KU LEUVEN

FACULTY OF MEDICINE

CENTER FOR BIOMEDICAL ETHICS AND LAW

Mystery Shopping in Primary Care: A Systematic Review

Master's thesis proposed to achieve the Master of Family Medicine by

PETERS Stijn

VERPLANCKE Tibo

Faculty of Medicine

Center for Biomedical Ethics and Law

Promotor: Prof. Dr. BORRY Pascal

Supervisor: Dr. VERBEKE Kamiel

Leuven, 2022 – 2025

©Copyright by KU Leuven

Deze masterproef is een examendocument dat niet werd gecorrigeerd voor eventueel vastgestelde fouten. Met uitzondering van het door de wet toegestaan gebruik, is zonder voorafgaande schriftelijke toestemming van de auteur(s) het overnemen, kopiëren, gebruiken of realiseren van deze uitgave of gedeelten ervan verboden. Voor aanvragen tot of informatie in verband met het overnemen en/of gebruik van gedeelten uit deze publicatie, wendt u zich tot de KU Leuven, Faculteit Geneeskunde, Campus Gasthuisberg O&N2, Herestraat 49 – bus 400, 3000 Leuven.

Voorafgaande schriftelijke toestemming van de promotor(en) is eveneens vereist voor het aanwenden van de in dit afstudeerwerk beschreven (originele) methoden, producten en programma's voor commercieel nut en voor de inzending van deze publicatie ter deelname aan wetenschappelijke prijzen of wedstrijden.

KU LEUVEN

FACULTY OF MEDICINE

CENTER FOR BIOMEDICAL ETHICS AND LAW

Mystery Shopping in Primary Care: A Systematic Review

Master's thesis proposed to achieve the Master of Family Medicine by

PETERS Stijn

VERPLANCKE Tibo

Faculty of Medicine

Center for Biomedical Ethics and Law

Promotor: Prof. Dr. BORRY Pascal

Supervisor: Dr. VERBEKE Kamiel

Leuven, 2022 – 2025

KU LEUVEN

FACULTY OF MEDICINE

COVER LETTER

Stijn Peters & Tibo Verplancke
Center for Biomedical Ethics and Law, KU Leuven
Kapucijnenvoer 7 Box 7001, 3000 Leuven
tibo.verplancke@student.kuleuven.be
stijn.peters@student.kuleuven.be

Leuven, 01.05.2025

Dear Editor or Reviewer,

Please find hereby attached the manuscript for our systematic review entitled “Mystery Shopping in Primary Care: A Systematic Review”, which we hope you will find valuable and consider for publication.

To address the growing need for robust and ethically sound quality assessment tools in primary care, our review synthesizes empirical studies on the use of unannounced standardized patients (mystery shoppers) in general practice. We focus on both the methodological strategies employed and the ethical considerations reported. Our aim is to offer practical insights that can inform future research and policy decisions in quality improvement, and to explore whether mystery shopping may be a useful educational tool in Belgian general practice.

We believe our manuscript makes a meaningful contribution to the literature by highlighting how mystery shopping is currently implemented in primary care, identifying recurring methodological patterns, and evaluating how ethical challenges are managed across studies. With regard to methodological patterns, we looked at portrayed clinical scenarios, training, data collection, and whether or how detection by healthcare providers was assessed. We also critically examined how ethical challenges such as consent, deception, and debriefing were addressed.

There has been no duplicate publication or submission elsewhere, all authors have read and approved the manuscript, there are no ethical problems or conflicts of interest.

We would like to thank you in advance, Editor or Reviewer, for your time and consideration of this manuscript.

Yours sincerely,

Stijn Peters and Tibo Verplancke
3rd year Advanced Master’s Degree in General Medicine
Faculty of Medicine
KU Leuven

TABLE OF CONTENTS

Title page III

Cover letter IV

Table of contents V

Abstract 1

Nederlandstalige samenvatting 2

Introduction 3

Methods 4

Results 6

Discussion 15

Conclusion 19

Acknowledgments 20

References 22

Appendix I. **Detailed search strategy** i

Appendix II. **Research Questions** ii

Appendix III. **Overview of included articles (N = 43)** iii

Appendix IV. **Systematic Data Extraction sheet** vi

Appendix V. **Ethical approval and Safeguards** vii

Appendix VI. **Informed consent** xii

ABSTRACT

Background

Mystery shopping using unannounced simulated patients (USPs) has emerged as a research method for observing clinical behavior in routine practice. In primary care, it is increasingly studied as a model for identifying opportunities to improve care quality. Although previous reviews have examined USP use broadly, a focused synthesis of how methodological rigor and ethical safeguards are implemented in primary care, and how these findings can inform quality improvement in general practice, remains limited and fragmented.

