Doctoral Programme in Clinical Research
Faculty of Medicine
University of Helsinki
Dissertationes Universitatis Helsingiensis
288/2025

De-implementation of low-value care

Aleksi Raudasoja

DOCTORAL DISSERTATION

To be presented for public discussion, with the permission of the Faculty of Medicine at the University of Helsinki, in the Haartman Institute (Hall 1), University of Helsinki, Haartmaninkatu 3, on August 29, 2025, at 1:00 PM.

Helsinki 2025

Supervised by

Professor Kari Tikkinen, MD, PhD
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Raija Sipilä, MD, PhD
Finnish Medical Society Duodecim
Helsinki, Finland

Docent Jorma Komulainen, MD, PhD
Finnish Medical Society Duodecim
Helsinki, Finland

Reviewed by

Assistant Professor Liina-Kaisa Tynkkynen, PhD
Finnish Institute for Health and Welfare
Helsinki, Finland

Associate Professor Sarah Alderson, MD, PhD
University of Leeds
Leeds, United Kingdom

Opponent

Professor Victor Montori, MD, PhD
Mayo Clinic
Rochester, Minnesota, USA

Publisher: Helsingin yliopisto
Series: Dissertationes Universitatis Helsingiensis 288/2025

ISBN 978-952-84-1175-8 (print) ISSN 2954-2898 (print)
ISBN 978-952-84-1174-1 (online) ISSN 2954-2952 (online)
PunaMusta, Joensuu 2025

Contents

Abstract

As healthcare spending continues to rise, particularly in high-income countries, the concern over low-value care is growing. Low-value care refers to healthcare practices that do not provide meaningful benefits to patients. It causes direct harm to patients through adverse effects and costs. Additionally, low-value care presents an opportunity cost, as resources are no longer available for higher-value care and other critical needs, making it also a threat to the sustainability and quality of healthcare. Reliable estimates on the total prevalence of low-value care are lacking, but even 20-30% of care could be of low value to patients. De-implementation strategies aim to solve this problem by decreasing the use of low-value care and, as such, increasing the overall quality of care.

We aimed to evaluate the current state of de-implementation intervention research, the effectiveness of various de-implementation strategies, and the barriers to de-implementation.

First, we conducted a systematic scoping review of de-implementation randomized trials to map the current status of the literature and provide guidance for improving trial methodology. Our results, which included 227 de-implementation trials, indicated that the applicability of trial results is hindered by poor generalizability caused by context-specific and complex interventions, as well as a high risk of bias. To improve the quality of future research, we recommend simplifying intervention designs, increasing the number of clusters in cluster trials, and better controlling potential confounders.

Second, guided by the scoping review, we conducted a systematic review and meta-analysis of de-implementation trials in primary care. Including 140 randomized trials in the analyses, we found moderate certainty evidence that provider education combined with audit and feedback decreases the use of low-value care (odds ratio 0.73 [95% confidence interval 0.63 to 0.84], -23% [-13% to -32%] relative decrease). Intervention combinations that included patient education resulted in potentially larger relative reductions of about 30-34% in low-value care; however, the evidence was of low certainty. Single-strategy interventions such as provider education, audit and feedback, and patient education resulted in only slight reductions in the use of low-value care, with the evidence quality rated as very low to low. Thus, multi-strategy interventions may be needed to have a meaningful impact.

Third, we conducted a survey among primary care physicians in six high-income countries to evaluate their attitudes and obstacles to de-implementation. A total of 1,731 physicians responded (response rate 10.2%). Over 80% of physicians considered overtreatment and overdiagnosis as problems in healthcare, yet only about 50% felt the same about their own practice. The respondents perceived patient-related factors, such as patient requests and the expectation that something will be done, as important barriers to de-implementation, with over 80% rating them as at least moderately important. Similarly, physicians perceived lack of time and fear of medical errors as major barriers, with over 80% rating these as at least moderately important. Since multiple barriers are important to general practitioners and vary depending on the local context (e.g., countries), understanding these barriers locally is crucial when planning de-implementation strategies.

Overall, healthcare administrators could prioritize de-implementation strategies combining audit and feedback with provider education, which have the strongest evidence base, and should consider whether a specific intervention is applicable in their setting. Furthermore, both low-certainty evidence from the systematic review and the survey results emphasize the importance of including the patient perspective in de-implementation strategies. This might involve improving shared decision-making tools and educating patients about the harms of low-value care and the natural progression of disease. Future research should focus on providing replicable and applicable evidence for these interventions.

Tiivistelmä

Vähähyötyinen hoito on kasvava ongelma etenkin länsimaisessa terveydenhuollossa. Sillä tarkoitetaan hoitokäytäntöjä jotka eivät tuo potilaalle merkittävää terveyshyötyä. Potilaille aiheutetun terveyshaitan ja kustannusten lisäksi se on uhka terveydenhuollon kestävyydelle. Käytöllä on rajallisten resurssien vuoksi myös vaihtoehtoiskustannus, joka tarkoittaa että suuremmassa hoidon tarpeessa oleva saattaa jäädä hoitamatta. De-implementaatiolla viitataan toimiin vähähyötyisen hoidon vähentämiseksi.

Selvitimme väitöskirjan ensimmäisessä osassa de-implementaatiotutkimuksessa käytettyjä menetelmiä ja suunnittelimme systemaattiseen katsauksen ja meta-analyysin toteutusta. Löysimme 227 satunnaistettua de-implementaatiotutkimusta. Tutkimusten sovellettavuus oli heikkoa ja näin ollen interventioiden vaikutukset käytäntöön tuotuna epävarmoja, johtuen kontekstisidonnaisista ja monimutkaisista interventioista, sekä kohonneesta harhan riskistä. Tutkimuksen laadun parantamiseksi suosittelimme yksinkertaisempia interventioita, suurempia tutkimuskokoja ja tutkimustulosten yleistettävyyttä heikentävien sekoittavien tekijöiden huomiointia.

Toisessa osassa arvioimme erilaisten de-implementaatiostrategioiden vaikutuksia vähähyötyisen hoidon käyttöön. Kohtalainen tutkimusnäyttö osoitti että lääkäreiden koulutus yhdistettynä auditointiin ja palautteenantoon vähentää vähähyötyisen hoidon käyttöä noin 23 prosenttia (95% luottamusväli -32% to -13%). Interventiot jotka sisälsivät potilaiden koulutusta sekä lisäksi jonkin muun strategian johtivat 30-34% vähenemiseen vähähyötyisen hoidon käytössä, mutta tutkimusnäyttö oli heikompaa ja täten vaikutuksen suuruuteen tulee suhtautua suuremmalla varauksella. Yhden strategian interventioilla oli vähäiset vaikutukset vähähyötyisen hoidon käyttöön ja näytön laatu heikkoa. Siten useampaa strategiaa yhdistävät interventiot saattavat olla usein tarpeen merkittävän vaikutuksen saavuttamiseksi.

Viimeisessä osatyössä arvioimme kyselytutkimuksella perusterveydenhuollon lääkäreiden asenteita ja mielipiteitä de-implementaatioon liittyen. Toteutimme kyselyn kuudessa korkean tulotason maassa. Kyselyyn vastasi 1731 lääkäriä (vastausprosentti 10.2%). Yli 80 % lääkäreistä arvioivat ylidiagnostiikan ja ylihoidon merkittäviksi ongelmiksi oman maansa terveydenhuollossa, mutta vain 50 % omassa työskentelyssä. Yli 80 % lääkäreistä koki ajanpuutteen, työmäärän, virheiden pelon ja potilaaseen liittyvät tekijät vähintää kohtalaisen merkittäväksi esteeksi vähähyötyisestä

hoidosta luopumiselle. Suuri esteiden määrä ja niiden paikallinen vaihtelu alleviivaa paikallisen esteiden selvittämisen tärkeyttä vähähyötyisen hoidon karsimista suunniteltaessa.

Resurssit tulisi terveydenhuollossa kohdentaa ensisijaisesti interventioihin joista on vahvin tutkimusnäyttö. De-implementaation osalta se kohdistuu interventioihin jotka yhdistävät lääkärin koulutuksen ja palautteen annon. Luotettavien vaikutusten saavuttamiseksi tulisi ensisijaisesti käyttää interventioita jotka sopivat omaan kontekstiin ja ovat mahdollisia toteuttaa samaan tapaan kuin niitä on tutkittu. Sekä heikko tutkimusnäyttö systemaattisesta katsauksesta, että kyselytutkimus viittaavat siihen että, potilaat tulisi huomioida osana de-implementaation suunnittelua. Tämä tarkoittaa esimerkiksi työkaluja jaettuun päätöksentekoon, sekä potilaan informointia vähähyötyisen hoidon haitoista ja sairauksien luonnollisesta kulusta.

List of original publications

1. Raudasoja AJ, Falkenbach P, Vernooij RWM, Mustonen JMJ, Agarwal A, Aoki Y, Blanker MH, Cartwright R, Garcia-Perdomo HA, Kilpeläinen TP, Lainiala O, Lamberg T, Nevalainen OPO, Raittio E, Richard PO, Violette PD, Komulainen J, Sipilä R, Tikkinen KAO. Randomized controlled trials in de-implementation research: a systematic scoping review. Implement Sci. 2022 Oct 1;17(1):65. doi: 10.1186/s13012-022-01238-z. PMID: 36183140; PMCID: PMC9526943.
2. Raudasoja A, Tikkinen KAO, Bellini B, Ben-Sheleg E, Ellen ME, Francesconi P, Hussien M, Kaji Y, Karlafti E, Koizumi S, Ouahrani E, Paier-Abuzahra M, Savopoulos C, Spary-Kainz U, Komulainen J, Sipilä R. Perspectives on low-value care and barriers to de-implementation among primary care physicians: a multinational survey. BMC Prim Care. 2024 May 9;25(1):159. doi: 10.1186/s12875-024-02382-9. PMID: 38724909; PMCID: PMC11084097.
3. Raudasoja A, Parpia S, Mustonen JMJ, Vernooij RWM, Falkenbach P, Aoki Y, Barchuk A, Blanker MH, Cartwright R, Crowder K, Garcia-Perdomo HA, Gutschon R, Halme ALE, Kilpeläinen TP, Kuitunen I, Lamberg T, Lang E, Matos J, Nevalainen OPO, Nordlund NK, Pourjamal N, Raittio E, Richard PO, Violette PD, Komulainen JT, Sipilä R, Tikkinen KA. Effectiveness of different de-implementation strategies in primary care: a systematic review and metaanalysis. BMJ Med 2025 (tentatively accepted, in minor revision))

Abbreviations

CI	Confidence interval
DL	DerSimonian and Laird approach
EPOC	Effective Practice and Organization of Care
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
ICC	Intra cluster correlation coefficient
MID	Minimally important difference
OR	Odds ratio
RCT	Randomized controlled trial

1 Introduction

As healthcare spending increases and the world population ages, reducing the inefficient use of healthcare resources is essential. Low-value care has been defined as: 1) care that has no benefit or minimal benefit to the patient, 2) harms outweigh the benefits, 3) is not cost-effective, 4) is less cost-effective than alternatives, or 5) unwanted care (1). Low-value care harms both patients and society. First, this care often leads to harmful side effects and can be costly to the patient. Second, it creates an opportunity cost, as patients in greater need of treatment may be deprived of care; and third, it wastes limited financial and environmental resources.

Low-value care exists as practices are often adopted without a high-quality evidence base. Moreover, adapting to new evidence is typically a slow and challenging process. Perhaps even more so when the goal is to reduce something that was previously considered to provide high value for patients. Several psychological factors might explain this phenomenon, including fear of malpractice, patient pressures, and uncertainty in care decisions (2,3). Previous research has emphasized a narrow set of barriers at a time. Consequently, it remains unclear which determinants are most critical to the de-implementation and use of low-value care.

De-implementation strategies aim to reduce the use of low-value care in healthcare. The reduction in low-value care use may occur in four ways: 1) abandoning, 2) reducing, 3) replacing, and 4) restricting care (4). Previous research suggests that interventions employing several strategies simultaneously have a greater impact (5-9). However, robust estimates of the effectiveness of different types of de-implementation interventions and assessments of evidence certainty are lacking.

This study aimed to assess the primary care physicians' attitudes toward low-value care and the barriers to de-implementation. Additionally, we sought to evaluate the effectiveness of various de-implementation interventions and their evidence quality.

2 Literature review

2.1 Low-value care

Unnecessary care, overtreatment, and overdiagnosis are closely related to low-value care. Overdiagnosis refers to diagnosing medical conditions that would never harm the patient or expanding diagnostic criteria to include ordinary life experiences (10). It may be caused by the overdetection or overdefinition of disease. Overtreatment, or medical overuse, means using medical practices that have no benefits for patients when considering the benefits, harms, and patient preferences (11).

The European Commission report (2025) defines low-value care as follows: “From a health system perspective, low-value care encompasses overuse, misuse, and underuse of healthcare services (for example, prevention, diagnostics, treatment, medication). Overuse and/or misuse comprise the delivery of harmful, ineffective, inappropriate, or not cost-effective healthcare services. Underuse refers to healthcare services not provided or used despite being necessary. Low-value care can lead to negative consequences for patients, their caregivers, the healthcare workforce, the health system as a whole and the wider environment.” (12) While the new definition is more comprehensive, it also includes the underuse of medical practices for which implementation, not de-implementation, is the answer. In this thesis, we align with the definition outlined in the first paragraph of the introduction, which does not include undertreatment.

In US healthcare, estimates for total healthcare spending waste range from 20% to 27% (13, 14). Estimates that include a few specific low-value practices already account for 3-8% of US healthcare spending, suggesting substantial waste from low-value services alone in the US healthcare system (13-15). In a large US survey, physicians estimated that 21% of medical care is unnecessary (3).

Estimates of total low-value care prevalence are lacking from other countries. However, several examples of prevalent specific low-value practices exist within European healthcare systems. A German study estimated that about 31% of patients with dementia receive low-value treatments (16). Low-value treatment was associated with a lower quality of life and increased hospitalizations. In Finland, up to 57% of the elderly received potentially inappropriate medication over the past year (17). In the

Netherlands, 52% of general practitioners ordered lumbar spine x-rays in the past year, and about 8% ordered at least 10 lumbar x-rays (18). Since most low-value practices are not measurable (19), the current total prevalence of low-value care may be underestimated.

Furthermore, register data may contain missing information and inaccurate diagnostic codes, which further decreases the reliability of estimates. Individual patient charts may provide more accurate estimates of the prevalence of low-value care. For example, a US study evaluated the prevalence based on patient chart reviews and found that over 50% of the antibiotic prescriptions were unnecessary, a substantially larger number than estimates based on register data (20).

Low-value care usually refers to medical practices proven ineffective or harmful to patients. It overlooks two potentially even larger problems for healthcare. First, most medical care is based on low-quality or no reliable evidence (21, 22). While these practices may not necessarily fall under the category of low-value care, a substantial share is later revealed to provide no benefit to patients (23). Second, low-value care can also encompass interventions that offer insufficient value to an individual patient when considering the balance of benefits, harms, and the patient's values and preferences. Misrepresenting potential harms and benefits or overstating the certainty of evidence may lead patients to accept care they might have declined had they understood the trade-offs involved more clearly.

2.2 De-implementation

De-implementation refers to strategies aimed at reducing the use of low-value care. It is part of quality improvement, which refers to systematic approaches to enhancing the quality of healthcare (24). Additionally, de-implementation aligns with the principles of value-based healthcare, which seeks to provide the maximum value given the available resources (25). Limited resources mean that every action incurs an opportunity cost. This implies that when low-value care is provided, it may restrict the use of higher-value care.

Contrary to de-implementation, implementation refers to strategies that increase appropriate, usually guideline-concordant care. Both implementation and de-implementation aim to change clinician or patient behavior and care decisions. However, adding and removing may involve different behavioral processes. A recent systematic analysis found that the strategies used in de-implementation differ from those in implementation (26). Whether these different approaches result in the highest impact remains unclear.

De-implementation is not just about using a specific strategy to change behavior; it also requires making decisions about which practices should be reduced and planning the intervention. To facilitate the development, evaluation, and scaling of de-implementation efforts, “The Choosing Wisely de-implementation framework” outlines five stages for de-implementation projects: 1) identifying potential areas for de-implementation, 2) identifying local priorities, 3) exploring barriers to de-implementation and finding potential interventions, 4) evaluation, and 5) spreading effective de-implementation programs (27).

2.2.1 Identifying and measuring low-value care practices

Identification

The first step is to identify which practices are of low value. Primarily, the information should come from high-quality randomized trials or systematic evidence synthesis of those trials. If a practice provides no benefit to patients or is harmful, it may be judged as low-value care. Similarly, deprescribing trials may offer evidence that a practice is of low value. If patients who stop a particular treatment do as well or better than those using it, the treatment is then considered low value for those patients. Clinical practice guidelines and Choosing Wisely initiatives provide recommendations on practices to

reduce. These recommendations may be helpful in “identifying the potential areas for de-implementation” (27, 28).

With a low-quality evidence base, a large practice variation between healthcare providers may indirectly suggest unnecessary care (29). Practice variation may also stem from patient preferences, differing patient populations, or indicate undertreatment, so it should be interpreted with caution. Nevertheless, limited resources necessitate prioritization, and reducing practices based on low-quality evidence is generally a better choice than restricting care with a high certainty of benefit.

Furthermore, the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) provides guidance recommends considering the patients’ values and preferences (30). If most patients would opt to avoid a practice given its harms and benefits, a recommendation against that practice is appropriate. However, values and preferences are often uncertain and can vary significantly from one patient to another. Therefore, identifying a low-value practice often occurs during patient interactions when the individual's values and preferences can be taken into account.

De-implementation often involves prioritizing which practices should be reduced. Besides the prevalence of low-value practices, the harms and costs associated with the intervention may affect the decision on which practices to de-implement.

Measuring low-value care use

Measuring the use of low-value care is important for determining local priorities and evaluating whether local de-implementation activities have achieved their intended goals. Two factors often limit measurement. First, low-value care is frequently non-measurable, limiting the possibility to accurately compare the prevalence of different low-value practices (19). Healthcare might completely lack register data to help with measurement. In these cases, options are limited to using individual patient records or collecting data separately (e.g., by surveying patients or care providers).

Secondly, some practices may hold significant value for some individuals while being of little value for others. For example, antibiotic treatment might be of low value for a patient with unilateral otitis media and no fever. However, the benefits could outweigh the harm (considering values and preferences) for a patient with otitis media in both

ears and a high fever (31). Understanding baseline risks (such as the risk of cardiovascular events when initiating blood pressure medication) and prognosis (like the natural progression of symptoms in mild-to-moderate depression) could aid in determining whether the treatment is of low value. However, these patient characteristics may not be present in the registry data, restricting the ability to accurately assess low-value care usage.

Whether it is cost-efficient to collect accurate data from individual patient records depends on the availability and precision of the total volume outcome (e.g., antibiotic use on all patients with otitis media) and the baseline prevalence of low-value care. If the majority of the total volume of care is low value, the accuracy may suffice. Furthermore, with a high baseline prevalence and significant resources allocated to specific care, utilizing individual patient records might be worthwhile despite the increased measurement costs.

Measuring unintended consequences

De-implementation may have unintended consequences. In a US trial regarding feedback to high quetiapine prescribers, a 2-year follow-up showed that low-value quetiapine prescriptions decreased by 17%. However, the evaluation also revealed that guideline-concordant prescribing (potentially appropriate care) decreased by about 14% (32). Alongside appropriate care, health outcomes (such as hospitalizations or other morbidity outcomes) should typically be considered in the careful evaluation of de-implementation interventions (33).

2.2.2 Drivers for low-value care use, and determinants for de-implementation

Clinical guidelines recommend medical practices based on low-quality evidence

Even most recommendations in clinical practice guidelines are based on low-quality evidence or expert opinion (21, 22). A US study found that about 50% of US and European cardiology guideline recommendations were based only on expert opinion (22). Furthermore, around 40 % of recommendations were supported by either one small RCT or observational evidence. This reliance on low-quality evidence often leads to the adoption of low-value practices, as many of these recommended interventions are later found to provide no benefit to patients. A systematic analysis revealed that out of 3000 randomized controlled trials published in three leading medical journals, 396 trials resulted in medical reversal (23). Medical reversal refers to new high-quality evidence contradicting current clinical practice by showing no benefit compared to no treatment or standard of care. Part of scientific discovery is that new and more effective practices replace some previously effective practices. This is a different phenomenon from medical reversal and does not justify implementing medical practices based on low-quality evidence.

