Peer Comparison Feedback and Professional Norms to Reduce Low-Value Care

Tobias Müller, Ph.D.[1] Raf Van Gestel, Ph.D.[2] Michael Gerfin, Ph.D.[3]

Study Protocol & SAP

This document contains the following items:

1. Study Protocol
2. Statistical Analysis Plan (SAP)

[1] First & corresponding author: Bern University of Applied Sciences, Institute of Health Economics & Policy, Bern (Switzerland), E-Mail: tobias.mueller@bfh.ch
[2] Erasmus School of Health Policy and Management & Erasmus School of Economics, Rotterdam (Netherlands), E-Mail: vangestel@eshpm.eur.nl
[3] University of Bern, Department of Economics, Bern (Switzerland), E-Mail: michael.gerfin@unibe.ch

1. Study Protocol

Background

For decades now, health spending has outpaced income growth in most health care systems around the globe. While substantial medical progress has been made, growing evidence shows that the health care sector is plagued by major inefficiencies. Numbers from the OECD suggest that up to one-fifth of health spending in the developed world are wasteful at best, or harmful at worst.[1]

Different campaigns and reforms have been launched to address the problem. "Choosing Wisely" in the United States and "Smarter Medicine" in Switzerland are just two promising examples of platforms that provide information to patients and physicians about commonly used medical procedures with little health benefits. Despite positive intentions, research suggests that between 30-40% of patients receive care that is not consistent with present evidence-based clinical guidelines.[2,3]

We present a pragmatic randomized controlled trial that evaluates the impact of a multi-faceted information intervention on the practice of low-value care among primary care physicians (PCPs) in Switzerland. The information intervention combines professional norms and peer comparison information and is constructed based on routinely collected insurance claims data. The intervention aims at reducing the use of low-value care services in two domains: vitamin D testing and the use of branded drugs when generic equivalents are available. In what follows, we lay out the rationale and design of the trial.

Methods and Design

Study Design

This primary objective of this trial is to assess the impact of a multi-faceted information intervention on the practice of low-value care among PCPs in Switzerland. To this end, we design a nationwide pragmatic randomized trial that uses routine health data collected by SASIS AG, the data warehouse of the industry association of the Swiss health insurers (santésuisse; details further below). The trial is based on a randomized parallel-group design allocating PCPs to either a control or one of the three intervention arms described below.

Prior to the trial, a contact letter was sent to all PCPs describing the general purpose of the trial, data privacy issues and the possibility to opt-out of the trial ("opt-out card"). After giving consent, PCPs in the intervention arms received a feedback report at the start of the trial. No information letter is sent to the PCPs in the control group.

Intervention and Control

PCPs randomized into one of three intervention groups will be sent an information letter containing professional norms on low-value care practice and peer comparison feedback on their health care expenditures or low-value care provision. PCPs in the control group receive no information letter. The letter will be sent out once at the beginning of the intervention period in November of 2020, and information on the primary endpoints described below will be gathered one year after the intervention. Figure 1 below outlines the precise timeline and milestones in the trial.

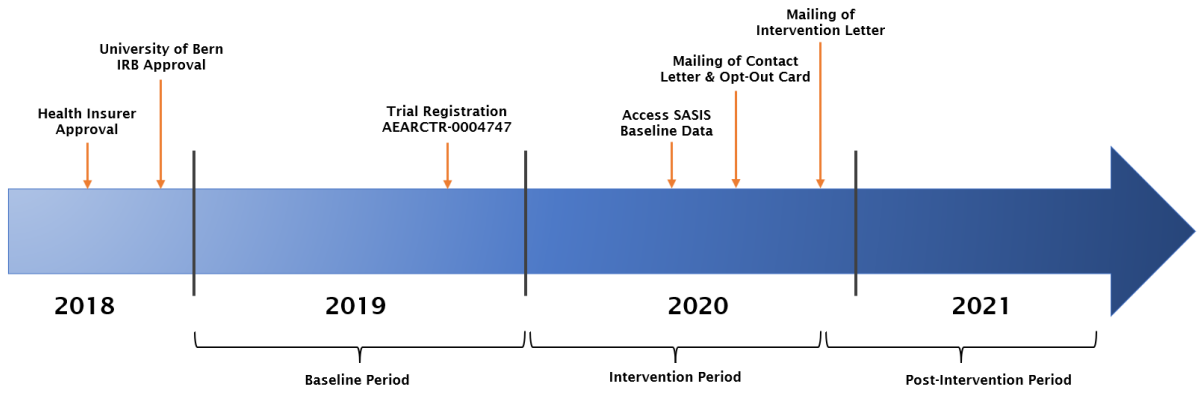


Figure 1: Trial Timeline

Inclusion and Exclusion Criteria

All health care providers who bill their services via mandatory health insurance in Switzerland can be identified based on a so-called central registration number ("ZSR"), a unique identifier. We apply the following eligibility criteria when selecting the PCPs for the trial:

- Eligibility is based on the following list of specializations: General practitioners ("Praktische Ärzte"), general medicine ("Allgemeinmedizin") or general internal medicine ("Allgemeine Innere Medizin")
- Exclusion of PCPs working in group practices
- Exclusion of PCPs with less than 100 patients per year
- Exclusion of high-cost outliers (santésuisse outliers)

We apply these exclusion restrictions to the data collected by SASIS AG in 2019 to construct the information intervention letters.