Objectives

This systematic review examined empirical studies on the use of USPs in primary care, focusing on study design, ethical safeguards, and contributions to quality improvement. Our aim was to inform policymakers and educators about the feasibility of mystery shopping as a tool for enhancing care quality. We conclude by proposing considerations for its educational implementation in Belgian general practice.

Methods

This preregistered systematic review was conducted following the PRISMA statement. We searched PubMed and Web of Science for empirical studies from the past 20 years, in English or Dutch, using USPs to assess primary care providers (PCPs) in real clinical settings. Two reviewers screened records independently. A supplementary Google Scholar search and snowballing were used to identify additional relevant studies. Data extraction focused on methodological and ethical elements, followed by a primarily deductive qualitative synthesis.

Results

Out of 155 unique records, 43 studies from diverse primary care settings were included. Relevant findings included low detection rates, inconsistent reporting on training, and limited integration of USPs into learning programs. Seventy-two percent of studies reported obtaining informed consent, while structured feedback and debriefing were rare.

Conclusion

Mystery shopping holds promise for improving care quality in general practice. Our review revealed wide methodological variation and ethical complexity, underscoring the need to integrate ethical safeguards and adapt study designs to local systems. These findings can guide policymakers and support future USP research and educational models. Further research should explore physician experiences and pilot voluntary feedback initiatives to promote improvement.

Keywords: Mystery shopping, Deception, Quality of care, Medical education, Unannounced standardized patients, Systematic review, Research methodology, Research ethics

NEDERLANDSTALIGE SAMENVATTING

Achtergrond

Mystery shopping met onaangekondigde gesimuleerde patiënten (USP's) is in opkomst als onderzoeksmethode om klinisch gedrag in de dagelijkse praktijk te observeren. In de eerstelijnszorg wordt het steeds vaker bestudeerd als een model om kansen voor kwaliteitsverbetering te identificeren. Hoewel eerdere reviews het gebruik van USP's breed hebben onderzocht, ontbreekt een gerichte synthese over de toepassing van methodologische zorgvuldigheid en ethische voorzorgsmaatregelen in de eerstelijnszorg. Ook is het onduidelijk hoe deze bevindingen kunnen bijdragen aan kwaliteitsverbetering in de huisartsgeneeskunde.

Doelstellingen

Deze systematische review onderzocht empirische studies over het gebruik van USP's in de eerstelijnszorg, met aandacht voor studieopzet, ethische beschermingsmechanismen en bijdragen aan kwaliteitsverbetering. Ons doel was om beleidsmakers en opleiders te informeren over de haalbaarheid van mystery shopping als instrument om de zorgkwaliteit te verbeteren. We sluiten af met enkele overwegingen voor de educatieve implementatie in de Belgische huisartsgeneeskunde.

Methodes

Deze vooraf geregistreerde systematische review werd uitgevoerd volgens de PRISMA-richtlijnen. We doorzochten PubMed en Web of Science naar empirische studies van de afgelopen 20 jaar, in het Engels of Nederlands, waarbij USP's werden gebruikt om eerstelijnszorgverleners (PCP's) in reële klinische settingen te beoordelen. Twee onderzoekers selecteerden de artikels onafhankelijk van elkaar. Een aanvullende zoekactie via Google Scholar en snowballsampling werd gebruikt om extra relevante studies te identificeren. De gegevensextractie richtte zich op methodologische en ethische aspecten, gevolgd door een voornamelijk deductieve kwalitatieve synthese.

Resultaten

Van de 155 unieke artikelen werden 43 studies uit uiteenlopende eerstelijnssettingen geïncludeerd. Relevante bevindingen waren onder meer lage detectiecijfers, inconsistente rapportering over training, en beperkte integratie van USP’s in opleidingsprogramma's. Tweeënzeventig procent van de studies meldde dat geïnformeerde toestemming werd verkregen, terwijl gestructureerde feedback en debriefing zeldzaam waren.

