAway listings lack a host identifier and are excluded.

The nights leg is not directly observed and is inferred from review counts. Following the standard approach (Garcia-López et al., 2020; Koster et al., 2021; Duso et al., 2024),

reviews per year are converted to booked nights using a 60% review rate and a 2.3-night average stay:[8]

$$\widehat{\text{Nights}}_i = \min\left(\frac{\text{Reviews}_i/\text{Years active}_i}{0.6} \times 2.3,\ \ 347\right) \tag{1.1}$$

where the cap at 347 corresponds to 95% annual occupancy. Table 1.2 summarizes the market and compares short-term revenue to long-term income.[9] Of the 5,441 active listings, 2,325 (43%) are professional. Entire-home three-room listings generate roughly € 23,100 in annual revenue against € 15,100 in annual rental cost, a break-even at 111 nights. One-room apartments break even at 97 nights and private rooms at 109. The 110-night threshold thus sits near the break-even point for most apartment types. The sensitivity check in Section 1.6 varies the cutoff from 90 to 140 nights and confirms the result is insensitive to this choice.

Table 1.2: Airbnb market structure and break-even with long-term rental income (2018 Q4)

			Airbnb			Rental comparison		
	N	Share	Price/night (€)	Nights	Revenue (€)	Area (sqm)	Annual rental cost (€)	Break-even
Active listings	5,441							
Professional	2,325	42.7%						
Entire home: 1 room	266	11.4%	78	209	16,382	36	7,609	97 nights
Entire home: 2 rooms	751	32.3%	85	191	16,166	62	11,061	131 nights
Entire home: 3 rooms	289	12.4%	136	169	23,112	89	15,137	111 nights
Private room	884	38.0%	52	190	9,828	31	5,625	109 nights

Notes: Sample: listings active in 2018 Q4. Professional: >110 booked nights/year or multiple listings. Avg. nights and revenue computed for single-listing professional hosts only. Revenue = price/night × nights. Rental cost = (avg. cold rent + operating costs) × avg. area × 12, restricted to 2018 listings. Operating costs: €2.96/sqm/month (Deutscher Mieterbund, Betriebskostenspiegel Cologne 2017). Private room rental cost imputed as one room's share of housing (≈ 30 sqm incl. communal space) at city-wide avg. rent; overstates opportunity cost, producing an upper-bound break-even. Break-even = annual rental cost / avg. Airbnb price per night.

[8] The 60% review rate is a widely used benchmark (Fradkin, 2015; Budget and Legislative Analyst's Office, 2015). The 2.3-night average stay follows Mindl and Arentz (2020) and sits between the commercial-lodging figure of 1.7 nights (Statistisches Jahrbuch Köln, Tabelle 517) and Airbnb's self-reported German average of about four nights. The threshold sensitivity check in Section 1.6.5 varies the professional cutoff in a way mechanically equivalent to varying the stay parameter from 1.7 to 2.8 nights. The result is stable across this range.

[9] Average nights and revenue are computed for single-listing professional hosts only. For multi-listing hosts the review-based proxy understates occupancy.

1.3.3 Short-term rental exposure: geography, size, and growth

To measure each neighborhood's exposure, I construct a stock density measure: professional Airbnb listings per 100 residential housing units:

$$\text{Density}_{jt} = \frac{\text{Professional Airbnb listings}_{jt}}{\text{Housing units}_{j,\text{year}(t)}} \times 100. \tag{1.2}$$

Neighborhoods are assigned to terciles on this density at the end of the sample (2018 Q4): T1 (low, 29 neighborhoods), T2 (middle, 29), and T3 (high, 28).

Figure 1.1 maps the terciles. T3 forms a contiguous cluster around the historic center (the Altstadt, Neustadt, and adjacent inner-city districts on both banks of the Rhine), the areas with the highest tourist traffic, the densest hospitality infrastructure, and the best transit access to landmarks. T1 neighborhoods lie in the suburban periphery. The pattern reflects the locational fundamentals of short-term rental demand.

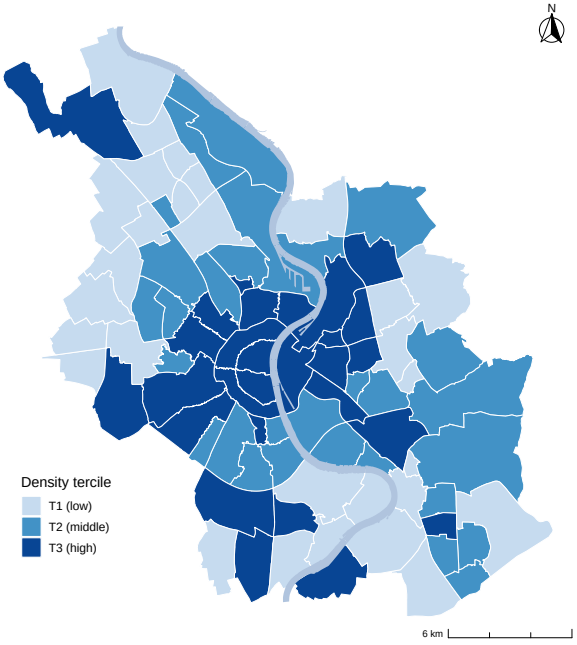


Figure 1.1: Airbnb density terciles across Cologne neighborhoods

Notes: Terciles based on end-of-sample density (professional Airbnb listings per 100 housing units, 2018 Q4). T1: 29 neighborhoods; T2: 29; T3: 28. The Rhine is shown in grey. Source: AirDNA, City of Cologne boundary data.

Figure 1.2 traces mean density by tercile. Airbnb activity was negligible before 2013 and grew rapidly but unevenly: mean T3 density rose from near zero to about 0.65 professional listings per 100 housing units by 2018 Q4, while T1 stayed near zero throughout. This geographic and temporal divergence is the variation the analysis exploits.

The third dimension is size. Cologne's Airbnb listings are overwhelmingly small: about 96% have three rooms or fewer,[10] against roughly 87% of rental listings. Equivalently,

[10]Airbnb room counts are approximated as bedrooms plus one, following the German convention that

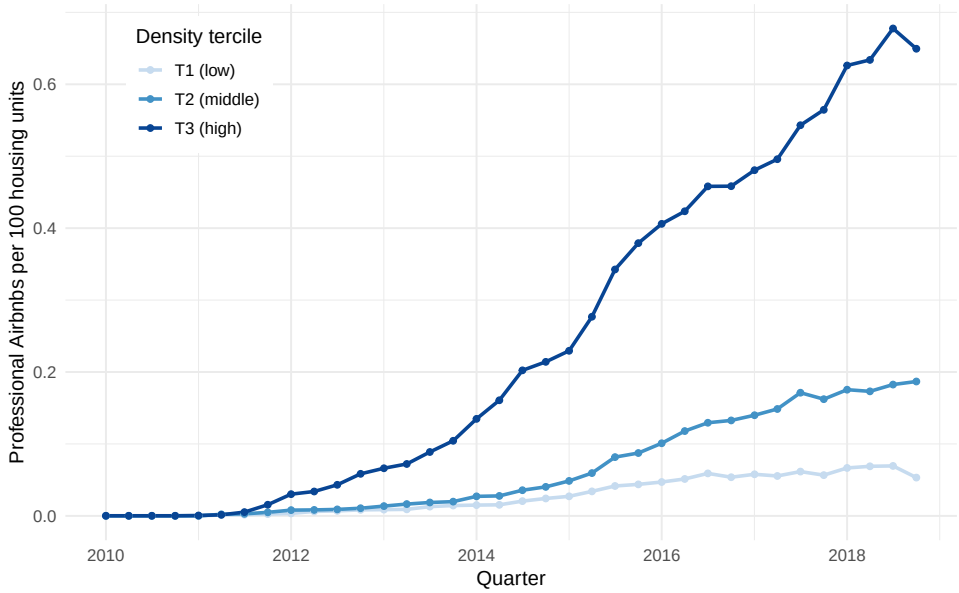


Figure 1.2: Mean Airbnb density by neighborhood tercile, 2010 Q1–2018 Q4

Notes: Sample: 2010 Q1–2018 Q4, all Cologne neighborhoods. Density = professional Airbnb listings per 100 housing units. Terciles assigned at end of sample (2018 Q4). Source: AirDNA, City of Cologne housing stock statistics.

the large-apartment (four or more rooms) share in the rental market is more than three times the Airbnb share (13% versus 4%), and the mismatch is sharpest at the bottom, studios and one-bedrooms. Within high-density neighborhoods, small apartments therefore face direct competition from professional Airbnb while large apartments do not. The observed distribution motivates a three-room cutoff that distinguishes apartments suitable for Airbnb use from those that are not. The cutoff is varied in a robustness check (Section 1.6).

The size concentration refines the density measure. Table 1.3 summarizes density within T3 under three denominators. Stock density (Equation 1.2) relates listings to the total stock: 0.42 per 100 units at the median T3 neighborhood, 1.46 at the 90th percentile, and 1.75 at the maximum (Altstadt-Nord). Because Airbnb concentrates in small apartments, two sharper measures help. The first scales listings by the small-apartment share of the surrounding district's stock,[11] used in the magnitude discussion (Section 1.7.1). The second expresses listings as a share of small-apartment rental ads in the same neighborhood and quarter. With three professional Airbnb listings per small-apartment ad at the median T3 neighborhood and nearly five at the 75th percentile, it captures competitive pressure on the directly exposed segment most sharply.

The main specification uses stock density for assignment, as it is less sensitive to short-run fluctuations in rental-ad volume. Defining terciles on the small-apartment-adjusted density reclassifies four of 86 neighborhoods at the boundaries, with no consequence for the result (Section 1.6). The geographic concentration motivates restricting estimation

counts the living room as a room.

[11] Room-count breakdowns are published at district (*Stadtbezirk*) level only; the district share is applied to all neighborhoods within a district.

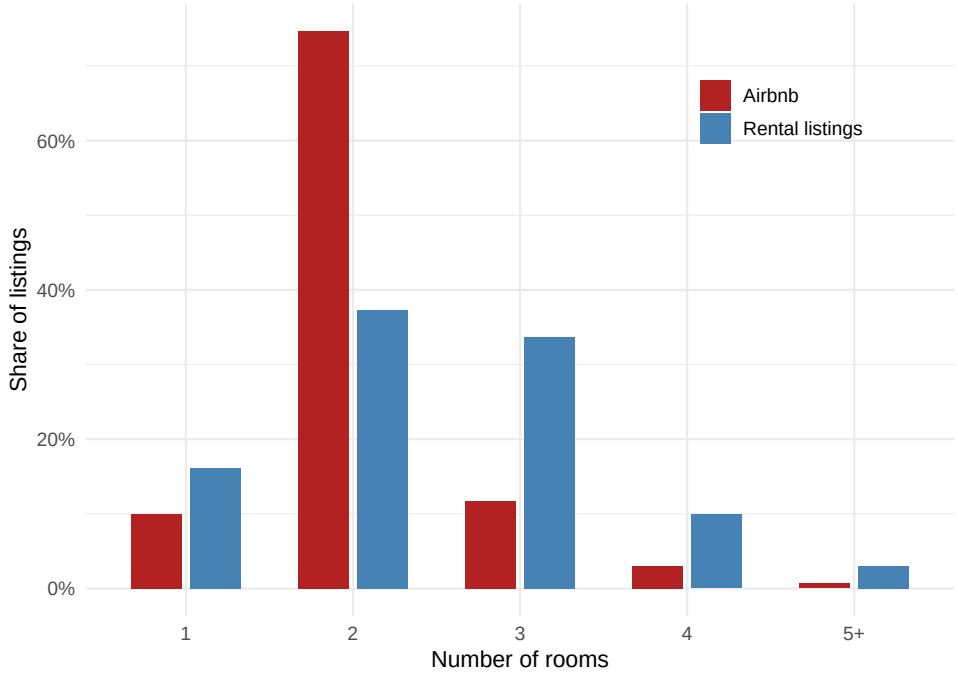


Figure 1.3: Room-size distribution: Airbnb listings vs. rental advertisements

Notes: Airbnb room count = bedrooms + 1. Airbnb listings: 2018 Q4 cross-section. Rental listings: full sample 2010 Q1–2018 Q4. The five-room category includes all listings with five or more rooms. Source: AirDNA, FUB/IGES via City of Cologne.

Table 1.3: Density distribution across T3 neighborhoods (2018 Q4)

	N	P10	P25	P50	P75	P90	Max
Stock density	28	0.31	0.35	0.42	0.84	1.46	1.75
Adjusted stock density	28	0.61	0.68	0.81	1.39	2.33	3.67
Flow density	27	110.6	169.0	300.0	482.7	618.0	785.7

Notes: Stock density = professional Airbnb listings per 100 housing units. Adjusted stock density = professional Airbnb listings per 100 housing units with three rooms or fewer; the small-apartment share of the housing stock is applied at the district level (range 0.43–0.66 across the nine districts; source: Statistisches Jahrbuch Köln 2018, Tabelle 304). Flow density = professional Airbnb listings per 100 small-apartment ($\leq$ 3 rooms) rental advertisements in the same neighborhood and quarter. One T3 neighborhood with zero small-apartment rental advertisements in 2018 Q4 is excluded from the flow measure.

to T1 and T3, dropping the middle tercile to sharpen the contrast: 67,719 observations across 57 neighborhoods.

1.3.4 Summary statistics

Table 1.4 reports summary statistics for the estimation sample: 67,719 advertisements across 57 neighborhoods over 2010 Q1–2018 Q4. Cold rent averages €9.97 per square meter, with a right-skewed distribution. The median is €9.50 and the interquartile range

runs from € 8.00 to € 11.30. Apartments average 2.4 rooms and 71 square meters, and building amenities vary widely across listings: 15% include an elevator, 42% a balcony or terrace, and 55% a basement.

1.4 Empirical Strategy

1.4.1 Hedonic model and estimand

I model rental prices with a semi-logarithmic hedonic specification. For listing i in neighborhood j and quarter t,

$$\ln\left(\frac{\text{rent}}{m^2}\right)_{ijt} = \alpha + \boldsymbol{X}'_{ijt}\boldsymbol{\rho} + \gamma_j + \lambda_t + \varepsilon_{ijt}, \tag{1.3}$$

where $\boldsymbol{X}_{ijt}$ is a vector of apartment characteristics, γ_j are neighborhood fixed effects, and λ_t are quarter fixed effects. Floor area is excluded from the control vector because it is itself a measure of apartment size and already enters the dependent variable as the per-square-meter denominator. Conditioning on it would attenuate the estimate rather than sharpen it. The semi-log form yields coefficients interpretable as approximate percentage changes and is standard in the short-term rental literature (e.g., Barron et al., 2020; Koster et al., 2021). Equation (1.3) is the baseline. The triple difference-in-differences below nests it by adding the treatment interactions.

The design exploits the three sources of variation documented in Section 1.3: professional Airbnb activity concentrates in specific neighborhoods (geography), targets small apartments of three rooms or fewer (size), and grows sharply after mid-2015 (time). I interact indicators for all three:

$$\begin{aligned}\ln\left(\frac{\text{rent}}{m^2}\right)_{ijt} = {} & \alpha + \beta_1\, \mathit{High}_j + \beta_2\, \mathit{Post}_t + \beta_3\, \mathit{Small}_i \\ & + \beta_4\, \mathit{High}_j \times \mathit{Post}_t + \beta_5\, \mathit{Small}_i \times \mathit{Post}_t + \beta_6\, \mathit{High}_j \times \mathit{Small}_i \\ & + \underbrace{\beta_7\, \mathit{High}_j \times \mathit{Small}_i \times \mathit{Post}_t}_{\text{Airbnb effect}} + \boldsymbol{X}'_{ijt}\boldsymbol{\rho} + \gamma_j + \lambda_t + \varepsilon_{ijt},\end{aligned} \tag{1.4}$$

where $\mathit{High}_j = 1$ if neighborhood j is in the top tercile of professional Airbnb density, $\mathit{Post}_t = 1$ after the treatment date, and $\mathit{Small}_i = 1$ if listing i has three rooms or fewer. The coefficient of interest is β_7, which is the differential rent change for small apartments in high-density neighborhoods, net of common time trends (λ_t), differential geographic trends (β_4), and differential size trends (β_5).

The neighborhood and quarter fixed effects absorb the High_j and Post_t main effects. The five remaining terms (Small_i, $\mathit{High}_j \times \mathit{Post}_t$, $\mathit{Small}_i \times \mathit{Post}_t$, $\mathit{High}_j \times \mathit{Small}_i$, and the triple interaction) are estimated explicitly. In particular, $\mathit{High}_j \times \mathit{Small}_i$ is not absorbed by the neighborhood fixed effects. Apartment size varies within a neighborhood, so the high-density size gap is a separately estimated coefficient.

Table 1.4: Summary statistics, estimation sample

	Mean	SD	P25	Median	P75
Cold rent (€/sqm)	9.97	3.03	8.00	9.50	11.30
Living area (sqm)	71.4	32.5	50.0	67.0	85.0
Rooms	2.42	0.97	2.00	2.00	3.00
Barrier-free	0.007	0.084			
Publicly subsidized	0.043	0.203			
Elevator	0.150	0.357			
Basement	0.546	0.498			
Central heating	0.526	0.499			
Gas heating	0.005	0.070			
Balcony/terrace	0.424	0.494			
Garden	0.175	0.380			
Garage	0.308	0.462			
Parking space	0.052	0.222			
Parquet floor	0.088	0.283			
First occupancy	0.100	0.300			
New build	0.042	0.200			
Built before 1918	0.043	0.203			
Built 1918–1945	0.032	0.175			
Built 1946–1960	0.098	0.298			
Built 1961–1975	0.174	0.380			
Built 1976–1989	0.067	0.249			
Built 1990–2005	0.116	0.321			
Built 2006–2009	0.034	0.182			
Built 2010–2015	0.089	0.285			
Construction year not reported	0.345	0.476			
Fitted kitchen	0.335	0.472			
Furnished	0.027	0.162			

Notes: Estimation sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4; $N = 67{,}719$ advertisements across 57 neighborhoods. Cold rent per square metre is reported in levels. Indicator variables report the share of listings. Year-of-construction rows report the share of listings in each construction-period class; the final row collects listings without a reported construction year.

1.4.2 Identifying assumption

The triple DiD requires that, absent Airbnb, the difference in rent trends between small and large apartments in high-density neighborhoods would have evolved like the same difference in low-density neighborhoods:

$$E\Big[\Delta Y_{jt}^{S} - \Delta Y_{jt}^{L} \;\Big|\; High_j = 1\Big] = E\Big[\Delta Y_{jt}^{S} - \Delta Y_{jt}^{L} \;\Big|\; High_j = 0\Big] \qquad \forall\, t \geq t^{*}, \tag{1.5}$$

where ΔY_{jt}^{S} and ΔY_{jt}^{L} denote rent changes for small and large apartments. This is weaker than the standard parallel-trends requirement in two ways: it nets out any differential geographic trend, and any differential size trend. Formally, the triple difference requires a single parallel-trends assumption on the small–large rent gap, which must evolve in parallel across high- and low-density neighborhoods, rather than separate assumptions for the geographic and the size comparison (Olden and Møen, 2022).

Each control group absorbs a distinct confound. Large apartments in high-density neighborhoods absorb any shock that moves rents across all sizes together, such as tourism-driven amenity capitalization, gentrification, or a local demand boom. Small apartments in low-density neighborhoods absorb any city-wide trend specific to the small-apartment segment. The triple difference removes both. Identification fails only if an unobserved shock changed rents on small apartments *specifically* in high-density neighborhoods, at the same time as Airbnb's growth, a shock that is simultaneously size-specific, density-specific, and timed to 2015. Section 1.6.4 examines these candidates. Each is absorbed by the design, ruled out by the institutional record, or bounded empirically.

1.4.3 Specification, fixed effects, and inference

Taking Equation (1.4) to the data requires settling the fixed-effects structure, the treatment date, the group assignment, and the level of inference. This section states each choice and its rationale. All four are revisited later, either as a reported variant or as a robustness check.

The preferred specification estimates Equation (1.4) with neighborhood and quarter fixed effects and the full set of lower-order interactions. Section 1.5 reports alongside it a variant that replaces the additive fixed effects and interactions with fully saturated neighborhood × quarter, neighborhood × size, and quarter × size fixed effects. The two estimates lie within one standard error of each other.

The treatment date is 2015 Q3, the quarter professional Airbnb crossed meaningful scale in Cologne (Section 1.2). The estimate is insensitive to shifting it by one quarter in either direction (Section 1.6). Tercile membership is fixed at a single date so that groups do not change over the sample. End-of-sample density (2018 Q4) is used because pre-period densities are too uniformly near zero to separate neighborhoods reliably.

Neighborhood fixed effects absorb time-invariant differences across neighborhoods such as building stock, centrality, and local amenities. Quarter fixed effects absorb trends common to the whole city such as macroeconomic conditions, mortgage rates, and seasonal

fluctuations. I cluster standard errors at the neighborhood level, the level at which treatment is assigned (Abadie et al., 2023). Twenty-eight of the 57 neighborhoods are treated, a moderate cluster count. Section 1.6 confirms significance with randomization inference and a wild cluster bootstrap, neither of which relies on cluster-count asymptotics.

1.4.4 Parallel trends and balance

The identifying assumption is about trends, and can be assessed over the pre-treatment period. Figure 1.4 plots the small–large rent gap, in logs, separately for high- and low-density neighborhoods. Before 2015 Q3 the two gaps move together, while afterward the high-density gap diverges upward. The differenced-trend test confirms that the differenced gap has a pre-period slope of -0.0015 log points per quarter ($p = 0.25$), and over the 2013–2015 window, $+0.006$ per quarter ($p = 0.27$), neither distinguishable from zero. The high-density small–large gap does trend upward in 2013–2015, but the low-density gap drifts with it. The third difference nets out the common movement.

These tests support the assumption but do not establish it: a pre-trend test that fails to reject is not proof of parallel trends (Roth, 2022). Section 1.6 therefore bounds how large a pre-trend violation would have to be to overturn the result, rather than asserting the trends are flat.

Table 1.5 reports pre-treatment normalized differences between high- and low-density neighborhoods, separately by size. Most characteristics are well balanced ($|ND| < 0.25$). The exceptions track the structural difference between inner-city and peripheral stock: small apartments in high-density areas have fewer garages and fewer rooms, large apartments less garden access, and the year-built composition tilts toward pre-war and early post-war stock in high-density areas and toward the 1961–1975 cohorts elsewhere. These are level differences, absorbed by the neighborhood fixed effects and the year-built controls.

1.5 Results

This section presents the main estimate (Section 1.5.1), traces its dynamics around the treatment date in an event study (Section 1.5.2), resolves it by treatment intensity (Section 1.5.3), and translates it into euro magnitudes (Section 1.5.4).

1.5.1 Main estimates

Table 1.6 reports the triple difference-in-differences estimates from Equation (1.4) across three control configurations. In the preferred specification with structural controls (column 2) the coefficient of interest is $\hat{\beta}_7 = 0.059$ ($p < 0.01$). The estimate indicates that small apartments in high-density neighborhoods experienced rents roughly 5.9% higher after 2015 Q3 than the counterfactual implied by large apartments in the same neighborhoods and small apartments in low-density neighborhoods.

Table 1.5: Pre-treatment balance: normalized differences

	Small (≤ 3 rooms)			Large (4+ rooms)		
Variable	T3	T1	ND	T3	T1	ND
Structural						
Rooms	2.083	2.363	-0.397	4.209	4.238	-0.072
Barrier-free	0.000	0.000	—	0.000	0.000	—
Publicly subsidized	0.033	0.078	-0.196	0.036	0.084	-0.205
Elevator	0.131	0.141	-0.029	0.157	0.136	0.061
Basement	0.544	0.569	-0.051	0.571	0.582	-0.023
Central heating	0.517	0.570	-0.107	0.512	0.557	-0.089
Gas heating	0.002	0.005	-0.039	0.004	0.011	-0.085
Balcony/terrace	0.367	0.437	-0.143	0.477	0.467	0.021
Garden	0.135	0.213	-0.206	0.221	0.337	-0.260
Garage	0.232	0.406	-0.380	0.400	0.507	-0.217
Parking space	0.035	0.066	-0.142	0.051	0.069	-0.072
Parquet floor	0.070	0.082	-0.046	0.137	0.096	0.131
First occupancy	0.090	0.066	0.088	0.179	0.123	0.156
New build	0.032	0.028	0.022	0.100	0.048	0.199
Year of construction (share by class)						
Built before 1918	0.051	0.013	0.215	0.095	0.022	0.313
Built 1918–1945	0.036	0.015	0.137	0.058	0.021	0.191
Built 1946–1960	0.123	0.035	0.331	0.091	0.035	0.234
Built 1961–1975	0.135	0.233	-0.257	0.095	0.254	-0.428
Built 1976–1989	0.052	0.103	-0.191	0.029	0.087	-0.249
Built 1990–2005	0.097	0.173	-0.222	0.086	0.138	-0.165
Built 2006–2009	0.028	0.041	-0.070	0.063	0.050	0.056
Built 2010–2015	0.059	0.046	0.057	0.192	0.109	0.234
Construction year not reported	0.419	0.341	0.161	0.291	0.285	0.014
Potentially endogenous (all controls)						
Fitted kitchen	0.355	0.290	0.138	0.314	0.259	0.121
Furnished	0.019	0.014	0.041	0.015	0.014	0.005

Notes: Treated (T3) and Control (T1) cell means by apartment size, for the pre-treatment period 2010 Q1–2015 Q2. Normalized differences computed following Imbens & Rubin (2015), who suggest $|ND| < 0.25$ as a balance threshold. Apartment size enters as a binary indicator (rooms ≤ 3 vs. 4+); area is the denominator of the dependent variable and is not balanced separately. Year-of-construction class enters the hedonic controls as a multi-level factor. Each class share, including the share with no reported construction year, is balanced as a binary indicator over all listings.

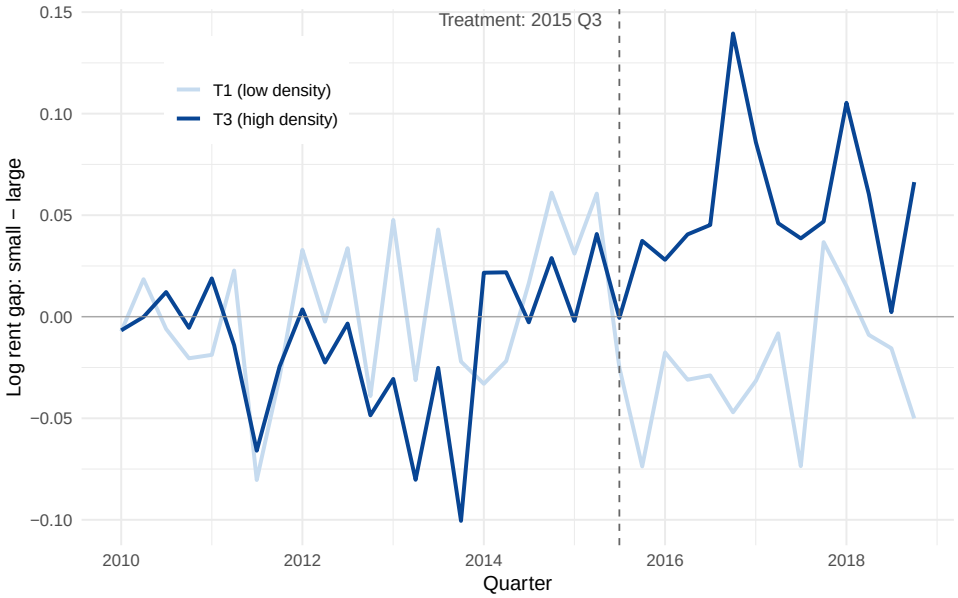


Figure 1.4: Within-tercile size gap in log rent, 2010 Q1–2018 Q4

Notes: Each line plots the difference in mean log cold rent per sqm between small ($\leq$ 3 rooms) and large (4+ rooms) apartments within a tercile. Dashed line: 2015 Q3 treatment. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. Source: FUB/IGES via City of Cologne.

The estimate moves with the control set in the expected direction. With no apartment controls (column 1) it is 0.088. Structural controls bring it to 0.059, and adding potentially endogenous amenities and demographics (column 3) yields 0.051. The step from column 1 to column 2 reflects composition, as structural controls adjust for the denser, older small-apartment stock of central neighborhoods. The further attenuation in column 3 is as expected. Amenities such as a fitted kitchen or furnished status can themselves respond to the same market pressure that drives short-term rental activity, a bad-controls problem in the sense of Angrist and Pischke (2009). The structural specification avoids the ambiguity by retaining only predetermined building characteristics. In the preferred specification the two-way interactions are individually insignificant, isolating the effect in the triple.

A variant that replaces the additive fixed effects and interactions with fully saturated neighborhood $\times$ quarter, neighborhood $\times$ size, and quarter $\times$ size fixed effects (column 4) yields 0.051, within one standard error of the preferred estimate. The additive specification with explicit interaction terms nonetheless remains the preferred one, because its variance matrix stays stable across the full robustness grid. With the full control vector, the cluster-robust variance matrix of the fully saturated variant is not positive semi-definite at 57 clusters, the same finite-cluster degeneracy that motivates the design-based inference in Section 1.6.[12]

[12]The saturated variant is therefore reported at the structural-controls tier.

Table 1.6: Effect of short-term rentals on small-apartment rents

Dependent Variable:	log(rent/sqm)			
	No controls	Structural	All controls	Saturated FE
Model:	(1)	(2)	(3)	(4)
Variables				
High × *Small* × *Post*	0.088***	0.059***	0.051***	0.051***
	(0.016)	(0.014)	(0.013)	(0.014)
High × *Post*	-0.054***	-0.019	-0.019	
	(0.015)	(0.013)	(0.013)	
Small × *Post*	-0.019	0.002	0.004	
	(0.012)	(0.012)	(0.011)	
High × *Small*	-0.032**	-0.011	-0.011	
	(0.015)	(0.012)	(0.012)	
Small	0.008	0.020***	0.016**	
	(0.008)	(0.007)	(0.007)	
Fixed-effects				
Neighborhood	Yes	Yes	Yes	
Quarter	Yes	Yes	Yes	
Neighborhood-Quarter				Yes
Neighborhood-Small				Yes
Quarter-Small				Yes
Fit statistics				
Observations	67,719	67,719	67,195	67,660
R^2	0.438	0.498	0.514	0.522
Within R^2	0.002	0.110	0.139	0.103

Notes: Triple difference-in-differences with fixed city-wide treatment date 2015 Q3. Standard errors clustered at the neighborhood level (57 clusters), reported in parentheses. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. Living area excluded from controls: it is itself a size measure, already enters the outcome as the per-square-meter denominator, and is correlated with the size indicator *Small* (≤ 3 rooms) along which the effect is identified (Section 4.1). Controls: (1) none; (2) structural (year-built class, barrier-free, publicly subsidized, elevator, basement, central heating, gas heating, balcony/terrace, garden, garage, parking space, parquet, first occupancy, new build) — predetermined apartment characteristics and amenities that cannot plausibly respond to Airbnb activity within the sample window; (3) structural plus potentially endogenous amenities (fitted kitchen, furnished) and neighborhood-level demographics (population, new construction). Column (4) replaces the additive fixed effects and manual lower-order interactions of column (2) with fully saturated neighborhood×quarter, neighborhood×size, and quarter×size fixed effects (structural controls); its estimate lies within one standard error of column (2). Significance: $^{***}\,p < 0.01$; $^{**}\,p < 0.05$; $^{*}\,p < 0.1$.

1.5.2 Event study

To trace the dynamics and examine pre-trends, I replace the post indicator with quarter-specific coefficients:

$$\ln\left(\frac{\text{rent}}{m^2}\right)_{ijt} = \alpha + \sum_{\substack{k=-K \\ k\neq -1}}^{K} \delta_k \, \mathbf{1}[t - t^* = k] \cdot \mathit{High}_j \cdot \mathit{Small}_i + \mathbf{\Gamma}' \boldsymbol{Z}_{ijt} + \gamma_j + \lambda_t + \varepsilon_{ijt}, \quad (1.6)$$

where t^* is the 2015 Q3 treatment quarter, k indexes quarters relative to it, and $k = -1$ is the omitted base period. Each δ_k measures the triple-difference effect in quarter k, relative to the base period. The quarter effects enter interacted with size and with density tercile, so λ_t in Equation (1.6) stands for quarter-by-size and quarter-by-density effects that absorb the period-specific size and geographic trends. $\boldsymbol{Z}_{ijt}$ collects the apartment controls and the time-invariant $\mathit{High}_j \times \mathit{Small}_i$ interaction, the one lower-order term these fixed effects do not absorb. The endpoints at $k = \pm 8$ bin all further periods. The controls are the predetermined building-fabric subset of the structural tier. With thirty-five event-time coefficients in the unbinned version and 57 clusters, the full structural vector renders the clustered covariance matrix rank-deficient, the same finite-cluster budget that motivates the design-based inference in Section 1.6. The plotted coefficients move by at most 0.014 under the fuller vector.

Figure 1.5 plots the coefficients. Pre-treatment they fluctuate around zero with no systematic trend. A single positive deviation at $k = -6$ is isolated, bracketed by near-zero coefficients on either side. From the treatment quarter onward they shift to persistently positive values, rising over the following two years. Individual post-treatment quarters are noisy (each coefficient is a contrast against the single reference quarter, estimated with 57 clusters) but stay above the pre-treatment baseline throughout. The static estimate of 0.059 summarizes this shift as a time-average, and the plotted post-treatment coefficients average exactly that value.[13]

1.5.3 Dose-response

The binary high/low split treats all high-density neighborhoods alike. Resolving treatment by intensity shows the effect rises with exposure. The dose-response replaces the binary High indicator in Equation (1.4) with the continuous end-of-sample density and its lower-order interactions, keeping the full triple structure. Splitting the treated tercile into density thirds against the same low-density control gives a monotone gradient of 0.045, 0.058, and 0.067 (all $p < 0.01$) for the low-, middle-, and high-dose thirds, which the binary estimate pools to 0.059. A cubic in the continuous dose (Figure 1.6) traces the same shape: rising from zero, flattening toward a peak near the densest neighborhoods, with wide confidence bands in the thinly populated high-dose tail. The peak is the shape

[13] The unbinned event study over the full sample window ($k \in [-22, +13]$) is reported in Appendix Figure A.1. It is the input to the sensitivity analysis in Section 1.6.

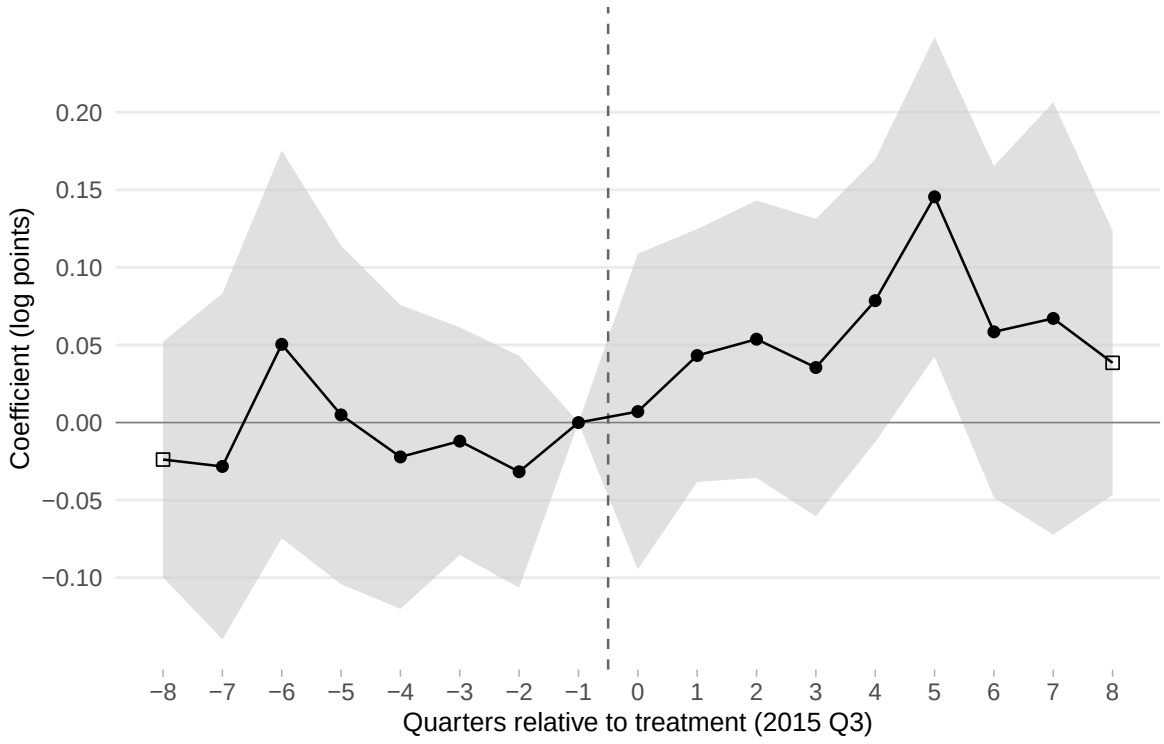


Figure 1.5: Event study: triple DiD coefficients by quarter relative to treatment

Notes: Coefficients $\hat{\delta}_k$ from Equation (1.6), fixed treatment date 2015 Q3, building-fabric controls (see text). Endpoints at $k = \pm 8$ bin all further periods. Reference period $k = -1$. Shaded area: 95% confidence intervals, standard errors clustered at the neighborhood level (57 clusters). Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. $N = 67{,}719$.

of a saturating response, not a sharp point estimate. It is not statistically distinguishable from the high-dose tercile coefficient.