Study Population

The population of PCPs amounts to 8,052 physicians in Switzerland in 2019 in the SASIS data (see Figure 2). After applying the above exclusion criteria, 4,782 PCPs are eligible for recruitment, and they received the general information letter including the option to opt out

of the trial. In total, 1,718 providers dropped out of the trial leaving 3,064 PCPs for randomization. Balanced assignment to one of the three intervention groups, or the control group, results in equal ex-ante group sizes of 766 PCPs.

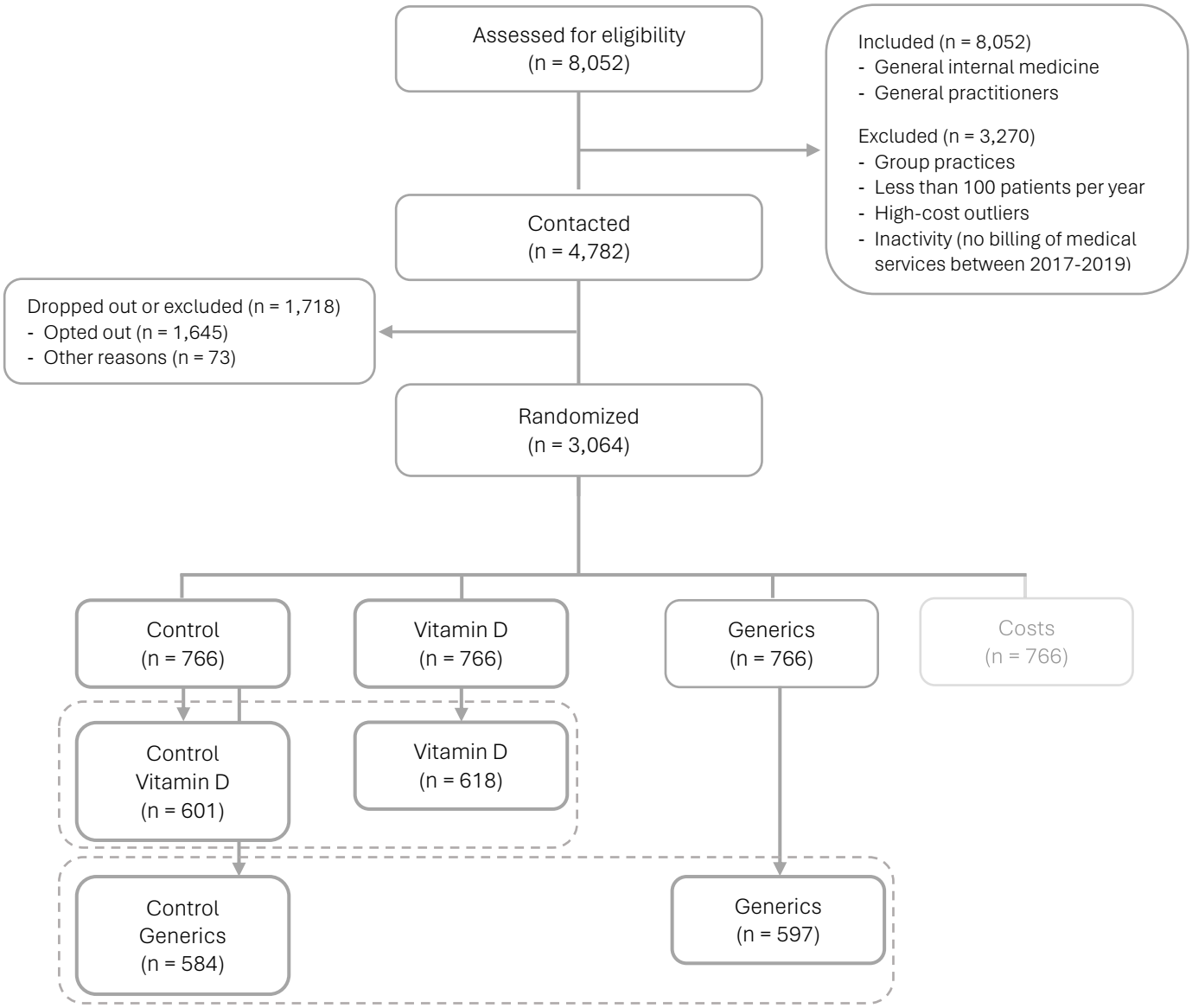


Figure 2: Trial profile

After the exclusion of PCPs with data errors in the number of vitamin D tests in 2020, PCPs with practice closures in 2021, missing value or negative values in endpoints, the vitamin D intervention arm consists of n = 618 PCPs and n = 601 in the control group. Similarly, the generic prescribing arm contains n = 597 PCPs and the control group n = 584. Overall, fewer observations remain in the generic prescribing arm due to a higher opt-out rate, more practice closures as well as more missing

values in endpoints because SASIS AG does not update generic rates and cost indices for physicians with less than 50 patients per year.