Conclusie

Mystery shopping biedt mogelijkheden om kwaliteitsverbetering in de huisartsgeneeskunde te versterken. Ons onderzoek toonde grote methodologische variatie en ethische complexiteit, en onderstreepte de noodzaak om ethische beschermingsmechanismen te integreren en studieopzetten aan te passen aan lokale systemen. Deze bevindingen kunnen als leidraad dienen voor beleidsmakers en toekomstige USP-onderzoeken en onderwijsmodellen ondersteunen. Verder onderzoek moet de ervaringen van artsen verkennen en vrijwillige feedbackinitiatieven testen ter bevordering van duurzame verbetering.

Trefwoorden: Mystery shopping, Misleiding, Kwaliteit van zorg, Medische bijscholing, Onaangekondigde gestandaardiseerde patiënten, Systematische review, Methodologie, Ethiek

INTRODUCTION

Modern primary healthcare faces increasing pressure to deliver evidence-based, patient-centered, and cost-effective care[1]. Yet most quality initiatives still rely on indirect data, such as exit surveys, written record audits or clinical vignettes[2]. These traditional methods of observation do not fully reflect how physicians and patients interact in real-life situations[1].

One approach that enables direct observation is the use of unannounced simulated patients (USPs), also known as mystery shoppers. First developed in the service industry, the method was introduced to healthcare in 1973 to evaluate how physicians perform under routine clinical conditions[3]. It is especially well suited for primary care, where patient presentations are often undifferentiated and diagnostic tools are limited. The method is now used across a range of first-contact settings, including general practice, community pharmacy, telephone triage, and emergency departments, and is applied in both high- and low-resource settings[2,4].

Because clinicians do not know they are being evaluated, USPs minimize the Hawthorne effect, capture authentic physician behavior, and standardize case mix[5,6]. It is considered the gold standard for measuring practitioner behavior due to its strong internal validity and its ability to reveal gaps that remain hidden in chart reviews[7]. The method also demonstrates immunity to social desirability bias and is effective for auditing patient experience[8].

However, its application is not without challenges. It is resource-intensive, requiring extensive training and coordination to ensure consistency across cases[9]. Consequently, the number of physicians who can be assessed is often restricted. Although most USPs go undetected, concerns about deception and reputational harm may lead to resistance among healthcare professionals[10,11].

Even when the USP method uncovers shortcomings, translation into improvement is often inconsistent. Most studies involving USPs in continuing medical education (CME) remain descriptive and rarely lead to structural change or sustained improvements in practice[12]. Additionally, delays in feedback undermine principles of continuous quality improvement, making it difficult for clinicians to act on findings in a meaningful timeframe[1]. Without mechanisms to integrate findings into professional development, supervision, or policy, the potential of mystery shopping remains underutilized.

Belgian general practice uses accreditation and peer-review groups (LOK/GLEM) as core tools for quality improvement, supported by centrally organized feedback reports. While widely implemented, these instruments rarely lead to measurable improvements and fail to distinguish between high- and low-quality practices. Objective, behavior-based metrics are lacking, leaving a gap in assessing real-world care[13]. Mystery shopping could help fill this gap by providing independent observations on access, safety, and patient-centeredness.

Covert observation raises ethical concerns around deception, consent, and trust. In low-risk settings, waiving consent may be justified if interactions are brief, anonymous, and followed by debriefing, while longer encounters may require opt-out options[14]. Ethical safeguards such as informed consent and debriefing must be implemented in a way that balances participant protection with study validity, acknowledging the variability and contextual sensitivity of ethical guidance in

deceptive research[15]. Provider acceptance depends on presenting mystery shopping as a tool for learning and improvement rather than for evaluation or punishment[1,16].

An earlier review of USPs assessing physicians in real-world practice covered studies up to 2003[3]. Since then, use has expanded and diversified, with increased focus on quality improvement. Reviews in pharmacy, CME, and low- and middle-income countries (LMIC) point to both methodological advances and persistent reporting gaps[4,8,17]. Many recent studies have not been systematically reviewed, particularly regarding methodological and ethical practices. There is also greater attention to ethical approval, informed consent, privacy, and protocol compliance, especially in research outside the scope of traditional clinical trials. A new synthesis is needed to assess how USPs are currently employed in primary care, how ethical standards are evolving. This is essential to determine whether mystery shopping holds potential as an educational tool in Belgian general practice.