This finding further supports the identification, because a dose-response of this form is difficult to reconcile with a confounder. A shock other than Airbnb would have to raise small-apartment rents in proportion to each neighborhood's Airbnb density.[14]

1.5.4 Magnitude

Evaluated at pre-treatment means for the treated cell (€ 9.96 per square meter and 61.3 square meters), the 5.9% effect is about € 0.59 per square meter per month, or roughly € 36 per month (€ 430 per year) for the average treated apartment.

Observed rents on small apartments in high-density neighborhoods rose from € 9.96 to € 12.38 per square meter between the pre- and post-treatment periods, an increase of 24.3%. The Airbnb-attributable 5.9 percentage points account for roughly one quarter of that growth. The remainder reflects rent growth shared with the control groups, city-wide appreciation and neighborhood-wide trends common to both apartment sizes.

[14]The gradient also holds under a nonparametric continuous-treatment estimator on small apartments (Callaway et al., 2024), reported in Appendix A.3. Its level is lower by construction (a continuous-dose double difference on small apartments rather than the triple) but the rising, saturating shape is the same.

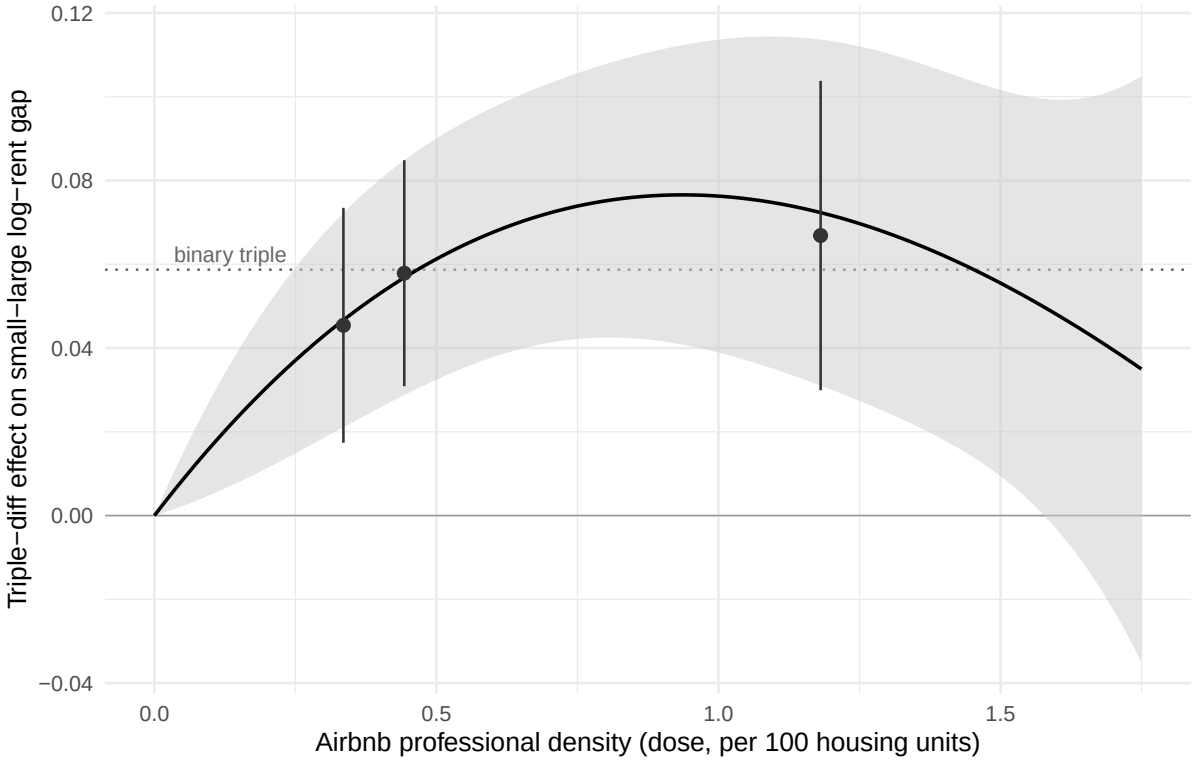


Figure 1.6: Dose-response: rent effect by Airbnb exposure intensity

Notes: Triple-difference rent effect as a function of end-of-sample professional Airbnb density, low-density (T1) neighborhoods set to dose zero. Cubic in dose with 95% confidence band. Tick marks along the horizontal axis show the observed neighborhood dose values. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. $N = 67{,}719$.

1.6 Robustness

The estimate rests on a research design rather than a single regression. This section examines its sensitivity to the assumptions and choices on which it depends: the parallel-trends assumption, the assignment of treatment across neighborhoods and over time, the level and structure of the standard errors, contemporaneous demand shocks, the thresholds that define the treatment and control groups, and the common treatment date. The estimate survives each.

1.6.1 Pre-trend sensitivity

The triple-difference design identifies β_7 under parallel trends. A pre-trend test has low power: failing to reject flat pre-trends is weak evidence that they are flat (Roth, 2022). I therefore ask directly whether the estimate survives a pre-existing linear trend in the small–large rent gap across terciles using the sensitivity analysis of Rambachan and Roth (2023), applied to the average effect over all post-treatment quarters that β_7 identifies.

The leading case restricts the differential trend to be exactly linear (Δ^{SD}, $M = 0$), the literal form of the concern that relative rents were already drifting upward before treatment. The slope of that line is not assumed to be zero. It is disciplined by the twenty-one pre-treatment coefficients. Every straight line consistent with their estimates

and sampling uncertainty is extrapolated through the post-treatment window, and the robust interval must cover the effect net of the least favorable of these lines (Appendix Figure A.3 visualizes this fan of admissible lines). This is the sensitivity counterpart to the pre-trend test above, because failing to reject a zero slope does not establish that the slope is zero. Under the $M = 0$ restriction the robust 95% confidence interval for the average effect over the post-treatment quarters still excludes zero, [0.001, 0.066], though with little margin. Slopes too small to detect in the pre-period accumulate over fourteen post-treatment quarters. The margin is thinner still for shorter averaging windows, where the effect itself is still building up. The two-year average survives ([0.017, 0.071]), the first-year average does not ([−0.004, 0.060]).

Beyond the linear case the bounds are uninformative. Under the smoothness restriction, significance is lost already at the smallest curvature examined, $M^* \approx 0.0005$ log points per quarter. Under relative magnitudes the robust interval includes zero at every $\bar{M}$. It does so already at $\bar{M} = 0$, where no violation is admitted at all, so the restriction fails on the imprecision of the event-study average before any admissible departure from parallel trends enters.[15] Two observations put these bounds in perspective. Under the smoothness restriction an admitted violation accumulates quadratically over the fourteen post-treatment quarters, so a breakdown value of this order is a mechanical feature of any effect of this size ($\hat{\beta}_7 = 0.059$) at this horizon rather than a property of this estimate, and the violations that matter economically in this setting are smooth ones. Confounders that could bend the small–large gap differentially across density terciles, such as gentrification or amenity growth, operate gradually, while the sharp candidates coinciding with 2015, the rent brake and the refugee inflow, are discussed directly in Section 1.6.4.

Reassigning treatment at random, across neighborhoods and dates, reproduces nothing like the estimate (Section 1.6.2), and the main results already show that the effect rises monotonically with exposure intensity (Section 1.5.3). Against this evidence, the one formal result the sensitivity analysis adds is that the estimate survives an exactly linear pre-trend of any slope.

1.6.2 Randomization inference and placebo tests

Treatment is assigned along two margins: the neighborhoods that are exposed and the date at which exposure begins. Randomization inference reassigns treatment along each margin and tests whether the reassignment reproduces the estimate.

Along the geographic margin, I hold the number of treated neighborhoods fixed at 28 and randomly relabel which 28 of the 57 neighborhoods are high-density, recomputing the saturated triple for each of 1,000 draws. The placebo distribution is centered on zero with a standard deviation of 0.018, and no draw reaches the observed estimate of 0.059.

[15] The event-study average is normalized against a single reference quarter and is therefore estimated far less precisely than the static coefficient; its conventional confidence interval spans [−0.026, 0.131]. The Δ^{SD} interval at $M = 0$ is narrower than that conventional interval because it also draws on the information in the twenty-one pre-treatment coefficients. The relative-magnitudes procedure does not use this information.

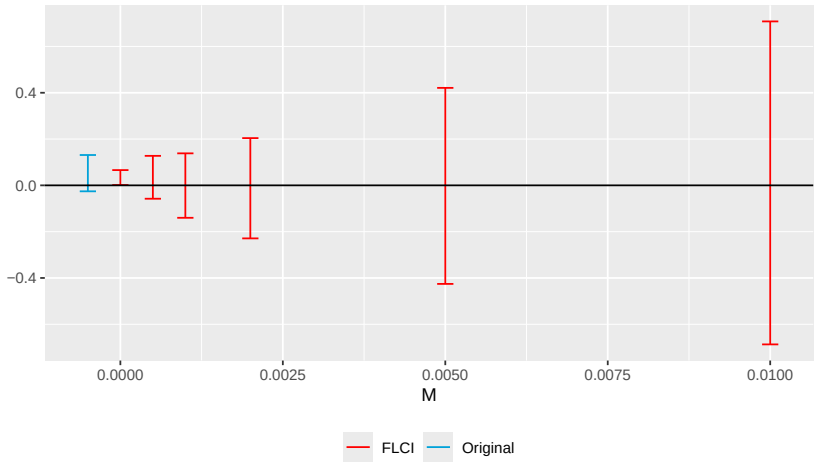


Figure 1.7: Pre-trend sensitivity under the smoothness restriction

Notes: Robust 95% confidence intervals for the average treatment effect over all post-treatment quarters (the estimand β_7 identifies) under the smoothness restriction $\Delta^{SD}(M)$ of Rambachan and Roth (2023), which permits the slope of the differential trend to change by at most M log points per quarter. The interval labelled *Original* imposes parallel trends; those indexed by M relax it. The Original interval spans zero because the event-study average is measured against a single reference quarter and is correspondingly imprecise. The robust intervals can be narrower than the Original because the fixed-length construction also draws on the pre-treatment coefficients. At $M = 0$ the pre-trend may follow any straight line and the robust interval still excludes zero, [0.001, 0.066]; the interval widens as larger departures from linearity are permitted. Estimated on the full unbinned event-study window. The relative-magnitudes (Δ^{RM}) sensitivity curve is reported in Appendix Figure A.2.

The estimate lies beyond the entire placebo distribution, for a randomization p-value of 0.001 (Appendix Figure A.4). The effect is specific to the neighborhoods in which Airbnb activity concentrated.

Along the time margin, I hold the geographic assignment fixed and place a fictitious treatment break in each of the fourteen interior quarters of the pre-period (2011 Q1–2014 Q2), estimating the triple on pre-treatment data alone so that the post-2015 effect cannot enter. The placebo estimates range from −0.010 to +0.016 (Appendix Figure A.5). The two largest fall in 2013 Q4–2014 Q1, a single isolated, individually insignificant deviation in the pre-period. To the extent it reflects a pre-existing linear drift, the smoothness analysis in Section 1.6.1 shows the estimate survives it. Even the larger of the two deviations is just above a quarter of the headline estimate. The genuine 2015 Q3 effect lies well beyond anything the pre-period produces.[16]

The estimate is not an artifact of which neighborhoods, or which date, is designated as treated.

1.6.3 Few clusters and spatial dependence

The design has 57 neighborhood clusters, but only 28 are treated, and they form a contiguous ring around the city center. With few treated clusters the cluster-robust standard

[16]The in-time test draws from a small set of overlapping pre-period windows and is therefore a placebo-distribution argument rather than an exact test. The geographic randomization is the load-bearing one.

error can overstate precision, and spatial correlation among adjacent neighborhoods can compound it.

The wild cluster bootstrap addresses the cluster count (Cameron et al., 2008). With restricted Rademacher weights over $B = 1{,}999$ replications, the bootstrap critical value for $|t|$ is 2.12, modestly above the asymptotic 1.96. The observed t-statistic of 4.05 exceeds it, and the bootstrap p-value is 0.001. The few-cluster correction widens the critical value only slightly.

A possible objection to the geographic randomization test of Section 1.6.2 is that its unrestricted relabelings are mostly spatially dispersed, unlike the actual contiguous treatment. A stratified variant therefore relabels high-density status only *within* each of the nine city districts, so every placebo assignment is as spatially concentrated as the actual one. The central district, whose five neighborhoods are all high-density and admit no relabeling, is excluded. The estimate on the remaining sample is 0.049, indistinguishable from the headline and slightly lower, consistent with the dose gradient, since the exclusion removes the densest neighborhoods. Within-district relabelings can retain part of the true assignment, which under a true effect pulls placebo estimates toward the observed value and makes the test conservative. The placebo distribution is nonetheless centered on zero, and no draw among 1,000 exceeds the estimate ($p = 0.001$; Appendix Figure A.6).[17]

For spatial dependence, I allow the errors to be correlated across neighborhoods using Conley standard errors based on the distances between neighborhood centroids (Conley, 1999). At a cutoff of three kilometers, a reasonable scale for local housing-market spillovers in a city roughly 30 kilometers across, the standard error is 0.0146, essentially the neighborhood-clustered value of 0.0145. Thus, short-range spatial correlation is mild and positive. The coefficient remains significant, and the full profile of the standard error across distance cutoffs is reported in Appendix Figure A.7.[18]

The estimate's significance does not depend on the number of clusters or on treating neighborhoods as spatially independent.

1.6.4 Rent brake, refugee inflow, and apartment composition

Three developments coincide with the treatment window and could, in principle, produce a size-specific rent effect in high-density neighborhoods that owes nothing to short-term rentals. They are the rent brake enacted in 2015, the refugee inflow of the same year, and drift in the composition of the advertised stock. Each is examined in turn. The design and the institutional record rule out the first, the demographic record the second, and the control vector bounds the third.

The rent brake took effect in July 2015, coinciding with the treatment date, which raises the concern that the estimated effect reflects the regulation rather than platform growth. Two facts rule this out. First, the triple difference removes any component of the

[17] Retaining the central district, whose treated status is then held fixed in every draw, is conservative for the same reason and yields $p = 0.002$.

[18] Clustering on the nine districts instead gives a comparable standard error (0.0154) but rests on too few clusters to be reliable, which is why the bootstrap is the appropriate tool for the few-cluster concern.

rent brake common to all apartment sizes within high-density neighborhoods, absorbed by the High × Post interaction, and any component common to small apartments citywide, absorbed by the Small × Post interaction. For the regulation to bias β_7, its effect would have to fall on small apartments in high-density neighborhoods specifically and have no counterpart in the control cells, a size-by-location-specific cap the statute does not impose. Second, the ordinance carried no enforceable force during the sample period. The Cologne designation was ruled void from its enactment, a defect that could not be cured retroactively, and a corrected ordinance took effect only in July 2020. A landlord who exceeded the cap between 2015 and 2018 therefore faced no enforceable penalty (Footnote 3 on page 10). The coincidence of dates is therefore uninformative about the source of the effect.

A further contemporaneous shock is the German refugee inflow of 2015. Asylum arrivals accelerated in the third quarter of 2015, coincident with the treatment date, and concentrated in single-adult households, which demand small apartments, and in central, transit-accessible neighborhoods. Both dimensions align with the treated cell of the triple difference. If high-density neighborhoods absorbed differentially more new residents after 2015, part of the size-specific effect attributed to Airbnb would reflect this demand shock. The data do not support the channel (Appendix Table A.1). Relative to low-density neighborhoods, high-density population is statistically indistinguishable from its pre-2015 path ($\hat{\beta} = 121$ residents, $p = 0.38$), and high-density unemployment falls rather than rises, the opposite of what a refugee inflow would imply ($\hat{\beta} = -0.34$ percentage points, $p = 0.08$). Adding time-varying neighborhood population and unemployment to the specification leaves the triple-difference coefficient essentially unchanged, at 0.058 (Appendix Table A.2).

A further concern is compositional drift in the advertised stock of small apartments in high-density neighborhoods. If the post-2015 stock reaching the rental market differs systematically in quality from the pre-2015 stock, the 5.9 percent estimate would partly reflect a change in the underlying product rather than a price effect on comparable units. Such drift could arise through renovation, selective withdrawal of lower-quality units to short-term use, or compositional turnover.

Table A.3 reports normalized differences in observable characteristics across the treatment boundary for the small-T3 cell, the small-T1 control cell, and the full Cologne rental sample. The post-2015 advertised stock is modestly newer everywhere. Elevators, balconies, first-occupancy units, and new builds rise by 0.10 to 0.27 standard deviations across all three samples, and the small-T1 control drifts comparably to or more than the small-T3 treated cell on every one of these characteristics. The structural controls in the preferred specification absorb this drift.

Furnished listings are the exception. Their share rises from 1.9 to 7.2 percent in the small-T3 cell against 1.4 to 3.2 percent in small-T1. Furnishing is not a fixed building characteristic but a margin the rental market can adjust on. Whether the differential drift reflects an Airbnb-induced response or an unrelated citywide trend cannot be settled from the composition table alone. The preferred specification excludes furnishing on the

standard logic of Angrist and Pischke (2009). Structural controls retain only variables that are unambiguously predetermined by building characteristics. The attenuation from 0.059 to 0.051 in column 3 (Table 1.6), once these adjustable characteristics enter the control vector, bounds how much of the estimate could operate through them.

1.6.5 Sensitivity to operational choices

The triple-difference estimate rests on several operational choices. These are the treatment date, the room-size cutoff separating small from large apartments, the Airbnb-density measure used to assign neighborhood terciles, the tercile split itself, and the nightly threshold that classifies a listing as professional. Varying each of these in turn leaves the coefficient of interest positive, significant, and of the same order as its baseline value of 0.059, within a band running from 0.045 to 0.065.

Moving the treatment date one quarter in either direction around the 2015 Q3 baseline moves the estimate only modestly. Dating the treatment to 2015 Q2 yields 0.055 and dating it to 2015 Q4 yields 0.065, against the baseline 0.059 (all $p < 0.01$ on the full sample of 67,719 observations; Table A.4). The spread of roughly 0.010 across the two dates lies within one baseline standard error (0.014).

The room-size cutoff defining *Small* is set at three rooms, matching the Airbnb size distribution, in which 96 percent of listings have three rooms or fewer. Tightening it to two rooms yields 0.045. Three-room apartments, which make up a large share of Airbnb-substitutable housing, then count as large and enter the control group, so the treated–control contrast is diluted from both sides. Widening it to four rooms yields 0.065, but the large-apartment control then collapses to fewer than 1,900 five-room observations, a residual category rather than a clean counterfactual (Appendix Table A.5).

Assigning terciles on average density over 2010 Q1 to 2018 Q4, rather than on the end-of-sample stock at 2018 Q4, gives 0.058 (Appendix Table A.6). No neighborhood switches between the treated and control terciles across the two measures. All twenty-two reclassifications involve the excluded middle tercile, fourteen between T1 and T2 and eight between T2 and T3.

Replacing the equal tercile split with a 60/20/20 split, which assigns the bottom 60 percent of neighborhoods to control, drops the middle 20 percent, and treats only the highest-density fifth, gives 0.057 with a slightly tighter standard error (Appendix Table A.7).

The professional threshold is moved from 110 booked nights per year to 90 and 140, equivalent to shifting the average-stay parameter in the booking proxy from 2.8 to 1.7 nights. The triple interaction is 0.059 at 90 nights and 0.056 at 140 nights, both within one standard error of the baseline (Appendix Table A.8).

Across all five exercises the coefficient of interest stays within roughly 0.045 to 0.065, with the baseline of 0.059 near the center of the band, and remains statistically significant throughout. The width of the band is driven almost entirely by the room-size cutoff and the treatment-date checks, and the lower endpoint of 0.045 at the at-most-two-room cutoff

lies about one baseline standard error below 0.059. The density-measure, tercile-split, and professional-threshold checks each move the estimate by 0.003 or less. The effect is therefore robust in sign, magnitude, and significance across every operational choice, with the sharpest sensitivity confined to the definition of the small-apartment margin itself.

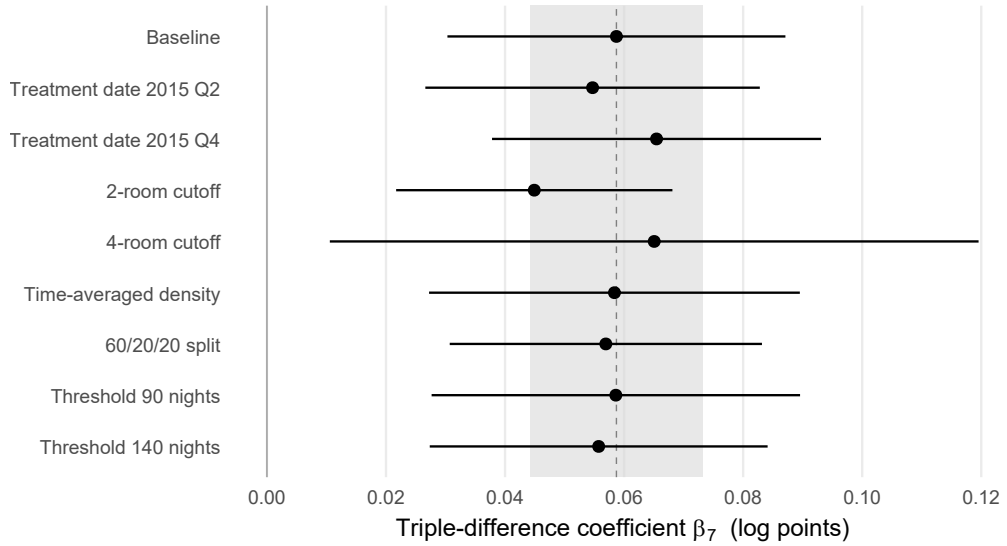


Figure 1.8: Sensitivity of the triple-difference estimate to operational choices

Notes: Each row plots β_7 under an alternative operational choice, with 95% confidence intervals from standard errors clustered on the 57 neighborhoods. The shaded band is the baseline estimate plus or minus one standard error. The dashed line marks the baseline of 0.059. Full regressions in Appendix Tables A.4–A.8.

1.6.6 Flexible entry

The headline design dates treatment at the common quarter 2015 Q3. A flexible-entry variant instead treats each high-density neighborhood at the quarter its professional Airbnb density first crosses the top-tercile lower threshold, with low-density neighborhoods never treated. Because entry timing is internal to the density measure rather than an external shock, and entry densities cluster tightly at the threshold (0.30 to 0.42, against a cutoff of 0.302), this is a robustness check on the design, not a staggered natural experiment.

I estimate the flexible variant with the Borusyak et al. (2024) imputation estimator, which accommodates the staggered timing and, unlike the Sun and Abraham (2021) interaction-weighted estimator, delivers a usable covariance matrix at this cluster count. At 57 neighborhood clusters with interacted fixed effects the Sun and Abraham covariance matrix is not positive semi-definite. The estimator is applied to the small–large gap in log rents within each neighborhood and quarter, without apartment-level controls. Differencing out the size margin in this way turns the triple into a standard difference-in-differences on the gap, with high-density neighborhoods as the treated group (Olden and Møen, 2022). The gap observations are weighted such that, under a common treatment

date, the gap regression reproduces the listing-level saturated triple exactly.[19] The two estimators target the same estimand, and their point estimates coincide (0.041 and 0.042).

The average effect of 0.042 is significant at the 5% level (standard error 0.018), of the same sign as, and statistically indistinguishable from, the static 0.059.[20]

Figure 1.9 plots the event-study path. Pre-treatment coefficients are flat: none of the eleven leads in quarters −12 to −1 is significant, and the entry quarter is insignificant (−0.019, $t = -0.8$). Post-treatment coefficients grow imprecise at longer horizons, where fewer neighborhoods have yet reached that lag, so the average effect rather than any single quarter is the informative summary.

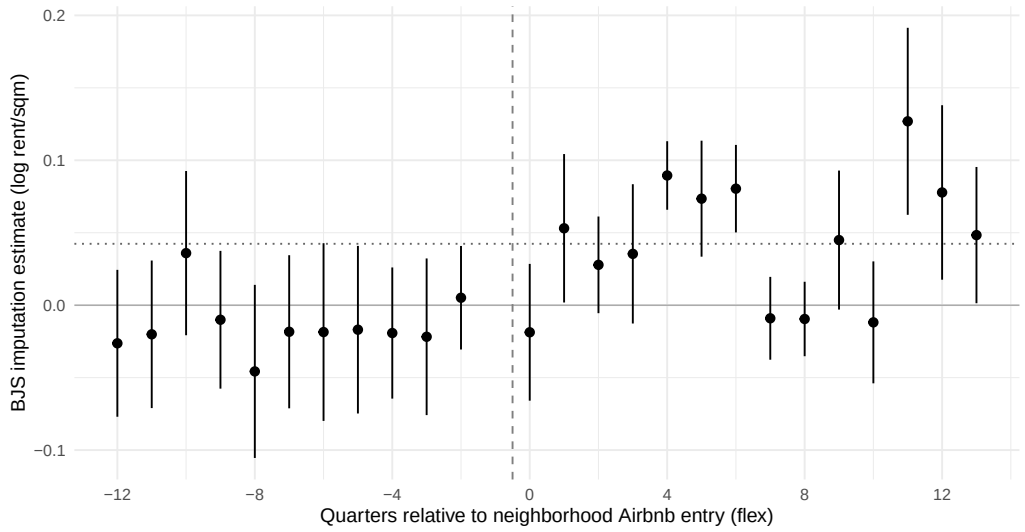


Figure 1.9: Flexible-entry event study, imputation estimator

Notes: Event-study coefficients on $High \times Small \times Post^{\text{flex}}$ from the Borusyak et al. (2024) imputation estimator, over event time in quarters from each neighborhood's treatment entry, the quarter its professional density crosses the top-tercile threshold. Vertical bars: 95% confidence intervals, clustered at the neighborhood level. Sample: low- and high-density neighborhoods, 2010 Q1–2018 Q4.

[19] The gap requires both apartment sizes present in a neighborhood-quarter. Cells lacking one, about 5% of observations and balanced across high- and low-density neighborhoods, drop out. A no-drop listing-level specification gives 0.040.

[20] The 0.018 standard error is from a two-stage estimate of the same specification (Gardner, 2022), which reproduces the imputation point estimate. A neighborhood cluster bootstrap gives 0.019 and the imputation estimator's analytic standard error is 0.014. All three routes reject zero at the 5% level. The gap to the static estimate is expected rather than anomalous. Re-dating treatment at neighborhood-specific entry already moves the conventional two-way fixed-effects triple to 0.053, and entry quarters carry mechanically low doses because entry is dated at the first threshold crossing, consistent with the dose-response gradient (Section 1.5.3). The flexible estimate's confidence interval, [0.007, 0.077], comfortably contains 0.059. A Callaway and Sant'Anna (2021) cross-check yields a point estimate of 0.050.

1.7 Discussion

1.7.1 Interpretation and Mechanism

The triple DiD design identifies the effect of short-term rental growth on the small-apartment segment of the housing stock within exposed neighborhoods. The preferred estimate of 5.9% corresponds to small apartments in high-density neighborhoods, relative to two control groups: large apartments in the same neighborhoods and small apartments in low-density neighborhoods. By construction, the design absorbs any effect that operates uniformly across apartment sizes within treated neighborhoods. Tourism-driven amenity capitalization (new cafés, retail, cultural venues that could themselves raise rents by making a neighborhood more attractive to long-term residents) is one example. So is any other shock that moves neighborhood rents across the board, whatever its origin. All such shocks load on the $High_j \times Post_t$ interaction. The remaining differential effect isolates Airbnb's impact on the directly exposed segment.

This segment-specific effect operates through the two channels introduced in Section 1.3.2. Every professional listing reduces long-term supply by one unit, raising rents in the affected segment through the standard rationing channel. A subset of listings additionally exerts demand-side pressure: when an entrepreneur acquires a unit through the rental or purchase market to operate it short-term, the bid exceeds the going long-term rent and raises the equilibrium price on comparable units. The host concentration pattern documented in Section 1.3.2 is consistent with both channels being present in the Cologne market. The large modal mass of single-listing hosts reads naturally as personal-property conversion within the supply channel, the concentrated tail as entrepreneurial assembly. The design does not decompose the two contributions. Both produce the same observable: a small apartment in professional short-term use and a higher rent on comparable long-term units. Decomposition would require an instrument that shifts one channel without the other, or transaction-level evidence on how each listing was assembled. Neither is available in this setting. The two channels share an upstream observable, professional short-term operating intensity, which is the lever the policy instruments discussed below in Section 1.7.2 act on. The policy implication therefore does not depend on which channel dominates the 5.9% estimate.

The event study supports the broader interpretation that Airbnb growth is the proximate driver of the segment-specific effect. Coefficients shift to positive values from the treatment quarter onward, with no slow build-up. This pattern argues against slow-onset alternatives such as gradual amenity capitalization or cumulative gentrification, and is consistent with both Airbnb channels, which respond directly to platform-scale entry.

The incidence of this price effect is broader than the subset of newly arriving tenants. Asking rents on new contracts determine the prices faced by any household moving within the Cologne market: couples forming a joint household, older couples downsizing after children leave, or tenants relocating for work. Reduced availability in the directly competing segment also slows turnover, thinning the market for all participants.

The magnitude is quantitatively plausible under existing estimates of rental demand,

evaluated against the supply channel alone. Mense (2025), working with German rental listings, estimates that a 1% increase in annual new supply lowers rents by 0.19%. The structural model in that paper implies a larger response of −0.79 when the supply shock is specific to the rental segment rather than to the overall housing stock.[21] Older experimental estimates from US housing allowance programs span 0.22 to 0.45 (Hanushek and Quigley, 1980; Friedman and Weinberg, 1981). Reading these elasticities in reverse, withdrawal of small apartments into short-term use should raise rents in the affected segment in proportion to the share of the small-apartment stock removed. In Cologne T3 neighborhoods, the small-apartment-adjusted density of professional Airbnb listings reaches 0.81% at the median and 2.33% at the 90th percentile (Table 1.3). Applied to these withdrawal shares, the elasticity range above implies rent effects of roughly 1–4% at the median T3 neighborhood and 3–11% at the upper percentiles. The 5.9% triple DiD estimate, averaged across all T3 neighborhoods and the full post-treatment period, falls inside this range, closer to the predicted magnitude at the upper percentiles. The arithmetic is approximate: the cited elasticities are estimated for aggregate construction shocks rather than segment-specific withdrawals, and substitution from the large-apartment segment is more limited than the literature's identifying variation. The calibration also uses the supply channel alone. The entrepreneurial channel, if operative, contributes additional rent pressure and would lower the implied burden on the supply elasticity. The order of magnitude lines up either way. The estimated effect does not require a demand response that is large by the standards of existing evidence on the rental market.

The magnitude is also in line with the international literature, particularly where studies isolate high-exposure market segments. City-wide averages are small (0.4% per standard deviation of Airbnb density in Boston (Horn and Merante, 2017), an elasticity of 0.018 across US ZIP codes (Barron et al., 2020), roughly 2% implied by the Los Angeles Home Sharing Ordinances (Koster et al., 2021)), but the same literature finds markedly larger effects where exposure concentrates: 3.1% in Boston's tourist destinations, up to 7% in Barcelona's high-activity neighborhoods (Garcia-López et al., 2020), and 1.3–2.4% per additional commercial listing in Berlin (Duso et al., 2024). At 5.9%, the present estimate sits in the upper range of these high-exposure estimates, as expected from its tight focus on the directly competing market segment, small apartments in the most affected neighborhoods.

1.7.2 Policy Implications

The result provides a quantified basis for evaluating policy instruments directed at short-term rental platforms. Two clarifications precede the policy discussion. First, any regulation of short-term rentals presupposes a normative aim, typically the protection of residential housing stock for long-term residents. Short-term rentals generate benefits for tourists and for hosts, and the analysis here does not attempt a full welfare calculation. The discussion below therefore concerns instrument choice conditional on the residential-

[21] See Mense (2025, footnote 2).

protection aim. Second, the Cologne result emerged under a permissive regulatory regime: the housing protection ordinance existed throughout the sample period but lacked the enforcement infrastructure (a registration system, platform disclosure obligations, meaningful fines) to constrain short-term rental activity in practice. The estimate therefore reflects the effect of largely unregulated platform growth in a tight housing market.

Several policy instruments have been proposed or implemented. The most obvious is an outright size-based restriction: prohibit or cap short-term rental of units below a given floor area or room count. Such rules impose definitional and administrative costs (who verifies apartment size? how are borderline cases adjudicated?) and create obvious substitution incentives for hosts to shift into marginally larger units. A night-based cap, by contrast, regulates the intensity of use rather than the characteristics of the unit. Operating intensity is the upstream lever shared by both channels identified in Section 1.7.1. The policy acts on the full segment-specific effect regardless of which channel dominates. The WohnStG NRW adopted this approach, permitting occasional home sharing while restricting professional short-term operation: up to 90 days per year of short-term letting is permitted, with usage beyond that threshold requiring a permit to use residential housing for non-residential purposes. This threshold is close to the break-even calculation reported in Section 1.3.2, at which short-term letting becomes more profitable than a long-term contract. If night caps bind, size-based restrictions become redundant.

A third instrument is a progressive tax on short-term rental revenue, rising with the number of nights let, which would raise the break-even threshold for conversion. It would be simpler to enforce than individual permit review once a registry and platform-disclosure obligations are in place, and would preserve home sharing at low intensities without subsidizing full-time commercial operation.

All three instruments share a common requirement: a functioning registration and enforcement infrastructure. Without it, compliance is voluntary and the regulation is largely decorative. Until recently, this was the situation in North Rhine-Westphalia: the WAG NRW provided no registration requirement, no platform disclosure obligation, and a maximum fine of €50,000. Anonymous host profiles made systematic detection of violations nearly impossible (Mindl, 2021). The 2021 WohnStG NRW and the digital permit registry that went live on July 1, 2022 provide the first genuine enforcement infrastructure. By end-2023, 4,294 Cologne short-term rental providers were registered, and 152 fine proceedings had been opened in the second half of 2023 alone.[22] The effects of this enforcement regime on rental prices fall outside the sample period of this study and are a natural subject for future research.

1.7.3 Limitations

Three caveats qualify the result. The rental data are asking rents on new contracts rather than transacted rents. In tight urban German markets the two are close, because excess

[22] Stadt Köln, *Halbjahresbericht zur Zweckentfremdung von Wohnraum*, January 2024. The WohnStG raised the maximum fine to €500,000 and obliges booking platforms to allow only listings displaying a valid registration number.

demand pushes realized rents toward listed asking rents (Duso et al., 2024), but some residual discrepancy is possible. I obtain the Airbnb data from AirDNA, a commercial provider that scrapes public Airbnb listing pages. This is the same source used by Duso et al. (2024) for Berlin, and the broader approach of reconstructing Airbnb activity from public-page scrapes is standard in the literature (Barron et al., 2020). Independent comparisons between scraped Airbnb data and records released by Airbnb under subpoena to the New York Attorney General show high agreement (Budget and Legislative Analyst's Office, 2015). For Cologne, I additionally compared the AirDNA data against two independent scrapes of the same public Airbnb pages (one commissioned by the Institute for Economic Policy at the University of Cologne and one provided by a private individual) and found the listing counts and price distributions broadly consistent across all three sources. The booking proxy based on a 60% review rate is conservative relative to alternative assumptions in the literature. The professional-threshold sensitivity check (Section 1.6) confirms that moderate variation in this assumption leaves the result intact.

A second sense in which the estimate is conservative concerns spillovers across the treated and control groups. If tenants priced out of small apartments in high-density neighborhoods bid on adjacent substitutes (small apartments in low-density neighborhoods, or larger apartments in the same high-density neighborhoods), the two control groups absorb part of the rent pressure that would otherwise show up in the treated group, attenuating the main estimate. Excluding the middle tercile buffers part of this concern, because the T2 neighborhoods adjacent to the treated core are the most natural first destination for displaced demand and belong to neither comparison group. The estimate is therefore best read as a lower bound on the size-specific rent effect.

The analysis identifies rent effects but does not conduct a welfare assessment. A complete welfare analysis would weigh tourist surplus, host earnings, hotel-industry displacement, and agglomeration effects from tourist spending against the rent burden documented here and against negative amenity effects on residents, such as noise and congestion from tourist traffic. Quantifying the first three components along the lines of Farronato and Fradkin (2022) requires structural assumptions and data beyond the scope of this paper.