Data Source

The data for this study is provided by SASIS AG, the largest data warehouse for administrative health insurance claims data including patient characteristics and data on the provision of health care services of physicians in the outpatient sector covering more than 95% of the Swiss population of approximately 8.5 million people in 2019.[4]

Access to data depends on the consent of all health insurers prior to the trial. The consortium of 60 insurers of santésuisse meets quarterly to vote on data approvals for research projects. In the case of this trial, all 60 health insurers approved the use of the SASIS data for the trial (without a dissentient vote).

Besides approval, SASIS AG charges a fee for data preparation and assistance throughout the trial. The fee varies by complexity and size of the project.

Time Lags in the Data

An important issue to consider is the time lag that occurs between the date of medical treatment (e.g., date of a drug prescription) and the corresponding "footprint" in the SASIS data. In fact, three months after medical treatment approximately 84% of medical services are recorded in the SASIS claims data. After 6 months, roughly 99% are.

For this trial, the delay in data recording means that only 84% (99%) of, e.g., vitamin D tests provided in October 2020 are covered in the data collected in January (April) 2021. Consequently, the potential reaction to the information intervention in 2021 will only be fully observable in the SASIS data in the summer of 2022.

Intervention Letters

Figure 3-5 below provide an example of a vitamin D information letter. The letter contains three main components:

- Page 1: Professional norms on vitamin D testing

The information letters start off with a general description of the medical progress that has been achieved and the cost developments in developed countries over the past four decades (see Figure 3 below). Then, OECD estimates about the potential waste in health care are given before presenting scientific evidence on the appropriateness of vitamin D testing. Articles from the *New England of Medicine* (Manson et al., 2016) and *JAMA* (Zhao et al., 2017) are referenced and an excerpt from the article of Petrilli et al. (2018)[5] in the *American Journal of Medicine* is explicitly cited in the letter:

> *"Therefore, in an effort to reduce unnecessary testing, the American Society for Clinical Pathology's Choosing Wisely recommendation states "Don't perform population-based screening for 25-OH-Vitamin D deficiency," noting that laboratory testing is appropriate in higher-risk patients when results will be used to institute more aggressive therapy (e.g., osteoporosis, chronic kidney disease, malabsorption, some infections, obese individuals)."*

The introductory page ends with the involved institutions (SASIS AG & University of Bern), the funding source (Swiss National Science Foundation) and points to the personalized feedback the PCPs will receive on the following pages.

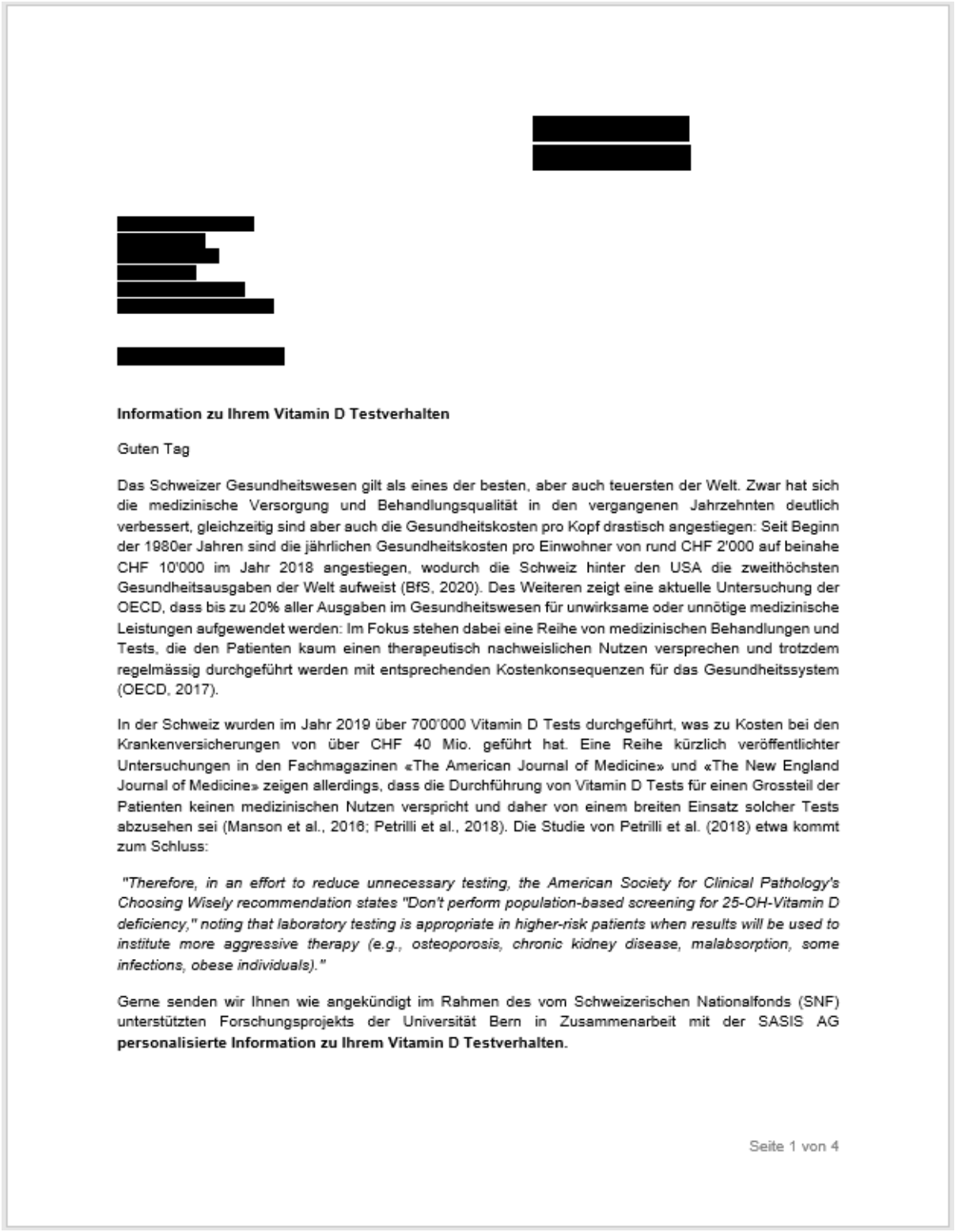
- Page 2: Peer comparison information on vitamin D testing