Our systematic review focuses on empirical studies using USPs in primary care. We assess methodological rigor, ethical safeguards, and whether findings led to quality improvement. We also examine how studies were designed and implemented, including scenario development, training, and feedback. By providing an overview of the feasibility of USP studies and their role in quality improvement, we aim to inform how these insights can be translated into the Belgian context. This includes supporting policymakers and educators in considering mystery shopping as a responsible and effective tool to enhance care quality.

METHODS

We carried out a systematic review of empirical studies that evaluated the clinical behavior of PCPs using USP methodologies. Our main objective was to examine the methodological approaches used in studies applying USP methods in primary care and to explore the ethical challenges associated with this type of research. In addition, we aimed to assess whether the use of mystery shopping techniques could have potential applications for CME and how such approaches might be practically implemented in Belgian general practice. These broader reflections are further elaborated in the discussion section of this paper.

We prioritized empirical studies over local quality improvement projects, which are often context-specific and less rigorously documented. Scientific studies provided clearer methods and more explicit attention to ethics, offering insights that were both robust and transferable.

This study was automatically approved by the Ethics Committee Research UZ/KU Leuven and the Education-Support Committee for Medical Ethics KU Leuven, Belgium (reference MP028810). We preregistered the systematic review protocol on the Prospero register (Registration number CRD42023482284) to ensure transparency and allow others to track our review process.

Search strategy

We developed a search strategy (see **Appendix I**) in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement[18]. Based on this search

strategy, we conducted a systematic literature search in PubMed and Web of Science (WoS). We searched for articles in Dutch or English, since these languages are known sufficiently by the authors. Only studies published from 2003 onward were considered, to ensure relevance to the current ethical and methodological landscape and to build upon earlier work of Rethans and colleagues[3]. The search was carried out in October 2023.

Eligibility criteria and study selection

To match our research objectives, we included empirical studies that assessed the clinical behavior of healthcare providers in real-world primary care settings. Eligible studies had to meet several criteria. First, the study population had to be situated in primary care, with a specific focus on physicians. Studies involving paramedical healthcare professionals were only included if their role was directly related to general practice. Studies focusing exclusively on secondary or tertiary care were excluded. However, studies where internal medicine physicians provided primary care were included, as this was sufficiently similar to general practice in accordance with our research focus. Second, studies had to use mystery shopping methodologies to evaluate physician behavior. We focused on research that assessed behavior covertly to reduce bias. Studies in which physicians were aware of being observed or recorded were excluded. Third, the setting had to reflect actual clinical practice. We included studies involving direct patient–physician interactions, whether through in-person consultations, telephone triage, or video consultations. Research conducted entirely in educational or simulated environments, such as objective structured clinical exams (OSCEs), was excluded. Finally, we focused on primary empirical research. We excluded reviews, theoretical papers, and secondary analyses that did not present original data.

Screening process

The selection process was conducted in several steps by two independent reviewers (TV and SP). Initially, titles were screened and duplicates removed. This was followed by abstract screening and subsequent full-text review of studies deemed relevant. At each step, studies were annotated as "yes," "no," or "doubt." A study was excluded at any given stage only if both reviewers marked it as "no." In cases of disagreement or uncertainty, the study was included in the next screening stage for further review. If uncertainty or disagreement persisted after the full-text review, a third reviewer (KV) was consulted to make the final decision. The entire selection process, including all annotations, was documented in a spreadsheet.

Recognizing the potential for relevant studies that could have been overlooked due to limitations in the primary search strategy, a supplementary search was conducted in Google Scholar to identify related articles meeting the inclusion criteria. For each study included following the previously described full-text analysis, related articles were identified using the "related articles" feature on Google Scholar, for which we considered the first nine results per study. The full texts of these related articles were subsequently screened for the keywords “myster,” “unann,” “standardi,” “covert,” and “simulat.” Both TV and SP reviewed these full texts and reached joint decisions regarding their relevance and alignment with the inclusion and exclusion criteria of the study.

Data extraction and synthesis

The studies included in this review were independently examined by the researchers, who extracted key topics of interest. To ensure comprehensive coverage, these topics were identified and structured by the research team based on an initial literature review on which a standardized data extraction form was developed to systematically collect data on the following topics: contextual information (including country and year of study), research objectives, participant consent, debriefing processes, types of data collected, ethical considerations, methods of deception used, participant burden, methods to increase credibility, flexibility versus standardization of methods, and any educational objectives within the studies. Specific sub-questions for each of these main topics are provided in **Appendix II**.