Overestimation of benefits

Physicians often overestimate the benefits of medical practices (34). A contributing factor to this overestimation may be selective reporting and publication bias (35, 36). However, a recent systematic review suggests that publication bias's impact on effect estimates' inflation could be diminishing (37). Clinicians' understanding of treatment effect estimates is also often poor (38), which gives guidelines and other evidence translators a critical role to help understand the treatment effects. The problem is that understandable information on benefits and harms is frequently lacking in clinical guidelines (39, 40).

Overestimation could also be caused by research waste, which probably means even most of the medical literature (41). Research waste includes mistargeted research questions, research designs that do not reliably answer the research question, publication bias, and biased or unusable research reports (41). The problem is not new, and already in 1994, a statistician, Doug Altman, concluded in his editorial piece, “we need less research, better research, and research done for the right reasons” (42).

Research waste is a harmful combination with confirmation bias. Fraudulent studies may be challenging to identify for clinicians (43). As some research areas are filled with fraudulent and heterogeneous results, it is easy to find research findings that confirm prior beliefs (44, 45). Therefore, if your prior beliefs suggest a benefit from a particular treatment, finding evidence to support those beliefs is often easy. Similarly, if a company wants to sell its medical product or if an individual clinician's salary depends on providing a specific treatment, it is easy to find some evidence to back the agenda.

Healthcare incentives

The use of low-value care may also be affected by how healthcare is organized at the system level. Supplier-induced demand occurs when an increase in the supply of healthcare services leads to greater consumption. In low-income countries, the increased consumption could often translate to improved access to appropriate care. However, this phenomenon is also present in high-income countries, suggesting that greater availability of services could lead to increased use, even if there is no actual need for more services. (46) Moreover, free healthcare may increase the health service utilization (47). Therefore, financial incentives to reduce healthcare utilization may be particularly weak in a system that rewards providers for increased utilization, and patients do not have to pay anything for the services they use.

Planning healthcare incentives is often a balancing between underuse and overuse. Paying hospitals for the number of hospitalization days incentivizes extending the length of stay. However, compensating hospitals per case may also fail to promote optimal care if hospitals reduce the length of stay excessively, leading to increased readmissions or higher resource use in other parts of the healthcare system (48).

Barriers and enablers for de-implementation

Understanding the barriers and enablers to de-implementation may help plan effective strategies (27, 28, 49). A scoping review provided six categories for determinants of low-value care use and de-implementation: patient, professional, outer context, inner context, process, and evidence determinants (28). Barriers and enablers to de-implementation could vary between different contexts, highlighting the need to assess the local characteristics. An intervention from another context may not lead to a similar

impact if it addresses a different issue, including different barriers and enablers to de-implementation (28). Table 1 lists potential barriers and enablers identified in the literature.

Table 1. Barriers and enablers to de-implementation

Category	Examples
Provider	**Barriers:** Lack of knowledge (2, 50-53) Fear of malpractice (3, 53, 54) Clinical uncertainty (51-53) Desire to meet patient expectations (2, 50, 51) Routines and habits (50) Lack of communication skills (50) Lack of trust to the guidelines (2) Perceived pressure from patients or other providers (2, 3) **Enablers:** Shared decision making (53)
Patient	**Barriers:** Lack of knowledge (50, 52) Patient's expectations (50, 52, 54) Requests for tests/treatments (3, 51, 54) **Enablers:** Patient awareness of low-value care (54)
Organization	**Barriers:** Lack of useful resources (50, 51, 54) Financial incentives (52) Protocols or norms (50, 53) Workload and lack of time (50-54) Unavailability of medical records (3) **Enablers:** Availability of patient education materials (53) Visible leadership commitment (53)
Other	**Barriers:** Lack of Credible evidence (54) Media information (50)

In a large US survey, the most cited reasons for overtreatment were fear of malpractice, patient pressure, and difficulties in accessing medical records (3).

Patient-related determinants

Several studies and a scoping review have highlighted patient-related factors as the most important or at least the most recognized determinants for de-implementation (51, 55, 56). A scoping review of de-implementation determinants found that patient expectations are the most often reported specific barrier to de-implementation (55). In a Dutch survey of primary care physicians, 76% of the respondents cited "maintaining a good relationship with the patient" as a driver for low-value care use (51). In German primary care, physicians rated patient expectations as the most important driver for overuse (56). Swedish researchers identified three main reasons for low-value care use in a focus group study: "uncertainty and disagreement about what not to do", "perceived pressure from others", and "a desire to do something for the patient" (2).

Physician-related determinants

Several studies have identified a lack of knowledge as an important physician-related barrier to de-implementation (2, 50-53). Many physician-related factors are related to physicians' attitudes, such as fear of malpractice, clinical uncertainty, or a defensive attitude (Table 1). Even when aware of low-value care practices, poor communication skills may hinder clinicians from decreasing the use of low-value care (50).

Organizational determinants

Creating a supportive organizational environment can facilitate the de-implementation process. For example, providing patient education materials or demonstrating visible leadership commitment could support clinicians in their efforts to reduce the use of low-value care (53). The organization may also act as a barrier to de-implementation. Clinicians may face challenges in opposing the organization's protocols, such as requiring pre-operative testing from low-risk patients. A lack of valuable resources or insufficient time to discuss with the patient could also hinder efforts to reduce low-value care usage (50, 51). However, a systematic review on changing the length of primary care appointments found only very uncertain evidence regarding practice change (57). In a more recent, large US cohort study, shorter primary care visits were associated

with increased inappropriate care, although the changes were small (58). Inappropriate antibiotic prescriptions declined by 0.11 percentage points for every 1-minute increase in appointment time.

An interview study identified three factors contributing to successful organizational change in healthcare: providing opportunities for influencing the change, giving preliminary notice to allow employees time to prepare, and offering a clear rationale for the change. Study participants perceived patient benefit as the most important reason for change (59).

Differences compared to implementation

Healthcare leaders in de-implementation may differ, at least in part, from those who drive implementation. De-implementation might be easier for those who questioned the practice from the beginning and never fully adopted it. However, some clinicians who are early adopters of new practices may also be the first to de-implement, especially if they are more open to new evidence, such as evidence suggesting the abandonment of current practices. Similar to implementation, the expectation of treatment effect drives the willingness to de-implement, and clinicians who expect no impact or a lower effect are more likely to reduce the use of a low-value practice. Nevertheless, motivational and financial factors may be more important in de-implementation than in implementation. (26, 60)

2.2.3 De-implementation strategies

There are several theoretical frameworks to categorize (de-)implementation interventions (61-63). The Effective Practice and Organization of Care taxonomy is probably the most widely used (64). The refined version divides the strategies into four main categories: professional, financial, organizational, and regulatory/system (64). Many of these strategies have been extensively studied for implementing new practices (26, 61). However, it remains unclear how well these strategies translate to effective de-implementation (26).

De-implementation research has primarily focused on strategies aimed at providers, patients, and, in some instances, health care units. A scoping review identified educational materials, educational meetings, audit and feedback, and reminders as the most common strategies in de-implementation research (65).

At a system level, policymakers could decide which practices receive governmental funding. Restricting funding and availability of care may be the most effective way of de-implementation, although often not possible strategy to use (66). A Canadian study evaluated the effectiveness of eliminating reimbursement for vitamin D testing and found about 90% decrease in testing volume (66).

Previous research suggests that de-implementation interventions should be designed to address the local context and barriers to change (27, 28, 49). For example, a routine behavior of offering a laboratory test along with a package of other tests, which is primarily driven by automatic cognitive processes and related to organizational protocols, may require different de-implementation strategies than cognitively slower, and more autonomous decisions like the use of antibiotics for upper respiratory tract infections. Furthermore, an intervention that includes context-specific methods and has been tested by others may require tailoring if conducted in a different setting or context.

There are several systematic reviews and meta-analyses on the effectiveness of implementation interventions; however, only a few exist on de-implementation. Although it is uncertain how well the evidence from implementation research translates to de-implementation, it may be used as indirect evidence for the effectiveness of de-implementation. Table 2 describes the different (de-)implementation strategies and current assessments of their effectiveness.

Table 2. Intervention categories and evidence on their effectiveness in implementation and de-implementation

Category	Effectiveness
Educational materials	**Implementation:** A systematic review of implementation trials measuring the effectiveness of printed educational materials found a median improvement of 4 percentage points in clinical practice (67).
Educational meetings	**Implementation:** A systematic review including 215 implementation trials on continuing educational meetings and workshops found 7 percentage points [95% CI 6.62 % to 6.97%] increase in compliance with desired practice (from 50% to 57% adherence) (68). The pooled estimates also included multi-component interventions (e.g., education combined with audit and feedback), which warrants caution in the interpretation.
Decision supports/aids	**Implementation:** A systematic review of implementation trials measuring the effectiveness of computerized decision support systems found a 5.8 percentage point [95% CI 4.0% to 7.6%] increase in compliance with desired care (69). Most of the trials included co-interventions.
Patient education/engagement	**De-implementation:** A systematic review of de-implementation interventions aimed at engaging patients in decision-making found approximately a 26% relative decrease in low-value care use. Most of the trials in this review evaluated some form of education on the benefits and harms for patients and/or shared decision making (70).
Audit and feedback	**Implementation:** A Cochrane review of 140 implementation trials found a median absolute increase of 4.3 percentage points (IQR 0.5% to 16%) in desired practice use (71). **De-implementation:** A systematic review of de-implementation trials on social norm feedback found an absolute 4% decrease in antibiotic use (72).
Local opinion leaders	**Implementation:** A Cochrane review of 18 implementation trials found a median 10.8% absolute improvement in compliance with desired practice (73).

Several systematic reviews have evaluated de-implementation interventions on narrow topics (5, 6, 74-79), and a few have aimed to include all de-implementation trials (7-9). A recent systematic review of 121 de-implementation trials found a median reduction of 15% in low-value care use after a single strategy intervention and a median reduction of 20% after interventions that included multiple strategies (9). Nevertheless, none of the previous systematic reviews was able to determine which strategies are most effective in reducing the use of low-value care.

Educational meetings and materials

Several de-implementation trials have measured the impact of educational materials on low-value care use. For example, a factorial trial (i.e. trial studying multiple interventions at the same time) from the UK tested feedback and educational messages in laboratory test reports (80). The educational messages outlined when the target laboratory tests were inappropriate and resulted in about a 10% relative reduction in inappropriate testing.

Other trials showed conflicting results (81, 82). A cluster trial conducted in four South American countries assessed the effectiveness of sending email links to online educational materials, including patient cases and feedback after completion (81). The trial found no impact on inappropriate antibiotic prescribing. An Australian factorial trial evaluated the effect of sending educational letters to primary care physicians and found approximately a 3% reduction in antibiotic prescribing (82). Other intervention arms that included feedback on individual prescribing rates led to a larger, approximately 10% reduction in antibiotic prescribing.

Educational meetings do not always lead to better results in low-value care use. An Indian trial measured the effectiveness of a 150-hour training program to improve the quality of curative care (83). They found no impact on the use of unnecessary medicines and antibiotics. However, they found a 14% relative increase in appropriate case management (appropriate care).

Considering the heterogeneity in effects, planning educational interventions should be careful and include an assessment of the barriers and facilitators to reducing low-value care. Aiming the intervention at the needs of the audience may be even more important than other intervention categories (e.g., audit and feedback).

Audit and feedback

Audit and feedback interventions involve measuring clinicians' practice use and providing feedback based on the results. Typically, the practice is compared to a social norm (e.g., comparison to the mean prescribing among colleagues working in similar settings). Measuring practice use usually requires registry data, although the audit and subsequent feedback can also be conducted by reviewing individual patient records and offering feedback based on that (84).

A systematic review evaluated the effectiveness of providing social norm feedback to clinicians regarding antibiotic prescribing and found an absolute decrease of 4% in antibiotic prescriptions (72). Their subgroup analysis suggested the largest effects from feedback that included a description of the health consequences of (low-value) antibiotic use.

Decision supports

Decision supports aim to change practice by providing clear pathways for patient care (clinical algorithms) and describing the benefits and harms of different treatment options. They are closely related to decision aids, which are more focused on supporting decision-making by enhancing understanding of the decisions being made. A decision aid may be intended for clinician use only, or it can be used in collaboration with a patient to promote shared decision-making (85). Implementation trials suggest that decision supports enhance clinical practice (69). However, there is less evidence on the effectiveness of decision supports in reducing low-value care use.

Patient education

Patient education can be provided face-to-face (86) or through the use of educational materials (87). It may, for instance, include information on the harms of low-value care, the natural progression of symptoms, and guidance on when to reconsult. A systematic review of engaging patients in decision-making (mostly involving trials on some form of patient education and shared decision-making) found an approximate 26% reduction in the use of low-value care (70).

Nudge strategies

Nudge theory was originally developed in behavioral economics to describe how human behavior can be modified through changes in the environment in which

decisions are made (i.e., changes in the “choice architecture”) (88). Similarly, clinicians have their own “choice architecture” for making clinical decisions. The building blocks include, for example, physicians’ knowledge and skills, financial incentives, organizational culture (e.g., senior support), and working environment (e.g., appointment length, available diagnostic tests).

A systematic review including 55 trials on nudge strategies to implement clinical guidelines found six intervention types (89):

1. Priming – subconscious cues, such as reminders or prompts in electronic health records, aiming to cause or help towards a certain choice
2. Salience/affect nudge – a prompt aiming to cause an emotional urge towards a specific choice (for example, with patient cases or personal consequences signs)
3. Norms and messenger nudge – prompting certain action due to guidance from respectful authorities or comparing performance to colleagues (i.e. social norm)
4. Default nudge – giving a default option for an action or restricting access to inappropriate choices
5. Commitment/ego nudge – for example, a written commitment from clinicians to use a specific practice
6. Incentives nudge – incentivizing a certain practice with financial or professional reward

The systematic review found a median effect for continuous outcomes of a standardized mean difference of 0.39 (IQR 0.22 – 0.45) and for binary variables of an odds ratio of 1.62 (IQR 1.13 - 2.76). Another systematic review of nudge interventions aimed at reducing unnecessary antibiotic prescribing in primary care suggested a likely smaller impact; however, the comparison is unreliable due to differences in reporting (90).

2.3 Trial methodology and evidence synthesis methods

Randomized trials

One of the major concepts in medicine over the last hundred years has been randomized trials. By randomizing patients, scientists have been able to differentiate the effect of medical interventions from confounding factors. In observational studies, adjusting for potential confounders, such as patient age and gender, helps provide more reliable estimates of an intervention's effect. Although the adjustments may lead closer to the true effect size for the intervention, the risk of false positives (and false negatives in some cases) is large. On average, observational studies, compared to randomized trials, might result in only minor increases in effect sizes (91). Nevertheless, likely mainly due to imbalances in prognostic factors, observational studies have a low probability of matching the effect sizes of randomized controlled trials (92), rendering them unreliable in estimating intervention effectiveness. However, observational studies play an important role in addressing research questions such as the rare harms of practices and care, prognosis, risk factors, or the prevalence of disease.

Cluster randomized trials

De-implementation trials often randomize at the cluster level. Cluster randomized trials are conducted to help avoid possible contamination. Contamination means that the study participants in different study arms are exposed to the intended intervention, and at least to some extent to the comparison. For example, if a sample of health personnel in a hospital unit receives an educational intervention, they may assist or educate the control group participants working in the same unit. In the worst-case scenario, by narrowing the difference between study groups, contamination could result in false negative results (93). On the other hand, cluster randomization usually results in smaller statistical power. Therefore, researchers should consider the advantages of adjusting for contamination, which may yield more accurate effect estimates, versus the greater statistical power that comes with individually randomized trials (94). An additional common problem in cluster trials is recruiting participants after randomization. This can undermine allocation concealment if recruiters know which study arms participants will enter. However, in (de-)implementation trials, participants are usually recruited before randomization, minimizing the risk of bias due to poor allocation concealment.

In healthcare quality improvement trials, such as de-implementation trials, the cluster randomization unit may be either a practitioner or a healthcare unit. In the first case, the cluster consists of the patients of one practitioner, while in the latter case, it consists of the patients in one unit. However, if the outcome is measured at the practitioner level, the practitioners working at the unit form the cluster. If units are randomized and outcomes are measured at the patient level, the trial involves two levels of clustering: firstly at the practitioner level and secondly at the patient level. Clustering should always be considered in the analysis and sample size calculations, as participants within the same cluster correlate (95).

The intra-cluster correlation coefficient (ICC) estimates between-cluster variance relative to within-cluster variance (i.e., between participants variance). When the between-cluster variance increases compared to intra-cluster variance, the ICC rises, causing the confidence intervals for the effect estimates to widen. A large correlation among individuals within a cluster, along with a correspondingly large ICC, can be especially problematic in small cluster randomized trials, as both a small number of clusters and a large ICC increase the likelihood of false positive results (96). Currently, very few trials utilize statistical methods that appropriately account for the small number of clusters in both sample size calculations and analyses of study results.

2.3.1 Risk of Bias

Assessing the risk of bias in individual studies is a central part of conducting systematic reviews and evaluating evidence certainty. Risk of bias refers to methodological characteristics that may influence the study results: “the risk that they will over-estimate or under-estimate the true intervention effect” (97). The Cochrane Risk of Bias tool (current version ROB2) is the most widely used instrument for assessing risk of bias. It includes five domains of bias: bias in the randomization process, bias due to deviation from intended interventions, missing data, bias in outcome measurement, and selection bias (98).

Randomization process

Bias in the randomization process may arise from randomization sequence generation and allocation concealment. The randomization sequence should be generated using a procedure that ensures the participants’ treatment arm is determined by chance (e.g., random number generator). Allocation concealment refers to the methods used to ensure that no one can influence which arm the participant is assigned (99) (e.g., on-site computer randomization).

Deviation from the intended interventions

Deviation from the intended interventions means whether study participants received the interventions specified in the protocol. A primary method to prevent this bias is blinding both study participants and intervention deliverers. In cluster-randomized trials, blinding study participants is uncommon. The Cochrane guidance for cluster randomized trials rates the risk of bias as high only if there are contextual deviations from the intended interventions (100). Nonetheless, knowledge of the intervention presents a potential source of performance bias, as participants in different arms may act differently just because they know their treatment arm, rather than due to the intervention itself (101).

Cluster randomized trials have three options for allocation concealment (95). First, they may recruit cluster participants (e.g., patients in a health center) before the randomization, thereby preventing them from foreknowing the allocation. Second, they may recruit all possible cluster participants. For example, if all patients from a health center are eligible and will be recruited, then the allocation can no longer affect the recruitment. Sometimes, not all possible cluster participants are recruited, and the

study has prespecified eligibility criteria. This may cause identification/recruitment biases, as the allocation could affect the identification of participants. For example, the intervention arms may be more likely to “identify” participants more willing to participate. Third, they may use blind recruitment. If the first two options are not possible, researchers may have recruiters blinded to the study allocation. Having personnel who are otherwise independent of the trial could be the most reliable option.

Missing data

Missing data, whether systematic or unsystematic, occurs when participants drop out of the study and cannot be followed up until its conclusion or when data might be lost in other ways (102). Systematic missing data can especially lead to bias in the results.

Bias in outcome measurement

Bias in outcome measurement may occur when outcome assessors are not blinded. Blinding should be considered concerning the outcome. If the outcome is objective, such as mortality, then blinding of outcome assessors is not necessarily required. (102)

Selection bias

Finally, selection bias may happen if researchers choose favorable outcomes after preliminary analysis, and instead of all measured outcomes, report only outcomes that support the prespecified hypothesis (102). In addition to reporting all data, researchers should follow a prespecified plan to analyze results.

2.3.2 Evidence certainty

Developed by Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group, a group of healthcare methodologists, guideline developers, clinicians and other stakeholders, the GRADE approach aims to assess the evidence certainty, in other words “confidence in effect estimates” (103). It includes an assessment of both the internal and external validity of the study results. This assessment has five domains: risk of bias, indirectness, inconsistency, imprecision, and publication bias.

According to the GRADE guidance, assessing the certainty of evidence for treatment effects from randomized trials begins with high evidence certainty. The assessment involves reviewing all criteria and determining whether any concerns warrant a downgrade in evidence certainty. If multiple criteria raise concerns, the decision must consider the overall context to determine the total downgrading of evidence certainty.