The second page shows the number of vitamin D tests in 100 patients (red bars) and the overall average among all PCPs (green bars) (see Figure 4 below). The same numbers are also given separately by male and female patients. The bottom graph shows the distribution of vitamin D tests, the location of the PCP (“You”; “Sie” in German), the group average (green line) and the 90th

percentile. The graphical feedback is accompanied by verbal explanations of the statistical information. In particular, the letter spells out the average number of tests of the PCP, mentions the average across all PCPs and explicitly states whether the corresponding PCP conducts fewer or more tests than the average PCP. Likewise, an interpretation of the distribution is given. PCPs are given the precise location in the distribution and a verbal statement tells them the share of PCPs who conduct fewer vitamin D tests. For example, the PCP in the example letter is told that he/she conducts more tests than 89% of PCPs in Switzerland in 2019.

- Page 3: Cost implications of the low-value care practice

The third page of the information letter complements the testing rates with the cost implications of vitamin D testing (see Figure 5 below). The letter shows the distribution of total costs from vitamin D testing as well as the location of the PCP, the group average (green line) and the 90th percentile. Verbal explanations spell out the total costs of testing for the PCP under consideration, the average costs of testing and the precise percentile of the PCP. PCPs are told whether their testing behavior yields above (below) average costs, and the letter also states the percentage of PCPs with lower testing costs (e.g., 77% of PCPs have lower vitamin D testing costs in the example letter below).



[redacted]

[redacted]

[redacted]

Information zu Ihrem Vitamin D Testverhalten

Guten Tag

Das Schweizer Gesundheitswesen gilt als eines der besten, aber auch teuersten der Welt. Zwar hat sich die medizinische Versorgung und Behandlungsqualität in den vergangenen Jahrzehnten deutlich verbessert, gleichzeitig sind aber auch die Gesundheitskosten pro Kopf drastisch angestiegen: Seit Beginn der 1980er Jahren sind die jährlichen Gesundheitskosten pro Einwohner von rund CHF 2'000 auf beinahe CHF 10'000 im Jahr 2018 angestiegen, wodurch die Schweiz hinter den USA die zweithöchsten Gesundheitsausgaben der Welt aufweist (BfS, 2020). Des Weiteren zeigt eine aktuelle Untersuchung der OECD, dass bis zu 20% aller Ausgaben im Gesundheitswesen für unwirksame oder unnötige medizinische Leistungen aufgewendet werden: Im Fokus stehen dabei eine Reihe von medizinischen Behandlungen und Tests, die den Patienten kaum einen therapeutisch nachweislichen Nutzen versprechen und trotzdem regelmässig durchgeführt werden mit entsprechenden Kostenkonsequenzen für das Gesundheitssystem (OECD, 2017).

In der Schweiz wurden im Jahr 2019 über 700'000 Vitamin D Tests durchgeführt, was zu Kosten bei den Krankenversicherungen von über CHF 40 Mio. geführt hat. Eine Reihe kürzlich veröffentlichter Untersuchungen in den Fachmagazinen «The American Journal of Medicine» und «The New England Journal of Medicine» zeigen allerdings, dass die Durchführung von Vitamin D Tests für einen Grossteil der Patienten keinen medizinischen Nutzen verspricht und daher von einem breiten Einsatz solcher Tests abzusehen sei (Manson et al., 2016; Petrilli et al., 2018). Die Studie von Petrilli et al. (2018) etwa kommt zum Schluss:

"Therefore, in an effort to reduce unnecessary testing, the American Society for Clinical Pathology's Choosing Wisely recommendation states "Don't perform population-based screening for 25-OH-Vitamin D deficiency," noting that laboratory testing is appropriate in higher-risk patients when results will be used to institute more aggressive therapy (e.g., osteoporosis, chronic kidney disease, malabsorption, some infections, obese individuals)."