The Cologne estimate is best read as a benchmark, a magnitude of what professional short-term rental growth does to rents in directly competing segments of a dense urban market when the policy environment is largely passive. The setting is particularly clean: a sharp mid-2015 onset of platform-scale activity, no binding rent regulation on new contracts, a housing protection ordinance not enforced against short-term rental activity, and pronounced spatial concentration of the platform together isolate the size-specific effect from confounding shocks. Other cities differ along dimensions that matter for the size of the effect: platform penetration, tourist volume, housing-stock composition, enforcement intensity, and concurrent demand shocks. The specific magnitude will therefore not translate directly.

1.8 Conclusion

This paper provides the first segment-specific estimate of the rental-price effect of short-term rental platforms for a German city. The triple difference-in-differences design exploits variation across neighborhoods, apartment sizes, and time. I estimate that Airbnb growth raised rents on small apartments in high-density neighborhoods of Cologne by 5.9%. At the mean rent of affected units this is €430 per year, accounting for roughly one quarter of the rise in the segment's average rent between the pre- and post-treatment periods. The event study shows a clear post-treatment break against a pre-period with no systematic trend, sensitivity bounds quantify the pre-trend violation required to overturn the estimate, and randomization inference in time and geography places it outside its placebo distribution.

The 5.9% estimate identifies Airbnb's effect on the small-apartment segment in high-density neighborhoods, net of neighborhood-wide channels absorbed by the design. Every professional listing withdraws one unit of long-term supply from the affected segment. Entrepreneurial demand bidding by hosts acquiring units through the long-term market adds a layered rent effect, the size of which the design does not quantify. Both channels respond to the same regulatory lever: registration requirements and night caps act on professional short-term operating intensity directly, and the policy implication of the estimate is robust to the channel mix. The estimate sets a concrete magnitude against which regulatory instruments can be measured.

Whether the digital permit registry and its associated enforcement regime in Cologne, in place since 2022, attenuate the rental-price effect documented here is a question the coming years of data will answer.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *Quarterly Journal of Economics*, 138(1):1–35.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton.

Barron, K., Kung, E., and Proserpio, D. (2020). The sharing economy and housing affordability: Evidence from airbnb. *Marketing Science*.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, 91(6):3253–3285.

Breidenbach, P., Eilers, L., and Fries, J. (2022). Temporal dynamics of rent regulations—the case of the German rent control. *Regional Science and Urban Economics*, 92:103737.

Budget and Legislative Analyst's Office (2015). Analysis of the Impact of Short-Term Rentals on Housing. Technical report, City and County of San Francisco Board of Supervisors.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. C. (2024). Difference-in-differences with a continuous treatment. Working Paper 32117, National Bureau of Economic Research, Cambridge, MA.

Callaway, B. and Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92(1):1–45.

Duso, T., Michelsen, C., Schaefer, M., and Tran, K. D. (2024). Airbnb and rental markets: Evidence from Berlin. *Regional Science and Urban Economics*, 106:104007.

Farronato, C. and Fradkin, A. (2022). The welfare effects of peer entry: The case of Airbnb and the accommodation industry. *American Economic Review*, 112(6):1782–1817.

Fradkin, A. (2015). Search, matching, and the role of digital marketplace design in enabling trade: Evidence from Airbnb. Working paper, Stanford University.

Franco, S. F. and Santos, C. D. (2021). The impact of Airbnb on residential property values and rents: Evidence from Portugal. *Regional Science and Urban Economics*, 88:103667.

Friedman, J. and Weinberg, D. H. (1981). The demand for rental housing: Evidence from the housing allowance demand experiment. *Journal of Urban Economics*, 9(3):311–331.

Garcia-López, M.-À., Jofre-Monseny, J., Martínez-Mazza, R., and Segú, M. (2020). Do short-term rental platforms affect housing markets? evidence from airbnb in barcelona. *Journal of Urban Economics*, 119:103278.

Gardner, J. (2022). Two-stage differences in differences. Working paper, arXiv:2207.05943.

Hanushek, E. A. and Quigley, J. M. (1980). What is the price elasticity of housing demand? *The Review of Economics and Statistics*, 62(3):449–454.

Horn, K. and Merante, M. (2017). Is home sharing driving up rents? evidence from Airbnb in Boston. *Journal of Housing Economics*, 38:14–24.

Koster, H. R. A., van Ommeren, J., and Volkhausen, N. (2021). Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles. *Journal of Urban Economics*, 124:103356.

Mense, A. (2025). The impact of new housing supply on the distribution of rents. *Journal of Political Economy Macroeconomics*, 3(1):1–42.

Mense, A., Michelsen, C., and Kholodilin, K. A. (2023). Rent control, market segmentation, and misallocation: Causal evidence from a large-scale policy intervention. *Journal of Urban Economics*, 134:103513.

Mindl, F. (2021). Wohnraumstärkungsgesetz – WohnStG: wird das Home-Sharing zukünftig seinem namen gerecht? Kölner Impulse zur Wirtschaftspolitik 4/2021, Institut für Wirtschaftspolitik, Universität zu Köln.

Mindl, F. and Arentz, O. (2020). Kurzzeitvermietung in Köln: Lokale wohnungsmarkteffekte durch plattformbasierte Kurzzeitvermietung. iwp Discussion Paper No. 1/2020.

Olden, A. and Møen, J. (2022). The triple difference estimator. *The Econometrics Journal*, 25(3):531–553.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–322.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

2 Debunking "Fake News" on Social Media: Immediate and Short-Term Effects of Fact-Checking and Media Literacy Interventions

This chapter is joint work with Lara Marie Berger (University of Cologne), Anna Kerkhof (University of Munich, ifo Institute, and CESifo), and Johannes Münster (University of Cologne). It has been published as:

> Berger, L. M., Kerkhof, A., Mindl, F., and Münster, J. (2025). Debunking "fake news" on social media: Immediate and short-term effects of fact-checking and media literacy interventions. *Journal of Public Economics*, 245, 105345. https://doi.org/10.1016/j.jpubeco.2025.105345

In accordance with Section 31 (1) of the doctoral regulations of the Faculty of Management, Economics and Social Sciences of the University of Cologne, the chapter is replaced by this bibliographic reference in the published version of the dissertation. The dissertation as submitted to and accepted by the Faculty contains the chapter and its appendix in full. The article is published open access.

Abstract

We conduct a randomized survey experiment to compare the immediate and short-term effects of fact-checking to a brief media literacy intervention. We show that fact-checking primarily affects the specific fake news it directly addresses, whereas media literacy helps to distinguish between false and correct information more generally, both immediately and around two weeks after the intervention. A plausible mechanism is that media literacy enables participants to critically evaluate social media postings, while fact-checking fails to enhance their skills as much. Our results promote media literacy as an effective tool to fight fake news, that is cheap, scalable, and easy-to-implement.

3 Second Opinions, Health Care Cost, and Outcomes in Elective Surgery: An Empirical Investigation

Abstract

We analyze a structured second-opinion program for elective surgeries in Germany using administrative claims data (2013–2019). The program institutionally separates diagnosis from treatment: physicians providing second opinions provide independent assessments but are banned from performing a treatment themselves. In orthopedics and neurosurgery, first and second-opinion giving physicians disagree in 56% of cases. The disagreement is strongly asymmetric: second physicians contradict 63% of surgery recommendations but only 14% of recommendations against surgery—consistent with physician-induced demand. Seventy percent of patients comply with second-opinion recommendations. Using propensity score matching and patient fixed-effects event-study designs, we estimate that compliance reduces monthly health expenditures by about €431–478 (approximately baseline mean), the number of outpatient visits by 0.61, and hospital days by 0.06 over 18 months, with no adverse effects on sick leave. Interestingly, the effects concentrate around patients advised against surgery, who save €641–704 monthly without health deterioration. For these patients, surgeries remain avoided throughout the 18 months following consultation, with symptoms managed instead through conservative therapies. Taken together, these findings provide evidence that structured second-opinion programs effectively curb overtreatment by separating diagnosis from treatment, translating theoretical predictions into realized health and spending outcomes.

3.1 Introduction

Health care constitutes a large and growing expert market, with global expenditure of USD 8.3 trillion in 2018, accounting for ten percent of global GDP, and projections indi-

This chapter is joint work with Yero Samuel Ndiaye (University of Cologne, Department of Economics) and Daniel Wiesen (University of Cologne, Department of Operations Management). We are grateful for valuable comments and suggestions from Anna Bindler, Michael E. Darden, Rudi Kerschbamer, Mathias Kifmann, Carolin Koll, Philipp Koll, Albert Ma, Johannes Münster, Axel Ockenfels, Leonard Rose, Matthias Sutter, and Tom Zimmermann, as well as from participants of the dggö annual conference 2023 Hannover, the dggö workshop of the committee on "Allocation and distribution", the Verein für Socialpolitik Field committee "Social sciences", and the Webinar Series on Credence Goods and Expert Markets. We are indebted to Simon Löser and Thorsten Göbbels for supporting our research project. We also thank Razi Farukh for initiating this project. All authors declare no conflict of interest.

cating it will outpace GDP growth until 2030 (OECD, 2019; WHO, 2020). In Germany, statutory health insurance (SHI) providers have reported rapid spending increases, leading to increased monthly premiums for policyholders. In particularly, in elective surgeries (Bundesamt, 2020). Notably, substantial regional differences in surgical interventions have been documented for over 80 years (Glover, 1938). Recent studies find that patients with comparable characteristics underwent disc surgery 13 times more often across different municipalities in Germany (Bertelsmann-Stiftung, 2013, 2017), and in Switzerland, non-physicians underwent common surgeries 33 percent more often than physicians and their families (Domenighetti et al., 1993).

Part of this variation may be explained by specific treatment patterns of individual physicians or highly specialized centers. However, a second explanation lies in financial incentive structures. In German hospitals, department heads receive bonus payments for meeting target numbers; regulators require minimum-volume for certain interventions as a prerequisite for reimbursement eligibility; and office-based physicians can generate additional revenue through inpatient beds when performing surgeries. Therefore, such variations are often interpreted as inefficiencies stemming from overtreatment, so-called physician-induced demand (McGuire, 2000). Paradoxically, patient perceptions do not reflect this concern. In a representative survey from Germany, 83.9 percent of surgical patients were satisfied with their treatment and would refer the clinic to others (Zich and Tisch, 2018).

Policymakers and insurers have increasingly turned to structured second-opinion programs as a potential remedy against overtreatment. In a traditional second opinion encounter, patients privately consult another expert who is also eligible to provide the treatment and therefore faces the same incentive structure. By contrast, structured second-opinion programs typically offer an independent assessment from a physician who is explicitly excluded from providing treatment. This institutional design separates diagnosis from treatment. Thereby, it fundamentally alters the expert's incentives by decoupling diagnostic payoffs from surgical margins (Bester and Dahm, 2018).

Despite long-standing international experience with such programs and increasing policy interest over the past decade, empirical evidence on their effectiveness remains piecemeal. Recent studies document substantial discrepancies between first and second opinions (Weyerstraß et al., 2020a), show that unconfirmed recommendations can defer knee-replacement surgeries (Kisch et al., 2023), and based on hypothetical cost comparisons suggest that treatments recommended by second physicians could substantially lower costs (Weyerstraß et al., 2020b). However, the effect of a second opinion also hinges on the patient's decision. While Weigl et al. (2021) find that programs can better align patients' preferences with guideline-based care, there remains a sizeable gap between the second physician's recommendation and the patient's final choice. To date, there is no longer-term comparison of actually realized outcomes between patients who comply with

the second physician's recommendation and those who do not.

Therefore, we investigate if structured second-opinion programs reduce overtreatment without adverse health outcomes. Specifically, we examine whether patients who comply with second-opinion recommendations experience different health expenditures, utilization, and health outcomes compared to those who do not comply, and whether these effects differ depending on the content of the recommendation.

We address this gap by analyzing healthcare outcomes for 572 patients over the 18 months following their participation in a second-opinion program run by a large German SHI provider from 2013 to 2019. The insurer provides a free second opinion within two weeks, regardless of whether surgery was initially recommended. Critically, the second physician is an experienced surgeon in the relevant field. However, she does not perform surgery even if she advises it. This institutional design separates diagnosis from treatment, potentially aligning the second physician's incentives toward diagnostic accuracy and truthful reporting rather than surgical volume. We observe the recommendations of both physicians for most patients and track their complete billing histories for 18 months after the second opinion. This setup enables us to systematically compare realized health outcomes between those who follow the second physician's recommendation and those who do not. We focus our analysis on the field of orthopedics and neurosurgery, which are most frequent in the program; where disagreement rates are highest; surgical procedures are well-established; and which are most prevalent in the literature and policy relevant. Using propensity score matching combined with patient fixed-effects event-study regressions, we estimate the effects of compliance on health expenditure, utilization, and health outcomes, accounting for rich medical histories and utilization patterns before the second opinion.

Our results reveal that structured second-opinion programs can predict treatment decisions and significantly reduce health expenditures without adverse health consequences. Among the 150 patients for whom we observe both first and second opinions in orthopedics and neurosurgery, physicians disagree in 56 percent of cases. This disagreement is strongly asymmetric: when the first physician recommends surgery, the second physician disagrees in 62.7 percent of cases; when the first physician advises against surgery, the second physician concurs in 85.7 percent of cases. This pattern is consistent with overtreatment in the form of physician-induced demand.

Second opinions predict patient decisions. Overall, 69.8% of patients comply with the second-opinion recommendation. When physicians disagree, nearly 80% of patients follow the second opinion rather than the first. Critically, compliance is higher for recommendations against surgery (79.9%) than for recommendations in favor of surgery (46.8%).

Compliance with second-opinion recommendations reduces health expenditures without worsening health outcomes. Over eighteen months, compliers experience €431 to €478 lower monthly total health expenditure (approximately equal to baseline mean monthly

spending), 0.61 fewer monthly physician visits, and 0.06 fewer monthly days spent in the hospital, with no significant effects on certified sick leave. These effects concentrate in the first six months post-consultation and then stabilize, with no evidence of catch-up spending or delayed health deterioration in the medium and long-term.

Effects are highly heterogeneous by recommendation content. Among the 100 patients advised against surgery, compliers save €641 to €704 monthly, visit physicians 0.65 fewer times per month, and avoid 0.11 to 0.12 hospital days monthly over 18 months. Short-term savings are substantial (€1,718 monthly in cost, 0.31 hospital days), with no offsetting increases later. By contrast, among the 37 patients whose second opinion recommended surgery, compliers incur expected surgical costs without offsetting long-term benefits, though small sample size limits inference.

Treatment pathway analysis reveals that avoided surgeries remain avoided. Compliers exhibit a 12 percentage point lower surgery probability across all post-second-opinion periods, with no rebound even twelve to eighteen months after consultation. Instead of undergoing surgery, compliers maintain conservative therapy spending near baseline levels (approximately €25 monthly), while non-compliers who proceed to surgery discontinue such care. Painkiller utilization does not differ between groups, indicating that surgical avoidance does not worsen symptom severity.

This study makes three primary contributions. First, we provide the first comprehensive longitudinal analysis linking second-opinion recommendations to realized health outcomes in a structured program that separates diagnosis from treatment. Unlike prior work that examines disagreement rates (Weyerstraß et al., 2020a), hypothetical cost savings (Weyerstraß et al., 2020b), or short-term treatment decisions (Kisch et al., 2023), we track actual health expenditure, utilization, and health outcomes over 18 months using complete administrative claims data. This allows us to test whether cost savings reflect genuine reductions in overtreatment or merely delayed necessary care.

Second, we provide field evidence testing theoretical predictions about separating diagnosis from treatment in credence goods markets. While Bester and Dahm (2018) show theoretically that separation can align expert incentives toward truthful reporting, and Greiner et al. (2017) provide laboratory evidence, our observational data on realized outcomes in high-stakes medical decisions complement experimental findings with real-world validation. Our results support the theoretical prediction that separation reduces overtreatment. In contrast to the laboratory evidence, we find no corresponding increase in undertreatment concerns.

Third, we advance the empirical literature on physician-induced demand and second opinions by documenting asymmetric disagreement patterns consistent with systematic overrecommendation of surgery. The concentration of disagreement when surgery is initially recommended (62.7% vs. 14.3% when surgery is initially discouraged), combined with evidence that avoided procedures remain avoided without health deterioration, sug-

gests that a substantial fraction of initially recommended surgeries may have been unnecessary. Moreover, we infer that the program's low uptake rate implies that second opinions function primarily through correction rather than deterrence, as physicians face minimal risk of verification when recommending surgery.

The remainder of the paper proceeds as follows. Section 3.2 reviews related literature on credence goods and second-opinion programs in healthcare. Section 3.3 describes the institutional design of the German healthcare market and second-opinion program and develops hypotheses. Section 3.4 presents the data and descriptive statistics. Section 3.5 outlines the empirical strategy. Section 3.6 reports results on compliance effects overall and by recommendation content, as well as treatment pathway mechanisms. Section 3.7 presents robustness checks including alternative observation windows and an instrumental variables approach. Section 3.8 discusses interpretation, policy implications, and limitations. Section 3.9 concludes.

3.2 Related Literature

3.2.1 Credence Goods in Economics

Patients rely on physicians' advice when making treatment decisions. They lack the expertise to assess the quality of care provided, both ex ante and ex post (Arrow, 1963). This *asymmetric information* between expert and consumer, is a characterizing feature of credence goods markets, leading to differences in prices or costs create incentives for opportunistic behavior (*moral hazard*) like overcharging or overselling of services (Darby and Karni, 1973; Emons, 1997).[0] A large literature studies how experts' fraudulent behavior varies under different information and incentive conditions across markets (e.g., Balafoutas et al., 2013; Beck et al., 2014; Lu, 2014). In health economics, this behavior manifests in overcharging and overtreatment or overdiagnosing, so called physician induced demand (e.g., Gruber and Owings, 1996; Kessler and McClellan, 2000; McGuire, 2000).

The theoretical literature on the effect of second opinion as a remedy to overtreatment (under varying assumptions of verifiability and liability) finds mixed results. Wolinsky (1993) finds that a fixed price scenario leads to multiple equilibria with ambiguous implications for the effect of second opinions on overtreatment. In this setting, the consumer cannot observe the type of treatment received which incentivizes the expert to overcharge the consumer. However, for the decision over whether to undergo surgery—the focus of our study—verifiability is a plausible assumption, which in turn rules out overcharging.[1] In Pesendorfer and Wolinsky (2003), consumers can judge which treatment they receive,

[0] See (Balafoutas and Kerschbamer, 2020) and Kerschbamer and Sutter (2017) for surveys.

[1] Treatment verifiability is a plausible assumption, since patients must give informed consent to surgery and the procedure typically leaves verifiable traces in ex post diagnostic imagery.

but not the effort invested in diagnosis, shifting moral hazard to the experts' diagnostic effort. Here, equilibria that feature positive diagnostic effort arise only in mixed strategies. Intuitively, if every expert always invests in an accurate diagnosis, the consumer gains nothing from paying for a second opinion and therefore never seeks one. However, without second opinions, experts no longer face the threat of being caught and thus reduce their diagnostic effort.

Testing the theoretical findings, laboratory experiments by Mimra et al. (2016) and Agarwal et al. (2019) find that allowing buyers to obtain a second opinion can raise market efficiency. However, whether efficiency gains evolve, starkly depend on search costs. The net effect reflects two countervailing forces: Reduced overtreatment raises efficiency by cutting treatment costs, whereas the consumers' search costs for a second opinion diminish these gains (Mimra et al., 2016). In contrast, a field experiment on computer repairs (Bindra et al., 2021) finds that if a mystery shopper announces that a second opinion has already been acquired it neither improves the repair success rate nor lowers the average price. Likewise, Gottschalk et al. (2020) report only small, statistically insignificant effects of announcing a second opinion in a field experiment on dental care. Both studies treat second opinions in a similar fashion: Customers request a service from a first expert, where one group of customers explicitly states that they will (or already have) consulted another provider. Here, the announcement of a second opinion serves as a disciplinary signal meant to constrain opportunistic behavior by the first expert. Hence, this design records only how the first expert reacts; it does not observe whether a second expert could overturn a mistaken initial diagnosis, so any welfare gains from such corrections remain unobserved.

Notably, none of the aforementioned literature considers the typical setting of a structured program, as in our case, where the second expert is institutionally excluded from performing the service, thereby separating diagnosis and treatment. This mechanism has also been suggested as a remedy against fraudulent behavior in the theoretical work of Bester and Dahm (2018). They argue that separation can achieve the first-best outcome even in settings where diagnosis requires costly effort and neither the diagnostic effort nor the diagnostic outcome is observable by the consumer. In their experiment, Greiner et al. (2017) contrast a baseline where one physician carries out diagnosis and treatment with a separation regime in which diagnosis and prescription are provided free by one physician and treatment by another. They find that separation is an effective way to reduce overtreatment, but this effect is partly offset by a rise in undertreatment.

We contribute to this literature in several ways. First, our data indicates that only a small fraction of insured individuals obtain a second opinion when surgery was recommended. This low uptake suggests that the simple availability of a second opinion may exert little disciplinary pressure, underscoring the importance of the analysis of cases in which diagnoses are actually corrected. Unlike for computer repairs (Bindra et al.,

2021), elective surgery involves high entry barriers for providers and severe consequences for patients which puts strict ethical limits on randomized experiments—especially when outcomes rather than mere recommendations are of interest (Gottschalk et al., 2020). Field data are therefore not a second-best. They are the only setting in which the separation mechanism can be evaluated at the stakes that matter: real surgical decisions, real health consequences, and an eighteen-month horizon that no laboratory or survey design can replicate. In addition, we look at a structured second-opinion program the way they are implemented by governments and insurance companies rather than the traditional way of *asking another expert*. Although, close parallels between the two approaches exist, there is one key difference: in the traditional notion both experts can provide the treatment and therefore face identical incentives. Whereas, in structured programs the second expert typically only diagnoses the patient and does not perform treatment; because of this refinement, our setting is particularly close to the literature on separating diagnosis and treatment in credence good markets.

We extend this literature by linking the theory to real-world data. Using comprehensive administrative claims of patients participating in a second-opinion program, we analyze the full decision chain for elective surgery: (i) agreement between first and second expert, (ii) the patient's short- and medium-term treatment choice, and (iii) health care utilization and outcomes during the 18 months following the second opinion. This scope allows us to focus on the corrective channel of second opinions in a setting where the incentive structure of the second expert is altered, such that it resembles the separation of diagnosis and treatment. Observational population data therefore provide the closest ethical test of theory and an important extension to the laboratory evidence reported in Greiner et al. (2017).

Our field evidence delivers what the laboratory cannot. We observe high disagreement rates between experts in real surgical decisions, where the first expert recommends surgery while the second expert refuses it. These findings are suggestive of erroneous or opportunistic treatment recommendations in credence goods markets—in the form of overtreatment. Further, divergent recommendations significantly affect patients' decision-making, reducing the number of surgeries. Interestingly, contrary to Greiner et al. (2017) we do not find evidence that canceled surgeries systematically raise the concern of undertreatment. For patients who decide against surgery after participating in the second-opinion program, we do not observe a significant worsening of their health outcomes within the following 18 months. In the majority of cases, avoided surgeries are not performed within the following 18 months and days of sick leave do not increase. Instead, symptoms are often managed with conservative therapy (e.g., physiotherapy). Finally, because a sizable minority still ignores a dissenting—or even concurring—second opinion, this shifts the focus on patient behavior.

3.2.2 Empirical Evidence on Second-Opinion Programs in Health Care

Despite ongoing policy debates to introduce second-opinion programs, the empirical analysis for second-opinion programs are surprisingly limited. In the medical literature, an early review reported short-term reductions in elective surgery and modest savings of U.S. insurer schemes due to second opinions, while emphasizing wide heterogeneity and weak study designs (Lindsey and Newhouse, 1990). A recent scoping review echoes the earlier finding, noting that robust evidence on following a first versus second opinion is still needed (Gattas et al., 2021).

A few studies show disagreement between first and second physicians across specialties (e.g., Gattas et al., 2021; Weyerstraß et al., 2020a; Ruetters et al., 2016; Weigl et al., 2021) often used as an argument for second opinion programs. Yet utilization rates remain modest. For example, survey data indicate that about 15% of adult Israelis obtain a second opinion (Shmueli et al., 2016), and telephone and hospital-patient surveys report 6–24% in Germany (Geraedets and Kraska, 2016). Because such surveys rarely apply the stringent criteria of structured second-opinion programs in classifying a second opinion, these figures are at best upper bounds. Claims data for Germany from 2019–2020 imply utilization rates well below 5% (Bruch et al., 2024), mirroring the rates observed in our sample.

Single-cohort studies of knee-arthroplasty patients (Weigl et al., 2021; Kisch et al., 2023) and a billing-based cost simulation for orthopedic indications (Weyerstraß et al., 2020b) provide some suggestive evidence that second-opinion participation is associated with lower surgical volume and lower projected costs.

Evidence on how second opinions shape realized costs is also very scarce. Two single-center studies track actual spending: A mandatory program at the Cornell/New York Hospital (1972–80; McCarthy et al. 1981) and the spine board at a São Paulo based center (Antonioli et al., 2023). The former relied on phone follow-ups without claims data; the latter covers only spine cases and loses patients who did not accept the board's advice. Both report fewer operations and lower costs, but their narrow scope limits generalization.

Three questions thus remain open: whether patients actually follow independent second opinions, whether following them changes realized costs and health outcomes rather than hypothetical ones, and whether avoided procedures stay avoided. Compliance is the hinge—without it, even a perfectly designed program cannot affect outcomes.

We contribute to this literature in two ways. First, we provide what is, to our knowledge, the first claims-based empirical evaluation of a structured surgical second-opinion program that links first and second recommendations to realized eighteen-month patient outcomes. Earlier work has either documented disagreement rates, simulated cost savings using billing data, or studied small mandatory programs without claims linkage. By

following patients across the entire decision process: (i) first recommendation, (ii) second recommendation, (iii) treatment choice, and (iv) eighteen-month claims data—we deliver realized-outcome evidence that managerial decisions about program design requires. Second, the longitudinal nature of the data lets us test directly whether observed cost savings reflect genuine reductions in overtreatment or merely deferred necessary care, a distinction that case-specific or short-horizon studies cannot make.

3.2.3 Healthcare Operations and Provider Decision-Making

A growing healthcare-operations literature documents that operational features of care delivery shift treatment intensity and patient outcomes well beyond what underlying clinical heterogeneity alone would predict (see Dai and Tayur (2020) for an overview). Freeman et al. (2017) and Freeman et al. (2021) show that gatekeeper decisions in maternity and emergency departments respond systematically to operational pressures (case mix, congestion, staffing), with measurable effects on clinical outcomes. Kuntz et al. (2019) demonstrate that separating and concentrating patient flows by complexity changes hospital performance. The common thread is that organisational design choices about who sees the patient, when, and under what constraints have first-order effects on what care is delivered.

Closest to our paper, Bobroske et al. (2022) study "provider discordance" in primary-care opioid prescribing. Using a nationwide US database of medical and pharmaceutical claims, they show that patients whose 30-day follow-up appointment is with a different clinician than the initial prescriber are roughly one third less likely to become long-term opioid users. They identify the effect using a fully controlled logistic specification, supplemented by an instrumental-variable analysis and propensity-score matching with the minimum-bias estimator—a methodological approach similar to ours. Their interpretation is operational: a simple managerial rule that re-routes follow-up patients to a different provider can shift long-term outcomes substantially.

We contribute to this stream of work in two ways. First, the discordance we study is not an incidental scheduling outcome but a deliberate program design feature: The second specialist is recruited, vetted, and contractually constrained by the insurer. Second, the second specialist is barred from acting as the treatment provider, which sharpens the underlying incentive contrast and operationalizes the separation mechanism studied theoretically by Bester and Dahm (2018). Our findings extend the discordance result of Bobroske et al. (2022) from primary-care prescription decisions to elective surgery—one of the largest discretionary cost categories in modern health systems—and link the mechanism explicitly to the operational design of the second-opinion program.

3.3 Institutional Design and Hypotheses

This section discusses the institutional design and derives the hypotheses. First, we discuss the structure of the German health care market and the institutional design of the second-opinion program analyzed (Section 3.3.1). We then map the assumptions and results of the literature onto the program (Section 3.3.2) to derive our hypothesis about expected outcomes (Section 3.3.3).

3.3.1 Second-Opinion Programs and the German Health Care Market

German Health Care Market Germany provides near-universal health coverage through a dual system of statutory and private health insurance. As of 2022, approximately 73.3 million residents (88.1%) were insured through SHI providers (VDEK, 2022). Traditionally, care provision is organized into two distinct sectors: outpatient (office-based physicians) and inpatient (hospitals). These sectors differ in their planning and reimbursement regimes: outpatient remuneration is fee-schedule based (called EBM), while hospitals receive flat reimbursement rates (called DRG). Patients typically have free choice among office-based physicians and specialists.

The Second-Opinion Program We analyze the second-opinion program of a large German SHI provider with more than three million enrollees across Germany, initiated in early 2013. The service is available to all of the SHI provider's enrollees. Initially, enrollees could only request second opinions for orthopedic conditions. Starting in 2016, additional specializations were gradually added.

To provide these opinions, the SHI recruited specialists in the relevant fields. Patients seeking a second opinion can access the list of experts and their CV. Likewise, physicians typically indicate on their hospital, practice, or private websites that they are second opinion experts in the SHI provider's program. Program coordinators ensure that the second expert is not affiliated with the same hospital as the first expert. Furthermore, the second physician must not be the surgeon performing the operation. Between 2013 and 2019, there were an average of 82 second-opinion requests per year, for a total of 572 requests.

Figure 3.1 illustrates the timeline for patients participating in the structured second-opinion program. First, patients must be diagnosed with a relevant condition. Then, they consult their initial treating physician, who provides a recommendation in favor of or against surgery. Patients can then initiate the second-opinion program by calling the SHI provider. During the call, medically trained staff provide preliminary counseling to determine if a second opinion is appropriate for the patient's specific case. If deemed appropriate, a personal consultation with a contracted second physician is scheduled within

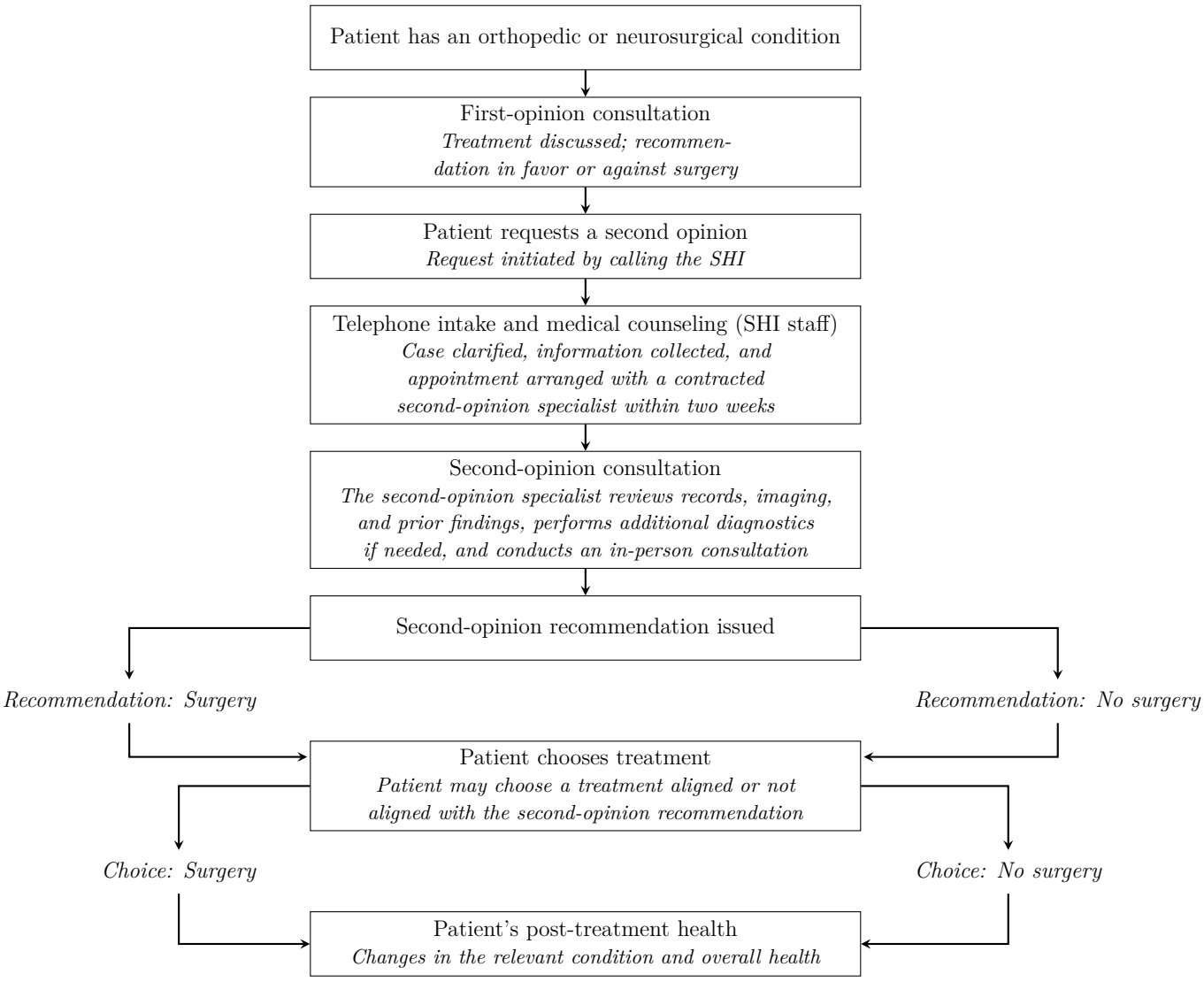


Figure 3.1: Timeline of the Structured Second-Opinion Program

two weeks. The specialist receives the patient's file for review. If further diagnostics, such as imaging, are required, the SHI provider pays for these procedures. The second physician then makes a treatment recommendation. The patient ultimately decides whether to undergo surgery. Therefore, requesting a second opinion imposes no obligation on the patient to follow it.

Other Second-Opinion Programs Second opinions have long existed in the German healthcare system as an informal practice, with patients consulting another physician regarding the same condition (Pieper et al., 2018). The 2015 Act to Strengthen Care in Statutory Health Insurance (GKV-Versorgungsstärkungsgesetz) established the legal right to obtain an independent medical second opinion for certain planned procedures.[2] Patients have the right to a second opinion when surgery is recommended for a procedure

[2]See Bundesministerium der Justiz und für Verbraucherschutz (2015) for the full text of the law.

listed in the Federal Joint Committee (G-BA) directive.[3] In such cases, the treating physician must inform the patient of their right to a second opinion before treatment. The specialist providing the second opinion must be an expert in the relevant specialty and be familiar with the current state of scientific research on the diagnosis, therapy, and possible alternatives for the procedure in question. This specialist may not perform the planned intervention. The National Association of Statutory Health Insurance Physicians (KBV) publishes a list of accredited specialists online.[4]

Alongside this mandatory framework, health insurance providers have offered voluntary second-opinion programs as discretionary benefits. In their 2016 systematic review, Pieper et al. (2018) identified 78 distinct programs across 59 of the 117 SHIs. They noted that only very small insurers did not offer second-opinion programs. These programs are concentrated in orthopedics and oncology and demonstrate substantial heterogeneity in program design and patient-facing transparency. A key distinction is who organizes and delivers the second opinion. Most programs either outsource case review to external service providers (58%) or refer patients to contracted providers (45%), while in-house delivery by insurer staff is rare (11%). Outsourced programs are usually file-based. Patient only submit medical documentation of their condition. In contrast, contract-partner models most often involve personal consultations. Furthermore, Pieper et al. (2018) report that the identity of second-opinion physicians is clearly disclosed in only 8% of programs. They conclude that, although second-opinion programs are common, they vary greatly in design, ranging from hotlines to specialist assessments.

3.3.2 Mapping the Literature to the Institutional Design

The literature on credence goods identifies two channels through which second opinions can improve patients' health outcomes. First, second opinions can, ex ante, deter fraudulent or low-effort diagnoses and treatments patients receive (*disciplinary channel*). Second, ex post, it can overturn erroneous initial recommendations when patients obtain a second assessment (*corrective channel*).

Disciplinary effects mean the threat of verification induces first physicians to invest in accurate diagnoses and report honestly. Such equilibria require consumers to seek second opinions with a sufficiently high probability. Otherwise, the first physicians face minimal detection risk and diagnostic effort collapses (Wolinsky 1993 and Pesendorfer and Wolinsky 2003). In this program, the SHI provider substantially reduces search costs that typically diminish efficiency gains from second opinions (Mimra et al., 2016). First, patients are matched with accredited specialists rather than searching independently. Second, consultations are scheduled within two weeks circumventing the long waiting times

[3] See Gemeinsamer Bundesausschuss (G-BA) (2024b) for the full list of procedures.
[4] See Kassenärztliche Bundesvereinigung (KBV) (2024) for the list of accredited second-opinion physicians.

common in the German outpatient system. Third, no out-of-pocket costs are incurred.