Indirectness

Indirectness refers to evaluating whether the studies address the study question; in other words, the evidence is “direct” in the sense that the included studies “directly compare the interventions which we are interested in” (103, 104). First, we need to consider indirectness in study populations. For example, if the question is whether community masking reduces the rate of respiratory infections, but studies are only conducted in healthcare settings, the assessment of evidence certainty would warrant a downgrade in evidence certainty. Similarly, interventions may introduce indirectness in study results. It is especially important for a complex intervention, such as de-implementation interventions, to consider whether the intervention will be conducted differently depending on the study context and whether it is even possible to replicate it in a sufficiently similar manner to reasonably expect comparable results (104). If the overall study results come from trials with several contexts and have consistent findings, it may increase the applicability (generalizability). Lastly, assessment should consider indirectness in study outcomes and the comparison group. GRADE approach recommends forming study questions for patient-important outcomes. However, an outcome of less importance to patients can be used to provide indirect evidence on the study question. For example, a change in blood pressure does not represent a patient-important effect, unlike mortality or stroke. Therefore, the GRADE approach would suggest rating the evidence certainty down by one level.

Inconsistency

Inconsistency of study results means that the available studies give variable results on a particular study question (105). When inconsistencies in study results are not explained by a priori hypothesis, authors should consider downgrading for inconsistency. Statistical heterogeneity alone is insufficient to fully evaluate the inconsistency of outcomes. For instance, a few large positive trials may result in inconsistent study results (high I2-values), but if they all suggest a meaningful effect in the same direction, a downgrade may not be necessary.

Imprecision

Imprecision refers to wide confidence intervals leading to difficulties in assessing the magnitude of the effect. GRADE guidance has three ways for rating imprecision. A minimally contextualized approach is usually used in systematic reviews. It recommends assessing confidence in whether the effect is truly present. Review authors should define a minimally important difference, and if confidence intervals cross that threshold, they should consider rating down for imprecision (106). If the point estimate is inside the thresholds for not providing at least a minimally important effect, the evidence certainty should be rated for the intervention having little to no impact. A partly contextualized approach aims to assess confidence in whether the effect falls within a prespecified range. In this case, authors should prespecify thresholds for small, moderate, and large effects and then assess the certainty of whether the true effect lies between the thresholds for one effect size category (107). A fully contextualized approach adds all critical outcomes (both benefits and harms) to deciding on the final evidence certainty.

Publication bias

Publication bias occurs when negative studies are not published while positive studies are. This situation can lead to an overestimation of effect size. The GRADE approach suggests downgrading the certainty of evidence if there is a high risk of publication bias (108).

Criteria to upgrade evidence certainty

After going through the criteria for downgrading, there are three domains to consider regarding whether the evidence certainty should be upgraded (109): :

1. *Large effect* – Authors should consider rating up the evidence certainty if the effect size suggests that confounding is unlikely to explain the correlation. For binary outcomes, a reasonable threshold (suggested by modelling studies) could be over a relative risk of 2.0 (or under 0.5).
2. *Dose-response* – Authors should consider rating up if there is a credible dose-response gradient (110). Five criteria may help in assessing credibility: 1) use of adequate analytic methods, 2) is confounding a likely explanation for the dose response gradient, 3) is there concerns of ecological bias, 4) is the gradient consistent across all studies, and 5) does indirect evidence support the dose response finding.
3. *Residual confounding is expected to reduce the observed effect* – Authors should consider rating up if all unaccounted confounding would likely decrease the effect estimate. The GRADE guidance includes a possible case example: “If, for instance, only sicker patients (in an observational study) receive an experimental intervention or exposure, yet they still fare better (than the control), it is likely that the actual intervention or exposure effect is even larger than the data suggest.”

2.3.3 Considering evidence certainty in complex intervention research

Complex intervention research is more likely to provide low certainty evidence compared to simpler intervention research (111, 112). Complex interventions are often context-dependent, and even the same type of interventions usually contain different factors that lead to inconsistent study results (113). Another common problem in complex intervention research is performance bias (i.e., knowledge of being in the intervention or control group could bias the results). Therefore, it is difficult to estimate how well the results will translate to clinical practice.

Previous research has suggested that GRADE criteria should be modified for complex interventions (113, 114, 115). Movsisyan et al. (2015) suggested that assessing the blinding of study participants should focus on blinding to the study hypothesis instead of blinding to the intervention (114). Montgomery et al. (2019) suggested that lack of blinding could often be "an essential aspect of the intervention", and therefore, review authors should consider whether it could lead to biased study results (115). This is in line with the Cochrane Risk of Bias tool for cluster-randomized trials (100), which suggests that studies should only be rated as high risk of bias if there are contextual deviations from the intended interventions (as described earlier in section 2.2 "*Deviation from the intended interventions")*.

Murad et al. (2017) suggested that inconsistency may not warrant downgrading if contextual factors have been explored via other research (process evaluation) (113). Although the intervention heterogeneity in complex intervention trials likely leads to some heterogeneity in study results, exploring the potential factors for inconsistency (such as intervention intensity) could be enough to assess the evidence certainty.

2.3.4 Summarizing evidence certainty and interpretation

There are four categories for evidence certainty: high, moderate, low, and very low. Evidence certainty describes how confident evaluators are that the true effect of an intervention lies close to the given point estimate. Table 2 provides the definitions for each category (116).

However, GRADE guidance recommends assessing imprecision of estimates in relation to minimally important difference (MID) thresholds (106). Therefore, instead of the accuracy of an estimate, the assessment of imprecision answers to how confident the evaluators are that the true effect lies on one side or another of the MID threshold (i.e., does the intervention provide a clinically meaningful impact).

Table 3. Evidence certainty levels and their definitions (reproduced from Balshem et al 2011 [116])

Evidence certainty	Definition
High	We are very confident that the true effect lies close to that of the estimate of the effect
Moderate	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low	We have very little confidence in the effect

2.3.5 Meta-analysis

Systematic reviews often include a meta-analysis to summarize results. Meta-analytic methods enable the combination of estimates from several trials measuring the same outcome into a single effect estimate. These pooled estimates are usually more precise because of the increased statistical power. Individual studies often have very specific study questions (117), so meta-analysis might allow for answering broader study questions that could be generalized to larger populations (118). However, broad study questions may also have downsides if they include study populations or interventions with substantial differences. In such cases, it becomes difficult to determine to whom and which type of intervention the pooled estimate applies. Meta-analysis may also provide misleading estimates if the underlying evidence is of low quality.

Two meta-analytic models exist: fixed-effects and random-effects models. The fixed-effects model assumes the true effect size is the same across all studies (119). This assumption is never true precisely, as study populations from two studies are always somewhat different. However, the fixed-effects model could be justifiable in rare cases where the study populations, interventions, and comparison groups are very similar. Furthermore, a random-effects model makes the estimates more generalizable to individuals outside the study population, and researchers should avoid using a fixed-effects model if their intention is to generalize the effect estimates (120).

Random-effects meta-analyses mostly use the DerSimonian and Laird approach (DL) to calculate the pooled estimates (121). However, the DL approach has a high probability of false positive results when the number of studies is low or the heterogeneity of treatment effects is high. In these situations, researchers could use the Hartung-Knapp adjustment, which decreases the rate of false positive findings (122).

3 Aims of the study

The overall aim of this thesis is to evaluate how to design effective de-implementation interventions that address key behavioral factors for practice change. The goals for the two systematic reviews (Studies I and II) were to summarize all de-implementation RCTs, assess the effectiveness of various de-implementation strategies in reducing low-value care, identify limitations and knowledge gaps in the literature, and provide guidance on methodological improvements. In the survey study (Study III), we aimed to explore primary care physicians' attitudes and barriers to de-implementation.

The specific aims for each study were:

Study I – a scoping review

1. To produce a broad picture of the current status of de-implementation trials.
2. To identify methodological limitations in de-implementation trials and provide recommendations for improvement.

Study II – a systematic review and meta-analysis

1. To evaluate the effectiveness of different de-implementation strategies in primary care and assess the evidence certainty.

Study III – a survey study

1. To assess the current opinion climate among primary care physicians regarding low-value care.
2. To explore primary care physicians' perceptions of the key barriers to de-implementation and potential differences between high-income countries.

4 Methods

4.1 Review methods (Studies I and II)

We preregistered our studies at Open Science Framework (study I - OSF hk4b2) and Prospero (study II - CRD42023411768) and followed PRISMA guidelines for systematic and scoping reviews (123, 124).

4.1.1 Data sources and searches

First, with an information specialist, we developed a comprehensive search strategy to include all randomized trials on de-implementation. A previous scoping review identified de-implementation-related terms (125), which we used as a basis for our search strategy. Additionally, we identified relevant articles from previous systematic reviews (8, 126) and updated the strategy with new index terms. Second, we searched MEDLINE and Scopus without language and date limitations. Third, we identified systematic reviews from the search results, searched their reference lists for potentially eligible articles, and added them to the selection process. Lastly, we identified protocols and post hoc analyses of de-implementation trials and followed them up to include the main articles in the selection process.

We searched for the scoping review until May 24, 2021, and updated it for the systematic review and meta-analysis on 10 July 2024.

4.1.2 Eligibility criteria

We included studies that compared de-implementation interventions to no intervention/usual care, placebo/sham intervention, or another de-implementation intervention. Furthermore, we included de-implementation trials involving any target group (such as physicians, nurses, or patients) and any medical intervention (including drug treatments, diagnostic tests, and surgery).

The scoping review included all randomized trials on de-implementation interventions, and the systematic review and meta-analysis was limited to primary care. The systematic review included only trials that measured low-value care use, as the scoping review also included trials measuring intentions to reduce the use of low-value care.

Since the mechanism for change might differ from de-implementation—where the goal is to reduce the use of new practices—we excluded deprescribing trials, which focus on discontinuing treatments already used by patients. We also excluded trials that only measure resource use (such as costs or visits) and those that compare one medical practice to another, like using a laboratory test versus not using one.

In the meta-analysis of Study II, we included trials that reported odds ratios, mean differences, arm-specific event rates, or mean values with variance estimates, as these could be pooled in the analysis (127). If trials did not provide enough data for meta-analysis, even after requesting additional data from the authors, we analyzed the trials descriptively.

4.1.2 Outcomes and variables

In the scoping review we collected following variables/outcomes: 1) study country, 2) year of publication, 3) unit of randomization allocation (individual vs. cluster), 4) the number of clusters, 5) was an intra-cluster correlation (ICC) used in sample size calculation, 6) duration of follow-up, 7) setting, 8) medical content area, 9) target group for intervention, 10) the number of study participants, 11) mean age of study participants, 12) the proportion of female participants, 13) intervention categories, 14) rationale for de-implementation, 15) goal of the intervention, 16) outcome categories, 17) reported effectiveness of the intervention, 18) conflicts of interest, 19) funding source, 20) risk of bias, 21) implementation theory used, 22) costs of the de-implementation intervention, 23) effects on total healthcare costs, 24) changes between baseline and after the intervention, and 25) tailoring the de-implementation intervention to study context.

Additionally, in the systematic review and meta-analysis, we collected 1) low-value care use, 2) total volume of care, 3) appropriate care (or guideline-concordant care), and 4) health and health-related outcomes (including mortality, morbidity, quality of life, and readmissions).

4.1.3 Study selection and data collection

For the study selection, data extraction, and assessment of bias risk, we created standardized forms with detailed guidance. Before the study selection and data collection, we piloted the forms to clarify the guidance where needed. Methodologically trained reviewers used the forms to assess the eligibility of study reports and data

extraction independently and in duplicate. In case of disagreement, reviewers discussed to resolve them, and if consensus was not reached, a third party (clinician-methodologist) was involved.

If data on low-value care use and total volume of care outcomes were missing, we contacted the study authors and added the data to the analysis.

4.1.4 Intervention categories and outcome hierarchy

In the scoping review, we refined the Effective Practice and Organization of Care (EPOC) taxonomy of health system interventions for de-implementation interventions in three steps. 1) We used the existing taxonomy to categorize interventions. 2) We identified limitations in the taxonomy and modified it through discussion and consensus-building. 3) We applied the refined taxonomy to categorize the interventions.

We defined five possible categories for de-implementation rationale by using a previous definition of low-value care, “care that is unlikely to benefit the patient given the harms, cost, available alternatives, or preferences of the patient” (1). The categories were: 1) evidence suggests little or no benefit from treatment or diagnostic test, 2) evidence suggests another treatment is more effective or less harmful, 3) evidence suggests more harms than benefits for the patient or community, 4) poor cost effectiveness, and 5) patient(s) do not want the intervention.

In the systematic review, we used a modified TIDieR checklist to extract intervention characteristics (128). Based on these data and the previous intervention taxonomy, we formed five groups for the meta-analysis: provider education, patient education, audit and feedback, decision support, and others. Provider education could include education for physicians, nurses, or other healthcare staff. We searched and assessed study protocols and other published reports to determine accurate categories.

For the effectiveness outcomes, we modified Kirkpatrick’s levels for educational outcomes (129) (Table 4). potentially leading to a decrease in the use of low-value care.

Table 4. Outcome categories for de-implementation effectiveness (reproduced from [Raudasoja et al. Randomized controlled trials in de-implementation research: a systematic scoping review. 2022])

Name	Rationale and definitions	Examples
Health outcomes	De-implementing clinical practice should improve (or at least have no negative effect on) health outcomes. Health outcomes can therefore be considered as measuring the safety of de-implementation.	Mortality, morbidity, quality of life, symptoms
Low-value care use	The primary aim of a de-implementation intervention is to reduce low-value care. Predefined low-value care use should therefore be (one of) the primary outcome(s) of de-implementation effectiveness. Typically, the definition of low-value care is based on diagnoses or clinical criteria that represent low-value care in combination with a specific clinical practice. Data is often gathered from individual patient records or administrative databases. Individual patient records usually contain more specific information on clinical decisions and may therefore yield more accurate information.	Antibiotic use for viral upper respiratory infections. Use of radiological imaging in patients with acute low back pain without “red flag” symptoms.
Appropriate care use	Can be used as an outcome when a medical practice can be either appropriate or inappropriate. For instance, in patients with respiratory infection, the use of antibiotics can be either appropriate or inappropriate. Change in appropriate care use measures unintended consequences of de-implementation, and can therefore be considered as a measure of the safety of de-implementation.	Antibiotic use for confirmed pneumonia. Use of radiological imaging in patients with low back pain and “red flag” symptoms.
Total volume of care	Total volume includes both appropriate and inappropriate care and is an indirect measure of low-value care. It may sometimes be justifiable to use in very large samples if it is impossible to differentiate between appropriate and inappropriate care, and if using individual patient records is not possible. Outcomes that are based on diagnoses often include both appropriate and inappropriate care and should therefore be considered as total volume care, not as low-value care outcomes.	Total use of antibiotics in upper respiratory tract infections. Use of radiological imaging in low-back pain.
Intention to reduce the use of low-value care	Intention is the first step to change, but it does not reliably describe actual change in the use of low-value care. Since intention can be measured earlier than other outcomes, it may sometimes be justified to use it as a preliminary assessment of the effectiveness of a de-implementation intervention. It is often used after educational interventions and when data is collected through surveys.	Intention to reduce the use of inappropriate antibiotics in upper respiratory tract infections. Intention to reduce the use of inappropriate radiological imaging in low-back pain.

4.1.5 Risk of bias and evidence certainty

We modified the Cochrane risk of bias tool for cluster randomized trials to give further guidance on complex intervention designs and to enhance interrater agreement (additional file 1, Study 1). Through discussion and consensus building, and considering previous literature (100, 130, 131), we included six criteria in the assessment: 1) randomization procedure, 2) allocation concealment, 3) blinding of outcome collection, 4) missing outcome data, 5) contamination, and 6) selective reporting. In the scoping review, we also included criteria for baseline imbalances, which were assessed as an issue of indirectness (when rating evidence certainty) rather than a risk of bias concern.

In the systematic review, we rated evidence certainty based on guidance provided by the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach (103). For low-value care use, two methodologically trained reviewers assessed the evidence certainty, considering GRADE guidance on risk of bias, indirectness, imprecision, inconsistency, and publication bias.

4.1.6 Scoping review analysis

In addition to risk of bias, we used seven criteria to assess trial quality: 1) number of clusters, 2) length of follow-up, 3) use of intra-cluster correlation in the sample size calculation, 4) tailoring, 5) theoretical background, 6) level of randomization, 7) reporting before and after the intervention. We compared these quality indicators in trials published on or before 2010 and after 2010 to identify potential improvements or declines in more recent studies. Based on these results, we used discussion and consensus building to create a list of recommendations for future de-implementation trials to enhance trial methodology. Authors provided feedback on these recommendations via email, and final versions were developed through in-person meetings.

4.1.7 Systematic review analysis

To provide one summary estimate, we converted effect estimates of continuous outcomes to odds ratios according to Cochrane guidance (127). If a study provided only arm-specific estimates for continuous outcomes, we first calculated a standardized mean difference (Hedges' g). Finally, we converted it to an odds ratio with corresponding 95% confidence intervals. To accurately adjust for baseline measurements, a trial should also report arm-specific changes with variance estimates, which are rarely reported in these trials. We therefore did not adjust for baseline, if it was not appropriately done by the study authors, but used only arm-specific means at follow-up to calculate the effect estimates. In case of cluster-randomized trials, if only arm-specific estimates were available, we adjusted for clustering by using the reported intra-cluster correlation (ICC). Furthermore, if a trial did not report intra-cluster correlation, we used the median ICC from all trials to calculate the effect size (for binary outcomes, practice-level ICC 0.105 and provider-level ICC 0.22; for continuous outcomes, practice-level ICC 0.305).

We used R statistics with the "meta" package to conduct the meta-analysis, including the creation of forest plots. If at least three studies existed for an intervention category with a comparison to no intervention, we conducted a random-effects meta-analysis using the DerSimonian-Laird approach. As using Hartung-Knapp adjustment decreases the risk of false positive findings (122), we decided to include it in the meta-analysis. We used means and standard deviations to summarize continuous outcomes, and proportions with confidence intervals for binary outcomes.

We estimated absolute effects by using the median control risk of low-value care use in all trials.

4.1.8 Sensitivity and subgroup analysis

We conducted subgroup and sensitivity analyses when at least three trials were available for the subcategory. We conducted a sensitivity analysis on trials with binary and continuous variables, as the ratio of means offers better interpretability compared to the standardized mean difference (SMD) (38). Furthermore, if the variance is low, even small absolute or relative differences translate to large effect sizes with SMD and corresponding unintentionally large effect sizes with odds ratios.

Since educational materials, such as the distribution of guidelines, may only have a limited effect on reducing low-value care, we conducted a sensitivity analysis on broader intervention categories that combine educational materials with other strategies. For example, interventions that included both audit and feedback and educational materials were combined with trials using only audit and feedback only, and control groups that only received educational materials were combined with those receiving no intervention.

In the last sensitivity analysis, we excluded trials with large (over 10%) baseline differences between intervention and control groups, and trials with small sample sizes and no baseline measurement (under 20 clusters or under 100 providers/participants). The sensitivity analyses were iteratively compared to the primary analysis, comparing each intervention type to no intervention.

We conducted subgroup analyses comparing high-intensity vs low-intensity interventions, tailoring vs no tailoring of the intervention, and theoretical background vs no theoretical background for the intervention. Our prespecified hypothesis was that these characteristics would increase the effectiveness of de-implementation.

We considered an intervention as tailored if the study context was assessed and considered in the intervention development, for example, if the authors conducted a survey to assess barriers to de-implementation before designing the intervention. We considered an intervention as having a theoretical background if the authors reported theoretical literature or otherwise described the use of theoretical literature for planning the intervention.

We divided provider education into educational materials (low intensity) and educational meetings (high intensity). We divided interventions combining audit and feedback and provider education into two categories based on the intervention intensity. High intensity interventions had either feedback given two times and educational meetings lasting one day or over in total, or feedback given three times or more and educational meetings lasting under one day.

4.2 Survey methods

Through the international Choosing Wisely community, we sought research collaboration in high-income countries that could participate in data collection and had a representative sample of primary care physicians available. Finally, we included six countries in the study: Austria, Finland, Greece, Italy, Japan, and Sweden. The survey had five sections: 1) background information, 2) familiarity with Choosing Wisely recommendations, 3) attitudes towards overdiagnosis and overtreatment, 4) barriers to de-implementation, 5) interventions, and possible facilitators for de-implementation. We used previous literature to explain the relevant concepts in the survey (10, 11). We defined overdiagnosis as "1) the diagnosis of a medical condition that would never cause any symptoms or problems, or 2) medicalizing ordinary life experiences through expanded definitions of diseases. Overdiagnosis can be caused by overdetection or overdefinition of disease." We explained overtreatment as "treatment for which there is no or little benefit to the patient, considering both the potential harm from and benefit of the treatment." We did not provide examples of overdiagnosis or overtreatment to avoid directing the focus of participants to specific topics.

We asked and received comments from two content experts and piloted the survey in Finland. The pilot included five primary care practitioners and one layperson from Finland. The survey was translated into local languages by professional translators. Local researchers reviewed the translations to confirm accurate wording.

As background information (section 1), we asked respondents to provide their age, work experience, specialization status, and gender.