The studies were reviewed in detail to ascertain whether they provided information relevant to our key questions or sub-questions, ensuring no critical data were overlooked. Following extraction by both independent reviewers, the findings were integrated, compared, and discussed within the research team. This allowed for a thorough analysis of the extracted fragments, and any patterns recurring across studies were identified.

A qualitative research approach was applied to the resulting dataset, enabling in-depth exploration and interpretation. To support our analysis, we employed a predominantly deductive research method, allowing the findings to align with our pre-established research questions while also remaining open to emerging patterns that recurred across studies. This way, pertinent topics were extracted systematically and will be thoroughly discussed in the results section of this paper.

RESULTS

Study selection

The study selection process is illustrated in **Figure 1**. In this study, a comprehensive selection process was undertaken to identify relevant literature for systematic analysis. Initial searches in PubMed and WoS yielded 121 and 65 articles, respectively. We filtered out 31 duplicates leaving 155 articles to be screened. After an initial review, 39 studies were excluded based on titles, and an additional 22 studies were excluded after abstract screening, with both primary reviewers (TV and SP) reaching consensus on these exclusions. Following full-text analysis, TV and SP agreed to exclude a further 43 studies, leaving 23 studies marked for further consideration due to ambiguity. A third researcher (KV) then conducted an additional assessment, leading to the exclusion of 20 more studies. This process resulted in 31 studies deemed suitable for systematic analysis. Further searches for related articles on Google Scholar identified three additional relevant studies. Later in the research process, an additional 9 studies were included through snowballing, resulting in a total of 43 studies selected for systematic analysis. All included articles are listed in **Appendix III**.

Excluded studies primarily involved training programs for students, lacked the use of mystery shopping, or were set in geographical contexts that diverged significantly in primary care structures compared to the Belgian context. Primary care clinics staffed by internists were generally included, as their practices closely aligned with the primary care framework in Western Europe. Certain

studies identified during our systematic review, such as Liao et al. (2018) retrieved via database search and Schwartz (2020) identified through snowball sampling[17,19], were based on datasets identical to other included studies (e.g., Zeng et al. (2022) and Schwartz et al. (2021), respectively)[20,21]. While these overlapping records were not counted as independent studies in our analysis, they were nonetheless thoroughly reviewed to comprehensively assess their methodological contributions and contextual nuances.