Despite these efforts to increase accessibility of the second opinion, patient uptake in this program and most others remains very low (2.7 requests per 100,000 enrollees per year in our sample).[5] This implies that the disciplinary effect is weak at best. Physicians recommending surgery face little risk that their patients will seek a standardized institutional review.[6] We therefore focus our analysis on the corrective channel.

Corrective effects arise when second opinions offer more accurate recommendations than initial assessments. The quality of the diagnosis depends on the incentives that drive the diagnostic efforts of the first and second physicians. Assuming both experts can provide treatment, as is standard in the literature on second opinions, they both face the same incentive structure. Therefore, second opinions may reduce the overall level of fraudulent advice, but systematic differences in the quality of first and second opinions do not occur. In most standardized programs, however, diagnostic compensation and treatment margins are separated, which can potentially align incentives toward truthful reporting (Bester and Dahm, 2018). In the German SHI context, primary care physicians operate under reimbursement systems with administratively set prices that aim for equal markups across procedures. However, economies of scale at specialized centers, bonus payments for meeting volume targets, and minimum-volume regulations likely create strong incentives for the first physician to recommend major procedures (surgery) over minor ones (conservative care). Second-opinion physicians face no such pressure. Their flat fee is independent of the recommendation and treatment choice.

3.3.3 Hypotheses Development

Next, we develop two hypotheses by mapping the literature onto our institutional context. The low program uptake implies a minimal detection probability for the first physicians. With rare consultations, the ex-ante deterrent effect of potential second opinions is negligible. Our empirical analysis considers the reception of a second opinion and examines the corrective channel. We do not hypothesize or estimate the population-level effects of program availability. However, we make predictions about the pattern of recommendations and the quality of diagnostics.

Treating physicians face asymmetric treatment margins. Due to economies of scale and volume incentives, surgery is most profitable. Therefore, theoretical and experimental literature suggests that physicians are likely to overrecommend surgery. In contrast,

[5] The low uptake rate is consistent with rates reported elsewhere (Geraedets and Kraska, 2016; Obermann et al., 2021). They argue that limited awareness among both enrollees and physicians is a key explanation.

[6] Survey evidence suggests that second opinions outside standardized programs are more common—often from another potential treatment provider: 24% of adults report having consulted a second expert at least once, and 16% report doing so within the past 12 months. Among hospitalized patients, 6% report a second opinion prior to admission (Geraedets and Kraska, 2016).

second-opinion physicians receive flat fees, regardless of their recommendations. Thus, no such incentive exists. The theoretical literature does not suggest that second-opinion physicians will overrecommend surgery. Conversely, the experimental literature suggests that separation may actually lead second opinions to underrecommend surgery.

Hypothesis 1 (Recommendation Asymmetry)**.** *Disagreement between the first and second physicians will be asymmetric. Disagreement will concentrate in cases where the first physician recommends surgery, but the second does not. Conversely, there will be high agreement when the first physician recommends against surgery.*

Theoretical and experimental literature on second opinions suggests that both primary and secondary physicians are equally likely to over-recommend surgery when they can both diagnose and treat. However, the aggregate probability of overrecommendation may fall. Bester and Dahm (2018) show that separation can sustain diagnostic effort and achieve near-first-best outcomes when diagnosis is compensated based on adequacy, independent of treatment provision. In our setting, however, compensation is a flat fee that does not depend directly on adequacy. Therefore, low diagnostic effort could persist if compensation does not depend on accuracy.[7] However, second physicians' reputation concerns or fear exclusion from the program if recommendations are bad, may be enough such that overall utility depends on the positively on the adequacy of the recommendation.

Hypothesis 2 (Diagnostic Accuracy)**.** *Second-opinion physician recommendations provide weakly more adequate treatment advice than first-opinion recommendations.*

3.4 Data and Descriptive Statistics

In this section we present the data. First, we describe the available data and the restrictions used to construct the analysis sample (Section 3.4.1). Next, we explain the construction of additional analysis variables and define the outcome measures (Section 3.4.2). Finally, we present descriptive statistics and discuss imbalances in the sample (Section 3.4.3).

3.4.1 Data and Analysis Sample

We analyze comprehensive administrative claims for all 572 enrollees who participated in a large German SHI provider's structured second-opinion program from 2013 to 2019. The dataset combines program records (first opinion reported by the patient, second opinion

[7] One might also think that second physicians will underrecommend surgery in the insurer's interest to avoid costly procedures. However, denying necessary operations might only shift or increase costs and harm an insurer's reputation in a competitive market where members can easily switch insurance providers. High-income members, who pay high premiums, are especially likely to value reliable second opinions.

documented, request date, and medical specialty) and demographics (gender, year of birth and death, insurance category indicating employment or benefit status, and approximate residence based on the first four digits of the postal code). It also includes complete claims covering inpatient and outpatient care. The first and second recommendations are classified as either recommending surgery (e.g., joint replacement for orthopedic cases) or discouraging surgery, with a small number of undocumented first opinions. Hospital records include ICD-10 diagnoses, procedure codes (called OPS), length of stay, and costs. Outpatient records include ICD-10 diagnoses, EBM service codes, and costs. The dataset also captures rehabilitation services, certified sick-leave episodes (with ICD-10 codes), prescription and administration histories with World Health Organization's (WHO) Anatomical Therapeutic Chemical (ATC) codes, therapeutic services, and aids/appliances (roughly comparable to DMEPOS in the U.S.). However, physician identifiers for second opinion experts and exact ICD-10 codes for the condition under review by the second physician are unavailable. Furthermore, during the first three program years (2013–2015), the initial treating physician's recommendation (first opinion) was not systematically documented in the program records. Therefore, for these early cohorts, we only observe the second opinion and subsequent utilization, not the corresponding first opinion.

Index Date and Observation Windows The analysis is based on the date of the second-opinion request. To be eligible for analysis, there must be a continuous 6-month look-back period before and an 18-month follow-up period after the second-opinion request. The six-month look-back period provides sufficient pre-second-opinion observations to characterize baseline health trajectories and pre-index utilization, while the 18-month follow-up period enables us to capture medium-term outcomes (e.g., deferred surgeries, rehabilitation, and recurrent visits) without excessive attrition.[8] Figure 3.2 reports exclusions due to insufficient observation periods: 254 observations were dropped, leaving 318 patients.

Specialty Focus The remaining patients span across 12 specialties (including "Other"). Often, cell sizes are too small for precise estimates (see Figure 3.2). For four reasons, we restricted the main analysis to orthopedics and neurosurgery. First, statistical power: these specialties comprise 74% of cases with adequate observation windows. Second, identifying variation: they exhibit the highest rates of physician disagreement (see Section 3.4.3). Third, clinical coherence: relevant ICD-10 indications overlap and involve elective procedures with clear surgical procedures. Fourth, these field are most prevalent and policy relevant. The literature documents pronounced regional variation (Bertelsmann-Stiftung,

[8]This design balances statistical power and outcome ascertainment against selection concerns that increase with longer windows. We assess sensitivity to the window choice in robustness checks using alternative follow-up periods (6 and 36 months). See Section 3.7.2

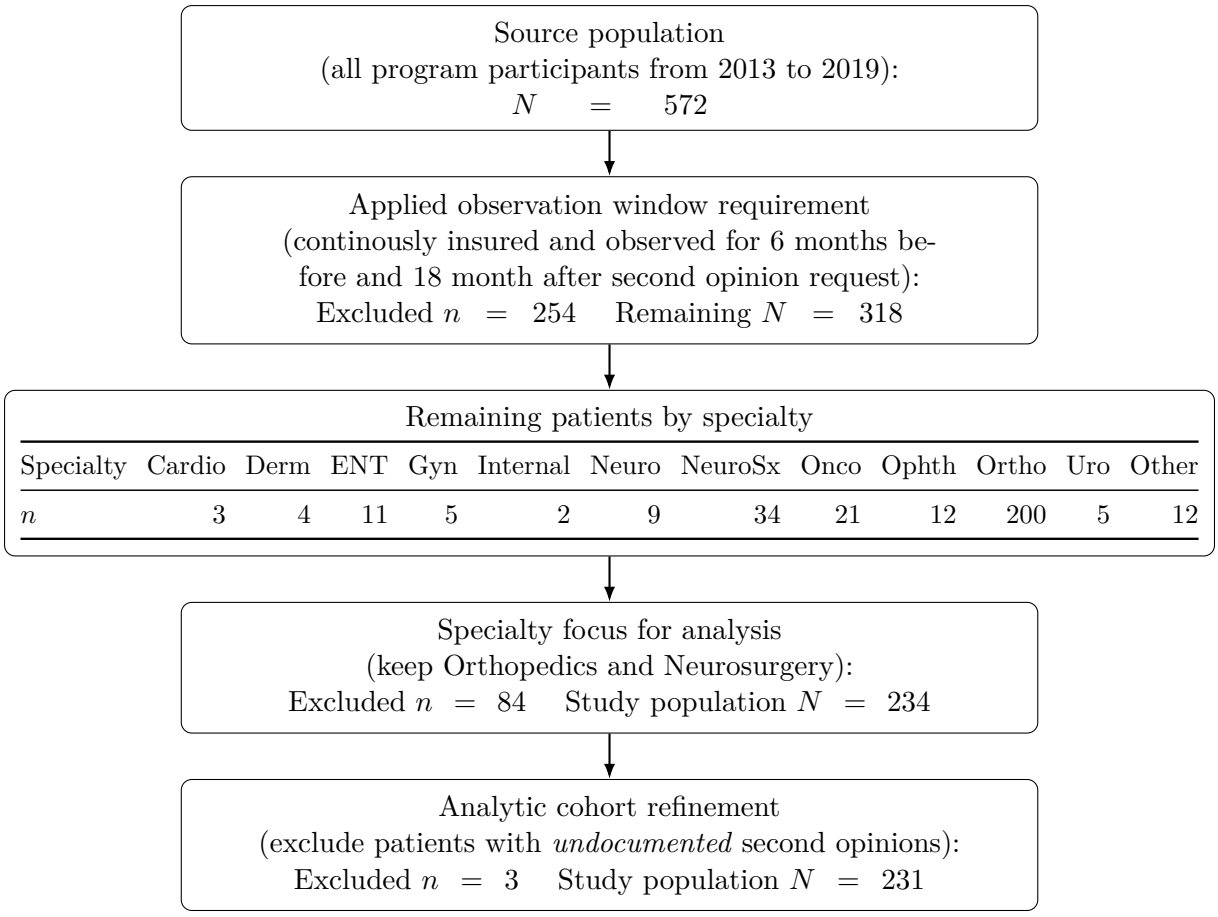


Figure 3.2: Construction of the Analysis Sample

2013; Domenighetti et al., 1993), most large insurance providers offer second-opinion programs for these conditions (Pieper et al., 2018), and survey evidence indicates especially high second opinion demand (Geraedets and Kraska, 2016; Obermann et al., 2021).

Finally, we excluded three patients for whom second opinions were *undocumented*, yielding a final study population of 231 enrollees.

3.4.2 Variables

Constructed Variables

Next, we define the study variables, which are not directly observed, but rather constructed from routine claims records. Specifically, we do the following: (i) infer the second opinion indication; (ii) build a surgery and compliance indicator; and (iii) compute Elixhauser comorbidities.

Second opinion indication Since the specific conditions and procedures under review are not documented, we infer each patient's indication from diagnoses observed during the 6-month prior to the second-opinion request. We develop a matching algorithm using a reference database of second-opinion-eligible conditions classified by specialty and

clinical plausibility (Appendix C.1.1 details our approach). The algorithm implements reliability-weighted matching that prioritizes high-confidence ICD-10 codes while accommodating multiple concurrent diagnoses. This approach successfully identifies plausible indications for 95.7% of patients. We then classify each patient into one or more body-region categories (*hand/foot, hip, knee, limb, shoulder, spine*), flag cases involving existing orthopedic implants, and assign remaining cases to *other*.

Surgery and Compliance We define a monthly surgery indicator for each patient based on documented inpatient and outpatient procedures.[9] We define patient compliance with the second expert's recommendation using the surgery indicator as follows: If surgery is recommended, compliance means undergoing surgery within 180 days of the second-opinion request. If the second expert advises against surgery, compliance means not undergoing surgery during the same period. See Appendix C.1.2 for a discussion of the threshold choice. Hereafter, patients who follow the second-opinion recommendation are classified as *compliers* and those who do not are classified as *non-compliers*.

Elixhauser Comorbidities Elixhauser comorbidity scores are constructed from ICD-10 diagnoses observed at least once in the six-month prior to the second-opinion request. The Elixhauser system identifies 31 comorbid conditions that independently predict mortality and healthcare utilization.

Residence Type Finally, four-digit postal codes are mapped to federal states and classified as either *urban* or *suburban/rural* according the Federal Ministry for Housing, Urban Development and Building.

Outcome Measures

A central challenge in evaluating second-opinion programs is that the medically optimal treatment for each patient is unobservable. Therefore, we examine outcomes across various domains to differentiate reductions in overtreatment from delays in necessary care. We do this using cross-group contrasts and tracking their evolution over time. We aggregate daily claims into 3-month periods indexed by t relative to the second-opinion month, where $t = 0$ comprises the month of the second-opinion request and the following two months, $t = -1$ the prior 3 months, etc. All outcomes are expressed as mean monthly values within each period. This aggregation makes medical histories more comparable across patients. We organize outcomes into three blocks: (i) health care measures capturing overall spending, utilization, and short-run health-related work absence; (ii) treatment-path measures that shed light on substitution between surgical and non-surgical management; and (iii)

[9] Appendix C.1.2 details the list of OPS and EBM codes that we consider.

labor-market outcomes capturing short-run recovery costs and longer-run labor-market attachment.

Health Care Measures To capture broad changes in health care utilization and overall recovery around the second opinion, we examine expenditure, medical utilization, and sickness absence. Total health expenditure is measured as mean monthly claims spending aggregated across all medical service categories and prescription drugs within each 3-month period. Utilization is summarized by mean monthly outpatient physician visits and inpatient hospital days over the same interval, which jointly capture the intensity of medical management and hospital-based treatment. As a complementary indicator of health-related work capacity, we also use certified sick-leave days. Because certified sick leave is systematically recorded only for individuals in employment, we analyze and interpret this outcome in an ever-employed subsample.

Treatment Path and Symptom Management We measure the use of prescribed painkiller-related medication, using WHO Anatomical Therapeutic Chemical (ATC) codes.[10] Analgesic use provides a proxy of symptom burden. To identify non-surgical treatment alternatives, we define conservative therapy as (i) non-physician therapeutic services (e.g., physiotherapy and occupational therapy), (ii) inpatient or outpatient rehabilitation, and (iii) conservative therapy-related medication.[11] To avoid conflating conservative management with early postoperative recovery care, we classify claims as conservative therapy only if no surgical procedure has been recorded within the preceding 90 days. Claims occurring within 90 days after the surgery are classified as postoperative rehabilitation.[12]

Labor Market We proxy labor-market attachment and medium-run economic consequences using insurance-status information. First, we construct wage-replacement payments by valuing certified sick-leave workdays among the employed at a wage proxy (median daily earnings); for absences exceeding six weeks, we apply a replacement rate of 70%, consistent with statutory sick-pay rules in Germany.[13] Second, we measure labor-

[10] See Appendix C.1.3 for the list of ATC codes that we include to identify painkiller-related medication.

[11] See Appendix C.1.3 for the list of ATC groups used to identify conservative therapy-related medication.

[12] The 90-day window aligns with common claims-based episode definitions for major orthopedic surgery, such as arthroplasty. In Germany, post-acute rehabilitation is typically approved for three weeks (Deutsche Rentenversicherung, 2026), and subsequent outpatient therapy is guided by the frequency recommendations in the Heilmittelkatalog (e.g., one to three treatments per week with a total of 18 sessions for relevant indication groups, see Gemeinsamer Bundesausschuss (G-BA) (2024a)). Physiotherapy centers often recommend a three-month episode (see for example Physiotherapie-Sendlinger-Tor, viewed January 2026). This is also consistent with international guidelines: Medicare's global surgical package uses a 90-day postoperative period for major procedures, and the CMS Comprehensive Care for Joint Replacement (CJR) model defines joint-replacement episodes as extending 90 days after discharge (Centers for Medicare & Medicaid Services, 2024, 2026).

[13] In Germany, employers continue to pay full wages during the first six weeks of absence due to illness. After six consecutive weeks, the SHI provider pays benefits equivalent to 70% of gross earnings (capped at 90% of net earnings).

market detachment using the share of months in unemployment. Third, we capture exits from the labor force using an early-retirement indicator that equals one if an enrollee is retired below the legislated regular retirement age (determined by year of birth) and zero otherwise. Together, these variables summarize short-run recovery-related income replacement and potential longer-run impacts on labor-market attachment.

3.4.3 Descriptive Statistics

Patients in the analysis sample requested a second opinion at an average age of 57 years, which is approximately 12 years older than the national and statutory health insurance population average (Statistisches Bundesamt (Destatis), 2024b).[14] The gender distribution is balanced (51.9% male and 48.1% female). Patients from urban areas constitute about 50.2%, patients from suburban constitute 47.6%, and patients from rural areas constitute 2.2% of our sample. Relative to national figures, urban and suburban patients are overrepresented (28.4% and 42.6% of the national population respectively) while rural patients (29% nationally) are underrepresented (Statistisches Bundesamt (Destatis), 2024a). Regarding insurance status, 40.2% of enrollees were employed, and 7.4% received insurance as dependents. This is compared to 64% of all German SHI enrollees with combined employment-based coverage. Pensioners comprise 35.9% of our sample versus 23.7% in the overall German SHI pool (Kassenärztliche Bundesvereinigung (KBV), 2023). Additionally, 9.1% received social benefits and 7.4% were in vocational training or attending university.

Figure 3.3 summarizes the recommendation patterns of the first and second opinion. As expected, the recommendations differ substantially. Among program participants, 129 out of 231 (55.8%) have received recommendation for surgery, while only 21 out of 231 (9.1%) received a recommendation against surgery from the first physician. We assume that similar recommendation patterns hold for the 81 cases (35.1%) where the first opinion is undocumented.[15] In contrast, only 77 out of 231 (33.3%) of patients received a recommendation in favor of surgery from the second physician while 154 out of 231 (66.7%) received a recommendation against surgery.

Accordingly, disagreement between the first and second physician accounts for 56% (84 out of 150) of the cases in which we observe both recommendations.[16] Importantly, disagreement is strongly asymmetric. When the first physician recommends against surgery, the second opinion concurs in 85.7% of cases (18 out of 21). By contrast, when the first

[14] The sample's mean age is comparable to that of prior second opinion cohorts (e.g., Weyerstraß et al. (2020a): mean 58.2, SD 15.1).

[15] The relatively large number of *undocumented* first opinions reflects the coding during the first three program years, as discussed in Section 3.4.1

[16] In non-orthopedic and non-neurosurgery cases, we observe substantially lower disagreement rates (e.g., 18% in oncology and approximately 37% across other specialties). However, the rates and patterns align with non-restricted sample in orthopedics and neurosurgery.

		Second Opinion		
		Surgery	No Surgery	Total
First Opinion	Surgery	48	81	129
	No Surgery	3	18	21
	Undocumented	26	55	81
	Total	77	154	231

Figure 3.3: Recommendations of First and Second Opinion

physician recommends surgery, the second opinion disagrees in 62.7% of cases (81 out of 129). This directional asymmetry is consistent with physician-induced demand and Hypothesis 1. We summarize:

Observation 1 (Disagreement is asymmetric and concentrates when surgery is initially recommended)**.** *Physician recommendation and disagreement patterns are strongly asymmetric. Second, physicians concur with recommendations against surgery in most cases (18 out of 21), but disagree with recommendations in favor of surgery in most cases (81 out of 129).*

Table 3.1: Patients' Treatment Choices by Physician Recommendations

First opinion	**Second opinion**	**Surgery rate**	***n***
No surgery	No surgery	0.278 (0.160)	18
No surgery	Surgery	0.333 (0.439)	3
Undocumented	No surgery	0.200 (0.086)	55
Undocumented	Surgery	0.667 (0.152)	21
Surgery	No surgery	0.185 (0.070)	81
Surgery	Surgery	0.417 (0.102)	48

Table 3.1 shows how patients made treatment decisions based on the recommendation pattern. Patients comply with the first opinion in less than one-third of cases (27.3%, 41 out of 150). Conversely, patients comply with the second opinion in more than two-thirds of cases (69.8%, 160 out of 231).[17] When the first and second opinions disagree, patients choose the first physician's recommendation in 20.2% of cases and the second

[17]Importantly, the compliance rates are not driven by restriction imposed on the sample discussed in Section 3.4.1. In the full sample of orthopedic and neurosurgical second-opinion request, 67.4% comply with the recommendation of the second physician.

physician's recommendation in 79.8% of cases. Notably, compliance is higher with any recommendation against surgery than with a recommendation in favor of surgery. If the second opinion recommends against surgery, patients comply in 79.9% (123 out of 154) of cases, and if the first opinion recommends against surgery, patients comply in 71.4% (15 out of 21) of cases. Conversely, if the second opinion recommends surgery, patients only comply in 46.8% (36 out of 77) of cases and 27.1% (25 out of 129) of cases, respectively. Interestingly, agreement between physicians does not strictly increase compliance with the second opinion. We summarize:

Observation 2 (Second opinions predict treatment decisions.)**.** *Patients are more likely to comply with the second physician's recommendation (69.8%) than the first physician's (27.3%). When physicians disagree, nearly 80% of patients follow the second opinion. Compliance is substantially higher for recommendations against surgery (79.9% for the second opinion and 71.4% for the first opinion) than for recommendations in favor of surgery (46.8% and 27.1%, respectively).*

Figure 3.4 summarizes mean monthly health outcomes and 95% confidence intervals by compliance status over the observation window. At baseline, mean monthly total health expenditures are €686 for non-compliers and €518 for compliers. Physician visits average around 2.7 for non-compliers and 3.2 for compliers per month. Patients spend 0.1 to 0.2 days in the hospital per month and certified sick leave averages 4 to 6 days per month among employed patients. Compliers and non-compliers experience a pronounced cost increase around the second opinion 3-month period (peaking at €991 to €1,084 for non-compliers and €669 to €847 for compliers in 3-month periods 0 to 1) before gradually returning toward baseline levels by nine to twelve months after the second opinion. Hospital days show more modest variation over time, while sick leave days and physician visits exhibits a declining trend for compliers and non-compliers. Confidence intervals overlap substantially across outcomes and periods, though means suggest baseline differences in cost levels and utilization patterns between compliers and non-compliers, motivating a balance check across groups.

Table C.1 in the Appendix summarizes the differences in baseline characteristics by compliance status. While compliers and non-compliers are similar in terms of demographics, age, residence, and Elixhauser comorbidities, they differ significantly along several dimensions. First, as documented in Observation 2, compliance is strongly associated with recommendation pattern: patients are substantially more likely to comply when the second opinion advises against surgery. Second, imbalances in relevant conditions emerge. Compliers more often have shoulder- and spine-related conditions (18.2% vs. 6.9% and 70.4% vs. 54.2% with standardized differences of about 0.3), while non-compliers more frequently have implant-related issues (15.3% vs. 5.0% with standardized difference -0.35). Third, compliers were slightly more frequent in 2015. Further, they exhibited higher pre-

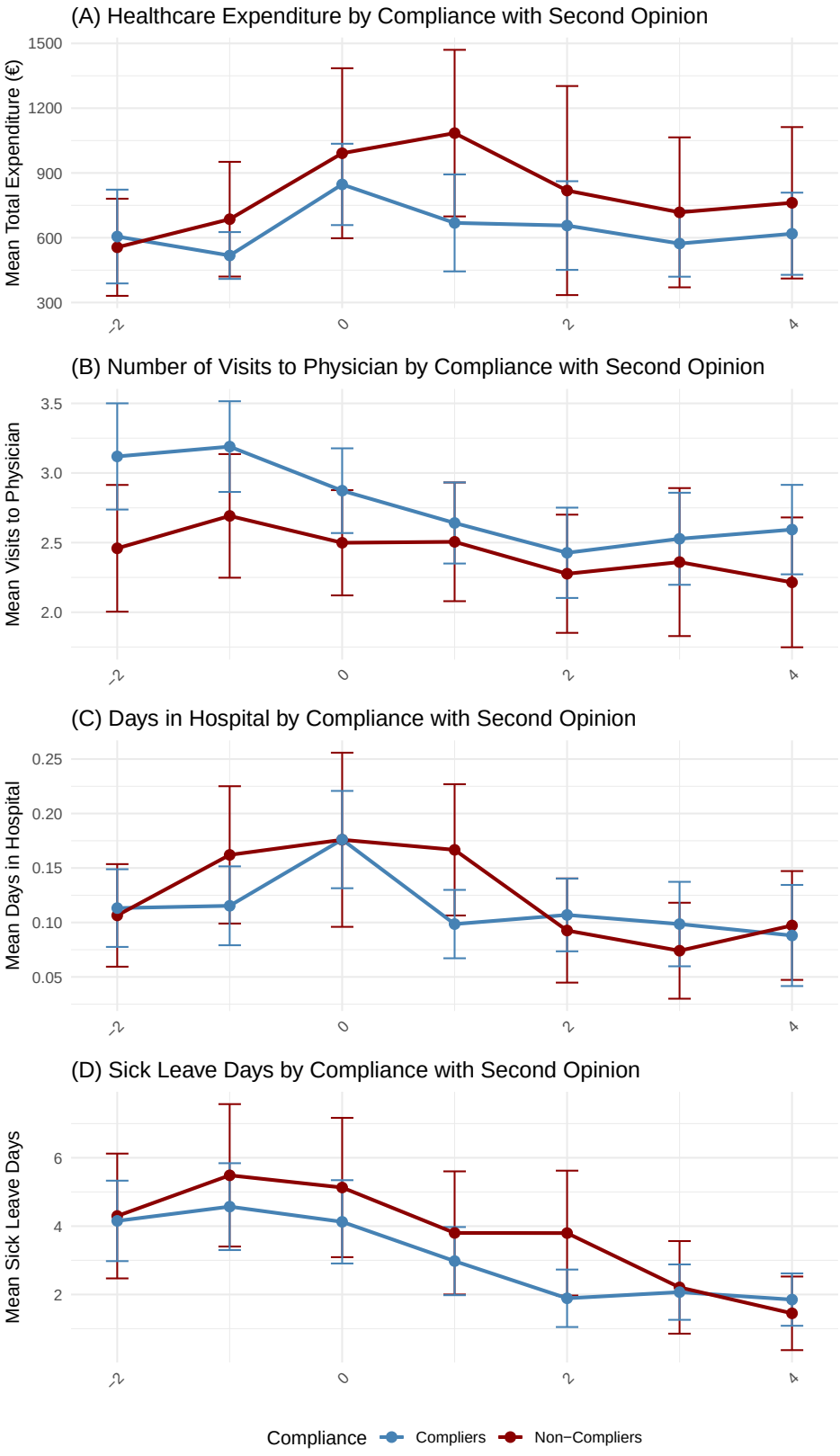


Figure 3.4: Health Outcomes by Compliance with Second Opinion.

period spending on painkillers and modestly more visits to the physician. Meanwhile, non-compliers show somewhat higher baseline surgery rates (6% vs. 3%). These imbalances suggest that selection into compliance is non-random and occurs along clinically meaningful dimensions.

3.5 Empirical Strategy

In this section, we present the empirical strategy. First, we discuss identification challenges (Section 3.5.1). Then, we present a matching procedure to address covariate imbalances (Section 3.5.2) and describe the event-study specification (Section 3.5.3). Finally, we discuss identifying assumptions (Section 3.5.4).

3.5.1 Identification Strategy

Due to two levels of self-selection, identifying the causal effect of structured second-opinion programs on health outcomes is challenging. First, patients self-select when deciding to request a second opinion. Second, conditional on receiving a second opinion, patients choose whether to comply with the treatment recommendation. Therefore, it is not possible to compare second-opinion recipients and non-recipients, or compliers and non-compliers, without addressing selection bias.

In an ideal setting, patients would be randomly assigned to receive a second opinion when the first physician recommends against or in favor of surgery. Comparing the post-second-opinion outcomes, Y_i, of patients with a second opinion ($SO = 1$) and those without access to a second opinion ($SO = 0$) would identify the average effect of receiving a standardized second opinion

$$\tau^{\text{SO}} = E[Y_i(SO = 1) - Y_i(SO = 0)].$$

To our knowledge, no structured second-opinion program has ever been implemented that randomly assigns patients to receive a second opinion. Therefore, no such data is available. Even if researchers or insurers attempted to implement such a program, randomization would face significant challenges. First, real-time integration into clinical consultations is required to assign second opinions immediately after the first consultation. However, insurance providers' data only records procedures that have been performed, not when surgery is discussed for the first time or what recommendations physicians make. Second, imposing or withholding second opinions from patients would lead to ethical and practical challenges. For example, German patients have the right to seek second opinions, which makes randomized denial problematic from both legal and ethical standpoints.

The natural alternative would be to compare participants to non-participants from the same insurer. However, this is infeasible for two reasons. First, we cannot observe

when surgery is initially discussed with non-participants. This makes it impossible to identify the date of the hypothetical second-opinion request. Any temporal alignment of outcomes would be arbitrary. Second, we do not observe the initial recommendations of non-participants. We cannot distinguish between those whose physicians recommended surgery and those whose physicians did not. Thus, it is impossible to match on initial treatment recommendation.

Instead, we base our analysis on program participation, taking into account both the recommendations and the timing of the request. We exploit variation in compliance with second-opinion recommendations. Let $C_i \in \{0, 1\}$ be a binary variable indicate compliance among second-opinion recipients.[18] We estimate the average effect of following second opinion advice among program participants

$$\tau^{\text{comply}} = E\big[Y_i(C_i = 1) - Y_i(C_i = 0) \,\big|\, SO = 1\big].$$

Notably, τ^{comply} is a treatment effect among program participants. It measures the value of following the second opinion rather than defying it, conditional on having sought one. It is not the population effect of offering the program, which would additionally depend on selection into participation, on take-up, and on any effects of the second opinion on non-compliers, none of which our design identifies. We therefore do not extrapolate our estimates to program-level effects and return to the policy implications, with this caveat in mind, in Section 3.8. Within this scope, the central identification challenge is selection into compliance: patients who follow the second opinion may differ systematically from those who do not. The remainder of this section addresses this selection.

3.5.2 Constructing a Balanced Sample Using Matching

To address selection into compliance, we implement a two-stage empirical strategy. First, we estimate propensity scores for compliance using pre-second-opinion covariates. Second, we use these scores to construct a matched sample of compliers and non-compliers who are balanced on observables.

In the first stage, we estimate the probability of compliance using only information available prior to the second opinion

$$\Pr(C_i = 1 \mid Z_i) = \Lambda(Z_i'\theta), \tag{3.1}$$

with $\Lambda(.)$ the logistic cumulative distribution function. The covariate vector Z_i i includes demographics (age, gender, and urban or rural residence), a comprehensive medical history (including all ICD-10 diagnoses from the six-month pre-period, affected body regions, orthopedic implants, and Elixhauser comorbidity scores), and detailed utilization mea-

[18] See Section 3.4.2 for the detailed definition of compliance.

sures. For utilization, we record total health expenditures, physician visits, hospital days, and surgery indicators separately for each of the two pre-second-opinion 3-month periods ($t = -2$ and $t = -1$), and we calculate trends by taking the difference between these 3-month periods. We also include the first and second physicians' recommendations (surgery, no surgery, or undocumented), insurance status (measured as the percentage of pre-period months in each category: employed, retired, family or other coverage, and unemployed), and medical specialty (orthopedics or neurosurgery). Thus, we capture individual predictors, health status, and disease severity before the second opinion.

We use elastic net penalized logistic regression with a penalty parameter $\alpha = 0.5$ (equal weighting of L_1 and L_2 penalties) to handle the high-dimensional covariate space of 74 variables (including trends and interactions). The penalty parameter λ is selected by 10-fold cross-validation with the area under the curve (AUC) as performance metric.

In the second stage, we use estimated propensity scores to construct a matched sample through full matching. We impose exact matching on the first and second physician recommendations and medical specialty. This ensures that we only compare patients who received the same pattern of expert advice in the same specialty. Importantly, by matching on recommendation patterns, we overcome the fundamental limitation that prevented valid comparisons with non-participants. Now, we can directly control for the initial clinical judgment (first physician's recommendation), which remains unobserved for enrollees who never entered the program.

Our distance metric uses the propensity score as the primary matching criterion, supplemented by a Mahalanobis distance component based on baseline total health expenditures, physician visits, hospital days, surgery frequency, and Elixhauser scores measured at $t = -2$ and $t = -1$. The Mahalanobis component ensures tight balance on our primary outcome variables. To exclude poor matches, we impose a 0.2 standard deviation caliper on the propensity score (Austin (2011), Rosenbaum and Rubin (1985)). We consider full matching, nearest-neighbor matching, and optimal matching, selecting the method that maximizes the number of observations and covariate balance. Section 3.6.1 reports the resulting matched sample; detailed comparisons across matching methods are provided in Appendix C.2.3.

3.5.3 Matched Event Study

We estimate weighted patient fixed effects event-study regressions on the matched sample. Compliers receive unit weight, while each non-complier receives the full-matching weight that reflects the number of compliers in his matched set, so that the reweighted non-compliers mirror the covariate distribution of the compliers.

$$Y_{ip} = \alpha_i + \sum_{p \in \mathcal{P},\, p \neq baseline} \Big[\delta_p \, \mathbb{1}\{p\} + \beta_p \, \mathbb{1}\{p\} \cdot C_i\Big] + \varepsilon_{ip}, \tag{3.2}$$

where Y_{ip} is the mean monthly outcome (e.g., total health expenditures, physician visits) for patient i in period p, and α_i are patient-fixed effects. p indexes periods that aggregate 3-month periods into clinically meaningful phases. The indicator $\mathbb{1}p$ equals one if the observation corresponds to period p. C_i is the binary compliance indicator, δ_p captures the time path of non-compliers relative to the *baseline period*, β_p captures the average treatment effect on the treated (ATT) of compliance in each period p for matched program participants. We use robust standard errors clustered at the patient level.

To reduce noise introduced by billing timing, we aggregate 3-month periods t into five mutually exclusive analysis periods p. First, p =*pre-baseline* corresponds to 3-month period $t = -2$ and captures the outcomes four to six months before the second-opinion request. This period is used to test for pre-trends. Second, p =*baseline* corresponds to 3-month period $t = -1$ and captures the outcomes in the three months before the second-opinion request. This is our omitted reference category. Third, p =*short-term* corresponds to 3-month period $t = 0$ and $t = 1$ and captures the outcomes in the first six months after the second-opinion request. Fourth, p =*medium-term* corresponds to 3-month periods $t = 2$ and $t = 3$ and captures the outcomes in the seven to twelve months after the second-opinion request. Fifth, p=*long-term* corresponds to 3-month periods $t = 4$ and $t = 5$ and captures the outcomes in months thirteen to eighteen after the second-opinion request.[19]

To improve compactness and address the limitations of small cell sizes on event-study power, we also estimate a collapsed specification

$$Y_{it} = \alpha_i + \delta \cdot \text{post}_{it} + \beta \cdot \text{C}_i \cdot \text{post}_{it} + \varepsilon_{it}, \tag{3.3}$$

where $\text{post}_{it} = \mathbb{1}\{t \geq 0\}$ now represents the average compliance effect across all post-second-opinion periods. We use robust standard errors clustered at the patient level. Both specifications are reported in all main results tables.

3.5.4 Discussion of Identifying Assumptions

Our empirical strategy relies on several identifying assumptions that we discuss here. First, we assume conditional unconfoundedness. That is, conditional on the propensity score and the exact matching variables, compliance is independent of potential outcomes

$$Y_i(C_i = 1),\, Y_i(C_i = 0) \;\perp\!\!\!\perp\; C_i \;\big|\; \hat{e}(Z_i), FO_i, SO_i, \text{Speciality}_i.$$

where $\hat{e}(Z_i) = \widehat{P}(C_i = 1 \mid Z_i)$ is the estimated propensity score. While this assumption cannot be formally tested, we argue that our rich set of covariates set (detailed medical histories, utilization patterns, trends in health status and recommendation patterns) and

[19] We also test for robustness with additional periods p in Section 3.7.

flexible estimation method allow us to capture the complex relationship. In Section 3.6.1, we provide details and demonstrate that propensity scores predict well with high performance metrics.

Secondly, we assume common support. The propensity scores must overlap sufficiently between compliers and non-compliers

$$0 < P(C_i = 1 \mid Z_i) < 1.$$

Figure C.2 in the Appendix shows that the propensity score distributions substantially overlap. After matching, no matched patient has a propensity score within 0.05 of either boundary.