For the fourth section (Table 5), we developed a list of potential barriers through discussions with an implementation expert, considering the previous literature (2, 3, 50, 55, 132). We categorized the barriers into three groups: individual, organizational, and patient-related barriers.

Table 5. Survey questions and response options

Question (questionnaire section)	Response options
Are you familiar with the Choosing Wisely recommendations? (Section 2)	I have never heard of them / I have heard of them / I have read a few / I have read many
Do you follow Choosing Wisely recommendations which are relevant in your own clinical practice? (Section 2)	Never / Rarely / Often / Always
In my practice / In my country's healthcare system / In other high-income countries, overdiagnosis is ____? (Section 3)	Not problem at all / A minor problem / A problem to some extent / A major problem
In my practice / In my country's healthcare system / In other high-income countries, overtreatment is ____? (Section 3)	Not problem at all / A minor problem / A problem to some extent / A major problem
How important each listed individual barrier is in your own clinical practice? (Section 4)	No importance / Small importance / Moderate importance / Major importance

Adapted from [Perspectives on low-value care and barriers to de-implementation among primary care physicians: a multinational survey, BMC Primary Care, 2024]

4.2.1 Sample

We initially aimed for a random sample of 2000 primary care physicians currently (or in the previous 24 months) working in primary care practice from each country. Italy (Tuscany), Sweden, Austria, and Japan decided to include all primary care physicians, resulting in larger samples than 2000. In the invitation emails, we described that only primary care physicians were targeted, and we further had an exclusion question to limit our participants to those who had worked in primary care within the previous 24 months.

Local medical societies were contacted to provide a list of emails for primary care physicians, or, if not possible, to send the survey invitation on their behalf. Sweden had an exception, and the emails were sent by a private company (IQVIA Solutions Sweden AB). We drafted the invitation emails in English and translated them into the local languages.

4.2.2 Survey procedure

One week before the initial invitation email, we sent the respondents information about the upcoming survey (heads-up email). After the invitation email including a personal link to the questionnaire, we sent two reminders at two-week intervals. We used SurveyMonkey in Finland, Sweden, Greece, and Japan to collect survey responses. Austria used Limesurvey, and Italy used a locally created platform. We collected the responses anonymously and described the use of the data in the invitation emails. Participants gave their informed consent by answering the survey.

4.2.4 Analysis

We summarized the survey responses by using descriptive statistics. In sections 2 to 4, we presented the proportion of respondents selecting different response options on the four-point Likert scale (Table 5). Furthermore, we calculated the proportion of respondents rating overdiagnosis as a bigger problem for themselves versus their country's healthcare, vice versa, or as big of a problem for themselves as for their country's healthcare.

For section four responses regarding the barriers to de-implementation, we divided the four-point scale into two categories: no/minor importance and moderate/major importance. We then presented the answers as a proportion of participants with an answer of moderate/major barrier to each specific barrier.

We conducted secondary analyses using R statistics with the "Survey" package. We formed a general linear model by using the country as stratum and work experience, gender, Choosing Wisely familiarity, and attitudes towards overdiagnosis and overtreatment as covariates. We tested differences between groups and the effects of covariates by using one-way ANOVA. Furthermore, we compared the country mean of each barrier to the country-specific means of all barriers, as there was a suspicion of response bias.

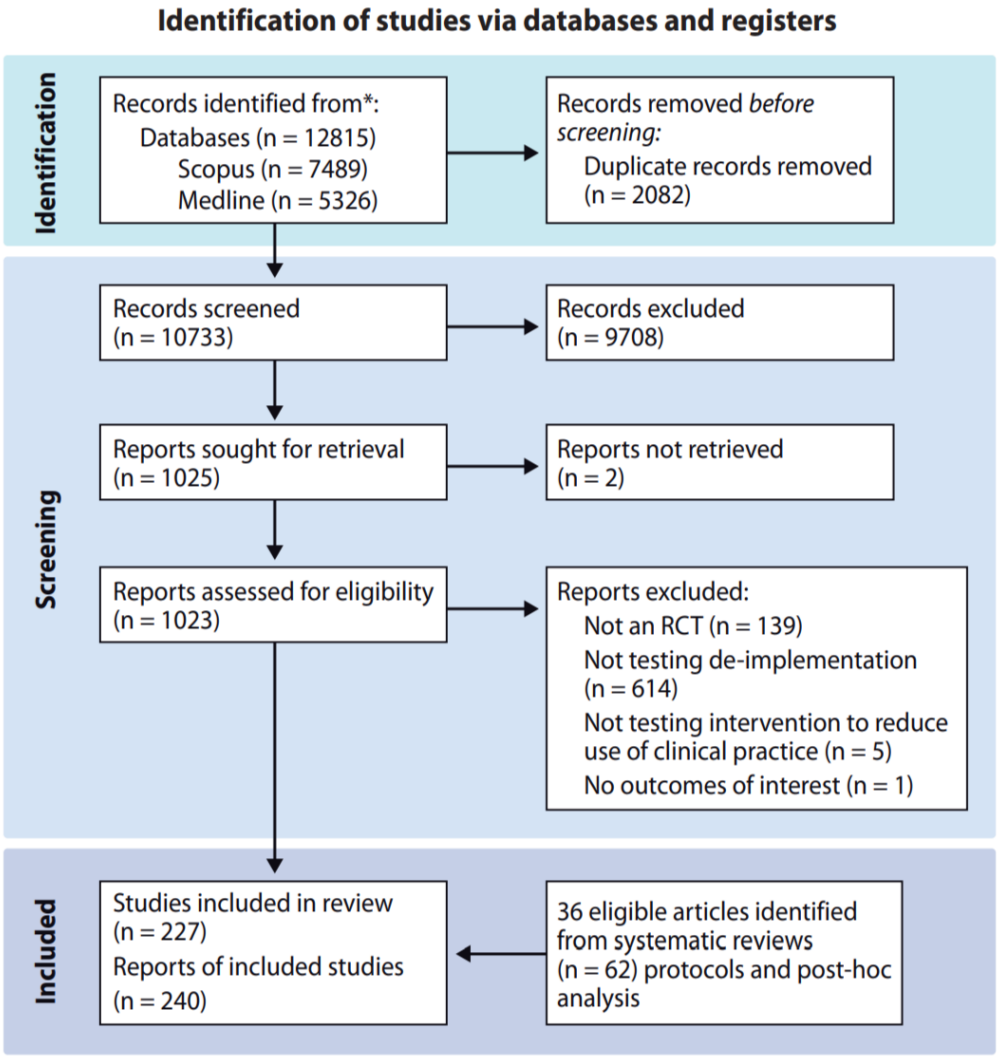
We evaluated the representativeness of the sample by comparing the age and gender distributions to the available true distributions in each country. Furthermore, we subjectively evaluated the representation (including age, amount of clinical work, work experience, and specialization status) by discussion with the local investigators.

5 Results

5.1 Scoping review (study I)

Of the 12,815 abstracts, we found 1,025 potentially eligible articles. After full text screening, we included 204 articles in the data extraction. Furthermore, we added 31 articles by hand-searching reference lists of systematic reviews and five by following up study protocols and other study reports. Finally, we included 240 articles of 227 studies in the data extraction (Figure 1).

Figure 1. Flow diagram of the search and screening process



5.1.1 Study characteristics

Of the 227 studies, almost half were conducted in North America (n = 101, 44%) and published after 2010. Of all trials, 145 (64%) were cluster-randomized trials, and the remaining 82 (36%) used an individually randomized design. Two-thirds (n = 149) were conducted in primary care, and 29% (n = 65) in secondary or tertiary care. Most of the trial interventions focused on reducing drug treatments (n = 163, 72%), with 108 (48%) of these trials targeting a decrease in antibiotic use. Of all trials, 42 (19%) aimed to replace low-value care with other medical practices (with potentially higher value for the patients).

The most cited reasons for de-implementation were “evidence suggests more harms than benefits for the patient or community” (n = 147, 64%), followed by “evidence suggests little or no benefit from treatment or diagnostic test” (n = 115, 50%).

Study outcomes

The most common outcome was total volume care, evaluated by 199 (87%) studies, low-value care use by 64 (28%) studies, patient health outcomes by 59 (26%), and intention to reduce use of low-value care by 15 (7%) studies. Authors reported appropriate care use in 34 (15%) trials. Of those, 16 found an increase in appropriate care, 16 no effect, and two a decrease in appropriate care.

Of all trials, 20 (9%) reported/measured intervention costs, and 45 (20%) costs for healthcare.

Intervention categories

Two-thirds of studies (n = 153, 67%) evaluated interventions consisting of two or more components (“multicomponent interventions”). The most frequently evaluated intervention types were educational materials (n = 101, 44%), educational meetings for providers (n = 99, 43%), and audit and feedback (n = 82, 36%).

Table 6 provides a detailed description of the study characteristics for all trials, and Figure 2 shows the number of trials evaluating each intervention type.

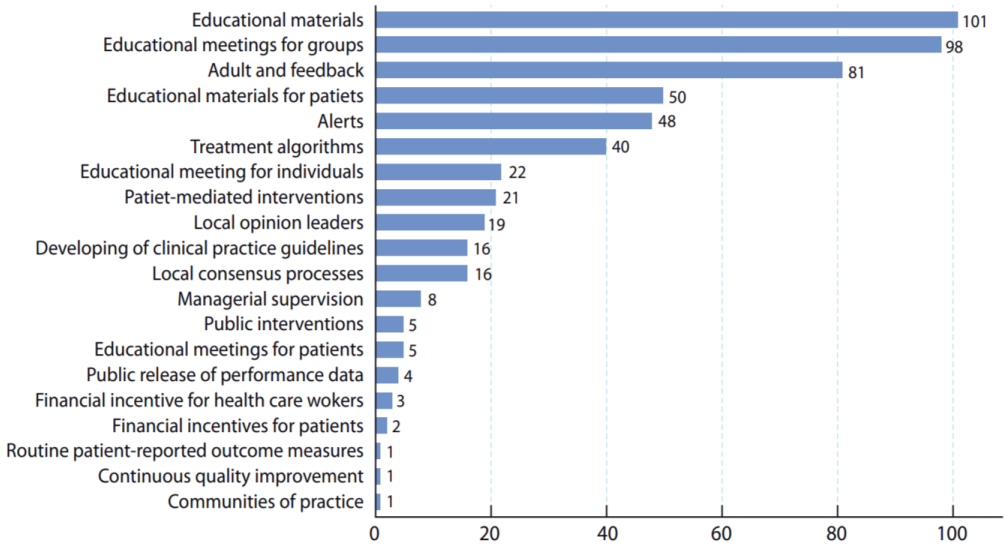
Table 6. Description of the included 227 randomized controlled trials: characteristics, aims, and outcomes (reproduced from [Raudasoja et al. Randomized controlled trials in de-implementation research: a systematic scoping review. 2022], 115)

Characteristics		Aim and rationale		Outcomes	
*Setting**	n (%)	*Aim**	n (%)	*Outcome categories**	n (%)
Primary care – Outpatient	149 (66%)	Abandon	0 (0%)	Health outcomes	58 (26%)
Primary care – Inpatient	3 (1%)	Reduce	225 (99%)	Low-value care use	63 (28%)
Secondary/tertiary care – Outpatient	28 (12%)	Replace	42 (19%)	Appropriate care use	34 (15%)
Secondary/tertiary care – Inpatient	40 (18%)	Unclear	2 (1%)	Total volume of care	194 (87%)
Other	22 (10%)	*Rationale**		Intention to reduce the use of low-value care	17 (7%)
Randomization unit		Evidence suggests little or no benefit from treatment or diagnostic test	115 (51%)	*Measured costs**	
Cluster	145 (64%)	Evidence suggests another treatment is more effective or less harmful	13 (6%)	Intervention costs	20 (9%)
individual	82 (36%)			Health care costs	45 (20%)
*Medical intervention**		Evidence suggests more harms than benefits for the patient or community	145 (64%)	*Reported effectiveness*	
Prevention	9 (4%)			(Some) desired effect	186 (82%)
Diagnostic imaging	29 (13%)	Cost effectiveness	70 (31%)	No desired effect	41 (18%)
Laboratory tests	28 (12%)	Patient(s) do not want the intervention	2 (1%)	*Theoretical basis and tailoring**	
Drug treatment	163 (72%)			Theory-based interventions	48 (21%)
Operative treatments	7 (3%)	Not reported/unclear	20 (9%)	Tailored interventions	40 (18%)
Rehabilitation	2 (1%)			*Intervention complexity***	
other	7 (3%)			Multicomponent	152 (67%)
*Target group**				Simple	84 (38%)
Public	5 (2%)				
Patients	42 (19%)				
Caregivers	17 (8%)				
Physicians	193 (85%)				
Nurses	37 (16%)				
Other	23 (10%)				

*1 trial could be categorized into several categories, and therefore the sum of percentages may be over 100%

**10 trials had multiple treatment arms and tested both simple and multicomponent interventions. Simple intervention was defined as having one intervention category with or without tailoring.

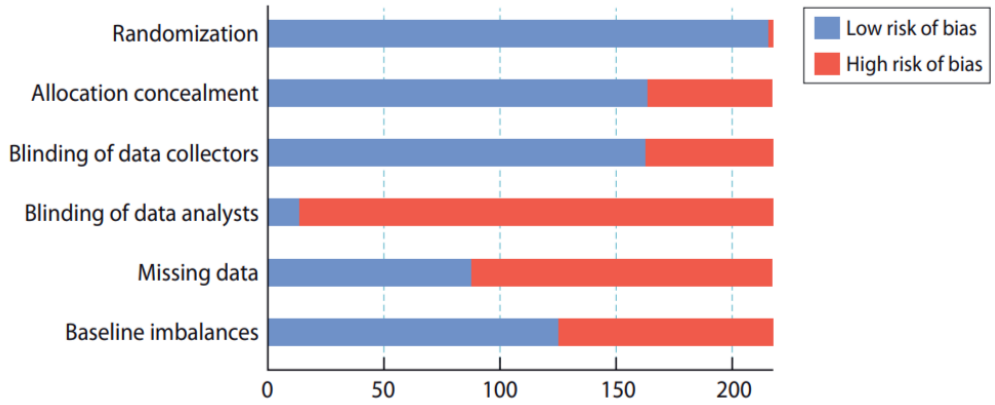
Figure 2. Number of trials in each intervention category



5.1.2 Risk of bias

Of all 227 trials, 226 (99%) had adequate allocation sequence generation, and 172 (75%) had adequate allocation concealment. Of all trials, 171 (76%) had adequate blinding of data collection, and 14 (6%) had blinded analysts. Almost half of the trials (n = 104, 45%) had insufficient reporting of missing data, 92 (40%) trials reported little missing data, and 33 (14%) had large amounts of missing data. Of all trials, 129 (56%) had no or little (under 10%) baseline imbalance. Figure 3 describes the risk of bias for these criteria.

Figure 3. Summary of risk of bias: judgements about each risk of bias domain are presented as numbers across all included studies.



5.1.3 Quality indicators

The median number of clusters was 24 (IQR 44); trials published before 2011 had a smaller number of clusters than trials published in 2011 or after (20 [IQR 26] vs 30 [IQR 42]). From 144 cluster randomized trials, 50 (35%) reported using ICC in sample size calculation (28% vs 39% in studies until and after 2010). Of all trials, 16 (7%) gathered outcome data immediately after the intervention, and 9 trials had unclear or unreported follow-up times. In the remaining studies, the follow-up time median was 274 days (IQR 183) (274 vs 335 days).

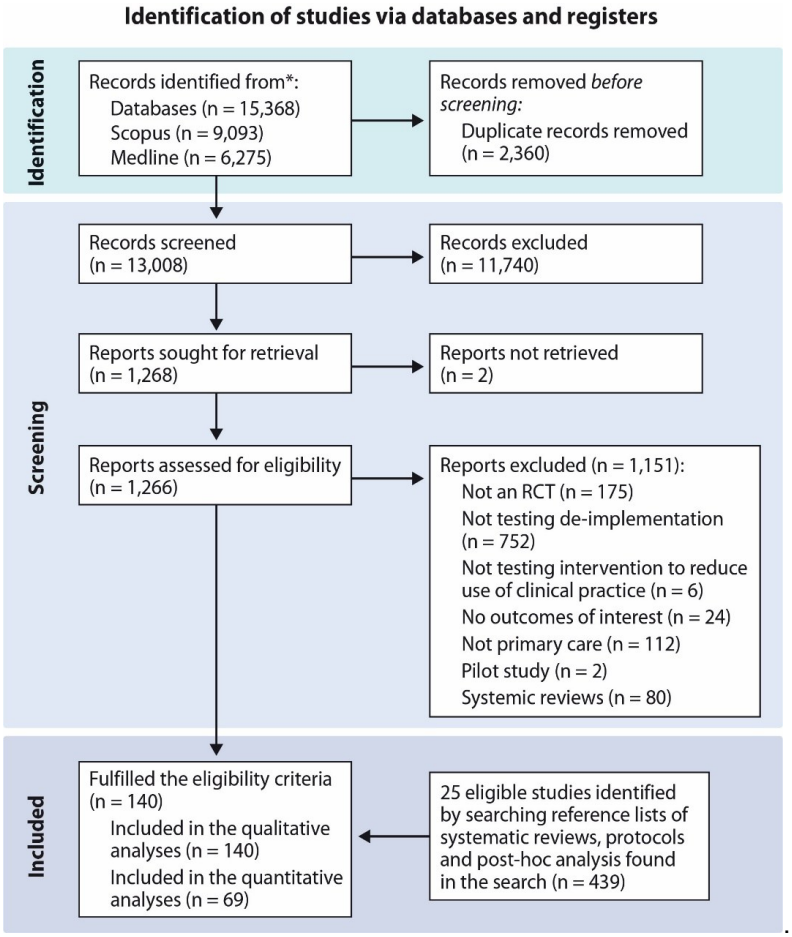
Out of all trials, 174 (76%) reported differences between baseline and follow-up or prevalence estimates for both time points (72% vs 80% in studies until and after 2010). Authors reported tailoring of the intervention in 41 (18%) trials (16% vs 19%). Theoretical frameworks were used to plan the de-implementation intervention in 49 (22%) trials (20% vs 23%). Among trials with provider-level outcomes, such as antibiotic prescriptions, 28 randomized at the patient level (13% vs 12%).

5.2 Systematic review and meta-analysis (Study II)

From 13,008 abstracts, we identified 1,266 potentially eligible articles. From these, we included 115 trials after full-text screening. Additionally, we manually searched reference lists from systematic reviews (found by our search) and followed up on protocols and secondary trial publications. From these, we added 25 trials, bringing

the total to 140 trials. Finally, after requesting missing data from study authors, we had 97 trials with sufficient data for meta-analyses.

Figure 4. Flow chart of the study selection process



5.2.1 Study characteristics

Out of the 140 trials, 109 (78%) were cluster trials, while 31 (22%) were randomized at the individual level. Among all trials, 91 (65%) compared de-implementation to no intervention or usual care, and 49 (35%) compared to another intervention. Authors reported low-value care outcomes in 37 (26%) trials, total volume of care in 122 (87%) trials, appropriate care in 17 (12%) trials, and health outcomes in 14 (10%) trials. About half (n=62) of the trials used binary outcomes, while the other half (n=78) used continuous outcomes. More than half of the trials (n=59, 53%) measured antibiotic use,

30 (27%) other drug treatments, 14 (13%) imaging, and 12 (11%) laboratory tests. In all trials, the median follow-up time was 365 days (IQR 180-365).

Table 7 presents the study characteristics of trials included in the meta-analysis, categorized by intervention. Specific study characteristics (including intervention details and estimated relative effect sizes) of all trials are in supplementary Tables 1-12 (Study II).