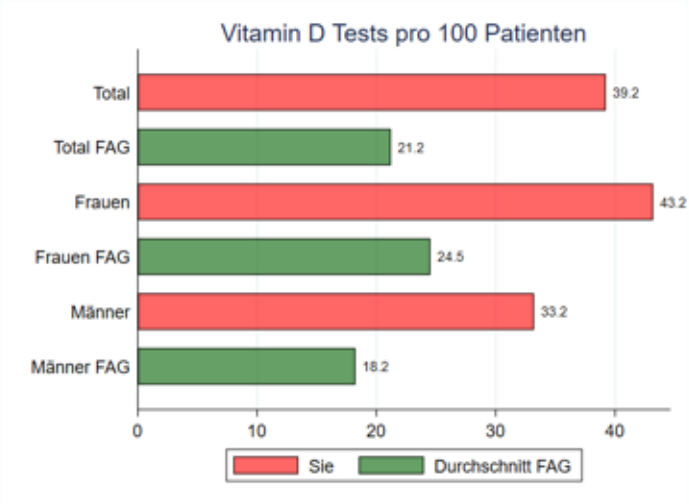
Gerne senden wir Ihnen wie angekündigt im Rahmen des vom Schweizerischen Nationalfonds (SNF) unterstützten Forschungsprojekts der Universität Bern in Zusammenarbeit mit der SASIS AG **personalisierte Information zu Ihrem Vitamin D Testverhalten.**

Seite 1 von 4

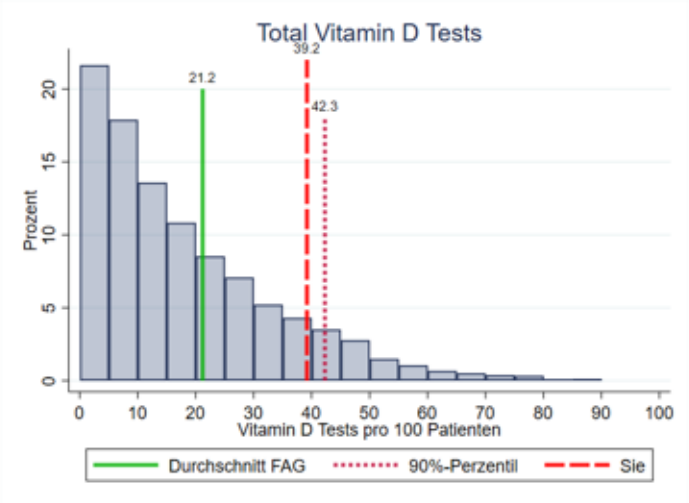
Figure 3: Page 1 of the information intervention showing scientific evidence on vitamin D testing

Information zu Ihrem Vitamin D Testverhalten für das Jahr 2019

Die nachfolgende Abbildung zeigt Ihr Vitamin D Testverhalten pro 100 Patienten (Total und nach Geschlecht der Patienten) sowie den Durchschnitt in Ihrer Facharztgruppe (FAG):



Sie haben im Jahr 2019 auf 100 Patienten **39.2** Vitamin D Tests durchgeführt. Damit liegen Sie **über** dem Durchschnitt von rund **21.2** getesteten Patienten in Ihrer Facharztgruppe. Insgesamt führen Sie im Schnitt **mehr** Vitamin D Tests auf 100 Patienten durch als Ihre Berufskolleginnen und -kollegen.



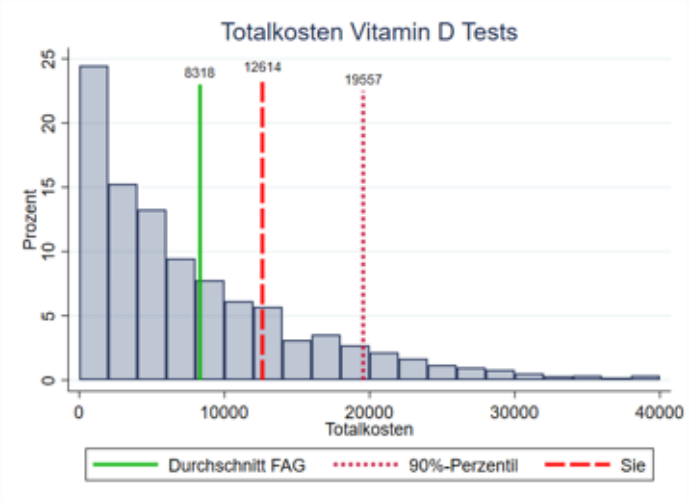
Seite 2 von 4

Figure 4: Page 2 of the information letter providing peer feedback on vitamin D testing.

Als zusätzliche Information zeigt ihnen die obenstehende Grafik die Verteilung der Vitamin D Tests auf 100 Patienten in Ihrer Facharztgruppe sowie Ihre Position in der Verteilung. Bei **39.2** Vitamin D Tests auf 100 Patienten haben rund **87.0%** der Leistungserbringer in Ihrer Facharztgruppe weniger Tests durchgeführt als Sie.

Informationen zu Ihren Vitamin D Testkosten für das Jahr 2019

Als zusätzliche Information finden Sie in der oben abgebildeten Grafik die Verteilung der totalen Vitamin D Kosten für das Jahr 2019 und Ihre Position in der Kostenverteilung. Die totalen Kosten für Vitamin D Tests beliefen sich in Ihrer Praxis auf **CHF 12614**, was **über** dem Durchschnitt von rund **CHF 8318** in Ihrer Facharztgruppe liegt. Insgesamt verursachen damit **77.0%** der Leistungserbringer in Ihrer Facharztgruppe **tiefere** Vitamin D Testkosten im Jahr 2019 als Sie.