Third, there are no spillovers. A patient's compliance decision and outcome do not depend on the compliance decisions of other patients, i.e., $Y_i(Ci, C_j) = Y(C_i)$. The program's low uptake rate of 2.7 per 100,000 enrollees annually makes patient-to-patient spillovers unlikely. The program operates nationwide, with participating first- and second-opinion physicians dispersed across practices and hospitals throughout Germany, which reduces the likelihood of spillovers at the physician level.

Fourth, we assume parallel trends given matching. In the absence of differential compliance, matched compliers and non-compliers would follow parallel outcome trajectories. We test this assumption by using the pre-baseline period ($t = -2$) as a placebo treatment. We demonstrate that there are no placebo differences across outcomes.

Fifth, we assume there are no time-varying unobserved confounders. Patient fixed effects control for any time-invariant heterogeneity in disease progression rates. After controlling for these effects, there are no unobservables affecting both compliance and outcomes over time. To mitigate this issue, we control for time-varying insurance status in robustness specifications.

To further show robustness, we consider different observation windows for outcomes, and implement an instrument variable approach (Section 3.7).

3.6 Results

We present our main findings in four steps. First, we describe the final sample that was constructed using a propensity score model and an exact matching procedure (Section 3.6.1). Second, we report the main effects of compliance on health expenditures, utilization, and sick leave (Section 3.6.2). Third, we examine whether effects are heterogeneous depending on the second-opinion recommendation (Section 3.6.3).Finally, we investigate the different treatment pathways induced by complying with the second opinion (Section 3.6.4).

3.6.1 Construction of the Matched Sample

In the first stage, the elastic net propensity score model retains approximately 30 non-zero coefficients when the cross-validation error is minimized.[20] The strongest predictors are clinically plausible. First, it includes the recommendation of a second physician and the body regions in question (e.g., shoulder, hand/foot, and implants). Second, it includes overall health patterns, specific comorbidities (notably, solid tumors and peripheral vascular disease), the number of surgeries undergone, the number of days spent in the hospital, and the number of physician visits. Finally, the share of pre-period months spent unemployed strongly predicts compliance. Trends in comorbidities and total health expenditures also contribute, but with smaller coefficients.[21]

The model performs well. The elastic-net model at $\lambda_{\min}$ achieves an AUC of approximately 0.86. It correctly classifies 84% of patients, improving on a majority-class benchmark of 68% accuracy by more than 16 percentage points. The more parsimonious $\lambda_{1\text{se}}$ specification similarly yields an AUC of approximately 0.86 with an accuracy of 82%, selecting a nearly identical set of covariates. Estimated propensity scores show substantial overlap between compliers and non-compliers (see Figure C.2 in the Appendix) and a quintile-based check confirms that the score is well-calibrated: the observed compliance rate rises monotonically from approximately 13% in the lowest propensity quintile to nearly 100% in the highest quintile (Table C.3 in the Appendix).

In the second stage, the combination of exact matching and propensity score yields the final sample containing 137 patients, including 94 compliers and 43 non-compliers. This corresponds to 59.3% of the study population, maintaining the same rate of compliers (69%). Full matching offers the best trade-off between covariate balance and sample size among the designs considered. Approximately 86.5% (64 out of 74) of covariates have absolute standardized mean differences below 0.2 after matching. The remaining imbalances mainly arise in insurance-status variables, where absolute differences are small but standardized mean differences are inflated by low cell counts (see Table C.5 in the Appendix). However, to ensure robustness, we report estimated second-opinion effects both with and without insurance status controls.[22]

3.6.2 Main Effects of Compliance

Table 3.2 shows the estimated effect of complying with second-opinion recommendations on monthly total health expenditure, physician visits, hospital days, and certified sick leave days. Columns (1)–(6) use the full matched sample of 137 patients (94 compliers, 43 non-compliers). Columns (7)–(8) restrict the sick leave analysis to the ever-employed

[20] The more parsimonious $\lambda_{1\text{se}}$ specification keeps 27.

[21] The full set of selected variables and coefficients and propensity score diagnostics is reported in Appendix C.2.2.

[22] Detailed balance tables and *Love plots* for all designs are reported in Appendix C.2.3.

subsample (76 patients), for whom certified sickness absence is reliably measured. For each outcome, we report specifications with and without insurance status controls. Period indicators trace non-compliers' trajectories relative to baseline ($p = baseline$), while interaction terms capture differential changes for compliers. The pre-second-opinion placebo coefficients (*complied* × *pre-baseline*) are statistically zero across all outcomes, validating parallel pre-trends.[23]

In the short-term, compliers exhibit markedly lower spending and utilization than non-compliers. Non-compliers' monthly total health expenditure increases by approximately €1,322 relative to the baseline period (column (1), $p < 0.01$). Compliers monthly total health expenditure of compliers increases by approximately €1,209 less than non-compliers ($p < 0.001$). Considering the mean monthly baseline expenditure of approximately €408.24, this implies that, in the first six months after the second opinion, the monthly expenditure of compliers increases only modestly, whereas the monthly expenditure of non-compliers spikes very sharply. Indeed, the difference is three to four times the baseline mean.

Compliers also have 0.85 fewer outpatient visit per month than non-compliers (column (3), $p < 0.1$) in the short-term. Hospital utilization shows a similar pattern. Non-compliers' spend 0.21 more days in the hospital per month than in the baseline period (column (5), $p < 0.01$). Conversely, compliers spend 0.21 fewer days per month than non-compliers ($p < 0.01$). Thus, hospital utilization remains close to its low pre-period level. Across these three outcomes, adding insurance-status controls leaves the dynamic coefficients virtually unchanged, which reinforces the idea that residual insurance imbalances are not the main driver of the results.

Beyond the short-term, the dynamic estimates suggest no systematic medium- or long-term divergence between compliers and non-compliers. In the medium-term (seven to twelve months after the second opinion) and long-term (thirteen to eighteen months after the second opinion), the effects of compliance on total monthly health expenditures, visits, and hospital days are small and statistically indistinguishable from zero. In sum, we observe a temporary reduction in intensive and costly care among compliers during the first six months, with no evidence of subsequent "catch-up" spending or utilization over the subsequent year and a half.

The dynamic coefficients for sick leave days show substantial declines in the medium and long-term for non-compliers, reflecting general return-to-work patterns after surgery and rehabilitation (column (7), $p < 0.1$). We find no differential effect for compliers. The magnitude of the estimates is sensitive to controlling for average employment intensity. In particular, the sick-leave coefficients shift somewhat once we include the employment-share control in column (8), which is natural because sick leave can only be recorded in months of employment. The pattern of medical spending and utilization remains

[23] Figure C.4 displays the weighted outcome trajectories in the matched sample.

Table 3.2: Health Outcomes by Compliance

	Total health expenditure		Visits to physician		Days in hospital		Sick leave days	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dynamic DiD								
pre-baseline	90.32	90.42	-0.15	-0.15	0.01	0.01	-0.88	-1.07
	(73.93)	(74.32)	(0.30)	(0.30)	(0.03)	(0.03)	(2.54)	(2.60)
short-term	1,321.98***	1,337.48***	0.35	0.33	0.21***	0.21***	-0.63	1.69
	(316.42)	(315.85)	(0.40)	(0.38)	(0.05)	(0.05)	(2.06)	(2.25)
medium-term	29.74	47.87	-0.51*	-0.54*	-0.03	-0.03	-6.10*	-2.35
	(51.67)	(60.85)	(0.31)	(0.31)	(0.02)	(0.02)	(3.44)	(3.73)
long-term	289.88	351.34	-0.16	-0.22	-0.01	-0.01	-12.22***	-8.46*
	(202.03)	(220.91)	(0.37)	(0.39)	(0.03)	(0.03)	(4.14)	(4.84)
complied × pre-baseline	-161.25	-161.01	0.03	0.03	-0.05	-0.05	0.53	0.55
	(117.66)	(118.07)	(0.37)	(0.37)	(0.04)	(0.04)	(2.93)	(2.95)
complied × short-term	-1,209.01***	-1,228.66***	-0.85*	-0.81*	-0.21***	-0.21***	-1.72	-3.42
	(330.66)	(330.89)	(0.45)	(0.44)	(0.06)	(0.06)	(2.37)	(2.53)
complied × medium-term	4.51	-17.92	-0.38	-0.34	0.01	0.01	-0.63	-1.84
	(108.57)	(111.91)	(0.38)	(0.38)	(0.04)	(0.04)	(3.86)	(4.19)
complied × long-term	-329.52	-388.18	-0.57	-0.50	-0.05	-0.05	4.69	3.78
	(220.79)	(235.17)	(0.44)	(0.46)	(0.04)	(0.04)	(4.56)	(5.20)
Static DiD								
complied × post	-430.72***	-478.11***	-0.61*	-0.57*	-0.06*	-0.06*	0.51	-0.89
	(156.42)	(175.53)	(0.34)	(0.34)	(0.03)	(0.03)	(3.07)	(3.62)
Insurance-status controls	No	Yes	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	685	685	685	685	685	685	380	380
Dep. Var. Mean ($p =$*baseline*)	408.24	408.24	2.84	2.84	0.10	0.10	12.25	12.25
No. Patients	137	137	137	137	137	137	76	76
No. Complier	94	94	94	94	94	94	53	53
No. Non-Complier	43	43	43	43	43	43	23	23

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), and *long-term* ($t \in \{4, 5\}$). Two columns per outcome report specifications without vs. with insurance-status controls for outcomes. Columns (7) and (8) use the ever-employed subsample. Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

unaffected. Overall, we find that compliance does not generate additional detectable gains in work capacity beyond the general improvement observed for non-compliers.

Finally, the *Static DiD* estimation provides a concise summary of the average effect of compliance across all post-second-opinion periods. We estimate a substantial reduction in average monthly total health expenditures for compliers in the post-second-opinion period (approximately $-€431$ without and $-€478$ with insurance-status controls, $p < 0.01$). These savings correspond to the mean pre-period monthly total health expenditures. For physician visits, the *static DiD* suggests an average monthly reduction of about 0.61 visits per month in the post-second-opinion period ($p < 0.1$). The number of days spent in the hospital decreases by an average of 0.06 days per month during the post-second-opinion period ($p < 0.1$). This is consistent with a persistent reduction in inpatient care. The static effects for sick leave days are small and insignificant. In summary, our findings are:

Result 1 (Compliance positively affect health measures without delayed health trade-offs)**.** *Over 18 months, complying with second-opinion recommendations reduces monthly health-care expenditure by* €431 *to* €478 *($p < 0.01$, more than total baseline expenditure), monthly outpatient physician visits by* 0.61 *($p < 0.1$), and monthly hospital days by* 0.06 *($p < 0.1$), with no significant effects on sick leave.*

(a) ***Short-term effects:*** *Compliers spend* €1,209 *less per month than non-compliers ($p < 0.01$), have* 0.51 *fewer outpatient visits per month ($p < 0.10$), and spend* 0.21 *fewer days per month in the hospital ($p < 0.01$).*

(b) ***Medium- and long-term effects:*** *Differentials in expenditure, visits, and hospital days are statistically zero, indicating persistent benefits without catch-up spending or delayed health deterioration.*

3.6.3 Heterogeneous Effects by Second-Opinion Recommendation

The average effects documented in Result 1 combine two clinical pathways: patients who avoid surgery following a recommendation against it and patients who undergo surgery following a recommendation in favor of it. Thus, they may mask important heterogeneity. The asymmetry in expert recommendations (Observation 2)—86% of unambiguous first opinions are in favor of surgery—suggest different effects. If second opinions prevent overtreatment, we should observe significant benefits among patients advised against surgery. Additionally, if second opinions also provide value by confirming and reassuring patients about necessary surgery, we should observe benefits among patients whose initial surgery recommendation is validated. To distinguish these channels, we estimate the effects separately for patients whose second opinion *discouraged surgery* and for those whose second opinion *recommended surgery*.

Table 3.3: Heterogeneity in Health Outcomes – Second Opinion: Surgery Discouraged

	Total health expenditures		Visits to physician		Days in hospital		Sick leave days	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dynamic DiD								
pre-baseline	34.06	34.06	-0.15	-0.15	0.02	0.02	0.19	0.19
	(77.76)	(78.16)	(0.35)	(0.36)	(0.03)	(0.03)	(3.03)	(3.04)
short-term	1,609.67***	1,635.41***	0.40	0.41	0.28***	0.28***	2.12	4.93**
	(342.55)	(338.49)	(0.46)	(0.44)	(0.05)	(0.05)	(1.85)	(2.25)
medium-term	48.91	80.96	-0.58*	-0.57	-0.01	-0.01	-2.11	1.92
	(57.50)	(71.43)	(0.34)	(0.35)	(0.02)	(0.02)	(3.04)	(4.26)
long-term	307.39	404.65	-0.24	-0.29	0.01	0.01	-10.02**	-6.06
	(236.18)	(262.40)	(0.41)	(0.45)	(0.02)	(0.03)	(4.81)	(6.15)
complied × pre-baseline	-145.23	-145.16	0.08	0.08	-0.05	-0.05	0.48	0.28
	(114.79)	(115.39)	(0.42)	(0.42)	(0.05)	(0.05)	(3.36)	(3.33)
complied × short-term	-1,718.31***	-1,748.16***	-0.97*	-0.94*	-0.31***	-0.31***	-5.68**	-7.75***
	(351.55)	(347.14)	(0.51)	(0.49)	(0.06)	(0.06)	(2.20)	(2.51)
complied × medium-term	-84.12	-113.70	-0.32	-0.27	-0.03	-0.04	-5.17	-6.75
	(109.67)	(117.17)	(0.42)	(0.42)	(0.04)	(0.04)	(3.54)	(4.62)
complied × long-term	-337.33	-420.28	-0.53	-0.43	-0.08*	-0.07*	2.61	1.39
	(258.32)	(276.49)	(0.49)	(0.52)	(0.04)	(0.04)	(5.24)	(6.46)
Static DiD								
complied × post	-640.64***	-704.09***	-0.65*	-0.59	-0.11***	-0.12***	-2.99	-4.61
	(166.41)	(187.91)	(0.37)	(0.37)	(0.03)	(0.03)	(3.22)	(4.35)
Insurance-status controls	No	Yes	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	500	500	500	500	500	500	285	285
Dep. Var. Mean (p =*baseline*)	395.75	395.75	2.88	2.88	0.10	0.10	11.48	11.48
No. Patients	100	100	100	100	100	100	57	57
No. Complier	78	78	78	78	78	78	47	47
No. Non-complier	22	22	22	22	22	22	10	10

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), and *long-term* ($t \in \{4, 5\}$). Two columns per outcome report specifications without vs. with insurance-status controls for outcomes. Columns (7) and (8) use the ever-employed subsample; column (8) additionally includes an employment-status control. Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

Table 3.3 shows the results for patients who were advised against surgery by the second physician. Non-compliers experience a significant short-term increase in total health expenditures and hospital utilization. Compared to the baseline period, short-term monthly total health expenditures for non-compliers increase by approximately €1,610 (column (1), $p < 0.01$), and hospital days increase by approximately 0.30 days per month (column (5), $p < 0.01$). In contrast, the monthly total health expenditures of compliers are €1,718 per month (column (1), $p < 0.01$) less than those of non-compliers, and hospital utilization is 0.31 days per month lower (column (5), $p < 0.01$). Thus, complying with the second opinion almost completely avoids the short-term increase in expenditures and hospital utilization observed for non-compliers.

In the first six months, following the second opinion reduces outpatient physician visits by approximately 0.94 visits per month (column (3), $p < 0.10$). Compliers also accumulate roughly six to eight fewer sick leave days per month in the short-term than non-compliers (columns (7) and (8), $p < 0.05$). In the medium and long-term, there are only small, typically statistically insignificant differences across outcomes for complying with a second opinion. There is only a modest residual difference in days spent in the hospital in the long-term. Therefore, when the second opinion advises against surgery, compliance is associated with significant short-term savings in spending and hospital utilization without any apparent adverse consequences within 18 months of the second opinion.

The corresponding static DiD estimates support this pattern. On average, complying with a second-opinion recommendation against surgery is associated with €641 to €704 lower total monthly health expenditures (columns (1) and (2), $p < 0.01$). Compliers also spend approximately 0.11 to 0.12 fewer days per month in the hospital (columns (5) and (6), $p < 0.01$) and visit outpatient physicians 0.65 fewer times per month (column (3), $p < 0.1$).

Table 3.4 shows the results for the 37 patients who received a recommendation in favor of surgery. Due to the small sample size, we interpret our estimates cautiously. Note that in this case, non-compliers are patients who *forgo* surgery when it is recommended by the second physician. For non-compliers, monthly total health expenditures, outpatient physician visit, and hospital days do not change significantly from the baseline mean. Sick leave days fall by approximately 10 days per month in the ever-employed subsample (column (7), $p < 0.01$). In the short run, the total health expenditures of compliers is €1,274 to €1,351 per month (columns (1) and (2), $p < 0.01$) higher than those of non-compliers. Additionally, compliers spend about 0.31 to 0.32 more days per month in the hospital than non-compliers (columns (5) and (6), $p < 0.01$). Sick-leave days increase by approximately 16 days per month (columns (7) and (8), $p < 0.01$). However, we find no effect on physician visits. We find an increase in the days spend in hospital (0.2 days per month, $p < 0.10$) and sick leave (17 days per month, $p < 0.10$) for compliers in the medium-term. In the long run, the effect on sick leave is marginally significant in one

Table 3.4: Heterogeneity in Health Outcomes – Second Opinion: Surgery Recommended

	Total Health Expenditures		Visits to physician		Days in hospital		Sick leave days	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dynamic DiD								
pre-baseline	364.57*	396.17*	-0.16	-0.05	-0.04	-0.03	-4.49	-5.18
	(209.98)	(220.39)	(0.44)	(0.49)	(0.06)	(0.06)	(4.21)	(4.52)
short-term	-80.53	-139.15	0.08	-0.08	-0.11	-0.13	-9.91***	-9.02**
	(137.67)	(160.67)	(0.81)	(0.69)	(0.12)	(0.13)	(3.31)	(3.95)
medium-term	-63.74	-173.89	-0.15	-0.58	-0.14	-0.17*	-19.54***	-16.73***
	(120.35)	(132.62)	(0.71)	(0.50)	(0.08)	(0.09)	(5.43)	(4.80)
long-term	204.50	89.23	0.24	-0.21	-0.13	-0.17	-19.65***	-16.65***
	(312.03)	(366.95)	(0.92)	(0.61)	(0.10)	(0.11)	(5.28)	(4.82)
complied × pre-baseline	-239.33	-270.93	-0.24	-0.35	-0.01	-0.01	-3.84	-3.15
	(410.42)	(419.18)	(0.81)	(0.84)	(0.08)	(0.09)	(6.87)	(7.14)
complied × short-term	1,273.86***	1,350.66***	-0.24	-0.12	0.31**	0.32**	16.96***	16.08***
	(323.28)	(353.70)	(0.97)	(0.89)	(0.14)	(0.14)	(3.86)	(4.43)
complied × medium-term	436.59	571.18	-0.67	-0.50	0.20*	0.21*	17.12*	17.75**
	(346.36)	(391.97)	(0.87)	(0.75)	(0.11)	(0.11)	(8.46)	(7.37)
complied × long-term	-291.43	-201.20	-0.75	-0.69	0.12	0.12	11.13	11.86**
	(336.94)	(379.23)	(1.09)	(0.88)	(0.12)	(0.12)	(7.26)	(5.28)
Static DiD								
complied × post	592.67**	695.32**	-0.44	-0.26	0.21**	0.22**	16.99***	16.73***
	(264.73)	(292.04)	(0.81)	(0.66)	(0.09)	(0.09)	(4.15)	(3.44)
Insurance-status controls	No	Yes	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	185	185	185	185	185	185	95	95
Dep. Var. Mean (p =*baseline*)	469.15	469.15	2.63	2.63	0.11	0.11	16.86	16.86
No. Patients	37	37	37	37	37	37	19	19
No. Complier	16	16	16	16	16	16	6	6
No. Non-complier	21	21	21	21	21	21	13	13

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), and *long-term* ($t \in \{4, 5\}$). Two columns per outcome report specifications without vs. with insurance-status controls for outcomes. Columns (7) and (8) use the ever-employed subsample. In columns (2), (4), and (6), the clustered VCOV is not numerically positive semi-definite; inference uses the eigenvalue adjustment described in Cameron and Miller (2015). Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

specification and insignificant in the other. We conclude there is only weak evidence of persistently higher sick leave days for compliers.

The static DiD specification yields a similar qualitative pattern. During the post-treatment period, the total health expenditure per month of compliers is €593 to €696 larger than for non-compliers (columns (1) and (2), $p < 0.05$). Compliers spend additional 0.21 to 0.22 days per month in the hospital (columns (5) and (6), $p < 0.01$), and take an additional 17 days of sick leave (columns (7) and (8), $p < 0.01$). As in the full sample, adding insurance-status controls leaves the compliance coefficients essentially unchanged.[24]

These patterns suggest that second opinions primarily generate value by discouraging surgery and thus potentially preventing overtreatment. Compliers in the surgery-discouraged subsample experience substantial, persistent savings of approximately €641 to €704 per month, 0.11 to 0.12 hospital days per month, and 0.65 physician visits per month without adverse long-term outcomes. We find no evidence of a positive effect of second opinions inducing surgery. The surgery-recommended subsample exhibits increased short-term surgical costs for compliers without offsetting long-term savings. However, this sample size ($n = 37$) may be too small for precise inference. We summarize:

Result 2 (Compliance benefits concentrate in avoiding discouraged surgery)**.** *Health effects of second opinion compliance are highly asymmetric by recommendation content. Second opinions generate welfare gains only by preventing unnecessary procedures (reducing overtreatment) rather than confirming appropriate ones.*

(a) ***When surgery is discouraged*** *($n = 100$)**:** Compliers save €641 to €704 monthly ($p < 0.01$), visit the physician 0.65 fewer times per month ($p < 0.10$), and avoid 0.11 to 0.12 hospital days monthly ($p < 0.01$) over 18 months. Short-term savings are substantial (more than €1,715 per month, more than 0.9 outpatient visits per month and 0.31 hospital days, $p < 0.01$), with no catch-up spending in medium and long-term, indicating genuine reductions in overtreatment rather than delayed necessary care.*

(b) ***When surgery is recommended*** *($n = 37$)**:** Compliers incur expected surgical costs (more than €1,270 per month, more than 0.3 hospital days per month, and more than 16 sick leave days per month in the short-term, $p < 0.01$) without offsetting long-term benefits. However, the small sample size (16 compliers) limits inference.*

[24] In this small subgroup (37 patients), the patient-clustered variance-covariance matrix (VCOV) becomes numerically non-positive semi-definite in the specifications with insurance-status controls (columns (2), (4), and (6)), a finite-sample/numerical issue that affects the covariance estimate (and thus standard errors) but not the coefficient estimates. Following Cameron and Miller (2015), we use an eigenvalue adjustment to restore a valid clustered VCOV. The corresponding point estimates are virtually identical to those without controls, and inference for the key short-run compliance interactions is unchanged.

3.6.4 Therapy Pathways

The results indicate that the large short-term reductions in spending and hospital utilization due to compliance are concentrated among patients who are advised against surgery and are not offset by worse outcomes later on. To understand whether patients simply manage pain, use alternative treatment or only postpone surgery, we analyze their medical treatment when advised against surgery. We examine three outcomes: painkiller utilization (as a proxy of symptom severity), conservative therapy spending (to detect treatment substitution), and monthly surgery probability (to assess longer-term catch-up).

Table 3.5 shows the results. Both compliers and non-compliers exhibit lower pain-related treatment in the three months prior to the baseline period. Specifically, painkiller spending is approximately €6.4 lower per month (columns (1) and(2), $p < 0.05$) and conservative-therapy costs are approximately €$11 - 7$ lower (columns (3) and (4), $p < 0.1$), relative to baseline means of around €11 and €25, respectively. These results are consistent with the idea that symptoms worsened in the period leading up to the second opinion, which prompted an increase in medication and conservative treatment closer to the consultation date. Importantly, the pre-period effects of compliance are small and statistically insignificant, indicating no evidence of differential pre-trends between compliers and non-compliers.

Immediately following the second opinion, spending on painkillers declines for both groups, reaching values substantially below baseline. We find no evidence of a differential compliance effect. Additionally, the static DiD estimate is close to zero, suggesting that compliance does not affect painkiller utilization.

By contrast, conservative therapy shows a clear divergence after the second opinion. Among non-compliers who proceed to surgery, conservative therapy costs fall sharply in the short-term (€22.3 per month, column (3), $p < 0.01$), while compliers spend €26.5 more on conservative therapy than non-compliers ($p < 0.01$). Taken together, these results imply that the spending of compliers on conservative therapy is slightly above baseline, whereas non-compliers largely discontinue such care once they undergo surgery. The medium- and long-term conservative therapy coefficients are noisy, and the static DiD estimate (approximately €9.8, column (3)) is positive but insignificant. This suggests that the difference in conservative care is not sustained beyond six months.

By definition, in this case, all non-compliers receive surgery within the short-term period, while compliers do not. This results in a large short-term increase in the monthly surgery probability of non-compliers relative to a baseline of roughly 3% in the baseline period. The short-term probability of surgery increases by approximately 29 percentage points per month (columns (5) and (6,) $p < 0.01$) for non-compliers. For compliers, the effect is the opposite in the short-term. Their surgery probability is zero in the short-term. Additionally, the medium and long-term, which are not mechanically determined

Table 3.5: Therapy Pathways by Compliance – Second Opinion: Surgery Discouraged

	Painkiller costs		Conservative therapy costs		Surgery	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dynamic DiD						
pre-baseline	-6.38**	-6.38**	-11.70*	-11.70*	0.00	0.00
	(2.57)	(2.58)	(6.09)	(6.12)	(0.01)	(0.01)
short-term	-3.25	-3.39	-22.29***	-21.08***	0.25***	0.25***
	(3.12)	(3.03)	(5.30)	(5.84)	(0.03)	(0.03)
medium-term	-5.26**	-5.36**	5.29	7.11	0.00	0.00
	(2.40)	(2.34)	(23.69)	(24.48)	(0.01)	(0.01)
long-term	-3.07	-2.61	-1.38	-6.14	0.02	0.03
	(3.07)	(3.45)	(8.31)	(11.50)	(0.02)	(0.02)
complied × pre-baseline	2.85	2.80	14.87	14.66	-0.01	-0.01
	(3.53)	(3.54)	(9.40)	(9.45)	(0.02)	(0.02)
complied × short-term	0.45	0.44	26.53***	25.48***	-0.29***	-0.29***
	(3.63)	(3.56)	(8.53)	(8.79)	(0.04)	(0.04)
complied × medium-term	-0.50	-0.37	18.13	18.24	-0.01	-0.01
	(3.40)	(3.43)	(37.34)	(37.69)	(0.02)	(0.02)
complied × long-term	-0.71	-1.43	6.95	13.73	-0.04	-0.05*
	(3.61)	(3.81)	(12.98)	(14.08)	(0.03)	(0.03)
Static DiD						
complied × post	-1.68	-1.81	9.77	11.89	-0.11***	-0.11***
	(2.28)	(2.20)	(16.18)	(17.00)	(0.02)	(0.02)
Insurance-status controls	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	500	500	500	500	500	500
Dep. Var. Mean (p =*baseline*)	11.29	11.29	25.09	25.09	0.03	0.03
No. Patients	100	100	100	100	100	100
No. Complier	78	78	78	78	78	78
No. Non-complier	22	22	22	22	22	22

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), and *long-term* ($t \in \{4, 5\}$). Two columns per outcome report specifications without vs. with insurance-status controls. Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

by this definition, show no evidence of merely postponed surgery. In fact, long-term effects suggest that, twelve to eighteen months after the second opinion, compliers have a 5 percentage point lower probability of surgery ($p < 0.10$) than non-compliers.

The static DiD estimate of approximately -0.11 surgeries per month ($p < 0.01$) illustrates this consistent disparity: throughout all post-second-opinion periods, those who comply are significantly less likely to undergo surgery than those who do not.

These patterns reveal the mechanisms behind cost savings: While compliers slightly increase the intensity of conservative therapy (€22 per month in the short-term, $p < 0.01$), non-compliers discontinue conservative care and undergo surgery (€27 per month in the short-term). Critically, avoided surgeries remain avoided. Compliers exhibit persistently lower surgery probabilities even twelve to eighteen months post-consultation (6 pp, $p < 0.10$), with no evidence of catch-up procedures. Painkiller utilization does not differ between groups, indicating that surgical avoidance does not worsen symptom severity.[25]

Result 3 (Surgeries remain avoided without health deterioration)**.** *Among patients advised against surgery, compliance prevents procedures persistently without symptom deterioration.*

(a) ***No catch-up surgery:*** *Compliers exhibit 12 percentage points lower surgery probability across all post-periods ($p < 0.01$), with no rebound even twelve to eighteen months later (-6pp per month, $p < 0.10$).*

(b) ***Conservative care continuation:*** *Compliers maintain their pre second opinion level of conservative therapy in the short-term (about €25 per month, $p < 0.01$), with no differential painkiller utilization, indicating successful symptom management without surgical intervention.*

3.7 Robustness

We assess robustness along three dimensions. First, we examine labor market outcomes to validate that health expenditure savings are not offset by an increase in societal costs due to workforce exit or prolonged absence from work (Section 3.7.1). Second, we vary observation windows. We use a six-month short-run sample to maximize power for immediate effects and a three-year long-run sample to test for delayed catch-up (Section 3.7.2). Third, we use an instrumental-variables design with second-opinion recommendations to address potential residual selection bias (Section 3.7.3).

[25] Corresponding estimates for the surgery-recommended subsample (Appendix C.3) show expected patterns but are underpowered given small sample size.

3.7.1 Labor Market Participation Effects

Labor market attachment provides an independent measure of functional capacity. If avoided surgeries merely postpone necessary treatment, compliers may experience deteriorating work capacity or premature exit from the workforce. We examine wage-replacement payments (short-term absence from work), unemployment (detachment from labor market), and early retirement (permanent exit from the workforce).

Table 3.6 presents the results for patients advised against surgery. Wage-replacement payments mirror the pattern of health outcomes: both groups receive lower wage-replacement payments after the second opinion, with those who comply showing a larger reduction in the short-term.[26]

By contrast, labor force participation outcomes show no systematic response. For unemployment and early retirement, we report only static DiD specifications. Overall, the evidence indicates that compliance does not worsen employment prospects or accelerate workforce exit. In fact, avoiding discouraged surgery is associated with modestly lower wage-replacement payments in the short-term, which is consistent with reduced recovery-related work absence.

3.7.2 Alternative Observation Windows

Next, we vary the main specification by constructing two alternative samples with different restrictions on the observation period. First, we created a short-run sample to maximize sample size and focus on short-run responses. In this sample, we restricted our observation to six months before and six months after the second opinion, rather than the 18-month post-second-opinion requirement. Second, we construct a long-run sample to assess longer-run dynamics and potential "catch-up" among compliers. In this sample, we include all patients who were observed for at least three years after the second opinion.

Short-Run Sample Table 3.7 presents the results for the short-run sample ($n = 216$). The estimates corroborate the short-run findings of the main sample. Monthly total health expenditures and days spent in the hospital per month rise sharply only for non-compliers in the six months after the second opinion. Quantitatively, short-run total health expenditures increase by approximately €1,179 to €1,186 per month for non-compliers. Monthly total health expenditures are €986 to €993 lower per month for compliers than non-compliers (column (1) and (2), $p < 0.01$). Hospital days increase by 0.19 to 0.20 days per month for non-compliers (column (5) and (6), $p < 0.01$). Compliers spend 0.17 to 0.18 fewer days per month in the hospital than non-compliers ($p < 0.01$). The effects of compliance on physician visits and sick leave remain statistically insignificant. Therefore,

[26] Table C.7 in Appendix C.4 reports results for the surgery-recommended subsample.

Table 3.6: Labor Market Outcomes by Compliance – Second opinion: Surgery Discouraged

	Wage-replacement payments	Unemployment	Early retirement
Model:	(1)	(2)	(3)
Dynamic DiD			
pre-baseline	6.19		
	(113.64)		
short-term	17.30		
	(88.63)		
medium-term	-199.31		
	(160.45)		
long-term	-488.73**		
	(235.15)		
complied × pre-baseline	12.38		
	(137.95)		
complied × short-term	-234.43**		
	(115.17)		
complied × medium-term	-246.01		
	(193.30)		
complied × long-term	18.43		
	(264.06)		
Static DiD			
complied × post	-160.20	-0.04	-0.07
	(169.01)	(0.06)	(0.07)
Pseudonym FE	Yes	Yes	Yes
Observations	500	500	500
Dep. Var. Mean (p =*baseline*)	674.11	0.09	0.04
No. Patients	100	100	100
No. Complier	78	78	78
No. Non-complier	22	22	22

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), and *long-term* ($t \in \{4, 5\}$). Unemployment and early retirement are reported only in the *static DiD*, as these outcomes are sparse and exhibit limited within-person variation over the observation window. Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

Table 3.7: Short-run Health Outcomes by Compliance

	Total Health Expenditures		Visits to Physician		Days in Hospital		Sick Leave Days	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dynamic DiD								
pre-baseline	12.50	15.10	-0.45	-0.42	0.01	0.01	-3.58*	-3.83*
	(74.02)	(73.57)	(0.33)	(0.33)	(0.04)	(0.04)	(2.13)	(2.22)
short-term	1,179.36***	1,185.90***	-0.29	-0.28	0.19***	0.20***	-4.30*	-2.90
	(179.93)	(186.84)	(0.40)	(0.40)	(0.04)	(0.04)	(2.50)	(2.71)
complied × pre-baseline	46.83	38.03	0.08	0.06	-0.03	-0.04	1.54	1.60
	(131.45)	(131.61)	(0.36)	(0.37)	(0.05)	(0.05)	(2.41)	(2.49)
complied × short-term	-986.45***	-993.14***	-0.39	-0.40	-0.17***	-0.18***	2.56	1.46
	(194.69)	(201.03)	(0.43)	(0.43)	(0.05)	(0.05)	(2.71)	(2.89)
Static DiD								
complied × post	-1,009.87***	-1,012.55***	-0.43	-0.42	-0.15***	-0.17***	1.79	0.73
	(188.77)	(197.10)	(0.35)	(0.36)	(0.04)	(0.05)	(2.87)	(3.01)
Insurance-status controls	No	Yes	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	648	648	648	648	648	648	324	324
Dep. Var. Mean (p =*baseline*)	373.82	373.82	3.07	3.07	0.09	0.09	13.63	13.63
No. Patients	216	216	216	216	216	216	108	108
No. Complier	146	146	146	146	146	146	80	80
No. Non-Complier	70	70	70	70	70	70	28	28

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), and *short-term* ($t \in \{0, 1\}$). Two columns per outcome report specifications without vs. with insurance-status controls. Columns (7) and (8) use the ever-employed subsample, since certified sick leave is only reliably recorded for individuals in employment. Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

we find no evidence of adverse effects.[27]

Long-Run Sample Table 3.8 presents the results for the long-run sample. The longer time frame substantially reduces the matched sample to $n = 77$ observations, resulting in noisier estimates. Nevertheless, the results provide no evidence of delayed offsetting increases for compliers. If anything, the pattern points in the opposite direction. In the second year, compliers spend 0.10 fewer days per month in the hospital (column (2), $p < 0.05$). In years two and three after the second opinion, compliers also visit their physician significantly less often (approximately 1.22 and 1.19 visits per month; columns (3)–(4), $p < 0.10$). The short-run effects remain qualitatively consistent with the baseline specification, albeit smaller in magnitude. The static DiD specification confirms that physician and days in hospital significantly decline for compliers in the three years after the second opinion ($p < 0.01$).[28]

3.7.3 Instrumental Variable Approach

In addition to our main selection-on-observables design, we check robustness using an instrumental-variables (IV) strategy that leverages variation in second-opinion recommendations. Specifically, we instrument an indicator for surgical avoidance (not undergoing surgery within 180 days post-consultation) with an indicator for receiving a second-opinion recommendation against surgery. The resulting Two-Stage Least Squares (2SLS) estimand is the Local Average Treatment Effect (LATE) for patients whose surgical decisions are shifted by second physicians' advice.[29]

The IV approach uses a static pre–post change framework. We collapse outcomes into mean monthly values over a 6-month pre-period ($t \in \{-2, -1\}$) and an 18-month post-period ($t \in \{0, 1, 2, 3, 4, 5\}$), then analyze changes $\Delta Y_i = \bar{Y}_i^{\text{post}} - \bar{Y}_i^{\text{pre}}$. Let $D_i = \mathbb{1}\{\text{no surgery within 180 days}\}_i$ denote surgical avoidance, and $Z_i = \mathbb{1}\{\text{Second physician discourages su}$ be the instrument. In the first stage,

$$D_i = \pi_0 + \pi_1 Z_i + X_i'\gamma + \nu_i \tag{3.4}$$

where X_i includes demographics, insurance status, specialty indicators, year fixed effects, and baseline utilization (pre-period costs, visits, hospital days, sick days, and Elixhauser

[27] Corresponding short-run estimates by second-opinion recommendation (*Surgery discouraged* and *surgery recommended*) are reported in Table C.8 and Table C.9 in the Appendix.