Table 7. Study Characteristics

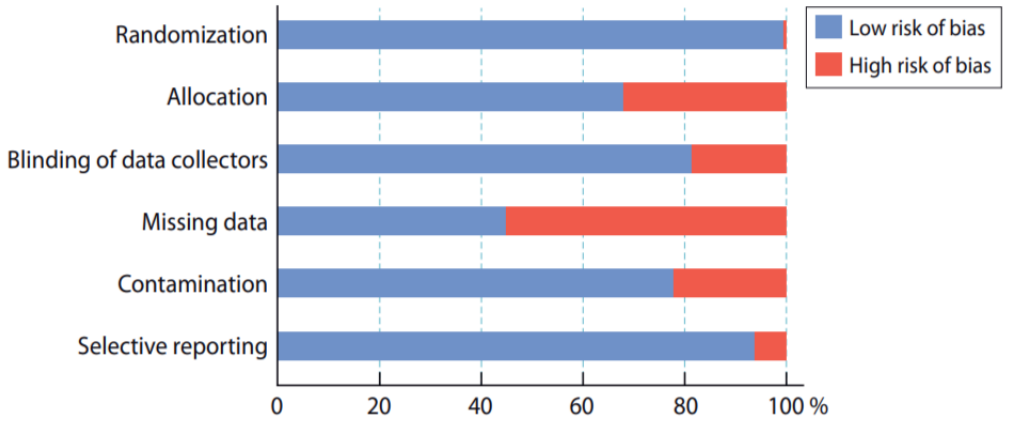
	Provider education (PrE)	Patient education (PaE)	Audit and Feedback (AF)	PrE + AF	PrE + PaE	PrE + PaE + AF	PrE + Decision support (DS)	PrE + PaE + DS
N of studies	11	4	6	20	10	5	4	3
N of randomized trials								
cluster	9	1	2	16	9	3	4	2
individual	2	3	4	4	1	2	0	1
Number of providers*	3944	18	9078	24191	1093	6106	121	143
N of trials with								
binary outcomes	7	4	2	12	9	1	4	2
continuous outcomes	4	0	4	8	1	4	0	1
Target intervention								
Antibiotics	5	4	1	12	5	4	2	3
Other drugs	4	0	3	5	3	0	1	0
Laboratory test	3	0	2	5	1	0	0	0
Other	0	0	1	2	3	0	1	0
Median follow-up days (IQR)	365 (165 to 518,5)	24 (7 to 62.5)	365 (297 to 639)	365 (180 to 365)	240 (120 to 365)	180 (180 to 365)	273 (180 to 639)	150

* Number of practices used from trials that did not report the number of physicians. In patient education trials, two trials randomized patients, and the total number of patients in the four trials was 18488.

5.2.2 Risk of bias

Random sequence generation was adequate in 139 of 140 trials (99%), and 95 (68%) had adequate allocation concealment. Of all trials, 114 (81%) had blinded outcome collectors or the data was collected from a database/registry; 63 (45%) had little or no missing data; 109 (78%) had low risk of contamination; and 131 (94%) had low risk of selective reporting. Figure 5 describes the overall risk of bias. The study-specific risk of bias assessment is reported in the supplementary Tables 1-12 in the original article.

Figure 5. Overall risk of bias

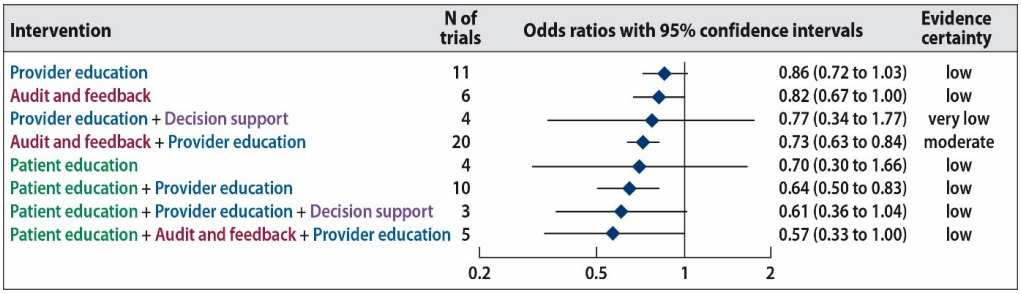


5.2.3 Impact on low-value care use

Provider education

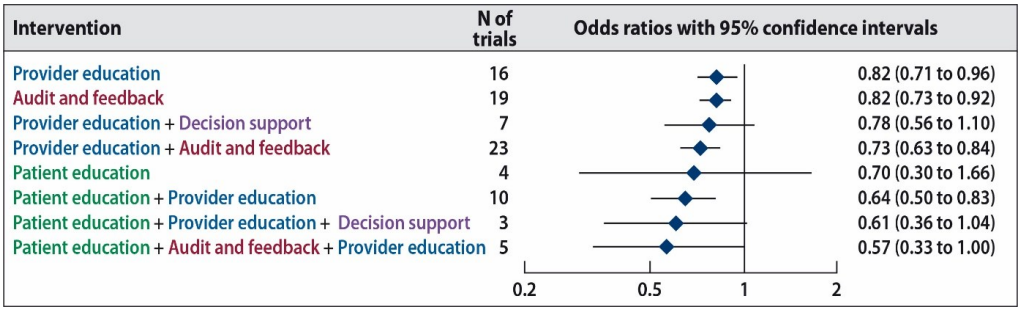
The results from 11 trials comparing provider education to no intervention suggested a small impact on low-value care (OR 0.86; 95% CI 0.72 to 1.03; I2 = 36%; low certainty; figure 6). The estimated absolute reduction was 32 fewer patients receiving low-value care (95% CI: 67 fewer to 6 more) per 1,000 patients. By including trials comparing provider education to educational material distribution, results were similar (16 trials, OR 0.82, 95% CI 0.71 to 0.96) (Figure 7).

Figure 6. Forest plot for meta-analyses of different de-implementation interventions



In addition, six trials more compared provider education to no intervention, but had missing data and could not be included in the meta-analysis. The results from those studies were similar to those of the included trials and would likely not have changed the point estimate, but would probably have increased precision (Supplementary Table 12, Study 2).

Figure 7. Forest plot of sensitivity analysis combining educational materials with other intervention types



We found no credible subgroup effects on theoretical background (vs no theoretical background), tailoring (vs no tailoring), or intensity (educational materials vs educational meetings). In sensitivity analyses, we found similar results for both trials with binary outcomes (7 trials, OR 0.93, 95% CI 0.76 to 1.13) and continuous variables (4 trials, ROM 0.90, 95% CI 0.75 to 1.10). A sensitivity analysis excluding small trials without baseline measurements and trials with large baseline imbalances also suggested a similar impact (OR 0.88, 95% CI 0.49 to 1.56).

Patient education

The results from four trials on patient education (vs no intervention) suggested a moderate impact on low-value care use (OR 0.70; 95% CI 0.30 to 1.66; I2 = 62%; low certainty; Figure 6). The estimated absolute reduction was 51 fewer patients receiving low-value care (95% CI: 162 fewer to 68 more) per 1,000 patients. One trial compared education on low-value care to education on viral infections. A sensitivity analysis excluding that trial suggested a slightly larger impact (OR 0.52, 95% CI 0.30 to 0.89).

Audit and feedback

The results from six studies on audit and feedback suggested a small impact on the use of low-value care (OR 0.82, 95% CI 0.67 to 1.00; I2 = 84%; low evidence certainty; Figure 6). The estimated absolute reduction was 41 fewer patients receiving low-value care (95% CI: 0 to 80 fewer) per 1,000 patients. By including trials that combined the distribution of educational materials with audit and feedback interventions or having educational materials for the control group, the results were similar (18 trials, OR 0.82, 95% CI 0.73 to 0.92) (Figure 7). The trials with continuous outcomes suggested smaller or no impact (ROM 0.92, 95% CI 0.80 to 1.07, 4 trials).

We excluded three trials from the meta-analysis because of missing data. The results in these trials were similar to those from the meta-analysis and would likely have increased the precision (Supplementary Table 12, Study 2).

Provider education combined with audit and feedback

The results from 20 trials on provider education combined with audit and feedback suggested a small impact on low-value care use (OR 0.73; 95% CI 0.63 to 0.84; I^2 = 86%; moderate certainty; Figure 6). The estimated absolute reduction was 73 fewer patients receiving low-value care (95% CI, 42 to 103) per 1,000 patients. By including trials that compared provider education combined with audit and feedback to educational material distribution, the results were similar (23 trials, OR 0.73, 95% CI 0.63 to 0.84) (Figure 7).

A subgroup analysis suggested a larger impact from higher-intensity interventions (OR 0.59 [95% CI 0.47 to 0.74] vs lower-intensity interventions OR 0.82 [95% CI 0.71 to 0.95]; p-value for the subgroup difference 0.006). A subgroup analysis on tailoring and theoretical background did not suggest credible differences. A sensitivity analysis of

trials with continuous outcomes suggested a smaller or no effect (12 trials, ROM 0.96, 95% CI 0.92 to 1.00). Trials with binary outcomes suggested similar effects (8 trials, OR 0.71, 95% CI 0.58 to 0.86).

We excluded eight trials from the meta-analysis because of missing data. The results in these trials were similar to those from the meta-analysis and would likely have increased the precision (Supplementary Table 12, Study 2).

Provider education combined with patient education

The results from ten trials on provider education combined with patient education suggested a moderate impact on low-value care use (OR 0.64; 95% CI 0.50 to 0.83; I^2 = 61%; low certainty; Figure 6). The estimated absolute reduction was 95 fewer patients receiving low-value treatment or test (95% CI, 43 to 137) per 1,000 patients.

A sensitivity analysis of trials with binary outcomes suggested a similar impact (9 trials, OR 0.68, 95% CI 0.52 to 0.88). A subgroup analysis on tailoring and theoretical background did not suggest credible differences.

We excluded two trials from the meta-analysis because of missing data. The results in these trials were similar to those from the meta-analysis and would likely have increased the precision (Supplementary Table 12, Study 2).

Provider education combined with audit and feedback and patient education

The results from five trials suggested a large impact on low-value care use (OR 0.57; 95% CI 0.33 to 1.00; low certainty; Figure 2). The estimated absolute reduction was 108 patients receiving low-value treatment or test (95% CI 0 to 185) fewer per 1,000 patients. A sensitivity analysis on trials with continuous outcomes suggested a similar impact (3 trials; ROM 0.91, 95% CI 0.72 to 1.15).

We excluded three trials from the meta-analysis because of missing data. The results in these trials suggested smaller impact and would probably have decreased the effect size (supplementary Table 12, Study 2).

Provider education combined with a decision support

The results from four trials suggested a slight impact on low-value care use, but the evidence was very uncertain (OR 0.77; 95% CI 0.34 to 1.77; I2 = 54%; very low

certainty; Figure 6). The estimated absolute reduction was 47 patients receiving low-value care (174 fewer to 120 more) per 1,000 patients. A sensitivity analysis including trials with educational materials as a control suggested similar results (OR 0.78; 95% CI 0.56 to 1.10; Figure 7).

Provider education combined with decision support and patient education

The results from three trials suggested a moderate impact on low-value care use (OR 0.61; 95% CI 0.36 to 1.04; I2 = 0%; low certainty; Figure 2). The estimated absolute reduction was 100 patients receiving low-value care treatment or test (179 fewer to 9 more) per 1,000 patients.

We excluded one trial from the meta-analysis due to missing data, which suggested a smaller impact on low-value care use (Supplementary Table 12, Study 2).

Others

The results from two trials measuring the impact of cost information on laboratory test forms suggested small to no differences in test ordering (ROMs 1.00, 95% CI 0.76 to 1.24; and 0.86, 95% CI 0.69 to 1.03) (133, 134). A US study measured the impact of removing inappropriate laboratory tests from the default menu and adding traffic lights to signal the appropriateness of laboratory tests on laboratory order forms. The results suggested a large reduction in low-value test ordering (ROM 0.48; 95% CI 0.35 to 0.60) (135).

5.2.4 Impact on appropriate care and health outcomes

Of the 111 trials, 11 measured appropriate care. In 10 trials, the authors found either no effect or an increase in appropriate care, while one trial reported a decrease in appropriate care use (136). Of all trials, six measured re-consultations or emergency department visits and suggested either a small decrease or no effect. In five trials measuring health outcomes, all outcomes differed and suggested little to no impact on health.

5.3 Survey (Study III)

We sent the survey to 16,935 primary care physicians, from which 1,731 responded (response rate 10.2%). Finally, 1505 had worked in primary care practice in the last 24 months and answered at least one question other than background information, and were included in the analysis. Study participant characteristics varied between countries but likely represented the local primary care physicians well. The respondent characteristics are reported in Table 8. Evaluations of sample representativeness in each country are in supplementary Tables 7-15 in the original article.

Of all 1,505 respondents, 53.3% had read at least a few Choosing Wisely recommendations, and of those, 72.1% answered that they often followed the recommendations (relevant to their practice). The familiarity varied between countries, ranging from 18.1% of respondents reading at least a few recommendations in Austria to 79% in Sweden.

Table 8. Respondent characteristics

	Total N = 1,505		Austria N = 238	Finland N =370	Greece N = 95	Italy N = 246	Japan N = 280	Sweden N =276
	%	n	%	%	%	%	%	%
Work experience (years)								
Under 5	17	258	17	30	4	24	10	7
5-10	21	316	22	19	13	15	30	23
11-20	26	385	28	23	40	9	38	25
21-30	20	298	14	14	35	28	14	26
Over 30	16	245	19	15	8	25	8	20
Gender*								
Male	53	803	52	39	54	61	72	47
Female	46	698	47	61	46	39	28	53
Specialization								
Specialist in FM	56	840	99	34	17	50	47	75
Specialist - other	19	289	1	9	80	28	34	3
Specializing/no specialization	25	376	0	56	3	22	19	21

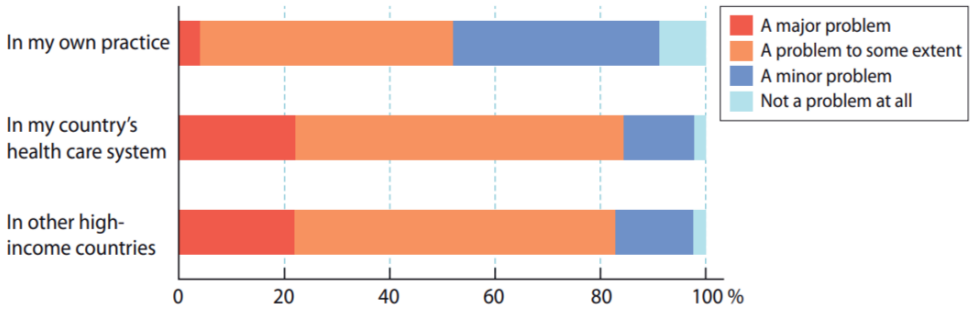
FM=family medicine

*In addition, one “other” response, and three respondents with a missing answer

5.3.1 Attitudes towards overdiagnosis and overtreatment

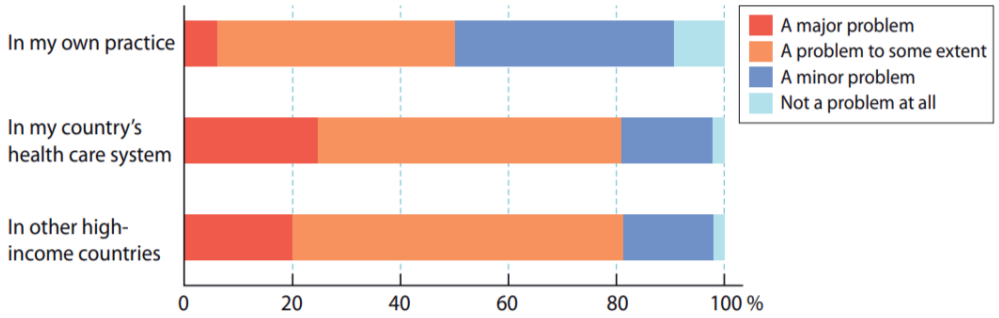
Of the 1,439 participants, 52% rated overdiagnosis as a problem in their own practice (problem to some extent/major problem), 85% in their country's healthcare, and 83% in other high-income countries (Figure 8). In a secondary analysis, 49% rated overdiagnosis as a bigger problem in their healthcare system than in their practice, and 2% as a smaller problem in their healthcare system than in their practice.

Figure 8. How big a problem is overdiagnosis___?



Of the 1,439 participants, 50% rated overtreatment as a problem in their own practice (problem to some extent/major problem), 81% in their country's healthcare, and 81% in other high-income countries (Figure 9). In a secondary analysis, 47% rated overtreatment as a bigger problem in their healthcare system than in their practice, and 2% as a smaller problem in their healthcare system than in their practice.

Figure 9. How big a problem is overtreatment___?

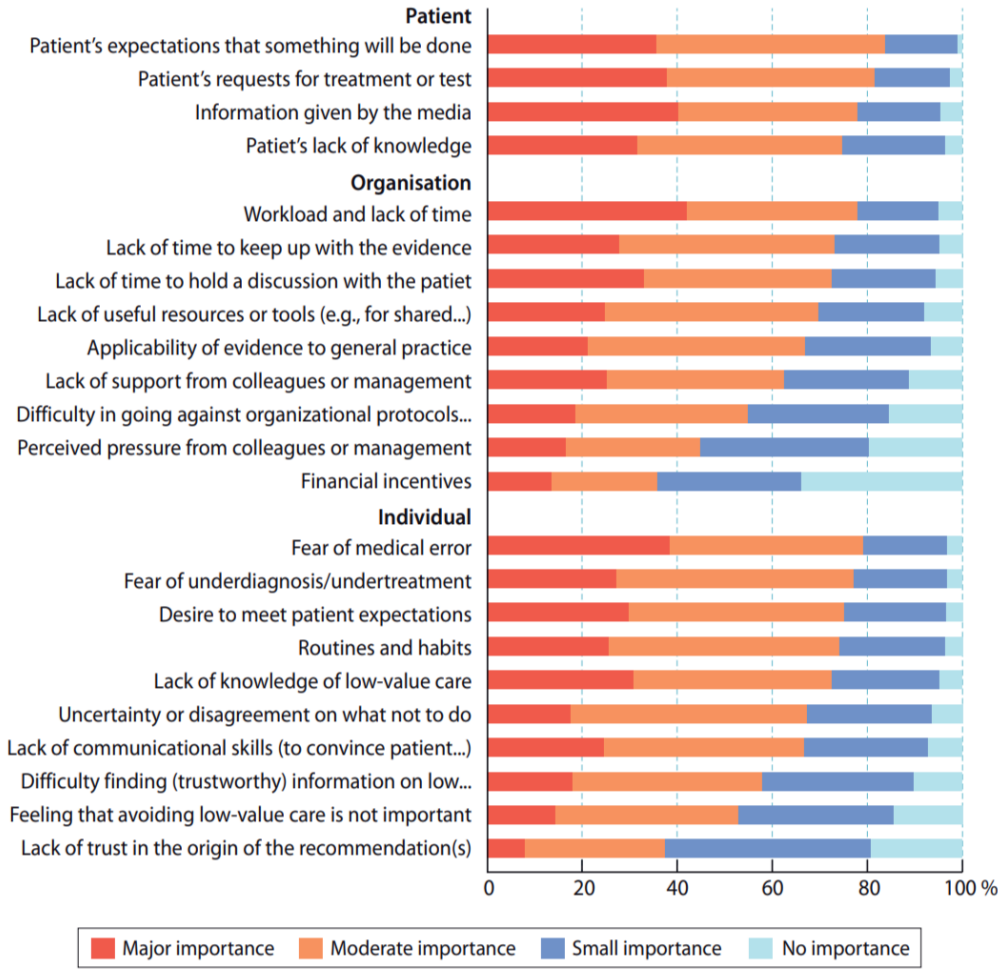


Country-specific results are in Article III, Figure 1.

5.3.2 Barriers to de-implementation

Of the 1,220 participants who responded to questions about barriers, 85% rated "patient expectations that something will be done" as having either moderate or major importance as a barrier to reducing their use of low-value care. The next three highest rated barriers were "*patients' requests for treatments and tests*" (83%), fear of medical error (81%), and workload and lack of time (81%). Figure 10 presents the specific answers, and Article III, Figures 2-4, show the country-specific proportions of moderate or major responses to all barriers.

Figure 10. Barriers to reducing use of low-value care



Reproduced from [Perspectives on low-value care and barriers to de-implementation among primary care physicians: a multinational survey, BMC Primary Care, 2024; Additional file 1]

5.3.3 Secondary analyses

The perception of overdiagnosis and overtreatment as a problem in their own practice (questions 9 and 12) was positively correlated with Choosing Wisely familiarity (η^2 = 0.016, $p < 0.001$). Men rated overdiagnosis and overtreatment as bigger problem vs women (mean 5.23 [scale 3-8] versus 5.04, η^2 = 0.003, p=0.014) Furthermore, countries had differences in their ratings (Austria mean 4.61, Finland 5.03, Greece 5.13, Italy 5.63, Japan 5.17, Sweden 5.20, η^2 = 0.064, $p <0.001$). The perception was not correlated with work experience (η^2 = 0.0004, p = 0.055).

A larger mean perception of all barriers was correlated with a larger perception of overdiagnosis and overtreatment as a problem in their own practice (η^2 = 0.053, $p < 0.001$) and with less work experience (η^2 = 0.003, p = 0.044). Participants from different countries showed substantial differences in their ratings (Austria: 2.68, Finland: 2.63, Greece: 3.12, Italy: 3.00, Japan: 3.12, Sweden: 2.83 [scale 1-4], η^2 = 0.147, $p < 0.001$). However, there were no differences between genders (η^2 = 0.005, p = 0.739), and no correlation to Choosing Wisely familiarity (η^2 = 0.00003, p = 0.078).