Wir danken Ihnen für die Kenntnisnahme und wünschen Ihnen einen schönen Tag!

Freundliche Grüsse

Prof. Dr. Michael Gerfin
Brief ohne Unterschrift

Figure 5: Page 3 of the information letter providing feedback on the cost implications of testing.

Intervention Arms

Health Care Costs

PCPs in this first treatment arm receive information on the health care expenditures they caused in comparison to GPs with a comparable risk pool of patients. Specifically, the measures we use are the (risk-adjusted) total health care costs per patient. We also consider indirect costs that arise from the doctor consultation including the costs from laboratory tests ordered by the physician. Moreover, we decompose these overall expenditures in subcategories of care such as lab and pharmaceutical expenditures and provide peer comparison information on these dimensions in the peer report. The different cost categories are risk-adjusted based on a two-step regression approach that accounts for a broad set of morbidity indicators, patient characteristics, location- and physician-specific factors (see section on primary endpoints below for details on risk-adjustment).

Vitamin D Testing

Recommendation: The "Choosing Wisely" campaign, a joint initiative by the American Board of Internal Medicine Foundation, and a series of medical organizations advocate for significant reductions in the use of low-value services. Vitamin D deficiency is prevalent in the adult population, and this leads to a high rate of Vitamin D deficiency testing. However, only few adults have "seriously low levels" of Vitamin D. Vitamin D testing is therefore only recommended for adults when this can be used for further (aggressive) therapy. The total laboratory cost of Vitamin D testing in Switzerland is estimated at $97.2 million in 2018 and an estimated 20% of the insured population received a Vitamin D test in that year. Since worldwide guidelines identify testing as unnecessary, and over-the-counter Vitamin D supplements are widely available, this may impose a large and wasteful cost on the healthcare system.[6] We summarize this argument in the intervention letter by citing the study by Petrilli et al. (2018)[7], this can be seen in the example letter above.

Measurements: SASIS AG, the data warehouse for administrative health insurance claims data, possesses coded information on laboratory testing, to the level of the specific test. We use claims data on the 25-Hydroxy-Vitamin-D test since this is widely recognized to be the most accurate test for testing Vitamin D deficiency. 25-Hydroxy-Vitamin-D test data are then used to construct the number of tests per PCP.

Peer Comparison Information: PCPs in the Vitamin D testing arm are given peer comparison information regarding the number of tests (per 100 patients) they prescribe to their patients. In addition, the information letter shows the distribution of the number of tests across all PCPs, the location of the corresponding doctor ("You") and the exact percentile, the 90th percentile and the overall group average.

Generic Prescribing

Recommendation: The "Choosing Wisely" campaign, a joint initiative by the American Board of Internal Medicine Foundation, and a series of medical organizations advocate for significant reductions in the use of low-value services. The use of brand-name drugs for which generic equivalents are available are a prime example of a low-value service that should be avoided. Despite their clinical equivalence to brand-name drugs, the market share of generic drugs is as low 18% (23%) in sales (volumes) in Switzerland in 2019 which is far below rates observed in other OECD countries (OECD averages of 25% in sales and 52% in volume) thus leaving substantial scope for savings in pharmaceutical expenditures.[8]

We summarize this argument in the intervention letter by citing a study by Patel *et al.* (2014):[9]

> *"Prescribing brand-name medications that have existing generic equivalents is a prime example of a low-value service. These medications are often more expensive than their generic equivalents, yet in most cases evidence suggests they are similar in effectiveness. A recent study of 20 popular multisource drugs found that in 2009, Medicaid spent an additional $329*

million that could have been saved by using existing generic equivalents instead of brand-name medications."

Measurements: SASIS AG regularly updates the list of generic drugs (respectively the list of "Pharmacodes") and based on this information they compute the share of generic drugs (sales and volumes) relative to all drug prescriptions for each health care provider. They also compute the generic rate among branded drugs with generic equivalents. The latter measure is used in this trial.

Peer Comparison Information: PCPs in the generic prescribing arm are given peer comparison information regarding the share of generics they prescribe to their patients (for branded drugs with generic equivalents). In addition, the information letter shows the distribution of the generic substation rate across all PCPs, the location of the corresponding doctor ("You") and the exact percentile, the 90th percentile and the overall group average. The same statistics are shown for the distribution of (risk-adjusted) drug costs per patient (see section on primary endpoints below for details).

Primary Endpoints

The primary endpoints in this trial are:

- Risk-adjusted average total costs per patient

Risk-adjustment is based on different morbidity criteria of the patient pool (age and gender, deductible rate, hospitalization in the previous year and pharmaceutical cost groups) as well as location- and provider-specific attributes. The adjusted costs are translated to a regression index which normalizes the index value to 100 for the average PCP in Switzerland. Consequently, PCPs with an index value above (below) the value of 100 have higher (lower) costs per patient than the average provider with a comparable patient structure and location conditions. The regression index is annually computed by santésuisse. For the trial, we get access to the regression index for the total costs for the years 2019-2021.