[28] Appendix Tables C.10 and C.11 report corresponding estimates split by second-opinion recommendation (*surgery discouraged* and *surgery recommended*); given the small subgroup samples (*surgery discouraged*: $n = 48$, Yes–Surgery: $n = 21$), these results are interpreted cautiously and primarily as supportive heterogeneity evidence.

[29] These are "compliers" in the IV sense.

Table 3.8: Long-run Health Outcomes by Compliance

	Total Health Expenditures		Visits to Physician		Days in Hospital		Sick Leave Days	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dynamic DiD								
pre-baseline	-81.21	-81.21	-0.76***	-0.76***	-0.03*	-0.03*	-2.27	-2.27
	(51.72)	(52.01)	(0.28)	(0.29)	(0.01)	(0.02)	(1.45)	(1.45)
short-term	806.66***	868.35***	0.87	0.82	0.17***	0.17***	-0.25	2.26
	(219.48)	(216.31)	(0.63)	(0.53)	(0.06)	(0.06)	(1.73)	(1.96)
medium-term	137.39	207.67	0.04	-0.04	0.01	0.02	-3.32	-0.09
	(121.24)	(142.97)	(0.41)	(0.45)	(0.03)	(0.03)	(2.87)	(3.09)
year 2	321.87	396.41	0.90	0.84	0.01	0.02	-5.35*	-2.28
	(233.18)	(257.14)	(0.62)	(0.54)	(0.03)	(0.03)	(3.08)	(3.80)
year 3	121.51	210.59	0.60	0.56	-0.01	0.00	-3.85	-0.70
	(133.09)	(164.79)	(0.44)	(0.47)	(0.02)	(0.03)	(2.77)	(3.18)
complied × pre-baseline	-25.79	-25.79	0.40	0.40	-0.01	-0.01	1.73	1.73
	(92.85)	(92.90)	(0.38)	(0.38)	(0.04)	(0.04)	(2.07)	(2.03)
complied × short-term	-836.04***	-896.79***	-1.14	-1.04*	-0.21***	-0.21***	-1.40	-3.86
	(232.75)	(229.95)	(0.70)	(0.61)	(0.07)	(0.06)	(2.29)	(2.54)
complied × medium-term	-82.49	-155.28	-0.83*	-0.75	-0.04	-0.05	-2.07	-4.55
	(157.09)	(173.70)	(0.49)	(0.52)	(0.04)	(0.05)	(3.76)	(3.99)
complied × year 2	-357.54	-428.14	-1.22*	-1.16*	-0.10**	-0.10**	0.63	-1.59
	(254.91)	(270.61)	(0.68)	(0.61)	(0.04)	(0.04)	(3.85)	(4.56)
complied × year 3	274.17	218.53	-1.19**	-1.15**	-0.05	-0.06	-1.82	-4.33
	(374.14)	(373.68)	(0.53)	(0.55)	(0.04)	(0.04)	(3.54)	(4.09)
Static DiD								
complied × post	-237.58	-299.35	-1.30***	-1.24***	-0.09***	-0.10***	-2.03	-4.66
	(183.33)	(190.27)	(0.47)	(0.43)	(0.03)	(0.03)	(2.75)	(3.27)
Insurance-status controls	No	Yes	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	462	462	462	462	462	462	240	240
Dep. Var. Mean (p =*baseline*)	360.44	360.44	2.66	2.66	0.11	0.11	9.05	9.05
No. Patients	77	77	77	77	77	77	40	40
No. Complier	57	57	57	57	57	57	31	31
No. Non-Complier	20	20	20	20	20	20	9	9

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), *year 2* ($t \in \{4, 5, 6, 7\}$), and *year 3* ($t \in \{8, 9, 10, 11\}$).Two columns per outcome report specifications without vs. with insurance-status controls. Columns (7) and (8) use the ever-employed subsample, since certified sick leave is only reliably recorded for individuals in employment. In columns (4), and (6), the clustered VCOV is not numerically positive semi-definite; inference uses the eigenvalue adjustment described in Cameron and Miller (2015). Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

scores). For outcome changes ΔY_i, the second stage estimates:

$$\Delta Y_i = \delta_0 + \delta_1 \widehat{D}_i + X_i'\beta + u_i \tag{3.5}$$

with heterogeneity robust standard errors.

This IV strategy relies on three standard assumptions. First, *relevance* requires that recommendations against surgery increase surgical avoidance ($\pi_1 \neq 0$). Second, *exclusion* requires that second-opinion recommendations affect outcomes only through surgical avoidance. Nevertheless, the exclusion restriction could be violated if second-opinion recommendations directly affect patients' health-related behavior or expectations (e.g., through reassurance or changes in care-seeking) independently of surgical avoidance. This channel cannot be fully ruled out. However, such effects are likely negligible by the standardized, non-treating nature of second-opinion consultations in our institutional setting. Third, *monotonicity* rules out defiers, i.e., patients who systematically undergo surgery when advised against it and avoid surgery when it is recommended; in a second-opinion context, systematic contrarian responses to expert advice are unlikely.

Table 3.9 presents the results. First-stage diagnostics confirm instrument strength: the second-opinion recommendation against surgery increases surgical avoidance by 0.29 percentage points (Panel (B), $p < 0.001$). The first-stage F-statistic of 17.25 is well above the conventional threshold of 10.[30]

In the second-stage, the IV point estimates align qualitatively and in magnitude with the main propensity-score results: surgical avoidance is associated with reductions in total health expenditures (approximately −€631 per month), near-zero effects on physician visits (−0.13) and hospital days (0.01), and sick leave days (0.39). However, across all outcomes, the 2SLS estimates are statistically insignificant, reflecting large standard errors.

For comparison, OLS estimates on the same static change specification are reported alongside the IV results. The similarity in magnitude between IV and OLS estimates is noteworthy. The IV estimate identifies the Local Average Treatment Effect (LATE) for the subpopulation of patients whose surgical decisions are causally affected by second-opinion recommendations, while the OLS estimate captures the average effect among all patients who avoid surgery (regardless of the reason). The close correspondence between these estimates (IV: −€631 versus OLS: −€755 for total expenditures) suggests that selection on unobservables is limited and that the second-opinion effect for patients at the margin—those whose decisions are swayed by second opinions—is comparable to the effect

[30] As a placebo check, we test whether the instrument predicts within-pre-period outcome changes (comparing months $t \in \{-6, -4\}$ to $t \in \{-3, -1\}$). Reduced-form coefficients are economically negligible and statistically zero across all outcomes: €21 for expenditure ($p = 0.91$), −0.06 visits ($p = 0.84$), −0.02 hospital days ($p = 0.66$), and −0.43 sick days ($p = 0.69$), ruling out pre-existing differential trends. See Table C.12 in the Appendix.

Table 3.9: IV Robustness Check: Effect of Surgical Avoidance on Health Outcomes

	Δ Total health expenditure	Δ Visits to physician	Δ Days in hospital	Δ Sick leave days
	(1)	(2)	(3)	(4)
Panel A: Second-stage estimates				
2SLS	−631.4	−0.13	0.01	0.39
	(464.1)	(0.69)	(0.07)	(2.44)
OLS	−755.4***	−0.11	−0.10***	−1.65·
	(155.1)	(0.20)	(0.02)	(0.86)
Panel B: First-stage diagnostics (2SLS only)				
SO: Surgery discouraged		0.292*** (0.070)		
First-stage F		17.25		
Observations	231	231	231	231
Controls	Yes	Yes	Yes	Yes

Notes: Panel A reports Two-Stage Least Squares (2SLS) estimates instrumenting surgical avoidance (indicator for no surgery within 180 days post-consultation) with second-opinion recommendation against surgery, alongside OLS estimates for comparison. Panel B reports first-stage coefficient and F-statistic (identical across all outcomes). Outcomes measured as changes in mean monthly values from 6-month pre-period ($t \in \{-2, -1\}$) to 18-month post-period ($t \in \{0, 5\}$). Control variables: federal state indicators, year of second opinion, baseline demographics (age, gender, urban residence), insurance status, medical specialty, Elixhauser comorbidity score, and pre-period utilization (total costs, physician visits, hospital days, sick leave days). Standard errors clustered at patient level. Full sample $n = 231$. $^{***}p < 0.01$, $^{*}p < 0.1$, $^{\cdot}p < 0.1$.

for the broader population of surgical avoiders. This pattern strengthens confidence in our propensity-score matched estimates, as it indicates that residual unobserved heterogeneity is unlikely to substantially bias the main results.

In sum, we interpret the imprecision of IV estimates as reflecting limited statistical power in the full study population ($n = 231$) rather than evidence against the qualitative patterns documented in the main analysis. The IV and OLS point estimates are of similar magnitude and direction for total expenditure, and the first-stage strength combined with the placebo test support the identifying assumptions. The IV estimates thus provide a conceptually distinct robustness check that, while underpowered for precise inference, is consistent with the propensity-score matched event-study findings.

3.8 Discussion

We find that a structured second-opinion program that institutionally separates diagnosis from treatment can reduce utilization and spending consistent with a reduction of overtreatment for elective surgery in orthopedics and neurosurgery. Over eighteen months following consultation, complying with second-opinion recommendations reduces monthly health expenditures by more than €430 ($p < 0.01$), outpatient physician visits by 0.60

($p < 0.10$), and hospital days by 0.06 ($p < 0.10$), with no significant effects on sick leave. These effects are economically substantial. The expenditure reduction exceeds baseline mean monthly spending.

The mechanism underlying these savings can be traced through three steps. First, disagreement between physicians is frequent and strongly asymmetric. Experts disagree in 56% of cases, with second physicians contradicting 63% of surgery recommendations but only 14% of recommendations against surgery—a pattern consistent with physician-induced demand. Second, patients respond to this information. The compliance rate is 80% when second opinions discourage surgery. Third, avoided procedures remain avoided: we find no evidence of catch-up surgery even twelve to eighteen months post-consultation. Instead of undergoing surgery, compliers continue conservative therapy and maintain stable painkiller utilization, with no claims-based evidence of worsening symptom management. These patterns persist in a three-year follow-up subsample, supporting the interpretation that second opinions reduce overtreatment rather than merely postponing warranted treatment.

Our findings are consistent with Bester and Dahm (2018)'s prediction that separating diagnosis from treatment aligns expert incentives toward truthful reporting. With program uptake of only 2.7 per 100,000 enrollees annually, first physicians face minimal verification risk, implying weak ex-ante disciplinary pressure. Welfare gains therefore arise from ex-post correction of individual cases rather than population-wide behavioral change among recommending physicians.

These findings carry direct policy implications. Where institutional design incentivizes overtreatment, structured second-opinion programs offer an effective remedy. However, the program's minimal utilization rate suggests substantial unrealized welfare gains. Why is uptake so low? Several barriers likely contribute. Patients may be unaware that structured programs exist or that their insurer covers consultations. Referring physicians have little incentive to mention second opinions, particularly if they benefit from surgical volume. The decision to seek a second opinion may also signal distrust, creating social costs patients wish to avoid. Finally, patients facing surgery recommendations may experience time pressure—real or perceived—that discourages additional consultations.[31] Policymakers should therefore prioritize awareness campaigns targeting both patients and referring physicians, with particular attention to reducing the social costs of seeking independent assessments.

Program design also matters. The institutional separation of diagnosis from treatment by prohibiting second-opinion physicians from performing the procedures they evaluate appears crucial for aligning incentives toward diagnostic accuracy rather than treatment

[31] Since our study period (2013–2019), information access to this program has likely improved through mandated patient notification, official provider directories, and insurer-supported intermediaries. Whether these changes have increased uptake remains an empirical question.

volume. Programs that merely refer patients to another surgeon who could perform the operation preserve the original incentive structure and may generate smaller benefits. To maximize impact, programs should target high-disagreement specialties like orthopedics and neurosurgery, where our evidence is strongest and where documented regional variation in surgery rates suggests substantial overtreatment. There, even modest increases in program uptake could generate substantial aggregate savings, provided patients comply with second-opinion recommendations.

These results extend existing evidence in several ways. We translate laboratory findings on credence goods and treatment separation (Mimra et al., 2016; Agarwal et al., 2019; Greiner et al., 2017) to a field setting with actual health stakes and real patient decisions. Unlike Greiner et al. (2017), who find that separation reduces overtreatment but increases undertreatment in the laboratory, we observe no corresponding undertreatment concern: patients who avoid discouraged surgery show no health deterioration over eighteen months. We confirm the hypothetical cost predictions of Weyerstraß et al. (2020b) using realized claims data rather than billing simulations, and extend the immediate decision analysis of Kisch et al. (2023) with longer-term health outcomes. Our findings also complement field experiments by Bindra et al. (2021) and Gottschalk et al. (2020), who find limited disciplinary effects from announced second-opinion intentions; we document that substantial corrective effects emerge when independent assessments actually occur.

Several limitations qualify these conclusions. First, selection into program participation limits population-level inference. We estimate the average treatment effect on the treated (ATT) among program participants, not the average treatment effect (ATE) on the eligible population. Because we cannot observe counterfactual non-participants with identical first-opinion recommendations, we cannot directly evaluate the program's full impact.Second, our outcome measures derive from administrative claims, which comprehensively capture utilization and expenditure but imperfectly proxy health status. Ideally, clinical assessments and patient-reported outcomes would complement claims-based measures. Third, our eighteen-month observation window may miss very long-term effects, though sensitivity analysis using a three-year subsample reveals no evidence of delayed deterioration and shows sustained health improvements. Fourth, missing physician identifiers preclude analysis of expert-level heterogeneity. We cannot examine which characteristics make some physicians more effective second-opinion consultants. Small sample sizes in some subgroups, particularly the surgery-recommended group ($n = 37$), limit statistical power for heterogeneity analysis. Lastly, selection into compliance may reflect unobserved patient characteristics despite our rich covariate controls. Compliers might have less severe underlying conditions, greater health consciousness, or better access to conservative care alternatives. While our propensity-score matching addresses observed differences and the instrumental-variables estimates yield similar magnitudes, we cannot definitively rule out residual confounding.

Generalizability merits careful consideration. Our results are specific to elective orthopedic and neurosurgical procedures where delays impose moderate costs but not acute health risks. This is precisely where second-opinion programs are most prevalent in practice and where policy interest concentrates, which enhances rather than limits the relevance of our findings. The results may not extend to time-sensitive conditions where treatment delays meaningfully affect survival or disease progression. Additionally, the German institutional context of statutory health insurance, administratively set prices, and universal coverage differs from market-based systems, and specific incentive structures vary internationally. Nevertheless, financial incentives as drivers of overtreatment appear widespread, as evidenced by the proliferation of second-opinion programs and related policy interventions across diverse healthcare systems internationally, suggesting our core findings may translate to other contexts.

These limitations point toward several productive directions for future research. Longer-term follow-up extending beyond three years would strengthen confidence that avoided surgeries remain clinically appropriate and detect any late-emerging health deterioration. Extension to other specialties and conditions would identify clinical domains where separating diagnosis from treatment generates the largest welfare gains. Identifying patient characteristics that predict compliance would enable targeted outreach to those most likely to benefit. To isolate the disciplinary channel that our low-uptake setting cannot assess, experimental variation in program accessibility such as randomized promotion campaigns could identify whether increased verification threats discipline first physicians' recommendations and enable estimation of population-level treatment effects. Natural experiments examining whether high second-opinion penetration rates in some markets discipline initial recommendations could provide complementary evidence. Finally, linking outcomes to physician identifiers would reveal which characteristics and practice patterns make physicians effective second-opinion consultants, informing recruitment and training.

3.9 Concluding Remarks

This study provides the first comprehensive longitudinal evidence on health outcomes from structured second-opinion programs in elective surgery. Using administrative claims data, we demonstrate that institutional separation of diagnosis from treatment effectively reduces overtreatment without adverse health consequences. Patients who comply with second-opinion recommendations against surgery save €641 to €704 monthly while maintaining stable health outcomes, with avoided procedures remaining avoided even twelve to eighteen months post-consultation. These findings validate theoretical predictions that decoupling diagnostic compensation from treatment margins aligns expert incentives toward truthful reporting.

We document the asymmetric disagreement patterns. Second physicians contradict 63%

of surgery recommendations but only 14% of non-surgery recommendation. This provides field evidence consistent with physician-induced demand in credence goods markets. Yet despite demonstrated clinical and economic benefits, program utilization remains below 3 per 100,000 enrollees annually, suggesting substantial unrealized welfare gains. Policy efforts should prioritize raising awareness among patients, particularly in high-disagreement specialties like orthopedics and neurosurgery.

Our results underscore a broader insight for credence goods markets: institutional design matters. When market forces alone cannot discipline expert opportunism, carefully structured interventions that separate diagnosis from treatment can increase efficiency.

References

Agarwal, R., Liu, C.-W., and Prasad, K. (2019). Personal research, second opinions, and the diagnostic effort of experts. *Journal of Economic Behavior & Organization*, 158:44–61.

Antonioli, E., Tavares Malheiro, D., Damazio Teich, V., Dias Paião, I., Cendoroglo Neto, M., and Lenza, M. (2023). Cost-effectiveness of a second opinion program on spine surgeries: an economic analysis. *BMC Health Services Research*, 23(1):1441.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, 53(5):941–973.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161.

Balafoutas, L., Beck, A., Kerschbamer, R., and Sutter, M. (2013). What drives taxi drivers? a field experiment on fraud in a market for credence goods. *The Review of Economic Studies*, 80(3):876–891.

Balafoutas, L. and Kerschbamer, R. (2020). Credence goods in the literature: What the past fifteen years have taught us about fraud, incentives, and the role of institutions. *Journal of Behavioral and Experimental Finance*, 26:100285.

Beck, A., Kerschbamer, R., Qiu, J., and Sutter, M. (2014). Car mechanics in the lab—-investigating the behavior of real experts on experimental markets for credence goods. *Journal of Economic Behavior & Organization*, 108:166–173.

Bertelsmann-Stiftung (2013). Regionale unterschiede und deren einflussfaktoren – schwerpunkt knieendoprothetik. Accessed June 14, 2021.

Bertelsmann-Stiftung (2017). Faktencheck rücken – rückenschmerzbedingte krankenhausaufenthalte und operative eingriffe. Accessed June 14, 2021.

Bester, H. and Dahm, M. (2018). Credence goods, costly diagnosis and subjective evaluation. *The Economic Journal*, 128(611):1367–1394.

Bindra, P. C., Kerschbamer, R., Neururer, D., and Sutter, M. (2021). On the value of second opinions: A credence goods field experiment. *Economics Letters*, 205:109925.

Bobroske, K., Freeman, M., Huan, L., Cattrell, A., and Scholtes, S. (2022). Curbing the opioid epidemic at its root: The effect of provider discordance after opioid initiation. *Management Science*, 68(3):2003–2015.

Bruch, D., Resch, O., Sehlen, S., Prediger, B., Schröter, F., Franzen, A., Ronckers, C., Neugebauer, E., and May, S. (2024). Inanspruchnahme von zweitmeinungen in deutschland – eine analyse von abrechnungsdaten der aok nordost. *Das Gesundheitswesen*.

Bundesamt, S. (2020). Pressemitteilung nr. 164 vom 12. mai 2020. Accessed June 14, 2021.

Bundesministerium der Justiz und für Verbraucherschutz (2015). Sozialgesetzbuch (SGB V), § 27b: Zweitmeinung bei planbaren eingriffen. `https://www.gesetze-im-internet.de/sgb_5/__27b.html`.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372.

Centers for Medicare & Medicaid Services (2024). Global surgery (mln907166). PDF. Accessed 2026-01-20.

Centers for Medicare & Medicaid Services (2026). Comprehensive care for joint replacement (cjr) model. Web page. Accessed 2026-01-20.

Dai, T. and Tayur, S. (2020). Healthcare operations management: A snapshot of emerging research. *Manufacturing & Service Operations Management*, 22(5):869–887.

Darby, M. R. and Karni, E. (1973). Free competition and the optimal amount of fraud. *The Journal of law and economics*, 16(1):67–88.

Deutsche Rentenversicherung (2026). Anschlussrehabilitation (ahb). Web page. Accessed 2026-01-20.

Domenighetti, G., Casabianca, A., Gutzwiller, F., and Martinoli, S. (1993). Revisiting the most informed consumer of surgical services: The physician-patient. *International Journal of Technology Assessment in Health Care*, 9(4):505–513.

Emons, W. (1997). Credence goods and fraudulent experts. *The RAND Journal of Economics*, pages 107–119.

Freeman, M., Robinson, S., and Scholtes, S. (2021). Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science*, 67(7):4209–4232.

Freeman, M., Savva, N., and Scholtes, S. (2017). Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science*, 63(10):3147–3167.

Gattas, S., Fote, G. M., Brown, N. J., Lien, B. V., Choi, E. H., Chan, A. Y., Rosen, C. D., and Oh, M. Y. (2021). Second opinion in spine surgery: a scoping review. *Surgical Neurology International*, 12.

Gemeinsamer Bundesausschuss (G-BA) (2024a). Heilmittel-richtlinie: Heilmittelkatalog. PDF. Accessed 2026-01-20.

Gemeinsamer Bundesausschuss (G-BA) (2024b). Zweitmeinung bei planbaren eingriffen. https://www.g-ba.de/themen/qualitaetssicherung/vorgaben-zur-qualitaetssicherung/zweitmeinung/.

Geraedets, M. and Kraska, R. (2016). Zweitmeinungen: Inanspruchnahme und nachfrage aus sicht der bevölkerung. *Gesundheitsmonitor*.

Glover, J. A. (1938). The incidence of tonsillectomy in school children. *Proceedings of the Royal Society of Medicine*, 31(10):1219–1236.

Gottschalk, F., Mimra, W., and Waibel, C. (2020). Health services as credence goods: a field experiment. *The Economic Journal*, 130(629):1346–1383.

Greiner, B., Zhang, L., and Tang, C. (2017). Separation of prescription and treatment in health care markets: A laboratory experiment. *Health Economics*, 26(S3):21–35.

Gruber, J. and Owings, M. (1996). Physician financial incentives and cesarean section delivery. *The RAND Journal of Economics*, 27(1):99–123.

Kassenärztliche Bundesvereinigung (KBV) (2023). Versichertenstruktur. https://www.kbv.de/infothek/zahlen-und-fakten/gesundheitsdaten/versichertenstruktur.

Kassenärztliche Bundesvereinigung (KBV) (2024). Zweitmeinung: Informationen und arztsuche. https://www.116117.de/de/zweitmeinung.php.

Kerschbamer, R. and Sutter, M. (2017). The economics of credence goods – a survey of recent lab and field experiments*. *CESifo Economic Studies*, 63(1):1–23.

Kessler, D. P. and McClellan, M. B. (2000). Is hospital competition socially wasteful? *The Quarterly Journal of Economics*, 115(2):577–615.

Kisch, R., Grill, E., Müller, M., Pietzner, J., and Paulus, A. C. (2023). Second opinion and time to knee arthroplasty: a prospective cohort study of 142 patients. *BMJ Open*, 13(10):e073497.

Kuntz, L., Scholtes, S., and Sülz, S. (2019). Separate and concentrate: Accounting for patient complexity in general hospitals. *Management Science*, 66(6):2482–2501.

Lindsey, P. A. and Newhouse, J. P. (1990). The cost and value of second surgical opinion programs: a critical review of the literature. *Journal of Health Politics, Policy and Law*, 15(3):543–567.

Lu, F. (2014). Insurance coverage and agency problems in doctor prescriptions: evidence from a field experiment in china. *Journal of Development Economics*, 106:156–167.

McCarthy, E. G., Finkel, M. L., and Ruchlin, H. S. (1981). Second opinions on elective surgery: the cornell/new york hospital study. *The Lancet*, 317(8234):1352–1354.

McGuire, T. G. (2000). Chapter 9 - physician agency. In Culyer, A. J. and Newhouse, J. P., editors, *Handbook of Health Economics*, volume 1 of *Handbook of Health Economics*, pages 461–536. Elsevier.

Mimra, W., Rasch, A., and Waibel, C. (2016). Second opinions in markets for expert services: Experimental evidence. *Journal of Economic Behavior & Organization*, 131:106–125.

Obermann, K., Brendt, I., and Müller, P. (2021). Die ärztliche zweitmeinung in der praxis. *Ärzte im Zukunftsmarkt Gesundheit*.

OECD (2019). Health spending set to outpace gdp growth to 2030. Accessed February 16, 2022.

Pesendorfer, W. and Wolinsky, A. (2003). Second opinions and price competition: Inefficiency in the market for expert advice. *The Review of Economic Studies*, 70(2):417–437.

Pieper, D., Heß, S., and Mathes, T. (2018). Bestandsaufnahme zu zweitmeinungsverfahren in der gesetzlichen krankenversicherung (gkv). *Das Gesundheitswesen*, 80(10):859–863.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Ruetters, D., Keinki, C., Schroth, S., Liebl, P., and Huebner, J. (2016). Is there evidence for a better health care for cancer patients after a second opinion? a systematic review. *Journal of Cancer Research and Clinical Oncology*, 142(7):1521–1528.

Shmueli, L., Shmueli, E., Pliskin, J. S., Balicer, R. D., Davidovitch, N., Hekselman, I., and Greenfield, G. (2016). Second medical opinion. *Medical Care*, 54(10):921–928.

Statistisches Bundesamt (Destatis) (2024a). Demografischer wandel in großstadtregionen. `https://www.destatis.de/DE/Themen/Querschnitt/Demografischer-Wandel/Aspekte/demografie-grossstadtregionen.html`.

Statistisches Bundesamt (Destatis) (2024b). Population by territory and average age. `https://www.destatis.de/EN/Themes/Society-Environment/Population/Current-Population/Tables/population-by-territory-and-average-age.html`. As at 14 June 2024.

VDEK (2022). Daten zum gesundheitswesen: Versicherte. Accessed February 16, 2022.

Weigl, M., Pietzner, J., Kisch, R., Paulus, A., and May, S. (2021). Effects of a medical second opinion programme on patients' decision for or against knee arthroplasty and their satisfaction with the programme. *BMC Musculoskeletal Disorders*, 22(1):954.

Weyerstraß, J., Prediger, B., Neugebauer, E., and Pieper, D. (2020a). Results of a patient-oriented second opinion program in germany shows a high discrepancy between initial therapy recommendation and second opinion. *BMC Health Services Research*, 20:406–411.

Weyerstraß, J., Prediger, B., Neugebauer, E., and Pieper, D. (2020b). Second opinions before surgery have the potential to reduce costs–an exploratory analysis. *Zeitschrift für Orthopädie und Unfallchirurgie*.

WHO (2020). Global spending on health: Weathering the storm. Accessed February 16, 2022.

Wolinsky, A. (1993). Competition in a market for informed experts' services. *The RAND Journal of Economics*, pages 380–398.

Zich, K. and Tisch, T. (2018). *Krankenhausqualität aus Patientensicht: Untersuchung auf Basis der PEQ-Daten der Weissen Liste*. Bertelsmann Stiftung.

A Appendix to Chapter 1: Short-Term Rental Platforms and the Small-Apartment Segment

A.1 Omitted Tables

Table A.1: Differential post-2015 demographic shock in T3

Dependent Variables:	Population	Unemployment rate
Model:	(1)	(2)
Variables		
High × *Post*	120.7	-0.341*
	(135.6)	(0.190)
Fixed-effects		
Neighborhood	Yes	Yes
Quarter	Yes	Yes
Fit statistics		
Observations	2,011	2,011
R^2	0.999	0.983
Within R^2	0.009	0.019

Notes: Each column reports a regression of the indicated neighborhood-year demographic variable on *High* × *Post*, with neighborhood and quarter fixed effects. Sample: unique neighborhood–quarter cells from the main estimation panel (T1 and T3 neighborhoods, 2010 Q1–2018 Q4; 57 clusters). Coefficients identify the average differential movement in T3 relative to T1 after 2015 Q3, net of time-invariant neighborhood differences and citywide trends. Standard errors clustered at the neighborhood level, reported in parentheses. Significance: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A.2: Sensitivity to time-varying neighborhood demographics

Dependent Variable:	log(rent/sqm)	
	Baseline	+ Demographics
Model:	(1)	(2)
Variables		
High × *Small* × *Post*	0.059***	0.058***
	(0.014)	(0.015)
High × *Post*	-0.019	-0.020
	(0.013)	(0.013)
Small × *Post*	0.002	0.002
	(0.012)	(0.012)
High × *Small*	-0.011	-0.011
	(0.012)	(0.012)
Small	0.020***	0.020***
	(0.007)	(0.007)
Fixed-effects		
Neighborhood	Yes	Yes
Quarter	Yes	Yes
Fit statistics		
Observations	67,719	67,719
R^2	0.498	0.498
Within R^2	0.110	0.110

Clustered (Neighborhood) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Static triple difference-in-differences with fixed treatment date 2015 Q3. Column 1 reproduces the baseline specification. Column 2 adds neighborhood-year population and unemployment rate, both varying at the neighborhood–year level. Table A.1 shows that neither demographic variable moved differentially in T3 relative to T1 after 2015, indicating no post-2015 demographic shock for the controls to absorb. Standard errors clustered at the neighborhood level (57 clusters), reported in parentheses. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. Structural controls: year-built class, barrier-free, publicly subsidized, elevator, basement, central heating, gas heating, balcony/terrace, garden, garage, parking space, parquet, first occupancy, new build — predetermined characteristics and amenities that cannot plausibly respond to Airbnb activity within the sample window. Living area excluded: correlated with the size indicator *Small* (≤ 3 rooms) along which the effect is identified (Section 4.1). Significance: $^{***}\,p < 0.01$; $^{**}\,p < 0.05$; $^{*}\,p < 0.1$.

Table A.3: Composition of advertised rental stock: pre vs post

	Small × T3		Normalized difference		
Variable	Pre	Post	Small × T3	Small × T1	Full sample
Structural					
Rooms	2.083	2.073	-0.013	0.046	0.077
Barrier-free	0.000	0.026	0.233	0.270	0.242
Publicly subsidized	0.033	0.025	-0.051	-0.180	-0.099
Elevator	0.131	0.191	0.163	0.134	0.157
Basement	0.544	0.500	-0.088	-0.068	-0.069
Central heating	0.517	0.493	-0.047	-0.015	-0.055
Gas heating	0.002	0.006	0.056	0.080	0.070
Balcony/terrace	0.367	0.461	0.192	0.224	0.209
Garden	0.135	0.138	0.010	0.005	0.035
Garage	0.232	0.248	0.039	-0.049	0.038
Parking space	0.035	0.056	0.104	0.095	0.107
Parquet floor	0.070	0.100	0.109	0.130	0.125
First occupancy	0.090	0.125	0.113	0.135	0.119
New build	0.032	0.055	0.113	0.124	0.109
Potentially endogenous (all controls)					
Fitted kitchen	0.355	0.387	0.067	0.001	0.040
Furnished	0.019	0.072	0.252	0.119	0.202

Notes: Means of pre- and post-treatment characteristics for small apartments (≤ 3 rooms) in T3 neighborhoods, and normalized differences (post − pre) across three samples: the treated cell (small × T3), the size-matched control cell (small × T1), and the full Cologne rental sample. Normalized difference computed as $(m_{\text{post}} - m_{\text{pre}})/\sqrt{(s^2_{\text{post}} + s^2_{\text{pre}})/2}$. Pre: 2010 Q1–2015 Q2. Post: 2015 Q3–2018 Q4.

Table A.4: Treatment date sensitivity: ±1 quarter

Dependent Variable:	log(rent/sqm)		
	2015 Q2	2015 Q3 (baseline)	2015 Q4
Model:	(1)	(2)	(3)
Variables			
High × *Small* × *Post*	0.055***	0.059***	0.065***
	(0.014)	(0.014)	(0.014)
High × *Post*	-0.022*	-0.019	-0.024
	(0.013)	(0.013)	(0.015)
Small × *Post*	0.009	0.002	-0.007
	(0.011)	(0.012)	(0.011)
High × *Small*	-0.012	-0.011	-0.011
	(0.012)	(0.012)	(0.012)
Small	0.017**	0.020***	0.022***
	(0.007)	(0.007)	(0.007)
Fixed-effects			
Neighborhood	Yes	Yes	Yes
Quarter	Yes	Yes	Yes
Fit statistics			
Observations	67,719	67,719	67,719
R^2	0.498	0.498	0.498
Within R^2	0.109	0.110	0.109

Notes: Static triple difference-in-differences with the treatment indicator *Post* redefined at each of the three dates shown. Column 2 reproduces the baseline (2015 Q3). Standard errors clustered at the neighborhood level (57 clusters), reported in parentheses. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. Structural controls: year-built class, barrier-free, publicly subsidized, elevator, basement, central heating, gas heating, balcony/terrace, garden, garage, parking space, parquet, first occupancy, new build — predetermined characteristics and amenities that cannot plausibly respond to Airbnb activity within the sample window. Living area excluded: correlated with the size indicator *Small* (≤ 3 rooms) along which the effect is identified (Section 4.1). The coefficient is stable over the broader 2015 Q1–2016 Q3 range (0.040–0.063); the 2015 Q3 baseline is conservative relative to all 2016 treatment dates. Significance: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A.5: Room-size cutoff sensitivity

Dependent Variable:	log(rent/sqm)		
	$\leq$ 2 rooms	$\leq$ 3 rooms (baseline)	$\leq$ 4 rooms
Model:	(1)	(2)	(3)
Variables			
High × *Small* × *Post*	0.045***	0.059***	0.065**
	(0.012)	(0.014)	(0.028)
High × *Post*	0.005	-0.019	-0.030
	(0.010)	(0.013)	(0.024)
Small × *Post*	-0.001	0.002	0.015
	(0.009)	(0.012)	(0.013)
High × *Small*	-0.008	-0.011	0.015
	(0.008)	(0.012)	(0.017)
Small	0.058***	0.020***	-0.010
	(0.006)	(0.007)	(0.011)
Fixed-effects			
Neighborhood	Yes	Yes	Yes
Quarter	Yes	Yes	Yes
Fit statistics			
Observations	67,719	67,719	67,719
R^2	0.507	0.498	0.497
Within R^2	0.126	0.110	0.108

Notes: Static triple difference-in-differences with fixed city-wide treatment date 2015 Q3; column 2 reproduces the baseline specification. Standard errors clustered at the neighborhood level (57 clusters), reported in parentheses. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. Structural controls: year-built class, barrier-free, publicly subsidized, elevator, basement, central heating, gas heating, balcony/terrace, garden, garage, parking space, parquet, first occupancy, new build — predetermined characteristics and amenities that cannot plausibly respond to Airbnb activity within the sample window. Living area excluded: correlated with the size indicator *Small* ($\leq$ 3 rooms) along which the effect is identified (Section 4.1). The $\leq$ 4 column's "large" comparison group collapses to 1,880 five-room observations (rooms are capped at five in data preparation) and is a residual bin rather than a clean counterfactual. Significance: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A.6: Density measure: end-of-sample vs. time-averaged

Dependent Variable:	log(rent/sqm)	
	End-period density (baseline)	Time-averaged density
Model:	(1)	(2)
Variables		
High × *Small* × *Post*	0.059***	0.058***
	(0.014)	(0.016)
High × *Post*	-0.019	-0.022
	(0.013)	(0.013)
Small × *Post*	0.002	-0.005
	(0.012)	(0.014)
High × *Small*	-0.011	-0.004
	(0.012)	(0.010)
Small	0.020***	0.026***
	(0.007)	(0.008)
Fixed-effects		
Neighborhood	Yes	Yes
Quarter	Yes	Yes
Fit statistics		
Observations	67,719	64,798
R^2	0.498	0.502
Within R^2	0.110	0.103

Notes: Static triple difference-in-differences with fixed treatment date 2015 Q3. Column 1 assigns terciles on end-of-sample density (2018 Q4); column 2 assigns terciles on mean density over 2010 Q1–2018 Q4. No neighborhood switches between T1 and T3 across the two measures. All reclassifications affect only the excluded T2 group. Standard errors clustered at the neighborhood level (57 clusters), reported in parentheses. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. Structural controls: year-built class, barrier-free, publicly subsidized, elevator, basement, central heating, gas heating, balcony/terrace, garden, garage, parking space, parquet, first occupancy, new build — predetermined characteristics and amenities that cannot plausibly respond to Airbnb activity within the sample window. Living area excluded: correlated with the size indicator *Small* (≤ 3 rooms) along which the effect is identified (Section 4.1). Significance: $^{***}\,p < 0.01$; $^{**}\,p < 0.05$; $^{*}\,p < 0.1$.