6 Discussion

6.1 Summary of findings

This thesis 1) provides methodological guidance to future de-implementation trials, 2) helps decision-makers to choose the de-implementation strategies with the highest potential impact on low-value care use, and 3) describes the attitudes on low-value care and barriers to de-implementation in primary care, essential knowledge for planning de-implementation strategies.

We found several methodological limitations in de-implementation trials, including small cluster randomized trials, potentially unreplicable intervention designs, and the use of indirect outcomes. The methodological limitations in de-implementation trials lead to unreliable effect estimates and, most importantly, to low applicability of de-implementation interventions. With simpler intervention designs, future research could help identify de-implementation strategies that are replicable. Larger trials, including multi-center studies with multiple contextual characteristics, would enhance the generalizability of the results. However, sometimes researchers may have to balance between locally developed interventions with potentially higher impact and simpler strategies that are applicable to other contexts.

We found moderate certainty evidence on the effectiveness of provider education combined with audit and feedback, leading to about 23% relative reduction in low-value care use. Single strategies—provider education, audit and feedback, and patient education—may slightly reduce low-value care, although the evidence is uncertain (low to very low evidence certainty). De-implementation strategies may have a larger impact when including patient education, as they led to about 30-35% reductions in low-value care use (low evidence certainty).

According to our multinational survey, primary care physicians perceive the problems of overdiagnosis and overtreatment to be bigger in their healthcare system than in their own practice. Furthermore, participants rated patient-related barriers higher than other barriers. It could make it more difficult to de-implement if the problem is mainly seen outside the primary care physicians' control. Other important barriers, such as fear of medical error, workload, and lack of time, should be acknowledged in the development of future healthcare. Multiple important barriers suggest that de-implementation may

be an even more complex issue than previously thought, and highlight the importance of assessing the local barriers when designing de-implementation strategies.

6.2 Implications

6.2.1 Producing applicable evidence on de-implementation interventions

We identified several methodological limitations in de-implementation trials, including a small number of clusters in cluster trials, potentially unreplicable study designs, and the use of indirect outcomes. These limitations, in addition to often high risk of bias, may lead to poor trustworthiness of study results. We provided a list of nine recommendations to address these issues in future research (Study I, Table 3).

First, we suggested that using complex intervention designs could often lead to unreplicable interventions. In addition to the lack of reporting of the intervention characteristics (137, 138), the complexity (including context-specific intervention characteristics) could make the intervention very difficult to replicate. Furthermore, de-implementation research lacks evidence on specific factors effective in reducing low-value care use. Identifying the factors and understanding the causal mechanisms of change would also require randomized trials on simple interventions as well as factorial trials measuring the effectiveness of (simple) intervention modifications.

Second, half of the studies measured the effect of educational sessions on low-value care use. As the characteristics of the educator may influence the effectiveness of an intervention, it can act as a confounder in the trial. The issues could be addressed by involving multiple educators or providing a protocol for education rather than delivering the educational sessions.

Third, small cluster randomized trials are at increased risk of false-positive results if the results are not adjusted using small-sample correction methods (96, 139). In our sample, the median number of clusters (in cluster RCTs) was 24, which is already at an increased risk of false-positive results. Furthermore, if contamination is likely, researchers should still weigh between the benefits of increased statistical power from individually randomized trials and more accurate point estimates from a cluster design (94).

Fourth, most of the studies used indirect outcomes, including total volume of care and intentions to reduce the use of low-value care. The total volume of care may include

both appropriate and inappropriate care, and as such, the changes in outcome may not accurately represent the actual low-value care of interest. Diagnosis-based outcomes may be more accurate measures for low-value care use. However, they are also usually indirect outcomes, and the accuracy may be limited by diagnostic shifting (i.e. the intervention [and trial characteristics] may induce providers to change their diagnostic practices while still making similar treatment decisions). If the resources do not allow for sufficiently accurate outcome measures, researchers may want to see both the change in total treatment/testing volume and diagnosis-based low-value care outcome. If diagnosis-based low-value care is lowering, it could be due to diagnostic shifting. But, if both are decreasing versus the comparison group and the diagnosis-based outcome is accurate enough, the change is likely to represent a decrease in actual low-value care use.

Fifth, a minority of the studies used a theoretical background and intervention tailoring in designing the interventions. In addition to potentially increased effectiveness, the theoretical background may help in identifying the factors leading to impact in de-implementation. Tailoring may lead to decreased applicability, but that could be countered with more pragmatic intervention designs. For example, instead of tailoring the intervention at the stage of intervention development, researchers could provide a protocol to the intervention participants on how to tailor the intervention to local needs. In this case, even though the final intervention is locally tailored, the intervention being tested can be replicated with the same details.

Considering the previous issues, designing a de-implementation intervention often involves balancing applicability and the effectiveness of an intervention. Simpler interventions would be easier to scale and could help with exploring the key mechanisms leading to successful de-implementation. But on the other hand, they may not be as effective as interventions focusing on the local barriers and facilitators with multiple de-implementation strategies. Is it worth doing a randomized trial if it does not provide generalizable evidence, or should these approaches be saved for local quality improvement? In that case, less resource-intensive methods of measurement might serve the purpose. However, randomized experiments can also serve as a tool for local quality improvement (140). This ensures that the measurement is accurate, even though the aim is not to provide applicable evidence for others.

6.2.2 Effectiveness of different de-implementation strategies

Previous research has already suggested that multifaceted de-implementation strategies may help achieve meaningful impact in reducing the use of low-value care (7, 8). A recent systematic review on de-implementation trials suggested that single-strategy interventions would reduce low-value care by a median of 15% (9), whereas de-implementation interventions including multiple strategies reduce low-value care by slightly more, 20%. The limitations in previous systematic reviews include a lack of comprehensive meta-analysis, and when meta-analysis has been conducted, the inclusion of heterogeneous interventions, which makes it difficult to estimate the potential effectiveness of different de-implementation strategies.

We used a modified TIDieR checklist to collect the intervention characteristics and formed categories for different de-implementation strategies. This allowed us to estimate the effect of specific de-implementation strategy as well as their combinations.

As our scoping review suggested, the evidence was overall of low certainty. The most certain evidence was for provider education combined with audit and feedback, which led to about 23% relative reduction in low-value care use (moderate certainty). The impact may be further increased with higher intensity interventions.

Considering whether a specific intervention is possible to replicate in a local setting may further help in choosing the most reliable strategies. For example, two trials measured the effect of sending feedback to clinicians with a high rate of low-value care use (136, 141). Both trials were large, multi-center studies that used low-cost and easily scalable/replicable interventions and found a decrease in low-value care use. Prioritizing these kinds of interventions over designing or adopting a locally developed intervention with more uncertain results would both save resources for better use and improve patient health by avoiding the harms associated with low-value care.

When the evidence is uncertain, and interventions require tailoring to the local characteristics (e.g. barriers and facilitators to de-implementation), the local effectiveness should be measured. Tailoring may increase the chances of achieving an impact, but it leads to more uncertain effects as the intervention changes. If the strategy does not achieve sufficient effect, the intervention can be modified or replaced by another intervention.

A previous systematic review on engaging patients in decision-making found about a 24% reduction in low-value care use (70). Most of the included trials measured the effect of patient education. Similarly, our results suggested that patient education could be the most effective single strategy to reduce low-value care use. Combining patient education with other strategies could decrease the use of low-value care by up to 35%, although the evidence was of low certainty. Patient education can be provided through educational materials, face-to-face education by healthcare providers, or media information campaigns. There may not be enough evidence to determine which type of information should be provided to patients. Nevertheless, information on the natural course of disease, benefits and harms, and alternative measures was tested by several trials, and all of those found large decreases in the use of low-value care (142-145).

Although improving patient health is the ultimate aim of healthcare, it was rare in our sample (in only 11% of all trials). All these trials suggested either a small improvement or no impact on patient health. Measuring appropriate care outcomes helps further assess the impact on the quality of care. Of the 11 trials using the outcome, only one trial suggested decreases in appropriate care (136). In this trial, the authors measured the impact of feedback on quetiapine use in older and disabled adults. In addition to the decrease in low-value care use, guideline-concordant prescribing decreased by 10% compared to the control group. Authors did not find effects on mortality and hospital use, suggesting no impact on patient health.

6.2.3 Barriers and attitudes to reducing low-value care

In our study, primary care physicians perceived multiple barriers as important for de-implementation. Previous studies have identified substantially smaller numbers of barriers in one study (2, 3, 50-56). It could be due to local differences in barriers, or also due to study methods not being able to find all important barriers. For example, a focus group may end up discussing only a small part of the process, leaving several barriers unidentified. The large number of important barriers may also explain why de-implementation is often so difficult, and it may take several years to abandon practices that are proven ineffective or harmful (147). Perhaps unsurprisingly, as the implementation of new practices may take a long time, one analysis suggests even 17 years from the time of the research finding a practice effective (148).

The barriers to de-implementation may not be much different than those of barriers to the use of evidence-based medicine (132). A systematic review of barriers to GPs' use of evidence-based medicine identified patients' preferences and expectations, the applicability of evidence to general practice, time pressure, and a lack of knowledge and skills as key barriers.

The overconfidence bias – that is, humans generally rate themselves above average – has been identified by several studies (149-151). In our study, only 2% of the respondents perceived overdiagnosis and overtreatment as a bigger problem for themselves than in their country's healthcare. Half of the respondents rated overdiagnosis and overtreatment as minor or no problem in their own practice. It may be difficult to change behavior if primary care practitioners do not see the problem in themselves. Clinicians often overestimate the benefits of medical treatments (15), which can complicate their assessment of the situation in their own practice. Nevertheless, these findings underscore the importance of incorporating the concept of low-value care into regular medical discourse, education, and healthcare. Furthermore, developing tools to understand the harms and benefits of medical treatments and tests should be a key priority. Discussing and evaluating the individual benefits and harms for a patient is very difficult if accurate and understandable estimates are lacking.

Primary care physicians rated time constraints as important barrier to reducing low-value care use, consistently across all countries. Future healthcare should acknowledge this issue when designing the system changes. Longer appointments

could facilitate discussions about the harms of low-value care with patients and, potentially, reduce resource use, leading to more cost-effective care.

Clinicians face a lot of uncertainty in their work, some of which may translate to defensive medicine (i.e. recommending treatments or tests with no benefits to a patient because of uncertainty and a desire to do something). In our results, the most important individual-related barriers were fear of medical error and fear of underdiagnosis and undertreatment. Providing education on why overtreatment and overdiagnosis are often higher risks for the patient than underdiagnosis/undertreatment could help with these issues. Furthermore, healthcare organizations should ensure that they have sufficient senior support for their younger physicians.

Barriers to de-implementation vary across different study contexts, which is why understanding the local situation is crucial. A strategy from one setting may not translate into impact in another setting. Understanding the complete process of change may require exploring the barriers and facilitators in a systematic way, including considering the previous literature on these issues. Our comprehensive list can help identify all potential barriers in the local setting.

Another approach to de-implementation could be strategies that do not require considering the local barriers and facilitators. With stricter regulations for new practices, we can avoid situations in which medical practices are later proven ineffective or harmful (23). Similarly, regulators could find ways to make current practices unavailable to the clinician. For example, after Canadian regulators eliminated reimbursement for vitamin D testing without appropriate indications, the total vitamin D testing volumes decreased by over 90% from the previous volumes (66).

6.3 Limitations

In Studies I and II, most of the limitations were related to the research methodology of the included original studies. These issues, mainly the high risk of bias, led to low to very low evidence certainty in all but one intervention category. Future studies should focus on pragmatic and replicable intervention designs as well as on using higher-quality trial methods.

However, some review methods should also be considered when interpreting the results. First, although we found substantially more randomized trials on de-implementation than the previous reviews (149 trials in the scoping review were not identified by previous reviews), the indexing of de-implementation trials was heterogeneous, and we may have missed some relevant studies. Second, we used the same risk of bias tool for both individually and cluster-randomized trials. Although we included specific criteria for both randomization methods when necessary, this may have led to some unintended differences in the risk of bias results.

Third, we combined the continuous and binary outcomes into a single odds ratio estimate. Converting continuous outcomes to odds ratios is dependent on both the point effect estimate and variance. This may lead to unintentionally large point estimates as well as small weights in the meta-analysis, especially with large trials (with less variance). Considering this issue, we provided a sensitivity analysis on binary and continuous outcomes separately if at least three trials were included from a specific intervention category for the outcome. Also, interpretation of the sensitivity analyses requires caution. The trials with continuous outcomes were generally larger-scale and tested less intense interventions. Finally, despite the limitations of this method, giving one estimate should help with the interpretation and avoid the difficulty in choosing the right estimate to use (binary/continuous).

Fourth, we did not include blinding of the study participants in the risk of bias assessment. The Cochrane risk of bias tool suggests judging a study as high risk of bias only if there are potential deviations from the interventions in the protocol due to the research context (e.g., the recruitment process or other study characteristics may cause participants to receive extra interventions or deviate from the protocol in different ways). Nevertheless, non-blinding may cause other biases in the study results in other ways than that. Known as “the Hawthorne effect”, a bias could arise from study participants performing better because of the knowledge of being observed.

Furthermore, knowledge of allocation may bias the results. For example, participants in the intervention arm may perform better because of the expectation of benefit from the intervention. However, if the criteria for blinding were included in the assessment, it may not have changed the evidence certainty. Most intervention categories were already downgraded due to risk of bias, and assessing the lack of blinding (or potential deviations from the intended interventions) would not likely have led to further downgrading.

The survey (Study III) also had some limitations. First, the recruitment was tailored to the local setting. This led to some differences in samples between countries. For example, country registers had varying strategies to include young doctors after graduation. The slow inclusion in some countries led to varying representation of young doctors in the country samples.

Second, the low response rate decreases the generalizability of our findings. As we were unable to get address information in all countries and we did not have funding for postal distribution, we used email lists to distribute the survey. Furthermore, we were unable to use the email lists ourselves; instead, the survey was distributed by the medical societies separately or through their regular newsletters. In Sweden, a private company holding email information for primary care physicians distributed the survey.

Third, we piloted the English version of the survey. Although the survey was translated into local languages by professional translators, this may decrease the comparability of results between countries. The local researchers, who were familiar with Choosing Wisely and related terms, revised the survey translation when needed. Participants may also have different cultural understandings of the questions, even if the translations were accurate. This may decrease the reliability of interpretation between country differences.

7 Conclusions

The overall results highlight the importance of including the patient perspective in de-implementation planning. Physicians' perceptions suggest that patients' preferences could play a key role in care decisions and interventions, including patient education, which may result in up to 34% relative decreases in low-value care use. However, the evidence was of low certainty, suggesting that future research should focus on finding more applicable evidence on patient education strategies.

We found moderate certainty evidence that provider education, combined with audit and feedback, could lead to a relative reduction of approximately 23% in low-value care. These interventions are often less resource-intensive than alternatives. Therefore, considering both the impact and the most certain evidence base, they could be prioritized by the healthcare authorities. Tools that help in self-evaluation and understanding the benefits and harms of treatments and tests (many of which are low value to patients) should be a key priority. This could also include tools for shared decision-making, which would help clinicians provide information to patients.

Most individual de-implementation trials produce unapplicable evidence. To change the course, de-implementation researchers should consider when a randomized trial is an optimal tool for evaluation. The primary aim of a randomized trial is to produce unbiased and generalizable evidence on a specific intervention. While RCTs will provide the most accurate estimates of impact, if the intervention is impossible to adapt similarly by others, and the setting does not allow for rigorous trial methods, they may not provide more value for decision-making. A randomized trial or research overall cannot answer all questions of interest.

Our findings suggest that primary care physicians perceive the issue of overdiagnosis and overtreatment as primarily driven by others rather than themselves. Patients, including their expectations, requests, and lack of knowledge, were seen as a major barrier to reducing low-value care. This potential externalization of responsibility may hinder de-implementation efforts. Nonetheless, future healthcare should also consider the perceived time constraints and clinical uncertainty as barriers to reducing low-value care use. Rather than increasing the time pressures through added responsibilities and time-consuming patient record systems, a clinician should have enough time to discuss with the patient and have well-organized senior support when facing uncertainty.

8 Acknowledgements

I did not think of becoming a researcher before I met my principal supervisor, Kari Tikkinen, in 2019. I took part in a Facebook conversation about evidence-based medicine. Kari contacted me and, after an exchange of a few words, asked if I was interested in joining his research projects. It's hard to say whether I would have taken a path anything like this had I not met Kari. I feel very fortunate and thankful that I met you, Kari. I can think of no better person, both professionally and personally, to guide me through this process. Thank you!

Next, I am also very thankful to Raija Sipilä for supervising my dissertation process and for guiding the start of my professional career in current care guidelines. Through several discussions and collaboration with Raija, I have learned a lot about evidence synthesis, work management, and teamwork in general. Currently, Raija is the head editor of the current care guidelines, and following her dedicated leadership feels like an excellent role model if I am to take that path in the future. Thank you!

Although Jorma had a somewhat smaller role supervising my dissertation, I am also very thankful for his guidance. When I came to the current care guidelines, I was often uncertain about taking the professional role of an editor and methodologist. Jorma's warm encouragement has meant a lot to me in those situations. I remember several discussions with Jorma where we have been arguing about different things in medicine. Jorma has always shown interest in hearing my thoughts and encouraged me to take my path. It has been truly valuable. Thank you!

I also want to thank the whole team that has collaborated with me on this project: Sameer Parpia, Jussi Mustonen, Robin Vernooij, Petra Falkenbach, Yoshitaka Aoki, Anton Barchuk, Marco Blanker, Rufus Cartwright, Kathryn Crowder, Herney Garcia-Perdomo, Rachel Gutschon, Alex Halme, Tuomas Kilpeläinen, Ilari Kuitunen, Tiina Lamberg, Eddy Lang, Jenifer Matos, Olli Nevalainen, Niko Nordlund, Negar Pourjamal, Eero Raittio, Patrick Richard, Philippe Violette, Benedetta Bellini, Eliana Ben-Sheleg, Moriah Ellen, Paolo Francesconi, Muuad Hussien, Eleni Karlafti, Shunzo Koizumi, Emir Ouahrani, Muna Paier-Abuzahra, Christos Sacopolous, and Ulrike Spary-Kainz. Also, many others have helped me along the way. I feel very fortunate to have been able to work with all of you! Thank you!

Furthermore, I would like to thank my pre-examiners, Docent Liina-Kaisa Tynkkynen and Associate Professor Sarah Alderson, for their valuable comments. They definitely improved the thesis. I also want to thank Professor Victor Montori for serving as my opponent. You are one of my biggest role models in medicine, and it is an honor to share this day with you.

Warm thanks also go to my family. Äiti ja Isä, uskon että kodista tulee iso osa siitä mitä olen tänä päivänä. Kiitos kaikesta tuesta ja huolenpidosta! Tuntuu hienolta nähdä miten olette ylpeitä siitä missä nyt olen.

Lastly, a very big thank you goes to my wife, Mirjam. Sharing the same experiences as a PhD researcher has been very valuable. You have been the first person to support me when I encountered any struggles with this project. Thank you! Life is beautiful with you!