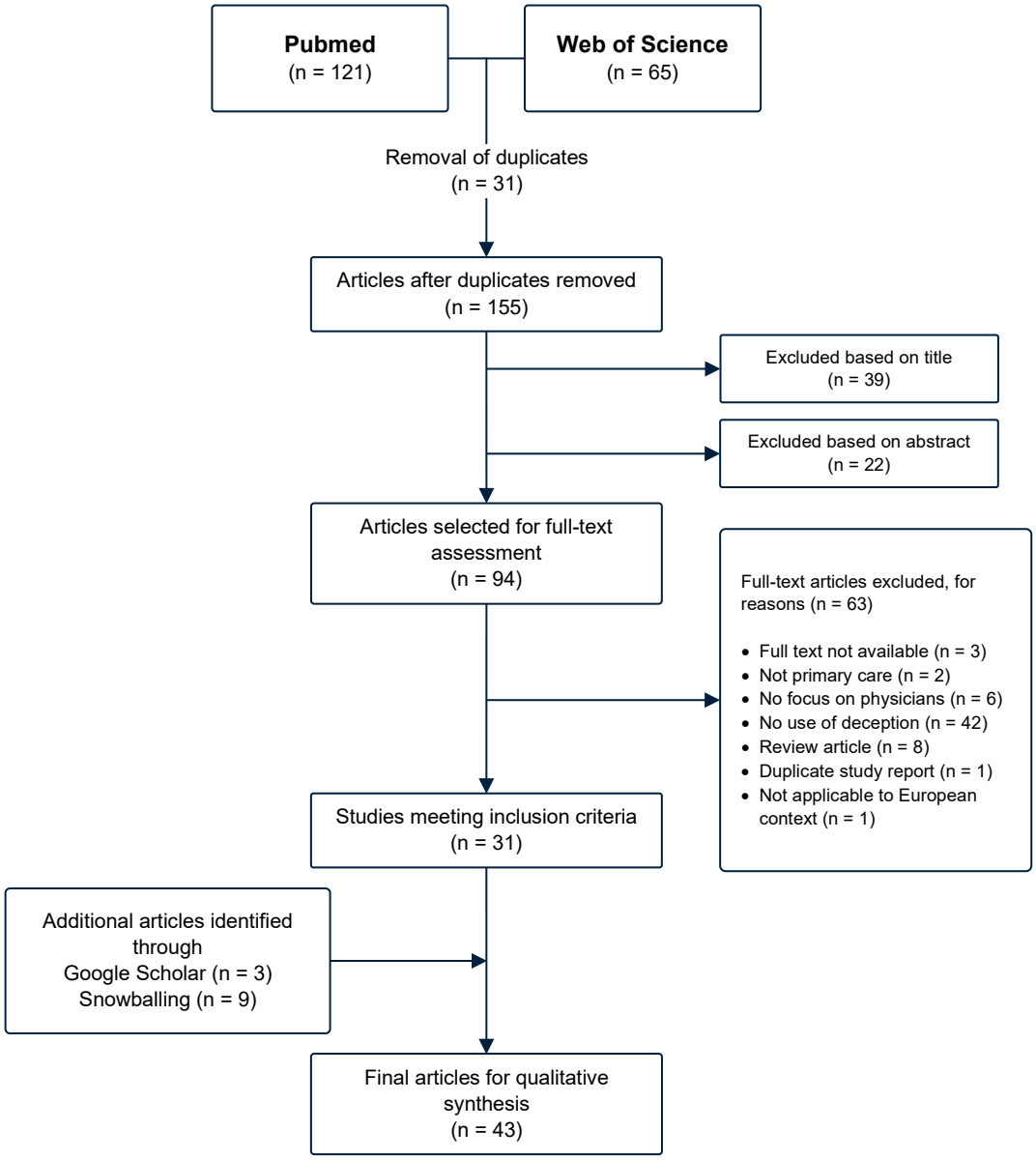


Figure 1. PRISMA Flow Diagram of Study Selection Process

Study characteristics

Among the 43 included studies, the majority (n = 30) used a cross-sectional design, capturing provider behavior at a single point in time. Thirteen studies applied a longitudinal design, typically involving multiple data collection points to assess changes or trends over time. Most studies were observational in nature (n = 24), meaning no intervention was introduced. Twelve studies employed an experimental design, usually through a randomized controlled trial (RCT), while seven used quasi-experimental methods involving non-randomized comparison groups. **Table 1** provides an overview of study designs along with other key characteristics. A detailed spreadsheet containing the full data extraction for all included studies is available in **Appendix IV**.

In terms of research aim, most studies focused on evaluating quality of care (n = 16) or current clinical practices (n = 12). Intervention assessment was the aim in nine studies, while six compared different assessment methods, such as USPs versus real patient exit interviews or patient surveys. Five studies focused on evaluating the USP method itself, examining its reliability, realism, or utility[22–26]. Only a small number of studies (n = 3) aimed to provide feedback to providers or were used as educational tools[27–29]. Several studies addressed multiple research objectives. For example, Van Onna et al. (2017), Feng et al. (2013) and Peremans et al. (2010) examined both guideline adherence and intervention evaluation[30–32]. Hornung et al. (2007) focused on both SP evaluation and guideline adherence[22]. Xie et al. (2023) investigated consultation evaluation and compared USPs with online real patient reviews[33]. Daniels et al. (2017) addressed USP evaluation alongside quality of care assessment[25].

The types of healthcare providers assessed varied widely. General practitioners (GPs) were evaluated either alone (n = 9) or in combination with internal medicine physicians (n = 9). Some studies focused exclusively on internal medicine physicians (n = 4) or nurses (n = 4), while others assessed GPs and nurses together (n = 1). Nine studies examined primary care providers (PCPs) without specifying their role, while five included a mix of PCPs. Two studies assessed both primary and secondary care providers.

With regard to setting, over half of the studies were conducted in urban primary care clinics (n = 22), which include both community health centers (CHCs) and other outpatient clinics. General practice was the setting in nine studies, and four were conducted in rural primary care facilities, such as village clinics and township health centers in China. A smaller number of studies focused on telemedicine (n = 5) or were conducted across mixed settings (n = 3), including combinations of facility types or care levels.