Table A.7: Group definition: equal terciles vs. 60/20/20 split

Dependent Variable:	log(rent/sqm)	
	Terciles 33/33/33 (baseline)	60/20/20 split
Model:	(1)	(2)
Variables		
High × *Small* × *Post*	0.059***	0.057***
	(0.014)	(0.013)
High × *Post*	-0.019	-0.021*
	(0.013)	(0.012)
Small × *Post*	0.002	0.011
	(0.012)	(0.009)
High × *Small*	-0.011	-0.004
	(0.012)	(0.009)
Small	0.020***	0.023***
	(0.007)	(0.005)
Fixed-effects		
Neighborhood	Yes	Yes
Quarter	Yes	Yes
Fit statistics		
Observations	67,719	67,766
R^2	0.498	0.500
Within R^2	0.110	0.104

Notes: Static triple difference-in-differences with fixed treatment date 2015 Q3. Column 1 uses traditional terciles (28 treated, 29 control neighborhoods); column 2 uses a 60/20/20 split assigning the bottom 60% of neighborhoods to control, dropping the middle 20%, and restricting treatment to the top 20% — the highest-density tail only (approximately 17 treated and 52 control neighborhoods). Standard errors clustered at the neighborhood level (57 clusters), reported in parentheses. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. Structural controls: year-built class, barrier-free, publicly subsidized, elevator, basement, central heating, gas heating, balcony/terrace, garden, garage, parking space, parquet, first occupancy, new build — predetermined characteristics and amenities that cannot plausibly respond to Airbnb activity within the sample window. Living area excluded: correlated with the size indicator *Small* (≤ 3 rooms) along which the effect is identified (Section 4.1). Significance: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table A.8: Professional threshold sensitivity: 90, 110, and 140 nights

Dependent Variable:	log(rent/sqm)		
	110-night (baseline)	90-night	140-night
Model:	(1)	(2)	(3)
Variables			
High × *Small* × *Post*	0.059***	0.059***	0.056***
	(0.014)	(0.016)	(0.014)
High × *Post*	-0.019	-0.021	-0.021
	(0.013)	(0.013)	(0.014)
Small × *Post*	0.002	0.002	0.005
	(0.012)	(0.013)	(0.012)
High × *Small*	-0.011	-0.002	-0.009
	(0.012)	(0.010)	(0.012)
Small	0.020***	0.021***	0.018**
	(0.007)	(0.007)	(0.007)
Fixed-effects			
Neighborhood	Yes	Yes	Yes
Quarter	Yes	Yes	Yes
Fit statistics			
Observations	67,719	66,151	66,449
R^2	0.498	0.503	0.500
Within R^2	0.110	0.105	0.107

Notes: Static triple difference-in-differences with fixed treatment date 2015 Q3. A listing is classified as professional if it is booked for more than the indicated number of nights per year or if its host operates multiple simultaneous listings. The 90-night and 140-night columns are mechanically equivalent to varying the average-stay parameter in the booking proxy (Equation 1.1) from 2.8 nights to 1.7 nights, bracketing the plausible range for Cologne. Terciles are reassigned under each definition based on end-of-sample density (2018 Q4); T1 and T3 group memberships agree across definitions for the large majority of neighborhoods. Standard errors clustered at the neighborhood level (57 clusters), reported in parentheses. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. Structural controls: year-built class, barrier-free, publicly subsidized, elevator, basement, central heating, gas heating, balcony/terrace, garden, garage, parking space, parquet, first occupancy, new build — predetermined characteristics and amenities that cannot plausibly respond to Airbnb activity within the sample window. Living area excluded: correlated with the size indicator *Small* (≤ 3 rooms) along which the effect is identified (Section 4.1). Significance: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

A.2 Omitted Figures

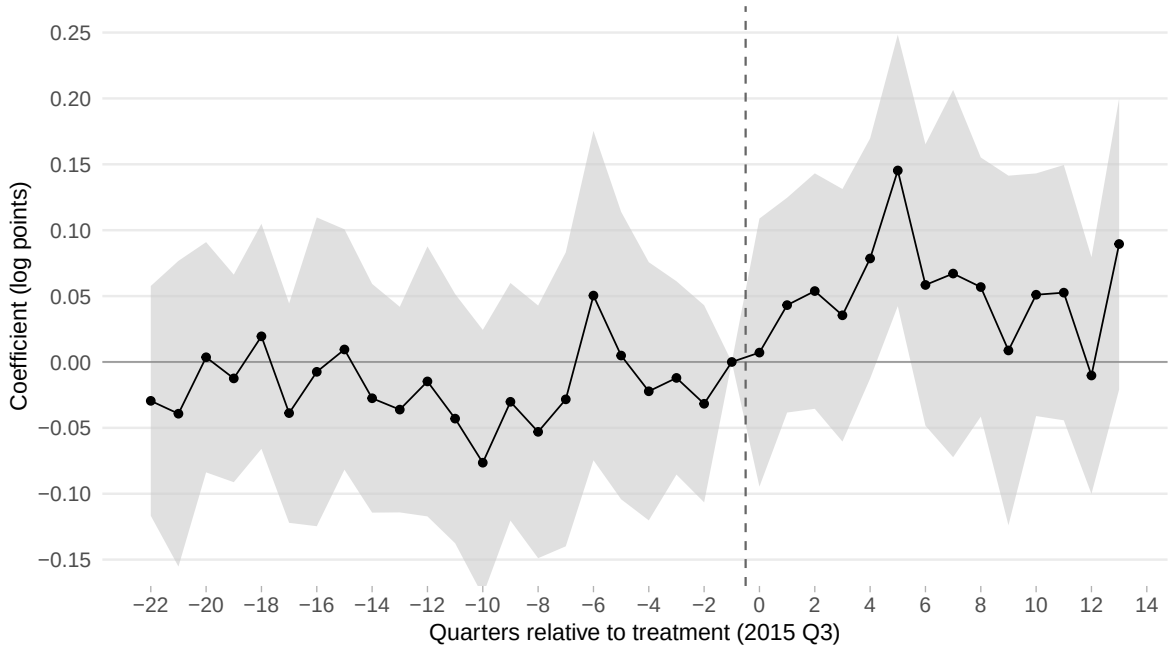


Figure A.1: Event study over the full sample window

Notes: Unbinned event-study coefficients $\hat{\delta}_k$ from Equation (1.6) over the full sample window, $k \in [-22, +13]$, fixed treatment date 2015 Q3, building-fabric controls (see Section 1.5.2), reference period $k = -1$. 95% confidence intervals from standard errors clustered at the neighborhood level (57 clusters). This unbinned series is the input to the pre-trend sensitivity analysis in Section 1.6.1. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4. $N = 67{,}719$.

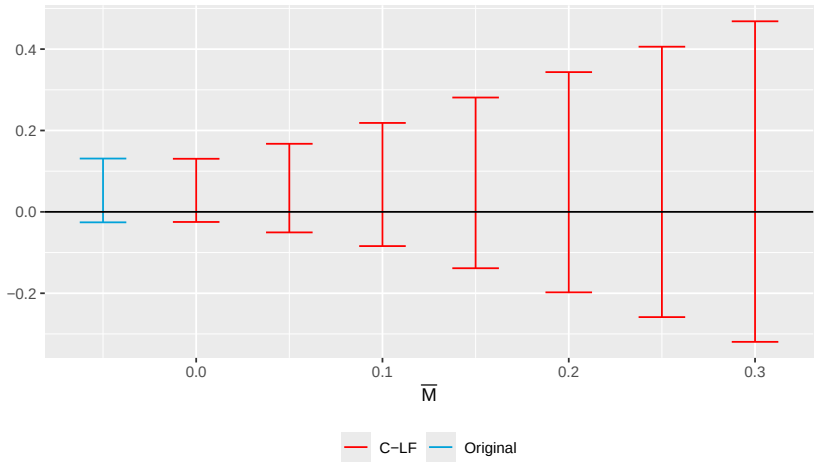


Figure A.2: Pre-trend sensitivity under the relative-magnitudes restriction

Notes: Robust 95% confidence intervals for the average treatment effect over all post-treatment quarters under the relative-magnitudes restriction $\Delta^{RM}(\bar{M})$ of Rambachan and Roth (2023), which bounds the post-treatment violation of parallel trends by $\bar{M}$ times the largest pre-treatment deviation. The interval labelled *Original* imposes parallel trends. The robust interval includes zero at every $\bar{M}$, including $\bar{M} = 0$: the relative-magnitudes restriction has no informative range for this long-horizon average, which is estimated against a single reference quarter (Section 1.6.1). Estimated on the full unbinned event-study window; the smoothness (Δ^{SD}) result appears in Figure 1.7.

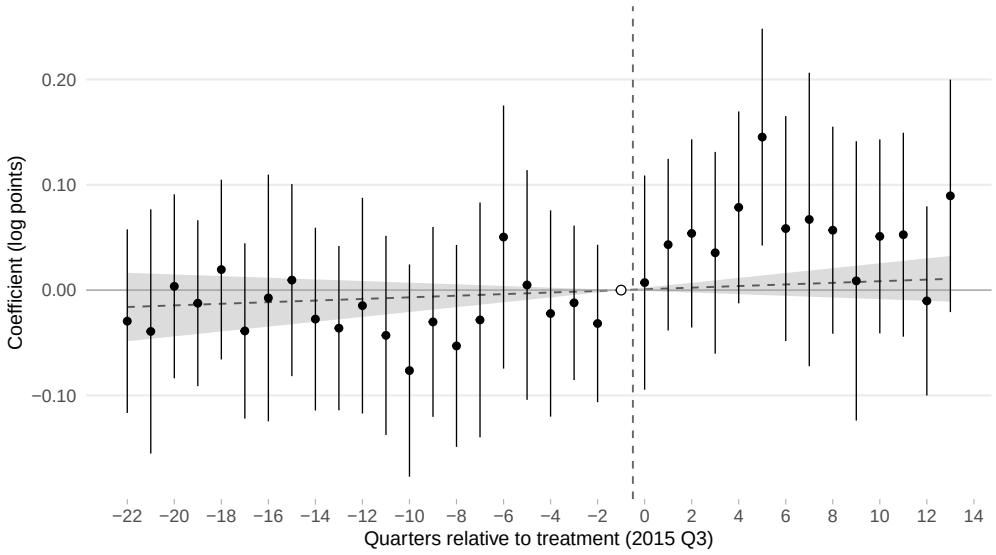


Figure A.3: Admissible linear differential trends under $\Delta^{SD}(M = 0)$

Notes: Unbinned event-study coefficients with 95% confidence intervals (standard errors clustered at the neighborhood level), together with the fan of linear differential trends through the reference period $k = -1$ that the pre-treatment coefficients cannot rule out. The shaded region spans slopes within the 95% confidence interval of the precision-weighted linear fit to the twenty-one pre-treatment coefficients (0.0008 log points per quarter, CI $[-0.0008, 0.0023]$; the dashed line marks the point estimate). Extrapolated to the average post-treatment quarter, the least favorable admissible line accounts for up to 0.017 log points. The figure illustrates the mechanism of the $M = 0$ restriction; the formal robust interval in Section 1.6.1 is the fixed-length confidence interval of Rambachan and Roth (2023).

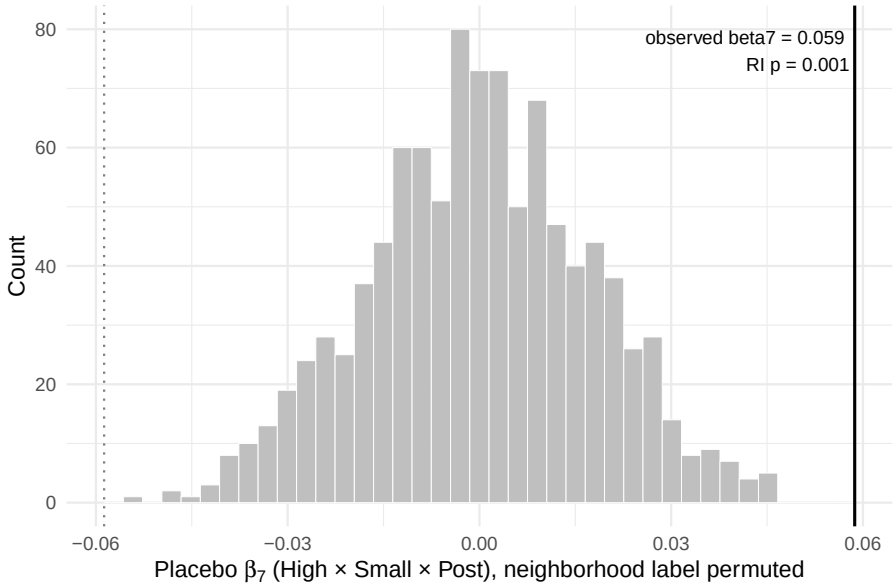


Figure A.4: Geographic randomization inference: placebo distribution of the triple-difference estimate

Notes: Distribution of the saturated triple-difference coefficient over 1,000 random relabelings of which 28 of the 57 neighborhoods are high-density, holding the number treated fixed. The solid line marks the observed estimate (0.059); the dashed line marks its negative. The randomization p-value is the share of placebo estimates at least as large in absolute value as the observed estimate (0 of 1,000). Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4.

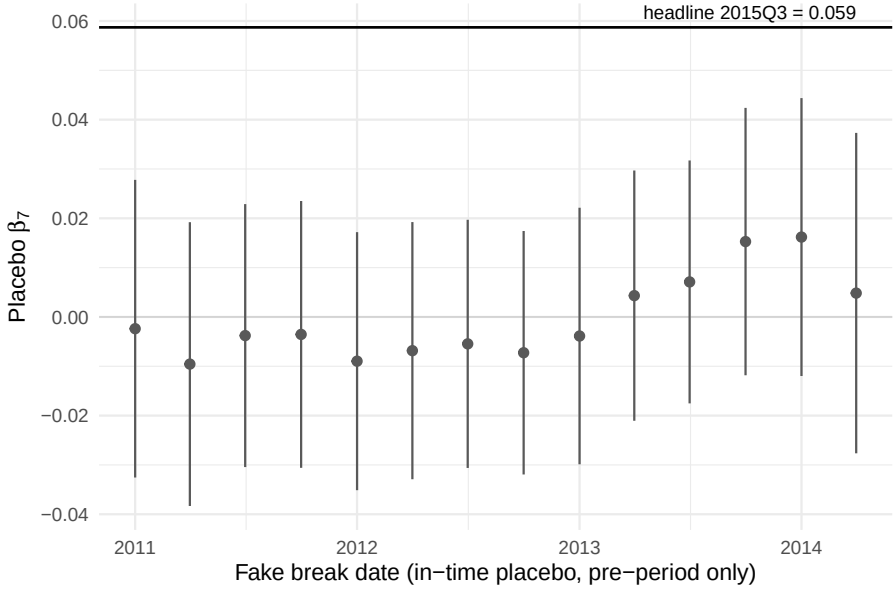


Figure A.5: In-time placebo: triple-difference estimate at fictitious treatment dates

Notes: Placebo triple-difference estimate with 95% confidence interval for a fictitious treatment break placed in each interior quarter of the pre-period (2011 Q1–2014 Q2), estimated on the pre-treatment sample (2010 Q1–2015 Q2) so the post-2015 effect cannot enter. The solid line marks the headline estimate (0.059). Placebo estimates rise toward the treatment date, tracking the mild 2013–2015 differential drift; all remain far below the headline and individually insignificant. Sample: T1 and T3 neighborhoods.

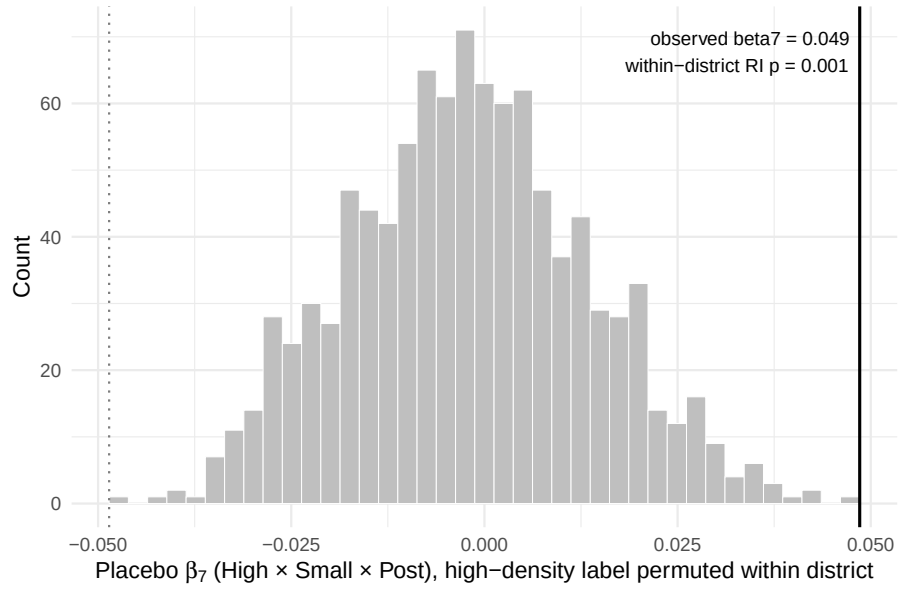


Figure A.6: Within-district randomization inference: placebo distribution under spatially constrained relabeling

Notes: Distribution of the saturated triple-difference coefficient over 1,000 relabelings of high-density status *within* each city district (*Stadtbezirk*), holding the number treated per district fixed, so that treatment remains spatially concentrated in every draw. The central district (Innenstadt), whose neighborhoods are all high-density and admit no relabeling, is excluded; the solid line marks the observed estimate on the remaining sample (0.049). No placebo draw exceeds it ($p = 0.001$). Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4.

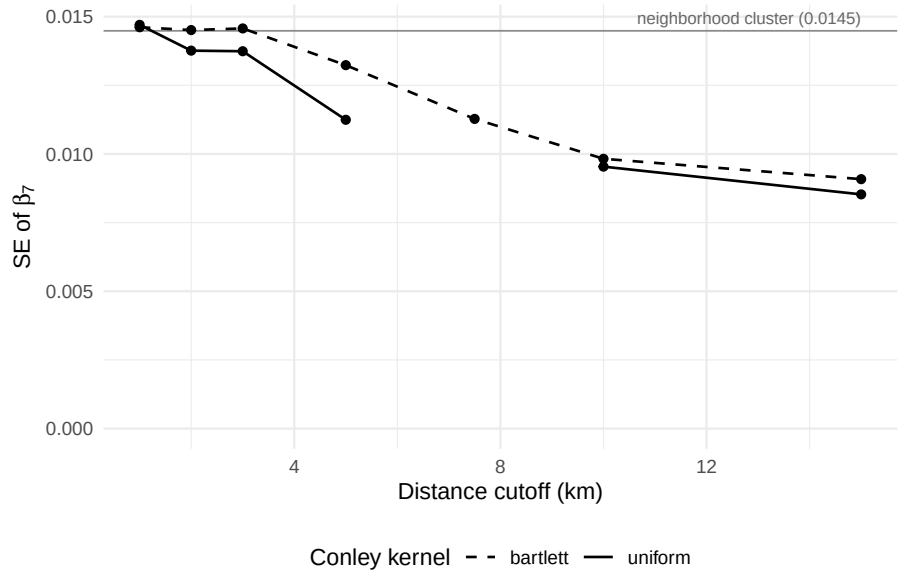


Figure A.7: Conley standard error of the triple-difference estimate by distance cutoff

Notes: Conley standard errors for the saturated triple-difference coefficient, Bartlett kernel, computed from neighborhood-centroid distances. As the cutoff approaches zero the Conley standard error reproduces the neighborhood-clustered value of 0.0145; at short cutoffs of one to three kilometers it is approximately 0.0146, reflecting mild positive short-range spatial correlation. The decline at larger cutoffs is a mechanical artifact: beyond roughly seven kilometers the Bartlett kernel weight matrix on two-dimensional distances is no longer positive semi-definite and the estimated variance is biased downward; it does not indicate negative spatial dependence. The coefficient remains significant at every cutoff. Sample: T1 and T3 neighborhoods, 2010 Q1–2018 Q4.

A.3 Dose-response under a continuous-treatment estimator

Section 1.5.3 establishes the dose-response within the triple-difference design. This appendix verifies the gradient under the nonparametric continuous-treatment difference-in-differences estimator of Callaway et al. (2024), which imposes no functional form on the dose. The estimator requires a panel of persistent dose-carrying units; I collapse the small-apartment sample to a two-period (pre/post) neighborhood panel, with dose equal to end-of-sample professional Airbnb density and low-density (T1) neighborhoods assigned dose zero.[1]

This is a continuous-dose double difference on small apartments: only the low-density control differences out confounders, and the large-apartment within-neighborhood control of the triple is absent. Its level is therefore expected to lie below the triple, and the informative comparison is the shape. Figure A.8 traces the average treatment effect by dose, $ATT(d)$: the effect rises monotonically from about zero at the treated-tercile floor to a plateau of roughly 0.04 in the densest central neighborhoods, is significantly positive above a dose of about 0.84, and is nowhere significantly negative. The slope summary $ACRT(d)$ is imprecise and sensitive to the dose scale, so the $ATT(d)$ curve is the informative object. The rising, saturating shape matches the regression dose-response of Figure 1.6. The estimator's overall average of 0.017 is a dose-distribution-weighted summary of this curve, not an estimate of the triple-difference effect, and is not comparable to the 0.059 headline.

[1] Assigning the low-density group dose zero differences out its small realized exposure (about 8% of the treated group's), which makes the estimate conservative under the maintained hypothesis that Airbnb raises rents. Two neighborhoods with fewer than ten small-apartment advertisements per period are trimmed for precision; 55 neighborhoods remain (28 low-density, 27 high-density).

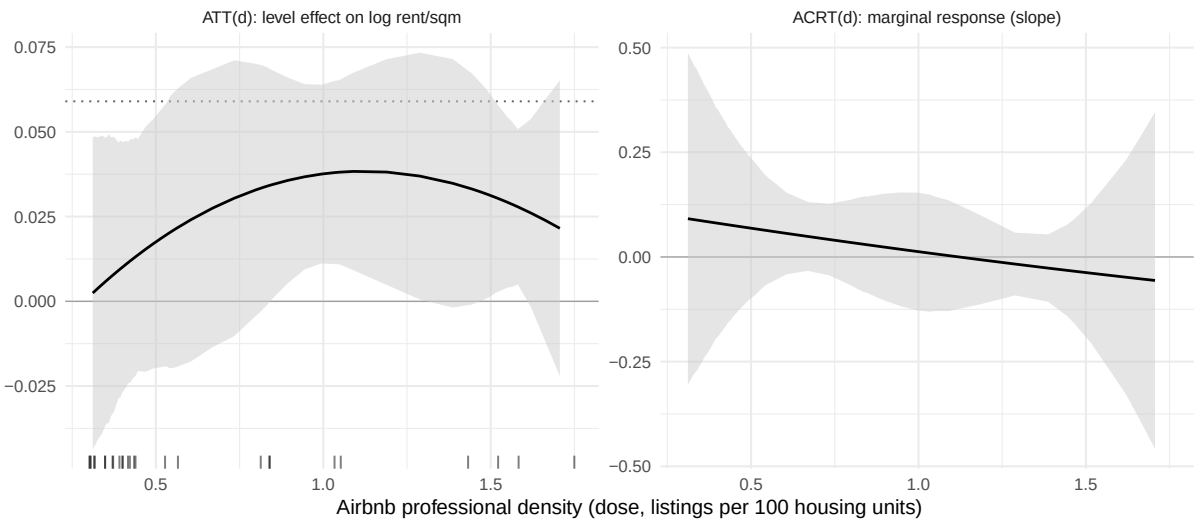


Figure A.8: Continuous-treatment dose-response on small apartments

Notes: Left panel: average treatment effect $ATT(d)$ as a function of end-of-sample professional Airbnb density under the Callaway et al. (2024) continuous-treatment estimator, applied to a two-period (pre/post) neighborhood panel of small-apartment log rents with low-density (T1) neighborhoods at dose zero. Right panel: the corresponding average causal response $ACRT(d)$. Shaded areas show uniform 95% confidence bands, tick marks along the horizontal axis the observed neighborhood dose values, and the dotted line the 0.059 triple-difference estimate for reference. Sample: 55 neighborhoods (28 low-density, 27 high-density), 2010 Q1–2018 Q4.

B Appendix to Chapter 2: Debunking "Fake News" on Social Media

Chapter 2 is replaced by a bibliographic reference in the published version of the dissertation (see Chapter 2). Its appendix is therefore not reproduced here. The appendix material is available with the published article:

> Berger, L. M., Kerkhof, A., Mindl, F., and Münster, J. (2025). Debunking "fake news" on social media: Immediate and short-term effects of fact-checking and media literacy interventions. *Journal of Public Economics*, 245, 105345. https://doi.org/10.1016/j.jpubeco.2025.105345

C Appendix to Chapter 3: Second Opinions, Health Care Cost, and Outcomes in Elective Surgery

C.1 Data Construction and Variable Definitions

C.1.1 Identifying Second-Opinion Indications

Administrative records report the second-opinion provider's specialty and the consultation date, but not the clinical indication. We therefore infer an index diagnosis from insurance claims by searching diagnoses recorded in the second-opinion month and the six preceding months. We retain claim-month observations with non-missing ICD codes in this window and match observed three-character ICD categories to a pre-defined specialty-specific candidate list (orthopedics and neurosurgery). Each candidate ICD is mapped to a coarse body-region group (hip, knee, spine, shoulder, hand/foot, implants, other) and assigned a reliability tier (A/B/C) reflecting diagnostic specificity.[1]

For the study population ($n = 231$), we assign each patient a single "best" index diagnosis by selecting the highest available tier (A over B over C). We obtain a match for 221 patients (95.7%): 86.6% receive a Tier A best match, 6.5% Tier B, and 2.6% Tier C, while 4.3% remain unmatched. For most matched patients, the inferred indication concentrates in at most two body-region groups, consistent with second opinions typically focusing on a narrow anatomical problem.

The inferred index ICD is a claims-based proxy for the second-opinion indication. Accordingly, we use the coarser body-region indicator for descriptive case-mix characterization and as an input to propensity-score matching.

C.1.2 Surgical Procedures and Compliance

Surgery is defined using comprehensive procedure codes: OPS codes 5-01 through 5-05 and 5-76 through 5-86 for inpatient procedures; EBM codes 31121-31147, 36121-36147, and 36191-36198 for outpatient interventions. We construct a surgery indicator for each patient-month if one of these codes is documented. Compliance is then operationalized relative to the second expert's recommendation within 180 days of the second-opinion request. If the recommendation is in favor of surgery, compliance requires surgery within 180 days; if the recommendation is against surgery, compliance requires no surgery during that period.

Figure C.1 plots the timing of the first observed surgery after the second-opinion request and shows that first surgeries are concentrated in the first months following the request, with only a small right tail beyond the dashed 180-day threshold. This pattern

[1] Tier A is most specific, Tier B intermediate, and Tier C broad; tiers are used only to prioritize matches when multiple candidates are observed.

is consistent with elective scheduling delays and suggests that a 180-day window captures the large majority of surgeries plausibly linked to the index second-opinion episode. Using a longer window would therefore reclassify comparatively few additional cases while increasing the likelihood of capturing surgeries arising from subsequent decision making rather than the original recommendation.

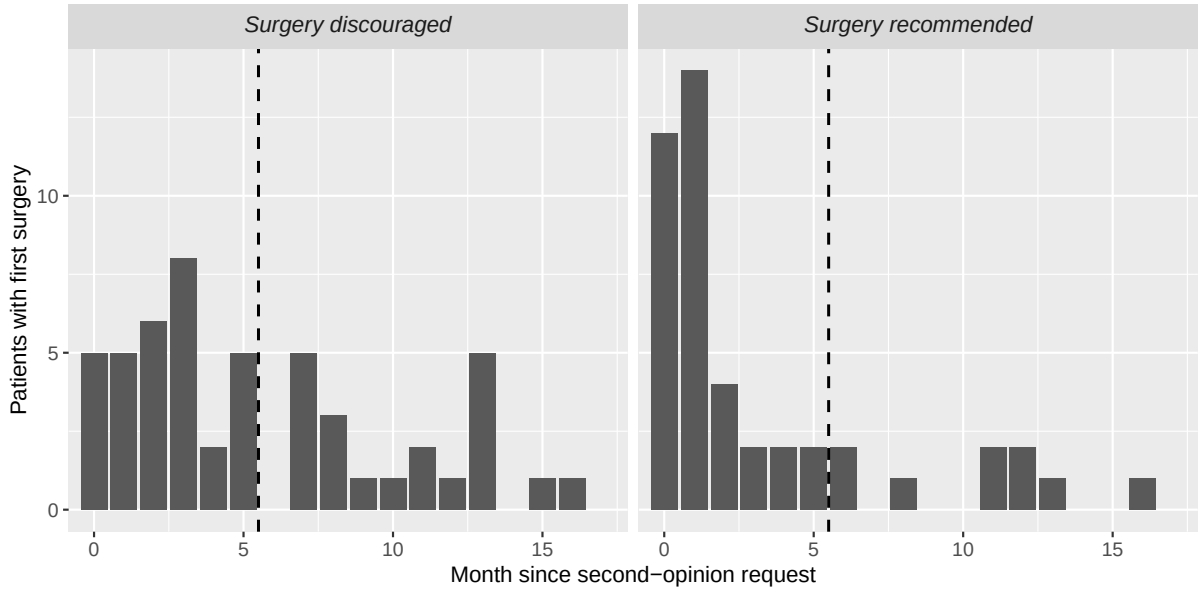


Figure C.1: Distribution of Time to First Surgery after the Second-opinion Request, by Recommendation; dashed line marks the 180-day compliance window.

C.1.3 Medication Classification

Identifying Painkiller-related Medication To identify painkiller-related medication costs, we define painkiller utilization based on the World Health Organization's Anatomical Therapeutic Chemical (ATC) classification at the group level. A medication record is coded as painkiller-related if its ATC group is `N02A` (opioid analgesics), `N02B` (other analgesics and antipyretics), `N02C` (anti-migraine preparations), `M01A` (non-steroidal anti-inflammatory and antirheumatic products), or `M03` (muscle relaxants).

Identifying Conservative-therapy-realted Medication Analogously, we define a medication record as conservative-therapy-related if its ATC group is `M01B` (anti-inflammatory and antirheumatic products in combination), `L04A` (immunosuppressants), or `H02` (systemic corticosteroids).

C.2 Sample Construction and Propensity Score Matching

C.2.1 Pre-Match Covariate Balance

We report group means with standard deviations in parentheses, mean differences between the groups, standardized differences and p-values in Table C.1.[2]

C.2.2 Propensity Score Estimation

We estimate the propensity score for compliance using a logistic regression with an elastic net penalty ($\alpha = 0.5$) on the full set of pre-second-opinion covariates: demographics, first- and second-opinion indicators, body regions, Elixhauser comorbidities (levels and changes across the two pre-second-opinion 3-month periods), pre-second-opinion utilization (levels and changes in costs, visits, surgery, hospital days), and pre-second-opinion insurance status (shares of the two pre-second-opinion 3-month periods spent employed, retired, in family coverage, in vocational training/university, or on unemployment/welfare). The penalty parameter λ is chosen by 10-fold cross-validation. Table C.2 reports all non-zero coefficients (log-odds) from the specifications at $\lambda_{\min}$ and $\lambda_{1\text{se}}$. A quintile-based calibration check shows a monotonic increase in the observed compliance rate, from roughly 13% in the lowest propensity-score quintile to almost full compliance in the highest quintile (Table C.3). Kernel density plots of the propensity score by compliance status show substantial overlap between compliers and non-compliers and look very similar for the $\lambda_{\min}$ and $\lambda_{1\text{se}}$ specifications (Figure C.2).

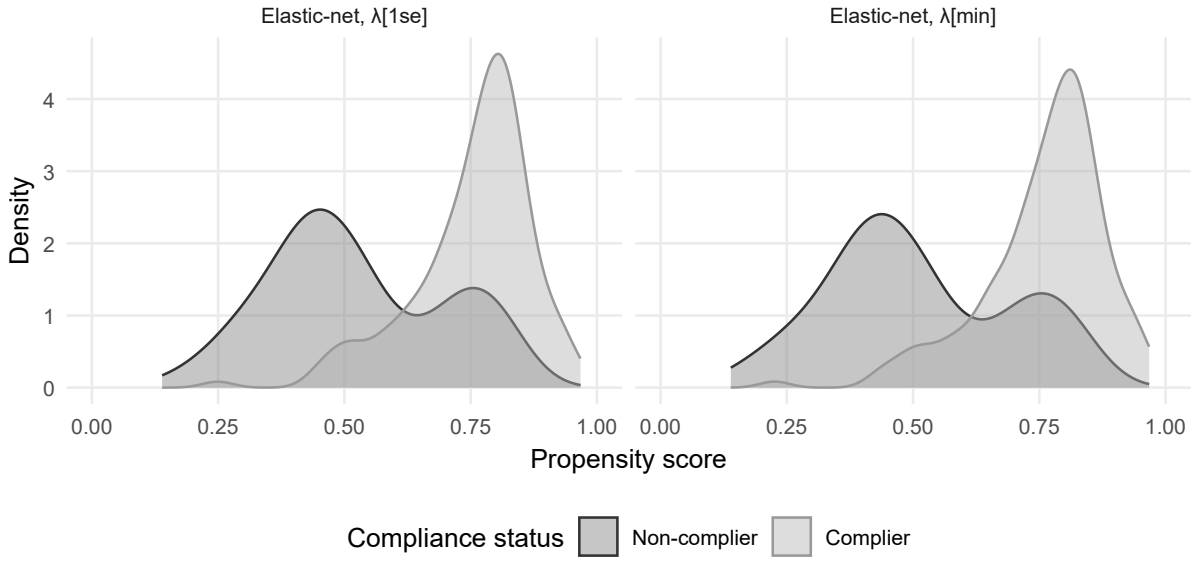


Figure C.2: Propensity Score Distributions for Compliers and Non-compliers under the Elastic-net $\lambda_{\min}$ and $\lambda_{1\text{se}}$ Specifications.

[2]Standardized differences are Cohen's h for binary variables and standardized mean differences for continuous variables. For inference, we report p-values from proportion/Fisher tests for binary variables and from Welch t-tests for continuous variables.

Table C.1: Baseline Characteristics by Compliance

Variable	Mean Non-Complier	Mean Complier	Difference	p-value	Std. Diff.
Demographics					
age	55.58 (17.89)	57.70 (14.72)	2.12	0.381	0.13
Female	51.4%	46.5%	-4.8 pp	0.589	-0.10
Urban	50.0%	49.7%	-0.3 pp	1.000	-0.01
Pre-SO comorbidity prevalence					
carit	8.3%	12.6%	4.2 pp	0.471	0.14
chf	4.2%	3.8%	-0.4 pp	1.000	-0.02
cpd	18.1%	19.5%	1.4 pp	0.938	0.04
dane	1.4%	1.3%	-0.1 pp	1.000	-0.01
depre	19.4%	28.9%	9.5 pp	0.174	0.22
hypc	5.6%	6.9%	1.4 pp	0.782	0.06
hypothy	8.3%	8.8%	0.5 pp	1.000	0.02
hypunc	43.1%	42.8%	-0.3 pp	1.000	-0.01
obes	18.1%	22.0%	4.0 pp	0.609	0.10
para	1.4%	2.5%	1.1 pp	1.000	0.08
pcd	2.8%	0.0%	-2.8 pp +	0.096	-0.33
pvd	15.3%	6.9%	-8.4 pp +	0.078	-0.27
rf	4.2%	3.8%	-0.4 pp	1.000	-0.02
rheumd	6.9%	6.9%	-0.0 pp	1.000	-0.00
solidtum	2.8%	6.3%	3.5 pp	0.350	0.17
valv	5.6%	4.4%	-1.2 pp	0.743	-0.05
Elixhause score	0.93 (4.33)	0.73 (4.14)	-0.20	0.745	-0.05
Insurance-status shares					
retired	33.3%	36.5%	3.2 pp	0.753	0.07
employed	35.9%	42.8%	6.9 pp	0.316	0.14
unemployed	13.0%	7.4%	-5.6 pp	0.215	-0.19
vocational university	8.1%	6.9%	-1.2 pp	0.755	-0.05
family & other	9.7%	6.4%	-3.3 pp	0.408	-0.12
Share of SO-relevant body regions					
hand or foot	11.1%	17.0%	5.9 pp	0.340	0.17
hip	20.8%	20.8%	-0.1 pp	1.000	-0.00
knee	36.1%	34.6%	-1.5 pp	0.940	-0.03
limb	2.8%	1.9%	-0.9 pp	0.648	-0.06
shoulder	6.9%	18.2%	11.3 pp *	0.041	0.35
spine	54.2%	70.4%	16.3 pp *	0.024	0.34
implants	15.3%	5.0%	-10.2 pp *	0.018	-0.35
other	1.4%	2.5%	1.1 pp	1.000	0.08
# of SO-relevant regions per patient	1.50 (0.93)	1.70 (0.98)	0.20	0.132	0.21
Shares by disagreement (FO/ SO)					
Yes/ No	20.8%	41.5%	20.7 pp **	0.004	0.45
Yes/ Yes	38.9%	12.6%	-26.3 pp ***	0.000	-0.62
No/ No	6.9%	8.2%	1.2 pp	0.953	0.05
No/ Yes	2.8%	0.6%	-2.1 pp	0.230	-0.18
Undocumented/ No	15.3%	27.7%	12.4 pp +	0.060	0.30
Undocumented/ Yes	9.7%	8.8%	-0.9 pp	1.000	-0.03
Annual program entry share					
Year 2014	8.3%	10.1%	1.7 pp	0.863	0.06
Year 2015	11.1%	23.9%	12.8 pp *	0.038	0.34
Year 2016	23.6%	20.8%	-2.9 pp	0.752	-0.07
Year 2017	41.7%	31.4%	-10.2 pp	0.173	-0.21
Year 2018	9.7%	11.3%	1.6 pp	0.894	0.05
Mean monthly outcome measures pre-SO					
total health expenditures	620.87 (772.44)	561.65 (858.31)	-59.22	0.603	-0.07
visits to physician	2.58 (1.49)	3.15 (2.01)	0.58 *	0.016	0.33
days in hospital	0.13 (0.18)	0.11 (0.18)	-0.02	0.433	-0.11
sick leave days	6.78 (10.77)	6.02 (10.26)	-0.75	0.619	-0.07
costs painkillers	15.03 (26.39)	25.33 (60.09)	10.30 +	0.072	0.22
costs conservative therapy	27.24 (84.64)	41.39 (97.35)	14.15	0.264	0.16
surgery	0.06 (0.12)	0.03 (0.09)	-0.04 *	0.033	-0.32
replacement wage payment	577.59 (877.75)	516.71 (838.93)	-60.88	0.621	-0.07

Notes: Means (SD) for continuous; shares for binary; Welch t-test for continuous, prop/Fisher for binary. Signif.: + p≤0.10, * p≤0.05, ** p≤0.01, *** p≤0.001.