9 References

1. Verkerk EW, Tanke MAC, Kool RB, van Dulmen SA, Westert GP. Limit, lean or listen? A typology of low-value care that gives direction in de-implementation. Int J Qual Health Care. 2018 Nov 1;30(9):736-739. doi: 10.1093/intqhc/mzy100. PMID: 29741672; PMCID: PMC6307334.
2. Ingvarsson S, Augustsson H, Hasson H, Nilsen P, von Thiele Schwarz U, von Knorring M. Why do they do it? A grounded theory study of the use of low-value care among primary health care physicians. Implement Sci. 2020;15(1):93.
3. Lyu H, Xu T, Brotman D, et al. Overtreatment in the United States. PLoS One. 2017;12(9):e0181970 Published 2017 Sep 6.
4. Norton WE, Chambers DA. Unpacking the complexities of de-implementing inappropriate health interventions. Implement Sci. 2020;15(1):2.
5. Hiscock H, Neely RJ, Warren H, Soon J, Georgiou A. Reducing Unnecessary Imaging and Pathology Tests: A Systematic Review. Pediatrics. 2018;141(2):e20172862. doi:10.1542/peds.2017-2862
6. Hu Y, Walley J, Chou R, et al. Interventions to reduce childhood antibiotic prescribing for upper respiratory infections: systematic review and meta-analysis. J Epidemiol Community Health. 2016;70(12):1162-1170. doi:10.1136/jech-2015-206543
7. Cliff BQ, Avanceña ALV, Hirth RA, Lee SD. The Impact of Choosing Wisely Interventions on Low-Value Medical Services: A Systematic Review. Milbank Q. 2021;99(4):1024-1058. doi:10.1111/1468-0009.12531
8. Colla CH, Mainor AJ, Hargreaves C, Sequist T, Morden N. Interventions Aimed at Reducing Use of Low-Value Health Services: A Systematic Review. Med Care Res Rev. 2017;74(5):507-550. doi:10.1177/1077558716656970
9. Heus P, van Dulmen SA, Weenink JW, Naaktgeboren CA, Takada T, Verkerk EW, Kamm I, van der Laan MJ, Hooft L, Kool RB. What are Effective Strategies to Reduce Low-Value Care? An Analysis of 121 Randomized Deimplementation Studies. J Healthc Qual. 2023 Sep-Oct 01;45(5):261-271. doi: 10.1097/JHQ.0000000000000392. Epub 2023 Jun 29.
10. Brodersen J, Schwartz LM, Heneghan C, O'Sullivan JW, Aronson JK, Woloshin S. Overdiagnosis: what it is and what it isn't. BMJ Evid Based Med. 2018;23(1):1-3. doi:10.1136/ebmed-2017-110886
11. Morgan DJ, Brownlee S, Leppin AL, et al. Setting a research agenda for medical overuse. BMJ. 2015;351:h4534. Published 2015 Aug 25. doi:10.1136/bmj.h4534
12. European Commission. (2025). Identifying, measuring and reducing low-value care in the context of health system performance assessment. Publications Office of the European Union. https://doi.org/10.2875/68275
13. Berwick DM, Hackbarth AD. Eliminating waste in US health care. JAMA. 2012;307(14):1513-1516. doi:10.1001/jama.2012.362
14. Shrank WH, Rogstad TL, Parekh N. Waste in the US Health Care System: Estimated Costs and Potential for Savings. JAMA. 2019;322(15):1501-1509. doi:10.1001/jama.2019.13978
15. Brownlee S, Chalkidou K, Doust J, Elshaug AG, Glasziou P, Heath I, Nagpal S, Saini V, Srivastava D, Chalmers K, Korenstein D. Evidence for overuse of

medical services around the world. Lancet. 2017 Jul 8;390(10090):156-168. doi: 10.1016/S0140-6736(16)32585-5. Epub 2017 Jan 9.
16. Platen M, Fleßa S, Rädke A, Wucherer D, Thyrian JR, Mohr W, Scharf A, Mühlichen F, Hoffmann W, Michalowsky B. Prevalence of Low-Value Care and Its Associations with Patient-Centered Outcomes in Dementia. J Alzheimers Dis. 2021;83(4):1775-1787. doi: 10.3233/JAD-210439.
17. Paulamäki J, Jyrkkä J, Hyttinen V, Jämsen E. Prevalence of potentially inappropriate medication use in older population: comparison of the Finnish Meds75+ database with eight published criteria. BMC Geriatr. 2023 Mar 10;23(1):139. doi: 10.1186/s12877-022-03706-z. PMID: 36899320; PMCID: PMC9999502.
18. Kool RB, Verkerk EW, Meijs J, van Gorp N, Maessen M, Westert G, Peul W, van Dulmen S. Assessing volume and variation of low-value care practices in the Netherlands. Eur J Public Health. 2020 Apr 1;30(2):236-240. doi: 10.1093/eurpub/ckz245. PMID: 31950995; PMCID: PMC7183367.
19. Chalmers K, Badgery-Parker T, Pearson SA, Brett J, Scott IA, Elshaug AG. Developing indicators for measuring low-value care: mapping Choosing Wisely recommendations to hospital data. BMC Res Notes. 2018 Mar 5;11(1):163. doi: 10.1186/s13104-018-3270-4. PMID: 29506573; PMCID: PMC5836437.
20. Shively NR, Buehrle DJ, Clancy CJ, Decker BK. Prevalence of Inappropriate Antibiotic Prescribing in Primary Care Clinics within a Veterans Affairs Health Care System. Antimicrob Agents Chemother. 2018 Jul 27;62(8):e00337-18. doi: 10.1128/AAC.00337-18. PMID: 29967028; PMCID: PMC6105840.
21. Chong, M.C., Sharp, M.K., Smith, S.M. et al. Strong recommendations from low certainty evidence: a cross-sectional analysis of a suite of national guidelines. BMC Med Res Methodol 23, 68 (2023). https://doi.org/10.1186/s12874-023-01895-8
22. Fanaroff AC, Califf RM, Windecker S, Smith SC Jr, Lopes RD. Levels of Evidence Supporting American College of Cardiology/American Heart Association and European Society of Cardiology Guidelines, 2008-2018. JAMA. 2019 Mar 19;321(11):1069-1080. doi: 10.1001/jama.2019.1122.
23. Herrera-Perez D, Haslam A, Crain T, Gill J, Livingston C, Kaestner V, Hayes M, Morgan D, Cifu AS, Prasad V. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. Elife. 2019 Jun 11;8:e45183. doi: 10.7554/eLife.45183. PMID: 31182188; PMCID: PMC6559784.
24. Batalden PB, Davidoff F. What is "quality improvement" and how can it transform healthcare? Qual Saf Health Care. 2007 Feb;16(1):2-3. doi: 10.1136/qshc.2006.022046. PMID: 17301192; PMCID: PMC2464920.
25. Gray M. Value based healthcare BMJ 2017; 356 :j437 doi:10.1136/bmj.j437
26. Patey, A.M., Grimshaw, J.M. & Francis, J.J. Changing behaviour, 'more or less': do implementation and de-implementation interventions include different behaviour change techniques?. Implementation Sci **16**, 20 (2021). https://doi.org/10.1186/s13012-021-01089-0
27. Grimshaw JM, Patey AM, Kirkham KR, et al. De-implementing wisely: developing the evidence base to reduce low-value care. BMJ Quality & Safety 2020;29:409-417.

28. Kool T, Patey AM, Dulmen S, Grimshaw JM. How to Reduce Overuse in Healthcare: A Practical Guide. Wiley Blackwell. 7 September 2023. DOI:10.1002/9781119862758
29. Atsma F, Elwyn G, Westert G. Understanding unwarranted variation in clinical practice: a focus on network effects, reflective medicine and learning health systems. Int J Qual Health Care. 2020 Jun 4;32(4):271-274. doi: 10.1093/intqhc/mzaa023.
30. Schünemann H, Brożek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. The GRADE Working Group, 2013. Available from guidelinedevelopment.org/handbook.
31. Venekamp RP, Sanders SL, Glasziou PP, Rovers MM. Antibiotics for acute otitis media in children. Cochrane Database of Systematic Reviews 2023, Issue 11. Art. No.: CD000219. DOI: 10.1002/14651858.CD000219.pub5.
32. Sacarny A, Barnett ML, Le J, Tetkoski F, Yokum D, Agrawal S. Effect of Peer Comparison Letters for High-Volume Primary Care Prescribers of Quetiapine in Older and Disabled Adults: A Randomized Clinical Trial. JAMA Psychiatry. 2018 Oct 1;75(10):1003-1011. doi: 10.1001/jamapsychiatry.2018.1867
33. Maratt JK, Kerr EA, Klamerus ML, Lohman SE, Froehlich W, Bhatia RS, Saini SD. Measures Used to Assess the Impact of Interventions to Reduce Low-Value Care: a Systematic Review. J Gen Intern Med. 2019 Sep;34(9):1857-1864. doi: 10.1007/s11606-019-05069-5. Epub 2019 Jun 27. PMID: 31250366; PMCID: PMC6712188.
34. Morgan DJ, Pineles L, Owczarzak J, et al. Clinician Conceptualization of the Benefits of Treatments for Individual Patients. JAMA Netw Open. 2021;4(7):e2119747. doi:10.1001/jamanetworkopen.2021.19747
35. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med. 2008 Jan 17;358(3):252-60. doi: 10.1056/NEJMsa065779. PMID: 18199864.
36. Driessen E, Hollon SD, Bockting CL, Cuijpers P, Turner EH. Does Publication Bias Inflate the Apparent Efficacy of Psychological Treatment for Major Depressive Disorder? A Systematic Review and Meta-Analysis of US National Institutes of Health-Funded Trials. PLoS One. 2015 Sep 30;10(9):e0137864. doi: 10.1371/journal.pone.0137864. PMID: 26422604; PMCID: PMC4589340.
37. Turner EH, Cipriani A, Furukawa TA, Salanti G, de Vries YA. Selective publication of antidepressant trials and its influence on apparent efficacy: Updated comparisons and meta-analyses of newer versus older trials. PLoS Med. 2022 Jan 19;19(1):e1003886. doi: 10.1371/journal.pmed.1003886. PMID: 35045113; PMCID: PMC8769343.
38. Johnston BC, Alonso-Coello P, Friedrich JO, Mustafa RA, Tikkinen KAO, Neumann I, Vandvik PO, Akl EA, da Costa BR, Adhikari NK, Dalmau GM, Kosunen E, Mustonen J, Crawford MW, Thabane L, Guyatt GH. Do clinicians understand the size of treatment effects? A randomized survey across 8 countries. CMAJ. 2016 Jan 5;188(1):25-32. doi: 10.1503/cmaj.150430. Epub 2015 Oct 26. PMID: 26504102; PMCID: PMC4695351.
39. Morgott M, Heinmüller S, Hueber S, Schedlbauer A, Kühlein T. Do guidelines help us to deviate from their recommendations when appropriate for the

individual patient? A systematic survey of clinical practice guidelines. J Eval Clin Pract. 2020 Jun;26(3):709-717. doi: 10.1111/jep.13187. Epub 2019 Jun 17. PMID: 31206241.
40. Raittio E, Raittio L. Statements considering intervention effects in Finnish clinical practice guidelines: Recommending interventions with non-numeric effect-sizes or unspecified outcomes. J Eval Clin Pract. 2021 Aug;27(4):751-758. doi: 10.1111/jep.13455. Epub 2020 Jul 31. PMID: 32735367.
41. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009 Jul 4;374(9683):86-9. doi: 10.1016/S0140-6736(09)60329-9. Epub 2009 Jun 12. PMID: 19525005.
42. Altman D G. The scandal of poor medical research BMJ 1994; 308:283 doi:10.1136/bmj.308.6924.283
43. Tikkinen KAO, Guyatt GH. Understanding of research results, evidence summaries and their applicability-not critical appraisal-are core skills of medical curriculum. BMJ Evid Based Med. 2021 Oct;26(5):231-233. doi: 10.1136/bmjebm-2020-111542. Epub 2021 Mar 18. PMID: 33737327; PMCID: PMC8479738.
44. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol. 2015 Sep;68(9):1046-58. doi: 10.1016/j.jclinepi.2015.05.029. Epub 2015 Jun 6. PMID: 26279400; PMCID: PMC4555355.
45. Young, S.S. and Karr, A. (2011), Deming, data and observational studies. Significance, 8: 116-120. https://doi.org/10.1111/j.1740-9713.2011.00506.x
46. Léonard C, Stordeur S, Roberfroid D. Association between physician density and health care consumption: a systematic review of the evidence. Health Policy. 2009 Jul;91(2):121-34. doi: 10.1016/j.healthpol.2008.11.013. Epub 2009 Jan 16. PMID: 19150579.
47. van Dijk, C.E., van den Berg, B., Verheij, R.A., Spreeuwenberg, P., Groenewegen, P.P. and de Bakker, D.H. (2013), MORAL HAZARD AND SUPPLIER-INDUCED DEMAND: EMPIRICAL EVIDENCE IN GENERAL PRACTICE†. Health Econ., 22: 340-352. https://doi.org/10.1002/hec.2801

48. Saini V, Garcia-Armesto S, Klemperer D, Paris V, Elshaug AG, Brownlee S, Ioannidis JPA, Fisher ES. Drivers of poor medical care. Lancet. 2017 Jul 8;390(10090):178-190. doi: 10.1016/S0140-6736(16)30947-3. Epub 2017 Jan 9. PMID: 28077235.
49. French SD, Green SE, O'Connor DA, McKenzie JE, Francis JJ, Michie S, Buchbinder R, Schattner P, Spike N, Grimshaw JM. Developing theory-informed behaviour change interventions to implement evidence into practice: a systematic approach using the Theoretical Domains Framework. Implement Sci. 2012 Apr 24;7:38. doi: 10.1186/1748-5908-7-38. PMID: 22531013; PMCID: PMC3443064.
50. Van Dulmen SA, Naaktgeboren CA, Heus P, et al. Barriers and facilitators to reduce low- value care: a qualitative evidence synthesis. BMJ Open 2020;10:e040025. doi:10.1136/ bmjopen-2020-040025
51. Kool RB, Verkerk EW, Winnemuller LJA, et al. Identifying and de- implementing low- value care in primary care: the GP's perspective—a cross- sectional survey. BMJ Open 2020;10:e037019. doi:10.1136/ bmjopen-2020-037019
52. Verkerk EW, van Dulmen SA, Westert GP. To do or not to do programme collaborators, et al. Reducing low-value care: what can we learn from eight de-implementation studies in the Netherlands? BMJ Open Quality 2022;11:e001710. doi: 10.1136/bmjoq-2021-001710
53. Gangathimmaiah V, Drever N, Evans R, et al. What works for and what hinders deimplementation of low-value care in emergency medicine practice? A scoping review. BMJ Open 2023;13:e072762. doi: 10.1136/bmjopen-2023-072762
54. Leigh, J.P., Sypes, E.E., Straus, S.E. et al. Determinants of the de-implementation of low-value care: a multi-method study. BMC Health Serv Res **22**, 450 (2022). https://doi.org/10.1186/s12913-022-07827-4
55. Augustsson H, Ingvarsson S, Nilsen P, von Thiele Schwarz U, Muli I, Dervish J, Hasson H. Determinants for the use and de-implementation of low-value care in health care: a scoping review. Implement Sci Commun. 2021 Feb 4;2(1):13. doi: 10.1186/s43058-021-00110-3. PMID: 33541443; PMCID: PMC7860215.
56. Pausch M, Schedlbauer A, Weiss M, Kuehlein T, Hueber S. Is it really always only the others who are to blame? GP's view on medical overuse. A questionnaire study. PLoS ONE. 2020;15(1):e0227457. doi: 10.1371/journal.pone.0227457
57. Wilson AD, Childs S, Gonçalves-Bradley DC, Irving GJ. Interventions to increase or decrease the length of primary care physicians' consultation. Cochrane Database of Systematic Reviews 2016, Issue 8. Art. No.: CD003540. DOI: 10.1002/14651858.CD003540.pub3. Accessed 27 May 2024.
58. Neprash HT, Mulcahy JF, Cross DA, Gaugler JE, Golberstein E, Ganguli I. Association of Primary Care Visit Length With Potentially Inappropriate Prescribing. JAMA Health Forum. 2023;4(3):e230052. doi:10.1001/jamahealthforum.2023.0052
59. Nilsen, P., Seing, I., Ericsson, C. et al. Characteristics of successful changes in health care organizations: an interview study with physicians, registered nurses and assistant nurses. BMC Health Serv Res **20**, 147 (2020). https://doi.org/10.1186/s12913-020-4999-8

60. van Bodegom-Vos L, Davidoff F, Marang-van de Mheen PJ. Implementation and de-implementation: two sides of the same coin?. BMJ Quality & Safety 2017;26:495-501.
61. Nilsen, P. Making sense of implementation theories, models and frameworks. Implementation Sci 10, 53 (2015). https://doi.org/10.1186/s13012-015-0242-0
62. Powell, B.J., Waltz, T.J., Chinman, M.J. et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implementation Sci 10, 21 (2015). https://doi.org/10.1186/s13012-015-0209-1
63. Mazza D, Bairstow P, Buchan H, Chakraborty SP, Van Hecke O, Grech C, Kunnamo I. Refining a taxonomy for guideline implementation: results of an exercise in abstract classification. Implement Sci. 2013 Mar 15;8:32. doi: 10.1186/1748-5908-8-32. PMID: 23497520; PMCID: PMC3606141.
64. Powell, B.J., Waltz, T.J., Chinman, M.J. et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implementation Sci 10, 21 (2015). https://doi.org/10.1186/s13012-015-0209-1
65. Ingvarsson, S., Hasson, H., von Thiele Schwarz, U. et al. Strategies for de-implementation of low-value care—a scoping review. Implementation Sci **17**, 73 (2022). https://doi.org/10.1186/s13012-022-01247-y
66. Henderson J, Bouck Z, Holleman R, et al. Comparison of Payment Changes and Choosing Wisely Recommendations for Use of Low-Value Laboratory Tests in the United States and Canada. JAMA Intern Med. 2020;180(4):524–531. doi:10.1001/jamainternmed.2019.7143
67. Giguère A, Zomahoun HT, Carmichael P-H, Uwizeye CB, Légaré F, Grimshaw JM, Gagnon M-P, Auguste DU, Massougbodji J. Printed educational materials: effects on professional practice and healthcare outcomes. Cochrane Database of Systematic Reviews 2020, Issue 8. Art. No.: CD004398. DOI: 10.1002/14651858.CD004398.pub4. Accessed 27 May 2024.
68. Forsetlund L, O'Brien MA, Forsén L, Mwai L, Reinar LM, Okwen MP, Horsley T, Rose CJ. Continuing education meetings and workshops: effects on professional practice and healthcare outcomes. Cochrane Database of Systematic Reviews 2021, Issue 9. Art. No.: CD003030. DOI: 10.1002/14651858.CD003030.pub3. Accessed 27 May 2024.
69. Kwan J L, Lo L, Ferguson J, Goldberg H, Diaz-Martinez J P, Tomlinson G et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials BMJ 2020; 370 :m3216 doi:10.1136/bmj.m3216
70. Sypes, E.E., de Grood, C., Whalen-Browne, L. et al. Engaging patients in de-implementation interventions to reduce low-value clinical care: a systematic review and meta-analysis. BMC Med **18**, 116 (2020). https://doi.org/10.1186/s12916-020-01567-0
71. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, O'Brien MA, Johansen M, Grimshaw J, Oxman AD. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database of Systematic Reviews 2012, Issue 6. Art. No.: CD000259. DOI: 10.1002/14651858.CD000259.pub3