- Number of vitamin D tests in 100 patients

PCPs must order vitamin D tests from external laboratories in Switzerland. Identification of laboratory tests is based on unique tariff codes in the so-called analysis list ("Analyseliste"). Vitamin D tests (25-Hydroxy-Vitamin-D (25-OH-D)) have the tariff code 1006.00. We get access to the number of vitamin D tests and the number of patients per PCP for the years 2017-2021.

- Share of generics prescribed for branded drugs with generic equivalents

The generic rate is computed as the share of generics prescribed to patients for branded drugs for which generics are available. Santésuisse annually updates the generic rate based on ATC codes for all prescription drugs that can be reimbursed via mandatory health insurance. We get access to the generic share by PCP for the years 2019-2021.

Other outcomes are the regression index for drug, physician, laboratory and physiotherapy and appliances costs for the years 2019-2021. Additional outcomes are the number of CTs (Tarmed 39.4020-39.4090); Ultrasound scans (Tarmed 39.3220-39.3290); MRIs (Tarmed 39.5020-39.5190); x-rays (Tarmed 39.0100-39.0380); ECGs (Tarmed 17.001 & 17.002); mammographies (Tarmed 39.1307 & 39.1308); colonoscopies (Tarmed 19.1010-19.1400); glucose tolerance tests (Tarmed 00.2080); urine tests (Tarmed 39.6710-39.6730) per PCP for the years 2017-2021. In addition, we have access to the number of laboratory tests that are conducted within PCP offices.

We use these other outcomes for sensitivity checks and to evaluate potential spillover effects of the information intervention.

Privacy

To meet the ethical and legal standards common in pragmatic randomized trials and to ensure anonymity of providers throughout the entire field study, we closely follow the successfully implemented study protocol by Hemkens et al. (2016).[10] SASIS AG is going to provide anonymized physician-level data for the years 2017-2019 to the project team. In a first step, an independent

statistician at SASIS AG assigns anonymous identifiers (IDs) to all physicians before sharing the data with the team of researchers. Throughout the entire study, perfect concealment of individual providers is assured as the researchers only have access to anonymized provider-year data. In a next step, the researchers prepare the peer comparison information letters and randomly assign the physicians to the different treatment groups and a control group. In a final step, digital copies of the social comparison letters are sent back to SASIS AG, and the anonymous IDs are linked back to the actual address information of the physicians. Employees at the location of SASIS not involved in the study then address, pack, and mail the information letters to the PCPs in the different treatment groups.

2. Statistical Analysis Plan (SAP)

Average Treatment Effects

We use a multiple linear regression model to evaluate the average impact of the randomized information intervention based on data in the post-intervention period 2021. To be specific, we estimate regression models of the form:

$$y_i = \beta_0 + \beta_1 d_i + \beta_2 x_{i1} + \ldots + \beta_p x_{ip-1} + \varepsilon_i$$

Where y_i is the primary endpoint, $d_i$is a binary indicator that takes on the value one when PCP i is part of the intervention group and zero for PCPs in the control group. β_1 is our main coefficient of interest and is interpreted as the average difference in primary endpoints between the intervention and control groups. The errors, ε_i, are assumed to be normally distributed. Control variables x_{i1} to x_{ip-1} are included to account for potential imbalances between the intervention and control groups, and to improve statistical accuracy. The control variables include PCP gender, age, dispensing status, work region (canton), share of female patients, and share of old/young patients. Note that due to random assignment, the estimated parameter $\widehat{\beta_1}$ is an unbiased estimate of average treatment effect (ATE) of the intervention.

Pre-intervention balance tests are reported, in which Bonferroni-Holm corrections are applied to account for multiple hypothesis testing.[11,12]

Effect Heterogeneity

We analyze effect heterogeneity based on the causal forest. The causal forest is a data-driven approach to estimate heterogeneity in causal effects in experimental studies.[13-15] The causal forest algorithm ensures that hypothesis tests about the magnitude of differences in treatment effects between subgroups in the data are valid without pre-specifying groups. In more intuitive terms, the method enables researchers a) to detect subgroups in the data that systematically vary in their

response to treatment/intervention without the need to pre-specify these groups and b) to conduct valid hypothesis tests about differences in the treatment effect between these subgroups. In this trial, we are interested in identifying PCP subgroups that vary in their response to the information intervention and test for significant differences between these subpopulations.

The causal forest builds on Breiman's famous random forest algorithm.[16] However, instead of splitting the data by minimizing a mean-squared error criterion, sample partitioning is based on maximizing heterogeneity in treatment effects between subgroups or "terminal leaves". Like the standard random forest where all units in a subgroup receive the same outcome prediction, units who fall in the same subgroup have the same treatment effect. Hence, treatment effects vary between subgroups but test points within a subgroup have the same treatment effect.

Formally, our goal is to estimate conditional average treatment effects (CATEs) defined as:

$$\tau(x) = E(Y(1) - Y(0)|X_i = x)$$

Where $Y(W)$ denotes the potential outcome for a test point x (i.e., a single observation) and $W \in \{0,1\}$ is a binary treatment indicator which equals one (zero) for observations in the intervention (control) group. $Y(1)$ is the potential outcome that observation x would have had when treated. $Y(0)$ is the potential outcome that the same observation x would have had when not receiving the treatment (i.e., the counterfactual).