Table C.2: Elastic-net Propensity Score Model: Non-zero Coefficients

Variable	$\lambda_{\min}$	λ_{1se}
solidtum lag1	1.240	1.113
SO (=No)	1.155	1.102
SO-region (=implants)	-1.155	-1.045
pcd lag1	-1.058	-0.982
pcd lag2	-1.045	-0.981
pvd lag1	-0.940	-0.854
mean monthly surgery lag1	-0.615	-0.646
valv lag2	-0.631	-0.521
mean monthly surgery lag2	-0.618	-0.470
SO-region (=shoulder)	0.415	0.377
Year of SO^3	0.412	0.375
carit lag1	0.263	0.210
trend comorbidities	-0.246	-0.232
SO-region (=hand or foot)	0.222	0.174
hypunc lag2	0.209	0.178
Year of SO^5	-0.195	-0.170
SO (=Yes)	-0.187	-0.190
dane lag1	-0.167	-0.012
SO-region (=spine)	0.120	0.112
mean monthly unemployed lag1	-0.112	-0.087
mean monthly unemployed lag2	-0.109	-0.083
# of SO-regions	0.061	0.055
mean monthly visits to physician lag2	0.050	0.046
mean monthly days in hospital lag1	-0.045	-0.006
specialisation (=Orthopedics)	-0.027	NA
mean monthly visits to physician lag1	0.007	0.003
mean monthly sick leave days lag1	-0.006	-0.004
mean monthly costs painkillers lag2	0.003	0.003
mean monthly costs conservative therapy lag2	0.001	0.000
age	0.000	NA
trend total health expenditures	0.000	0.000

Table C.3: Compliance by Propensity Score Quintile

Propensity score quintile	$n_{\lambda_{\min}}$	Compliance rate $\lambda_{\min}$ (%)	$n_{\lambda_{1se}}$	Compliance rate λ_{1se} (%)
Quintile 1	47	12.8	47	14.9
Quintile 2	47	63.8	47	66.0
Quintile 3	46	82.6	46	78.3
Quintile 4	47	83.0	47	83.0
Quintile 5	47	97.9	47	97.9

Table C.4: Balance Summary for Full Matching by Elastic-net Propensity Score Specification.

PS model	Matched patients	Compliers	Non-compliers	Max. |SMD| (unadj.)	Max. |SMD| (adj.)	Share |SMD
Elastic-net $\lambda_{\min}$	137	94	43	2.07	0.37	
Elastic-net λ_{1se}	143	98	45	1.99	0.33	

C.2.3 Matching Diagnostics

We compare covariate balance across six matching designs that combine three algorithms (nearest-neighbor, optimal, and full matching) with two propensity-score specifications (Elastic-net at $\lambda_{\min}$ and λ_{1se}). All designs impose exact matching on the first- and second-opinion recommendation (*FO*, *SO*) and physician specialty, use the Elastic-net-based propensity score as the main distance (with a 0.2 standard-deviation caliper), and add a Mahalanobis component in pre-second-opinion costs, utilization measures, and the Elixhauser comorbidity score. Matching greatly reduces imbalance relative to the raw sample: across designs, the maximum absolute standardized mean difference (SMD) over all covariates declines from about 2.0 (propensity score distance) in the unmatched sample to at most 0.33–0.37 after matching, and the large majority of covariates achieve the standard threshold of $|\text{SMD}| < 0.2$ (Table C.4). For visualization, Figure C.3 shows *Love plots* for the two full-matching designs. Since propensity-score distance is a function of all covariates, it is excluded from these plots.

Our preferred specification uses full matching with the $\lambda_{\min}$ propensity score. This design retains 137 patients (94 compliers and 43 non-compliers), roughly two thirds of the baseline sample, and provides a good compromise between balance and sample size: 64 of 72 covariates (about 89%) have $|\text{SMD}| < 0.2$. Among the remaining covariates, some main outcomes remain slightly above our threshold: Mean monthly total health expenditures, and days in hospital in the 3-month period before the second opinion (lag 1), with $|\text{SMD}| \approx 0.36$, and ≈ 0.31, which we view as a modest residual imbalance, especially given that the corresponding lag-2 measures are well balanced.

The remaining imbalances are mainly in insurance-status. In the matched sample, compliers and non-compliers have broadly similar insurance profiles: the share of employed patients differs by 4 percentage points (0.525 vs. 0.484), retired patients by 4.2 points, family co-insured patients by 0.4 points, and unemployed patients by 3.2 percentage points

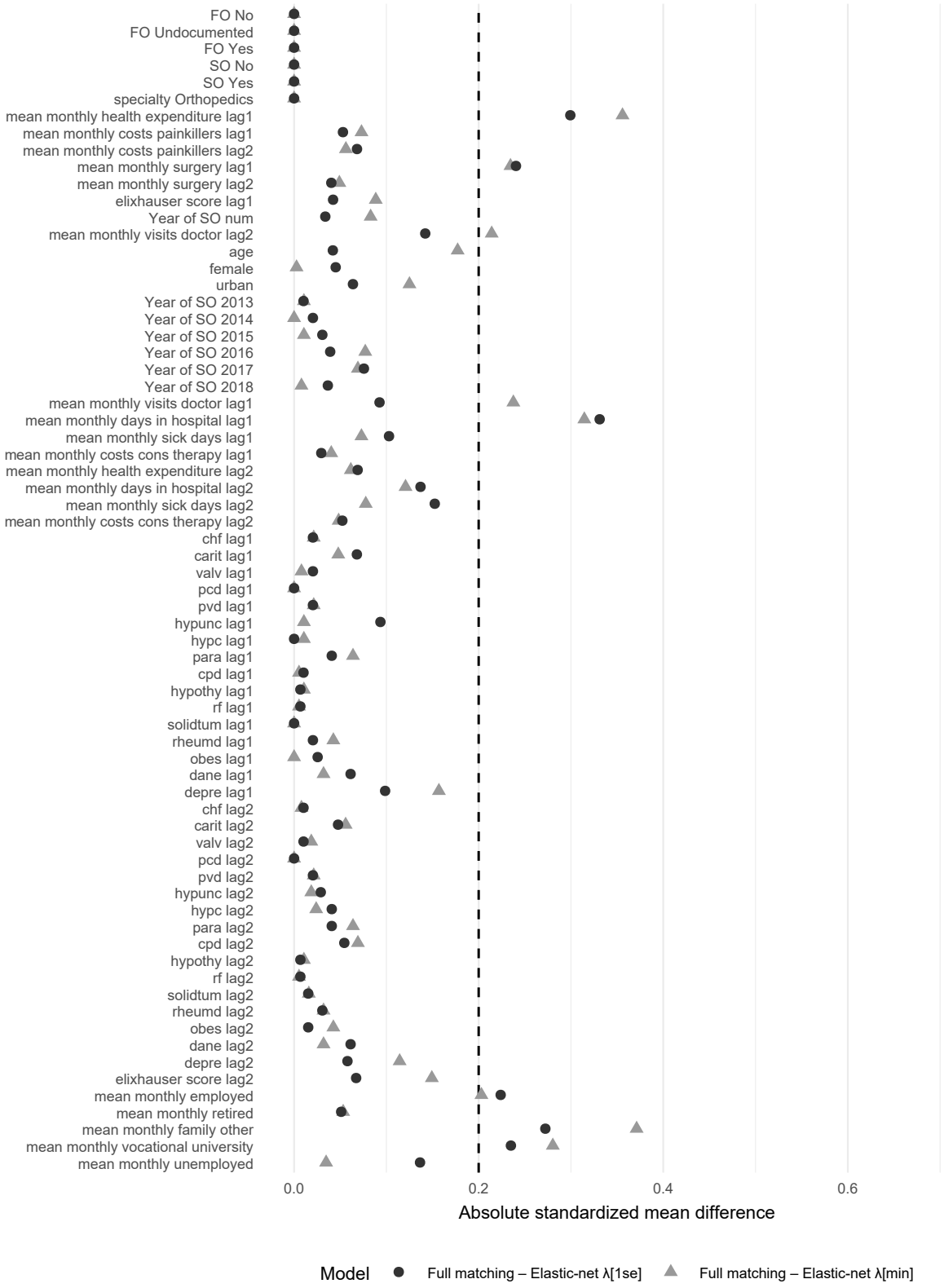


Figure C.3: Absolute mean standardized differences between compliers and non-compliers under the Elastic-net $\lambda_{\min}$ and $\lambda_{1\text{se}}$ specifications with full matching.

Table C.5: Insurance status in the preferred matched sample (full matching, elastic-net $\lambda_{\min}$).

Insurance status	Compliers	Non-compliers	Difference
Employed	0.525	0.484	0.040
Retired	0.298	0.256	0.042
Family/other	0.043	0.047	-0.004
Unemployed	0.092	0.124	-0.032

(Table C.5). Given the small cell sizes and the fact that these variables already enter both the propensity-score and matching distance, we view the residual differences in insurance-status as moderate. Nonetheless, we include insurance status as additional regressors in all outcome regressions.

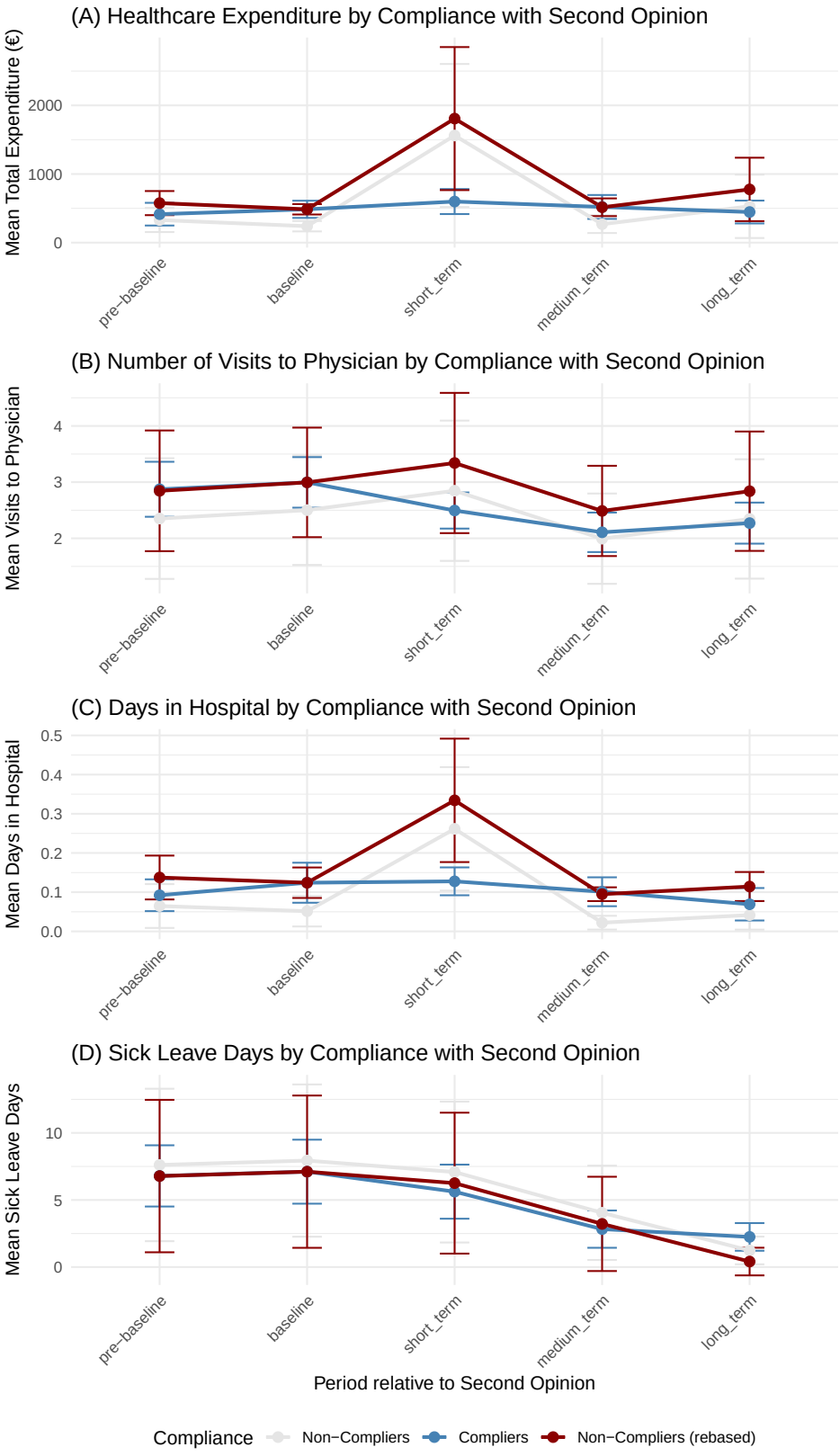


Figure C.4: Health Outcomes by Compliance in Matched Sample.

C.3 Treatment Pathways and Mechanisms

Table C.6 presents treatment pathway outcomes for the 37 patients whose second opinion recommended surgery. Given the small sample size (16 compliers, 21 non-compliers), we interpret these estimates cautiously and view them primarily as complementary evidence to the main analysis.

Unlike the surgery-discouraged subsample, we observe no significant pre-baseline changes in painkiller or conservative therapy costs, suggesting less pronounced symptom worsening in the period immediately preceding the second opinion. The pre-period compliance interactions are also statistically zero, confirming no differential pre-trends between compliers and non-compliers.

Following the second opinion, both groups exhibit substantial reductions in painkiller spending. Non-compliers' monthly painkiller costs decline by approximately €7 in the short-term (column (1), $p < 0.05$), with similar declines persisting through the medium- and long-term periods. Compliers show no differential change relative to non-compliers across any post-period, and the static DiD estimate is small and insignificant (€4.37). This pattern suggests that symptoms improve for both groups regardless of surgical compliance, possibly reflecting natural recovery or successful pain management through either surgical or conservative approaches.

Conservative therapy costs show a more complex trajectory. Non-compliers exhibit declining conservative care spending over time, with sharp reductions in the long-term of about €20 per month (column (3), $p < 0.01$). By contrast, compliers spend €35 more per month on conservative therapy than non-compliers in the long-term ($p < 0.05$). The static DiD estimate confirms this divergence: compliers spend €22 to €21 more per month on conservative therapy across all post-periods (columns (3) and (4), $p < 0.05$ and $p < 0.10$, respectively). This pattern suggests that patients who undergo surgery as recommended subsequently reduce conservative care, whereas those who decline surgery maintain or increase such treatment as an alternative management strategy.

Surgery probabilities follow the expected pattern by construction. Compliers exhibit a 0.21 percentage point increase in monthly surgery probability during the short-term period (columns (5) and (6), $p < 0.01$), consistent with compliance being defined as undergoing surgery within 180 days. Notably, the medium- and long-term compliance interactions are small and insignificant, providing no evidence that non-compliers subsequently "catch up" by undergoing delayed procedures. The static DiD estimate of 0.11 to 0.12 surgeries per month ($p < 0.01$) reflects the persistent difference between groups established in the short-term.

These patterns contrast with those observed in the surgery-discouraged subsample (Section 3.6.4). Here, both compliers and non-compliers experience symptom improvements (declining painkiller use), and the primary difference lies in treatment modality: compliers who undergo surgery subsequently reduce conservative care, while non-compliers who forgo surgery maintain higher conservative therapy spending as an alternative. The small sample size precludes strong conclusions, but the results are consistent with surgery and conservative therapy serving as substitutes for symptom management in this population.

Table C.6: Treatment Outcomes by Compliance – Second Opinion: Surgery Recommended

	Painkiller costs		Conservative therapy costs		Surgery	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dynamic DiD						
pre-baseline	-2.50	-2.69	-4.46	-4.99	0.07	0.07
	(3.10)	(3.07)	(8.65)	(8.90)	(0.05)	(0.05)
short-term	-7.16**	-6.98***	-11.54	-10.04	-0.02	-0.02
	(2.65)	(2.50)	(8.17)	(8.04)	(0.01)	(0.01)
medium-term	-6.41**	-5.71**	-7.55	-4.35	-0.01	-0.01
	(2.47)	(2.22)	(8.87)	(10.32)	(0.01)	(0.01)
long-term	-7.00**	-6.25**	-20.01***	-16.65**	0.03	0.02
	(2.93)	(2.62)	(7.23)	(7.29)	(0.03)	(0.03)
complied × pre-baseline	-14.78	-14.59	-12.42	-11.89	-0.07	-0.07
	(8.99)	(9.06)	(13.12)	(13.30)	(0.06)	(0.06)
complied × short-term	1.85	2.50	-3.23	-4.70	0.21***	0.22***
	(5.39)	(5.38)	(13.72)	(13.85)	(0.04)	(0.04)
complied × medium-term	-6.36	-4.74	16.91	13.90	0.06	0.07
	(6.56)	(6.33)	(19.02)	(20.63)	(0.05)	(0.05)
complied × long-term	-4.54	-2.98	35.32**	33.91**	-0.05	-0.04
	(5.45)	(4.81)	(16.07)	(16.59)	(0.04)	(0.04)
Static DiD						
complied × post	4.37	5.60	22.54**	20.54*	0.11***	0.12***
	(3.98)	(4.10)	(10.80)	(11.24)	(0.03)	(0.04)
Insurance-status controls	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	185	185	185	185	185	185
Dep. Var. Mean (p =*baseline*)	22.59	22.59	21.74	21.74	0.02	0.02
No. Patients	37	37	37	37	37	37
No. Complier	16	16	16	16	16	16
No. Non-complier	21	21	21	21	21	21

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), and *long-term* ($t \in \{4, 5\}$). Two columns per outcome report specifications without vs. with insurance-status controls. In columns (2), (4), and (6), the clustered VCOV of the *Dynamic DiD* is not numerically positive semi-definite; inference uses the eigenvalue adjustment described in Cameron and Miller (2015). Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

C.4 Labor Market Outcomes

Table C.7 presents labor market outcomes for the 37 patients whose second opinion recommended surgery. These estimates complement the main analysis by examining whether surgical compliance generates differential labor market consequences beyond the health utilization patterns documented in Section 3.6.3.

Wage-replacement payments exhibit a clear divergence between compliers and non-compliers following the second opinion. The pre-baseline compliance interaction is small and statistically insignificant (column (1)), confirming parallel pre-trends. After the second opinion, non-compliers, who forgo recommended surgery, experience substantial reductions in wage-replacement payments relative to baseline: €470 per month in the short-term, €971 in the medium-term, and €975 in the long-term (all $p < 0.05$). These declines suggest that non-compliers return to work or reduce certified absence as their conditions stabilize without surgical intervention.

By contrast, compliers who undergo surgery exhibit significantly higher wage-replacement payments than non-compliers. The differential is largest in the short-term (€698 per month, $p < 0.01$) and medium-term (€859 per month, $p < 0.10$), consistent with expected recovery periods following surgical procedures. The long-term differential diminishes to €680 per month and loses statistical significance, suggesting convergence as surgical recovery concludes. The static DiD estimate of €808 per month ($p < 0.01$) captures the average differential across all post-periods, reflecting the sustained recovery-related income replacement among compliers.

Unemployment and early retirement show no significant differential effects. The static DiD coefficients are small and statistically zero: -0.11 percentage points for unemployment and 0.10 percentage points for early retirement (columns (2) and (3)). These null findings indicate that surgical compliance neither accelerates workforce detachment nor precipitates premature labor force exit in this subsample.

The wage-replacement pattern contrasts with findings for the surgery-discouraged subsample (Table 3.6), where compliers who avoid surgery exhibit lower wage-replacement payments than non-compliers who proceed with discouraged procedures. Here, the relationship reverses: compliers undergoing recommended surgery incur expected recovery-related work absences, while non-compliers who forgo surgery maintain lower absence levels. Critically, the absence of effects on unemployment and early retirement suggests that neither surgical compliance nor non-compliance generates longer-run labor market scarring in this population. The elevated wage-replacement payments among surgical compliers reflect temporary recovery costs rather than permanent workforce detachment.

Table C.7: Labor Market Outcomes by Compliance – Second Opinion: Surgery Recommended

	Wage-replacement payments	Unemployment	Early retirement
Model:	(1)	(2)	(3)
Dynamic DiD			
pre-baseline	-149.88		
	(207.40)		
short-term	-469.64**		
	(216.26)		
medium-term	-971.16**		
	(439.59)		
long-term	-974.71**		
	(432.92)		
complied × pre-baseline	-124.24		
	(289.61)		
complied × short-term	698.18***		
	(237.86)		
complied × medium-term	858.67*		
	(486.40)		
complied × long-term	679.60		
	(474.20)		
Static DiD			
complied × post	807.60***	-0.11	0.10
	(309.80)	(0.15)	(0.10)
Pseudonym FE	Yes	Yes	Yes
Observations	185	185	185
Dep. Var. Mean ($p =$*baseline*)	674.99	0.11	0.06
No. Patients	37	37	37
No. Complier	16	16	16
No. Non-complier	21	21	21

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), and *long-term* ($t \in \{4, 5\}$). Unemployment and early retirement are reported only in the *static DiD*, as these outcomes are sparse and exhibit limited within-person variation over the observation window. Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

C.5 Alternative Observation Windows

Table C.8: Short-run Health Outcomes – Second Opinion: Surgery Discouraged

	Total health expenditures		Visits to physician		Days in hospital		Sick leave days	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dynamic DiD								
pre-baseline	7.39	5.16	-0.50	-0.46	0.04	0.04	-2.63	-2.91
	(87.34)	(86.96)	(0.41)	(0.41)	(0.05)	(0.05)	(2.70)	(2.88)
short-term	1,538.27***	1,553.92***	-0.21	-0.21	0.27***	0.28***	0.15	3.21
	(209.28)	(221.57)	(0.50)	(0.51)	(0.04)	(0.04)	(2.47)	(2.61)
complied × pre-baseline	59.91	64.06	0.22	0.19	-0.07	-0.07	1.03	1.00
	(155.72)	(158.16)	(0.44)	(0.44)	(0.05)	(0.06)	(2.94)	(3.11)
complied × short-term	-1,598.34***	-1,611.62***	-0.56	-0.56	-0.30***	-0.31***	-3.81	-6.45**
	(216.20)	(229.65)	(0.53)	(0.53)	(0.05)	(0.05)	(2.67)	(2.81)
Static DiD								
complied × post	-1,628.29***	-1,643.74***	-0.67	-0.65	-0.26***	-0.28***	-4.32	-6.86**
	(211.58)	(226.78)	(0.43)	(0.45)	(0.05)	(0.05)	(3.10)	(3.21)
Insurance-status controls	No	Yes	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	441	441	441	441	441	441	237	237
Dep. Var. Mean (p =*baseline*)	349.25	349.25	3.08	3.08	0.09	0.09	13.73	13.73
No. Patients	147	147	147	147	147	147	79	79
No. Complier	114	114	114	114	114	114	67	67
No. Non-complier	33	33	33	33	33	33	12	12

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), and *short-term* ($t \in \{0, 1\}$). Two columns per outcome report specifications without vs. with insurance-status controls. Columns (7) and (8) use the ever-employed subsample, since certified sick leave is only reliably recorded for individuals in employment. Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

Table C.9: Short-run Health Outcomes – Second Opinion: Surgery Recommended

	Total health expenditures		Visits to physician		Days in hospital		Sick leave days	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dynamic DiD								
pre-baseline	30.72	10.77	-0.25	-0.30	-0.10*	-0.10*	-5.19	-5.18
	(135.03)	(139.06)	(0.42)	(0.42)	(0.05)	(0.05)	(3.48)	(3.50)
short-term	-99.26	-71.78	-0.57*	-0.43	-0.09	-0.09	-11.85**	-11.85**
	(102.58)	(96.07)	(0.31)	(0.27)	(0.08)	(0.09)	(4.31)	(4.33)
complied × pre-baseline	0.22	20.16	-0.39	-0.35	0.10	0.10	0.89	0.87
	(234.09)	(238.54)	(0.60)	(0.60)	(0.07)	(0.07)	(4.96)	(4.93)
complied × short-term	1,193.38***	1,163.39***	0.22	0.11	0.28***	0.28***	20.00***	19.98***
	(238.13)	(221.32)	(0.45)	(0.43)	(0.09)	(0.10)	(4.92)	(4.92)
Static DiD								
complied × post	1,193.28***	1,153.38***	0.42	0.28	0.23***	0.23***	19.55***	19.51***
	(185.43)	(185.63)	(0.43)	(0.44)	(0.07)	(0.08)	(4.91)	(4.95)
Insurance-status controls	No	Yes	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	207	207	207	207	207	207	87	87
Dep. Var. Mean (p =*baseline*)	461.35	461.35	3.04	3.04	0.10	0.10	13.28	13.28
No. Patients	69	69	69	69	69	69	29	29
No. Complier	32	32	32	32	32	32	13	13
No. Non-complier	37	37	37	37	37	37	16	16

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), and *short-term* ($t \in \{0, 1\}$). Two columns per outcome report specifications without vs. with insurance-status controls. Columns (7) and (8) use the ever-employed subsample, since certified sick leave is only reliably recorded for individuals in employment. Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

Table C.10: Long-run in Health Outcomes – Second Opinion: Surgery Discouraged

	Total health expenditures		Visits to physician		Days in hospital		Sick leave days	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dynamic DiD								
pre-baseline	-97.17*	-97.17*	-0.83**	-0.83**	-0.02	-0.02	-2.32	-2.32
	(53.50)	(53.88)	(0.32)	(0.32)	(0.01)	(0.01)	(1.50)	(1.50)
short-term	949.86***	1,041.17***	1.06	1.01	0.19***	0.20***	1.48	4.14**
	(260.48)	(251.38)	(0.71)	(0.63)	(0.07)	(0.06)	(1.08)	(1.71)
medium-term	134.07	243.64	0.11	0.05	0.01	0.03	-0.44	2.76
	(139.69)	(174.04)	(0.46)	(0.53)	(0.03)	(0.03)	(1.22)	(2.50)
year 2	295.03	408.46	1.00	0.96	0.01	0.02	-2.67	0.30
	(264.47)	(298.51)	(0.70)	(0.62)	(0.02)	(0.03)	(1.69)	(3.56)
year 3	107.40	226.17	0.52	0.48	0.00	0.02	-1.04	2.02
	(150.99)	(193.88)	(0.46)	(0.49)	(0.02)	(0.03)	(1.07)	(2.73)
complied × pre-baseline	-16.75	-16.75	0.40	0.40	-0.02	-0.02	0.89	0.89
	(104.16)	(104.33)	(0.42)	(0.42)	(0.05)	(0.05)	(2.18)	(2.14)
complied × short-term	-1,087.98***	-1,177.99***	-1.47*	-1.37*	-0.26***	-0.26***	-4.95***	-7.56***
	(270.80)	(261.01)	(0.78)	(0.71)	(0.07)	(0.07)	(1.75)	(2.24)
complied × medium-term	-132.69	-242.63	-0.91	-0.84	-0.06	-0.07	-6.83**	-9.17**
	(175.37)	(203.33)	(0.55)	(0.60)	(0.04)	(0.05)	(2.71)	(3.45)
complied × year 2	-311.76	-416.05	-1.28	-1.22*	-0.09**	-0.10**	-2.94	-4.96
	(290.23)	(313.31)	(0.77)	(0.70)	(0.04)	(0.04)	(3.11)	(4.41)
complied × year 3	330.79	262.01	-1.13*	-1.09*	-0.07	-0.08*	-5.11*	-7.46*
	(434.40)	(436.66)	(0.57)	(0.58)	(0.04)	(0.05)	(2.74)	(3.87)
Static DiD								
complied × post	-292.03	-371.84*	-1.40***	-1.35***	-0.11***	-0.12***	-5.41***	-7.87***
	(211.87)	(219.11)	(0.51)	(0.49)	(0.03)	(0.03)	(1.98)	(3.04)
Insurance-status controls	No	Yes	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	366	366	366	366	366	366	192	192
Dep. Var. Mean (p =*baseline*)	365.20	365.20	2.70	2.70	0.11	0.11	9.05	9.05
No. Patients	61	61	61	61	61	61	32	32
No. Complier	49	49	49	49	49	49	27	27
No. Non-Complier	12	12	12	12	12	12	5	5

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), *year 2* ($t \in \{4, 5, 6, 7\}$), and *year 3* ($t \in \{8, 9, 10, 11\}$).Two columns per outcome report specifications without vs. with insurance-status controls. Columns (7) and (8) use the ever-employed subsample, since certified sick leave is only reliably recorded for individuals in employment. In columns (4), and (6), the clustered VCOV is not numerically positive semi-definite; inference uses the eigenvalue adjustment described in Cameron and Miller (2015). Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

Table C.11: Long-run in Health Outcomes – Second Opinion: Surgery Recommended

	Total health expenditures		Visits to physician		Days in hospital		Sick leave days	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dynamic DiD								
pre-baseline	16.52	16.52	-0.29	-0.29	-0.06	-0.06	-1.81	-1.81
	(210.53)	(213.05)	(0.67)	(0.68)	(0.05)	(0.05)	(6.64)	(6.73)
short-term	-70.43	-114.38	-0.31	-0.23	0.01	0.00	-14.60**	-13.54**
	(186.69)	(230.87)	(0.79)	(0.65)	(0.13)	(0.14)	(4.64)	(5.46)
medium-term	157.73	37.48	-0.35	-0.61	-0.02	-0.05	-27.24***	-23.00***
	(145.55)	(172.51)	(0.77)	(0.58)	(0.10)	(0.12)	(2.90)	(3.04)
year 2	486.26	349.99	0.31	0.03	0.01	-0.02	-27.57***	-22.80***
	(383.88)	(446.35)	(1.16)	(1.34)	(0.13)	(0.15)	(2.90)	(2.92)
year 3	207.96	43.90	1.07	1.08	-0.07	-0.12	-27.14***	-22.37***
	(245.44)	(351.32)	(1.48)	(1.77)	(0.09)	(0.11)	(3.33)	(3.11)
complied × pre-baseline	-81.17	-81.17	0.39	0.39	0.02	0.02	7.31	7.31
	(223.00)	(225.67)	(1.06)	(1.08)	(0.09)	(0.09)	(7.41)	(7.51)
complied × short-term	707.13**	751.08**	0.87	0.80	0.11	0.13	25.26***	24.20***
	(272.55)	(306.07)	(1.10)	(1.01)	(0.14)	(0.15)	(5.17)	(5.93)
complied × medium-term	224.98	345.22	-0.35	-0.09	0.08	0.11	34.49***	30.25***
	(323.56)	(339.52)	(0.96)	(0.83)	(0.16)	(0.17)	(8.13)	(8.28)
complied × year 2	-637.90	-501.63	-0.88	-0.60	-0.14	-0.11	28.90***	24.14***
	(391.94)	(453.47)	(1.24)	(1.41)	(0.14)	(0.16)	(3.80)	(3.84)
complied × year 3	-72.66	91.40	-1.54	-1.54	0.09	0.14	24.77***	20.00***
	(308.22)	(398.78)	(1.52)	(1.80)	(0.15)	(0.16)	(4.06)	(3.90)
Static DiD								
complied × post	95.97	229.96	-0.67	-0.43	0.03	0.05	24.70***	19.14***
	(249.16)	(297.16)	(1.09)	(1.13)	(0.09)	(0.11)	(3.64)	(5.68)
Insurance-status controls	No	Yes	No	Yes	No	Yes	No	Yes
Pseudonym FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	96	96	96	96	96	96	48	48
Dep. Var. Mean (p =*baseline*)	331.29	331.29	2.45	2.45	0.12	0.12	9.06	9.06
No. Patients	16	16	16	16	16	16	8	8
No. Complier	8	8	8	8	8	8	4	4
No. Non-Complier	8	8	8	8	8	8	4	4

Notes: Estimates are weighted least squares (WLS) from an event-study specification with individual (Pseudonym) fixed effects on matched samples using elastic-net ($\lambda_{\min}$) propensity scores with full matching. The dependent variables are measured in mean monthly values by hybrid period. The omitted reference period is *baseline* ($t = -1$); coefficients on period indicators report changes for non-compliers relative to *baseline*, and the interaction terms (*complied* × period) report differential changes for compliers relative to non-compliers (also normalized to zero in *baseline*). Hybrid periods are defined as *pre-baseline* ($t = -2$), *baseline* ($t = -1$), *short-term* ($t \in \{0, 1\}$), *medium-term* ($t \in \{2, 3\}$), *year 2* ($t \in \{4, 5, 6, 7\}$), and *year 3* ($t \in \{8, 9, 10, 11\}$).Two columns per outcome report specifications without vs. with insurance-status controls. Columns (7) and (8) use the ever-employed subsample, since certified sick leave is only reliably recorded for individuals in employment. In columns (2), (4), (6), (7), and (8) the clustered VCOV of the *Dynamic DID* is not numerically positive semi-definite; inference uses the eigenvalue adjustment described in Cameron and Miller (2015). Standard errors are clustered at the Pseudonym level. Signif. codes: *** 0.01, ** 0.05, * 0.10.

C.6 Instrument Variable Approach

Table C.12: Placebo Test: Effect of Instrument on Pre-Period Outcome changes

	Δ Total health expenditure	Δ Physician visits	Δ Hospital days	Δ Sick leave
	(1)	(2)	(3)	(4)
SO: Surgery discouraged	20.57	−0.06	−0.02	−0.43
	(190.23)	(0.29)	(0.04)	(1.08)
Observations	231	231	231	231

Notes: Reduced-form regressions test whether the instrument (second-opinion recommendation against surgery) predicts outcome changes within the pre-period. Outcomes measured as changes in mean monthly values between months $t \in \{-6, -4\}$ (early pre-period) and months $t \in \{-3, -1\}$ (late pre-period). All coefficients are statistically indistinguishable from zero ($p > 0.65$ for all outcomes), providing no evidence of pre-existing differential trends. Heteroskedasticity-robust standard errors in parentheses.

Eidesstattliche Erklärung

"I hereby affirm that I have written this dissertation independently and without the use of other aids and literature than those indicated. No other persons, apart from any co-authors listed in the thesis, were involved in the substantive preparation of this thesis. All passages that have been taken verbatim or analogously from published and unpublished works of others are marked as such. I affirm that this dissertation has not yet been submitted to any other Faculty or university for examination; that it has not yet been published – apart from the partial publications and incorporated articles and manuscripts indicated – and that I will not publish the dissertation before completion of the doctorate without the approval of the dissertation committee. I am aware of the provisions of these regulations. Furthermore, I hereby declare that I have read the Guidelines of the University of Cologne for Ensuring Good Scientific Practice and that I have observed them in carrying out the work underlying the dissertation and in writing the dissertation, and I hereby undertake to observe and implement the guidelines stated therein in all scientific activities. I affirm that the submitted electronic version corresponds completely to the submitted printed version. I affirm that to the best of my knowledge I have been truthful and have not concealed anything."

Köln, den 09.07.2026	gez. Felix Mindl
Ort, Datum	Unterschrift

Declaration on the Use of Generative AI

The author used Claude (Anthropic, versions from 2025 to 2026) and ChatGPT (OpenAI, versions 4.0 to 5.5) during the preparation of this work to assist with editing and assembling the LaTeX manuscript, identifying grammatical and typographical errors, and revising statistical code. After using these AI tools, the author reviewed and edited the content as needed manually and takes full responsibility.

Köln, den 09.07.2026 — gez. Felix Mindl

Ort, Datum — Unterschrift