72. Zeng Y, Shi L, Liu C, Li W, Li J, Yang S, Yang X, Huang Q, Yang L. Effects of social norm feedback on antibiotic prescribing and its characteristics in behaviour change techniques: a mixed-methods systematic review. Lancet Infect Dis. 2023 May;23(5):e175-e184. doi: 10.1016/S1473-3099(22)00720-4
73. Flodgren G, O'Brien MA, Parmelli E, Grimshaw JM. Local opinion leaders: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2019 Jun 24;6(6):CD000125. doi: 10.1002/14651858.CD000125.pub5.
74. Chaudhuri D, Montgomery A, Gulenchyn K, Mitchell M, Joseph P. Effectiveness of Quality Improvement Interventions at Reducing Inappropriate Cardiac Imaging: A Systematic Review and Meta-Analysis. Circ Cardiovasc Qual Outcomes. 2016;9(1):7-13. doi:10.1161/CIRCOUTCOMES.115.001836
75. Clyne B, Fitzgerald C, Quinlan A, et al. Interventions to Address Potentially Inappropriate Prescribing in Community-Dwelling Older Adults: A Systematic Review of Randomized Controlled Trials. J Am Geriatr Soc. 2016;64(6):1210-1222. doi:10.1111/jgs.14133
76. de Bont EG, Alink M, Falkenberg FC, Dinant GJ, Cals JW. Patient information leaflets to reduce antibiotic use and reconsultation rates in general practice: a systematic review. BMJ Open. 2015;5(6):e007612. Published 2015 Jun 3. doi:10.1136/bmjopen-2015-007612
77. Desai S, Liu C, Kirkland SW, Krebs LD, Keto-Lambert D, Rowe BH. Effectiveness of Implementing Evidence-based Interventions to Reduce C-spine Image Ordering in the Emergency Department: A Systematic Review. Acad Emerg Med. 2018;25(6):672-683. doi:10.1111/acem.13364
78. Fleming A, Browne J, Byrne S. The effect of interventions to reduce potentially inappropriate antibiotic prescribing in long-term care facilities: a systematic review of randomised controlled trials. Drugs Aging. 2013;30(6):401-408. doi:10.1007/s40266-013-0066-z
79. hen I, Opiyo N, Tavender E, et al. Non-clinical interventions for reducing unnecessary caesarean section. Cochrane Database Syst Rev. 2018;9(9):CD005528. Published 2018 Sep 28. doi:10.1002/14651858.CD005528.pub3
80. Thomas RE, Croal BL, Ramsay C, Eccles M, Grimshaw J. Effect of enhanced feedback and brief educational reminder messages on laboratory test requesting in primary care: a cluster randomised trial. Lancet. 2006 Jun 17;367(9527):1990-6. doi: 10.1016/S0140-6736(06)68888-0.
81. Urbiztondo I, Bjerrum L, Caballero L, Suarez MA, Olinisky M, Córdoba G. Decreasing Inappropriate Use of Antibiotics in Primary Care in Four Countries in South America-Cluster Randomized Controlled Trial. Antibiotics (Basel). 2017 Dec 14;6(4):38. doi: 10.3390/antibiotics6040038
82. Behavioural Economics Team of the Australian Goverment (BETA) . Nudge vs Superbugs: A behavioural economics trial to reduce the overprescribing of antibiotics. Canberra: Department of the Prime Minister and Cabinet, 2018.
83. Das J, Chowdhury A, Hussam R, Banerjee AV. The impact of training informal health care providers in India: A randomized controlled trial. Science (New York, N.Y.). 2016 Oct;354(6308):aaf7384. DOI: 10.1126/science.aaf7384
84. Zwar N, Wolk J, Gordon J, Sanson-Fisher R, Kehoe L. Influencing antibiotic prescribing in general practice: a trial of prescriber feedback and management

guidelines. Fam Pract. 1999 Oct;16(5):495-500. doi: 10.1093/fampra/16.5.495. PMID: 10533946.
85. Drug and Therapeutics Bulletin. An introduction to patient decision aids. BMJ. 2013 Jul 23;347:f4147. doi: 10.1136/bmj.f4147. PMID: 23881944.
86. Lee MHM, Pan DST, Huang JH, Chen MI, Chong JWC, Goh EH, Jiang L, Leo YS, Lee TH, Wong CS, Loh VWK, Lim FS, Poh AZ, Tham TY, Wong WM, Yu Y. Results from a Patient-Based Health Education Intervention in Reducing Antibiotic Use for Acute Upper Respiratory Tract Infections in the Private Sector Primary Care Setting in Singapore. Antimicrob Agents Chemother. 2017 Apr 24;61(5):e02257-16. doi: 10.1128/AAC.02257-16.
87. Macfarlane J, Holmes W, Gard P, Thornhill D, Macfarlane R, Hubbard R. Reducing antibiotic use for acute bronchitis in primary care: blinded, randomised controlled trial of patient information leaflet. BMJ. 2002 Jan 12;324(7329):91-4. doi: 10.1136/bmj.324.7329.91.
88. Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Penguin Books.
89. Yoong, S.L., Hall, A., Stacey, F. et al. Nudge strategies to improve healthcare providers' implementation of evidence-based guidelines, policies and practices: a systematic review of trials included within Cochrane systematic reviews. Implementation Sci **15**, 50 (2020). https://doi.org/10.1186/s13012-020-01011-0
90. Raban MZ, Gonzalez G, Nguyen AD, Newell BR, Li L, Seaman KL, Westbrook JI. Nudge interventions to reduce unnecessary antibiotic prescribing in primary care: a systematic review. BMJ Open. 2023 Jan 18;13(1):e062688. doi: 10.1136/bmjopen-2022-062688. PMID: 36657758; PMCID: PMC9853249.
91. Toews I, Anglemyer A, Nyirenda JL, Alsaid D, Balduzzi S, Grummich K, Schwingshackl L, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials: a meta-epidemiological study. Cochrane Database Syst Rev. 2024 Jan 4;1(1):MR000034. doi: 10.1002/14651858.MR000034.pub3. PMID: 38174786; PMCID: PMC10765475.
92. Soni PD, Hartman HE, Dess RT, Abugharib A, Allen SG, Feng FY, Zietman AL, Jagsi R, Schipper MJ, Spratt DE. Comparison of Population-Based Observational Studies With Randomized Trials in Oncology. J Clin Oncol. 2019 May 10;37(14):1209-1216. doi: 10.1200/JCO.18.01074. Epub 2019 Mar 21. PMID: 30897037; PMCID: PMC7186578.
93. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? BMJ. 2001 Feb 10;322(7282):355-7. doi: 10.1136/bmj.322.7282.355. PMID: 11159665; PMCID: PMC1119583.
94. Rhoads, C. H. (2011). The Implications of "Contamination" for Experimental Design in Education. Journal of Educational and Behavioral Statistics, 36(1), 76-104. https://doi.org/10.3102/1076998610379133
95. Karla Hemming, Monica Taljaard, Key considerations for designing, conducting and analysing a cluster randomized trial, International Journal of Epidemiology, Volume 52, Issue 5, October 2023, Pages 1648–1658, https://doi.org/10.1093/ije/dyad064
96. Clémence Leyrat, Katy E Morgan, Baptiste Leurent, Brennan C Kahan, Cluster randomized trials with a small number of clusters: which analyses should be

used?, International Journal of Epidemiology, Volume 47, Issue 1, February 2018, Pages 321–331, https://doi.org/10.1093/ije/dyx169
97. Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023). Cochrane, 2023. Available from www.training.cochrane.org/handbook.
98. Higgins JPT, Savović J, Page MJ, Elbers RG, Sterne JAC. Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023). Cochrane, 2023. Available from www.training.cochrane.org/handbook.
99. Dettori J. The random allocation process: two things you need to know. Evid Based Spine Care J. 2010 Dec;1(3):7-9. doi: 10.1055/s-0030-1267062. PMID: 22956922; PMCID: PMC3427961.
100. Sandra Eldridge, Marion K Campbell, Michael J Campbell, Amy K Drahota, Bruno Giraudeau, Barnaby C Reeves, Nandi Siegfried, Julian PT Higgins. Revised Cochrane risk of bias tool for randomized trials (RoB 2) Additional considerations for cluster-randomized trials (RoB 2 CRT). https://www.riskofbias.info/welcome/rob-2-0-tool/rob-2-for-cluster-randomized-trials
101. Catalogue of Bias Collaboration. Banerjee A, Pluddemann A, O'Sullivan J, Nunan D. Performance bias. In: Catalogue Of Bias 2019. https://catalogofbias/biases/performance-bias/
102. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng H-Y, Corbett MS, Eldridge SM, Hernán MA, Hopewell S, Hróbjartsson A, Junqueira DR, Jüni P, Kirkham JJ, Lasserson T, Li T, McAleenan A, Reeves BC, Shepperd S, Shrier I, Stewart LA, Tilling K, White IR, Whiting PF, Higgins JPT. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019; **366**: l4898.
103. Schünemann H, Brożek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. The GRADE Working Group, 2013. Available from guidelinedevelopment.org/handbook.
104. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G, Akl EA, Post PN, Norris S, Meerpohl J, Shukla VK, Nasser M, Schünemann HJ; GRADE Working Group. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011 Dec;64(12):1303-10. doi: 10.1016/j.jclinepi.2011.04.014. Epub 2011 Jul 30. PMID: 21802903.
105. Guyatt G, Zhao Y, Mayer M, Briel M, Mustafa R, Izcovich A, Hultcrantz M, Iorio A, Alba AC, Foroutan F, Sun X, Schunemann H, DeBeer H, Akl EA, Christensen R, Schandelmaier S. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency. J Clin Epidemiol. 2023 Jun;158:70-83. doi: 10.1016/j.jclinepi.2023.03.003. Epub 2023 Mar 9. PMID: 36898507
106. Zeng L, Brignardello-Petersen R, Hultcrantz M, Mustafa RA, Murad MH, Iorio A, Traversy G, Akl EA, Mayer M, Schünemann HJ, Guyatt GH. GRADE Guidance

34: update on rating imprecision using a minimally contextualized approach. J Clin Epidemiol. 2022 Oct;150:216-224. doi: 10.1016/j.jclinepi.2022.07.014. Epub 2022 Aug 4. PMID: 35934265.
107. Schünemann HJ, Neumann I, Hultcrantz M, Brignardello-Petersen R, Zeng L, Murad MH, Izcovich A, Morgano GP, Baldeh T, Santesso N, Cuello CG, Mbuagbaw L, Guyatt G, Wiercioch W, Piggott T, De Beer H, Vinceti M, Mathioudakis AG, Mayer MG, Mustafa R, Filippini T, Iorio A, Nieuwlaat R, Marcucci M, Coello PA, Bonovas S, Piovani D, Tomlinson G, Akl EA; GRADE Working Group. GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions. J Clin Epidemiol. 2022 Oct;150:225-242. doi: 10.1016/j.jclinepi.2022.07.015. Epub 2022 Aug 5. PMID: 35934266.
108. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams JW Jr, Meerpohl J, Norris SL, Akl EA, Schünemann HJ. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011 Dec;64(12):1277-82. doi: 10.1016/j.jclinepi.2011.01.011. Epub 2011 Jul 30. PMID: 21802904
109. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D, Kunz R, Brozek J, Montori V, Jaeschke R, Rind D, Dahm P, Meerpohl J, Vist G, Berliner E, Norris S, Falck-Ytter Y, Murad MH, Schünemann HJ; GRADE Working Group. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011 Dec;64(12):1311-6. doi: 10.1016/j.jclinepi.2011.06.004. Epub 2011 Jul 30. PMID: 21802902.
110. Murad MH, Verbeek J, Schwingshackl L, Filippini T, Vinceti M, Akl EA, Morgan RL, Mustafa RA, Zeraatkar D, Senerth E, Street R, Lin L, Falck-Ytter Y, Guyatt G, Schünemann HJ; GRADE Working Group. GRADE GUIDANCE 38: Updated guidance for rating up certainty of evidence due to a dose-response gradient. J Clin Epidemiol. 2023 Dec;164:45-53. doi: 10.1016/j.jclinepi.2023.09.011
111. Movsisyan A, Melendez-Torres GJ, Montgomery P. Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared with those of simple interventions. J Clin Epidemiol. 2016 Oct;78:22-33. doi: 10.1016/j.jclinepi.2016.03.014. Epub 2016 Mar 30. PMID: 27038850.
112. Fleming PS, Koletsi D, Ioannidis JP, Pandis N. High quality of the evidence for medical and other health-related interventions was uncommon in Cochrane systematic reviews. J Clin Epidemiol. 2016 Oct;78:34-42. doi: 10.1016/j.jclinepi.2016.03.012. Epub 2016 Mar 29. PMID: 27032875.
113. Murad MH, Almasri J, Alsawas M, Farah W. Grading the quality of evidence in complex interventions: a guide for evidence-based practitioners. Evid Based Med. 2017 Mar;22(1):20-22. doi: 10.1136/ebmed-2016-110577. Epub 2016 Dec 8. PMID: 27932400.
114. Movsisyan A, Melendez-Torres GJ, Montgomery P. Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. J Clin Epidemiol. 2016 Feb;70:191-9. doi: 10.1016/j.jclinepi.2015.09.010. Epub 2015 Sep 18. PMID: 26388040.
115. Montgomery P, Movsisyan A, Grant SP, Macdonald G, Rehfuess EA. Considerations of complexity in rating certainty of evidence in systematic reviews: a primer on using the GRADE approach in global health. BMJ Glob

Health. 2019 Jan 25;4(Suppl 1):e000848. doi: 10.1136/bmjgh-2018-000848. PMID: 30775013; PMCID: PMC6350753.
116. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr;64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015. Epub 2011 Jan 5. PMID: 21208779.
117. He J, Morales DR, Guthrie B. Exclusion rates in randomized controlled trials of treatments for physical conditions: a systematic review. Trials. 2020 Feb 26;21(1):228. doi: 10.1186/s13063-020-4139-0. PMID: 32102686; PMCID: PMC7045589.
118. Papakostidis C, Giannoudis PV. Meta-analysis. What have we learned? Injury. 2023 May;54 Suppl 3:S30-S34. doi: 10.1016/j.injury.2022.06.012. Epub 2022 Jun 12. PMID: 35753817.
119. Halme ALE, McAlpine K, Martini A. Fixed-effect Versus Random-effects Models for Meta-analyses: Random-effects Models. Eur Urol Focus. 2023 Sep;9(5):693-694. doi: 10.1016/j.euf.2023.10.023. Epub 2023 Nov 7. PMID: 37945406.
120. Tufanaru, Catalin MD, MPH, MClinSci (EBHC); Munn, Zachary PhD; Stephenson, Matthew PhD; Aromataris, Edoardo PhD. Fixed or random effects meta-analysis? Common methodological issues in systematic reviews of effectiveness. International Journal of Evidence-Based Healthcare 13(3):p 196-207, September 2015. | DOI: 10.1097/XEB.0000000000000065
121. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986 Sep;7(3):177-88. doi: 10.1016/0197-2456(86)90046-2. PMID: 3802833.
122. IntHout, J., Ioannidis, J.P. & Borm, G.F. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol **14**, 25 (2014). https://doi.org/10.1186/1471-2288-14-25
123. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Moher D. Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. J Clin Epidemiol. 2021;134:103-112. PMID: 33577987
124. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467–73.
125. Niven, D.J., Mrklas, K.J., Holodinsky, J.K. et al. Towards understanding the de-adoption of low-value clinical practices: a scoping review. BMC Med 13, 255 (2015). https://doi.org/10.1186/s12916-015-0488-z
126. Köchling A, Löffler C, Reinsch S, et al. Reduction of antibiotic prescriptions for acute respiratory tract infections in primary care: a systematic review. Implement Sci. 2018;13(1):47.
127. Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023). Cochrane, 2023. Available from www.training.cochrane.org/handbook
128. Hoffmann T C, Glasziou P P, Boutron I, Milne R, Perera R, Moher D et al. Better reporting of interventions: template for intervention description and

replication (TIDieR) checklist and
guide BMJ 2014; 348 :g1687 doi:10.1136/bmj.g1687
129. Yardley S, Dornan T. Kirkpatrick's levels and education 'evidence'. Med Educ. 2012;46(1):97–106.
130. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence–study limitations (risk of bias). J Clin Epidemiol. 2011;64(4):407–15.
131. Cochrane Effective Practice and Organisation of Care (EPOC). Suggested risk of bias criteria for EPOC reviews. In: EPOC resources for review authors; 2017. Epoc.cochrane.org/resources/epoc-resources-review-authors.
132. Zwolsman S, te Pas E, Hooft L, Wieringa-de Waard M, van Dijk N. Barriers to GPs' use of evidence-based medicine: a systematic review. Br J Gen Pract. 2012 Jul;62(600):e511-21. doi: 10.3399/bjgp12X652382. PMID: 22781999; PMCID: PMC3381277.
133. Chien AT, Lehmann LS, Hatfield LA, Koplan KE, Petty CR, Sinaiko AD, Rosenthal MB, Sequist TD. A Randomized Trial of Displaying Paid Price Information on Imaging Study and Procedure Ordering Rates. J Gen Intern Med. 2017 Apr;32(4):434-448. doi: 10.1007/s11606-016-3917-6. Epub 2016 Dec 2. PMID: 27913910; PMCID: PMC5377881.
134. Tierney WM, Miller ME, McDonald CJ. The effect on test ordering of informing physicians of the charges for outpatient diagnostic tests. N Engl J Med. 1990 May 24;322(21):1499-504. doi: 10.1056/NEJM199005243222105. PMID: 2186274
135. Martins CM, da Costa Teixeira AS, de Azevedo LF, Sá LM, Santos PA, do Couto ML, da Costa Pereira AM, Hespanhol AA, da Costa Santos CM. The effect of a test ordering software intervention on the prescription of unnecessary laboratory tests - a randomized controlled trial. BMC Med Inform Decis Mak. 2017 Feb 20;17(1):20. doi: 10.1186/s12911-017-0416-6. PMID: 28219437; PMCID: PMC5319139.
136. Sacarny A, Barnett ML, Le J, Tetkoski F, Yokum D, Agrawal S. Effect of Peer Comparison Letters for High-Volume Primary Care Prescribers of Quetiapine in Older and Disabled Adults: A Randomized Clinical Trial. JAMA Psychiatry. 2018 Oct 1;75(10):1003-1011. doi: 10.1001/jamapsychiatry.2018.1867
137. Candy B, Vickerstaff V, Jones L, et al. Description of complex interventions: analysis of changes in reporting in randomised trials since 2002. Trials. 2018;**19**:110. doi: 10.1186/s13063-018-2503-0
138. Negrini S, Arienti C, Pollet J, et al. Clinical replicability of rehabilitation interventions in randomized controlled trials reported in main journals is inadequate. J Clin Epidemiol. 2019;**114**:108–117. doi: 10.1016/j.jclinepi.2019.06.008
139. Kahan BC, Forbes G, Ali Y, et al. Increased risk of type I errors in cluster randomized trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. Trials. 2016;**17**(1):438. doi: 10.1186/s13063-016-1571-2
140. Horwitz LI, Kuznetsova M, Jones SA. Creating a Learning Health System through Rapid-Cycle, Randomized Testing. N Engl J Med. 2019 Sep 19;381(12):1175-1179. doi: 10.1056/NEJMsb1900856. PMID: 31532967.

141. Carney G, Maclure M, Patrick DM, Fisher A, Stanley D, Bassett K, Dormuth CR. A cluster randomized trial assessing the impact of personalized prescribing feedback on antibiotic prescribing for uncomplicated acute cystitis to family physicians. PLoS One. 2023 Jul 31;18(7):e0280096. doi: 10.1371/journal.pone.0280096. PMID: 37523381; PMCID: PMC10389722.
142. Macfarlane J, Holmes W, Gard P, Thornhill D, Macfarlane R, Hubbard R. Reducing antibiotic use for acute bronchitis in primary care: blinded, randomised controlled trial of patient information leaflet. BMJ. 2002 Jan 12;324(7329):91-4. doi: 10.1136/bmj.324.7329.91
143. Meeker D, Knight TK, Friedberg MW, Linder JA, Goldstein NJ, Fox CR, Rothfeld A, Diaz G, Doctor JN. Nudging guideline-concordant antibiotic prescribing: a randomized clinical trial. JAMA Intern Med. 2014 Mar;174(3):425-31. doi: 10.1001/jamainternmed.2013.14191
144. Little P, Stuart B, Francis N, Douglas E, Tonkin-Crine S, Anthierens S, Cals JW, Melbye H, Santer M, Moore M, Coenen S, Butler C, Hood K, Kelly M, Godycki-Cwirko M, Mierzecki A, Torres A, Llor C, Davies M, Mullee M, O'Reilly G, van der Velden A, Geraghty AW, Goossens H, Verheij T, Yardley L; GRACE consortium. Effects of internet-based training on antibiotic prescribing rates for acute respiratory-tract infections: a multinational, cluster, randomised, factorial, controlled trial. Lancet. 2013 Oct 5;382(9899):1175-82. doi: 10.1016/S0140-6736(13)60994-0
145. Altiner A, Brockmann S, Sielk M, Wilm S, Wegscheider K, Abholz HH. Reducing antibiotic prescriptions for acute cough by motivating GPs to change their attitudes to communication and empowering patients: a cluster-randomized intervention study. J Antimicrob Chemother. 2007 Sep;60(3):638-44. doi: 10.1093/jac/dkm254
146. Simula AS, Jenkins HJ, Hancock MJ, Malmivaara A, Booth N, Karppinen J. Patient education booklet to support evidence-based low back pain care in primary care - a cluster randomized controlled trial. BMC Fam Pract. 2021 Sep 7;22(1):178. doi: 10.1186/s12875-021-01529-2
147. Rampersaud YR, Canizares M, Zywiel MG, Leroux T, Gandhi R, Veillette C, Marshall W, Ogilvie-Harris D, Cram P, Coyte P, Mohamed N. Evaluation of Trends in Knee Arthroscopy from 2004 to 2019 in Ontario, Canada. NEJM Evid. 2022 May;1(5):EVIDoa2100036. doi: 10.1056/EVIDoa2100036. Epub 2022 Apr 26. PMID: 38319226.
148. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med. 2011 Dec;104(12):510-20. doi: 10.1258/jrsm.2011.110180. PMID: 22179294; PMCID: PMC3241518.
149. Cross, K.P. Not can, but will college teaching be improved?. New Directions for Higher Education, 1977: 1-15. https://doi.org/10.1002/he.36919771703
150. Svenson O. Are we more risky and more skillful than our fellow drivers? Acta Psychologica. 1981;47:143–148.
151. Pronin, E., Lin, D. Y., & Ross, L. (2002). The Bias Blind Spot: Perceptions of Bias in Self Versus Others. Personality and Social Psychology Bulletin, 28(3), 369–381. https://doi.org/10.1177/0146167202286008

10 Original publications