Given n training samples of the form (X_i, W_i, Y_i), where X_i are covariates/features, Y_i is the outcome and W_i is the binary treatment indicator. Single causal trees are grown based on the "honesty" principle. That is, the training data is split into two parts: one half is used to split the data, and another is used to estimate $\hat{\tau}(x)$. Growing a single causal tree involves the following steps:

1. Draw a random subsample of size s from $\{1, \dots, n\}$ without replacement and divide it into two disjoint sets of size $I = (s/2)$ and $J = (s/2)$. I is used to estimate $\tau(x)$ and J to split the data.

2. Split the data using the J sample. Splits are chosen to maximize the variance of $\hat{\tau}(x)$ for $i \in J$, i.e., the goal is to maximize heterogeneity in treatment effects between leaves/subgroups (see Athey & Imbens (2016) for details).
3. Estimate $\hat{\tau}(x)$ within each terminal leaf using only the data from the I sample, while ensuring that the terminal leaves contain a pre-specified minimum number of k observations of treated and control units. Estimation of $\hat{\tau}(x)$ for observation i in leaf L is based on the mean difference in outcomes between treated and control:

$$\hat{\tau}(x) = \frac{1}{\sum_{\{i|W_i = 1,\ X_i \in L\}} i} \sum_{\{i|W_i = 1,\ X_i \in L\}} Y_i - \frac{1}{\sum_{\{i|W_i = 0,\ X_i \in L\}} i} \sum_{\{i|W_i = 0,\ X_i \in L\}} Y_i$$

Conceptually, a causal forest is simply an ensemble of B trees grown by the above-described procedure, each of which produces an estimate $\hat{\tau_b}(x)$ based on observations $i \in I$. The causal forest then aggregates the estimates by averaging them.

The main reason to grow a forest instead of a single tree is to avoid overfitting and safeguard inference. It can be shown that the causal forest yields valid asymptotic confidence intervals for the true underlying treatment effect.

We grow forests based on 8,000 trees and use half the data to grow each single tree. We impose a minimum subgroup/leaf size of 20 observations, and the standard errors are constructed based on the bootstrap of little bags.[17] The set of splitting variables contains a binary indicator for physician gender, dispensing status, language region dummies and average patient age. In the vitamin D testing (generic prescribing) arm, we additionally use the pre-intervention laboratory cost index (drug cost index) and the number of vitamin D tests in 100 patients (generic rate).

REFERENCES

1. OECD. Tackling Wasteful Spending on Health. https://doi.org/10.1787/9789264266414-en
2. Grimshaw J, Eccles M, Tetroe J. Implementing clinical guidelines: current evidence and future implications. *J Contin Educ Health Prof*. 2004;24 Suppl 1(S1):S31-7. doi:10.1002/chp.1340240506
3. Slade SC, Kent P, Patel S, Bucknall T, Buchbinder R. Barriers to Primary Care Clinician Adherence to Clinical Guidelines for the Management of Low Back Pain: A Systematic Review and Metasynthesis of Qualitative Studies. *Clin J Pain*. 2016;32(9):800-816. doi:10.1097/AJP.0000000000000324
4. Janett-Pellegri C, Wildisen L, Feller M, et al. Prevalence and factors associated with chronic use of levothyroxine: A cohort study. *PLoS One*. 2021;16(12):e0261160. doi:10.1371/journal.pone.0261160
5. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet*. 2000;355(9209):1064-1069. doi:10.1016/S0140-6736(00)02039-0
6. Essig S, Merlo C, Reich O, Trottmann M. Potentially inappropriate testing for vitamin D deficiency: a cross-sectional study in Switzerland. *BMC Health Serv Res*. 2020;20(1):1097. doi:10.1186/s12913-020-05956-2
7. Petrilli CM, Henderson J, Keedy JM, et al. Reducing Unnecessary Vitamin D Screening in an Academic Health System: What Works and When. *The American Journal of Medicine*. 2018;131(12):1444-1448. doi:10.1016/j.amjmed.2018.06.025
8. OECD. Health at a Glance 2019. https://doi.org/10.1787/4dd50c09-en
9. Patel MS, Day S, Small DS, et al. Using default options within the electronic health record to increase the prescribing of generic-equivalent medications: a quasi-experimental study. *Annals of Internal Medicine*. 2014;161(10 Suppl):S44-52. doi:10.7326/M13-3001

10. Hemkens LG, Saccilotto R, Reyes SL, et al. Personalized prescription feedback to reduce antibiotic overuse in primary care: rationale and design of a nationwide pragmatic randomized trial. *BMC Infect Dis*. 2016;16:421. doi:10.1186/s12879-016-1739-0
11. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*. 1979;6(2):65-70. http://www.jstor.org/stable/4615733
12. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data, Second Edition*. MIT Press; 2010.
13. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci U S A*. 2016;113(27):7353-7360. doi:10.1073/pnas.1510489113
14. Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*. 2018;113(523):1228-1242. doi:10.1080/01621459.2017.1319839
15. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Statist*. 2019;47(2). doi:10.1214/18-AOS1709
16. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324
17. Sexton J, Laake P. Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis*. 2009;53(3):801-811. doi:10.1016/j.csda.2008.08.007