Geographically, most studies were conducted in North America (n = 17), followed by Asia (n = 10) and Europe (n = 10). Four studies were carried out in Africa, and two in Central America. At the country level, the United States accounted for the largest number of studies (n = 16), followed by China (n = 8) and the Netherlands (n = 5). Studies from China were predominantly published after 2015, while most U.S.-based studies were conducted earlier, with only four published more recently.

Table 1. Overview of study characteristics (N = 43)

Category	Response options[a]	N (%)
Design	Cross-sectional	30 (70%)
	Longitudinal	13 (30%)
	Observational	24 (56%)
	Quasi-Experimental	7 (16%)
	RCT / Experimental	12 (28%)
Aim	Quality Evaluation	16 (37%)
	Evaluating Current Practices	12 (28%)
	Intervention Assessment	9 (21%)
	Comparing Different Assessment Methods	6 (14%)
	USP Method Evaluation	5 (12%)
	Provider Feedback / Educational Tool	3 (7%)
Subject being assessed	GPs	9 (21%)
	Nurses	4 (9%)
	Internal medicine physicians	4 (9%)
	GPs and Internal medicine physicians	9 (21%)
	GPs and nurses	1 (2%)
	PCPs (unspecified)	9 (21%)
	Mix of PCPs[b]	5 (12%)
	Mix of primary and secondary care providers	2 (5%)
Setting	General Practice	9 (21%)
	Urban Primary Care Clinics[c]	22 (51%)
	Rural Primary Care Facilities[d]	4 (9%)
	Mixed Settings[e]	3 (7%)
	Telemedicine	5 (12%)

[a] Non -mutually exclusive
[b] Includes physicians (medical doctors, clinical officers), paramedical staff (nurses, midwives), and non-medical personnel (e.g. social workers, data clerks)
[c] Comprises primary care clinics and community health centers (CHCs).
[d] Includes both village clinics and township health centers, representing the lower and mid-tier levels of rural primary care in China.
[e] Mix between different primary care facilities and or hospitals.

GPs, general practitioners; PCP, primary care provider; RCT, randomized controlled trial; USP, unannounced standardized patient

Cases

During the analysis, it was found that 26 out of 43 studies used 1–4 case scenarios, with 2 cases (n = 9, 21%) and 1 case (n = 7, 16%) being the most common. Eight studies used 5–9 cases, two studies used 10–14 cases, and three studies included more than fifteen cases. Two studies did not report the number of cases: one used real patients[34], and the other did not specify the number of cases[35].

During the analysis, a total of 167 distinct and relevant clinical cases were identified. To facilitate the analysis, these cases were grouped into the following major medical disciplines: gastrointestinal tract disorders (n = 23 out of 167, 14%), musculoskeletal diseases (n = 19, 11%), patient-centered care (n = 19, 11%), cardiovascular diseases (n = 18, 11%), infectious diseases (n = 18, 11%), psychiatric disorders (n = 18, 11%), pulmonary diseases (n = 11, 7%), neurological disorders (n = 11, 7%), gender-related care (n = 9, 5%), diabetes (n = 7, 4%), oncology (n = 5, 3%), ophthalmology and otolaryngology (n = 4, 2%), and urology (n = 3, 1%).

Frequently encountered cases included depression (n = 11 out of 167, 7%), angina pectoris (n = 11, 7%), lower back pain (n = 10, 6%), diabetes (n = 7, 4%), asthma (n = 7, 4%), and diarrhea (n = 7, 4%). Urgent medical conditions were relatively uncommon, with the exception of stroke[36] and several cases of angina pectoris.

USP characteristics

The studies reported a wide range in the number of USPs sent. Ten studies involved 1–5 USPs, with employing five USPs being the most common (n = 4). Seven studies used 6–10 USPs, while four studies used 11–15 USPs. Seven studies employed 16–20 USPs, with 18 USPs being the most frequently used number in this group (n = 4). More than 20 USPs were used in four studies. In eleven studies, the exact number of USPs was not specified. The number of USPs used in each study is summarized in **Figure 2**.

Several studies reported specific eligibility criteria for USP participation. These included age (n = 11), ethnic background (n = 8), gender (n = 6), being in good health (n = 8), and other study-specific criteria (n = 13). Some studies included overlapping characteristics across these categories.

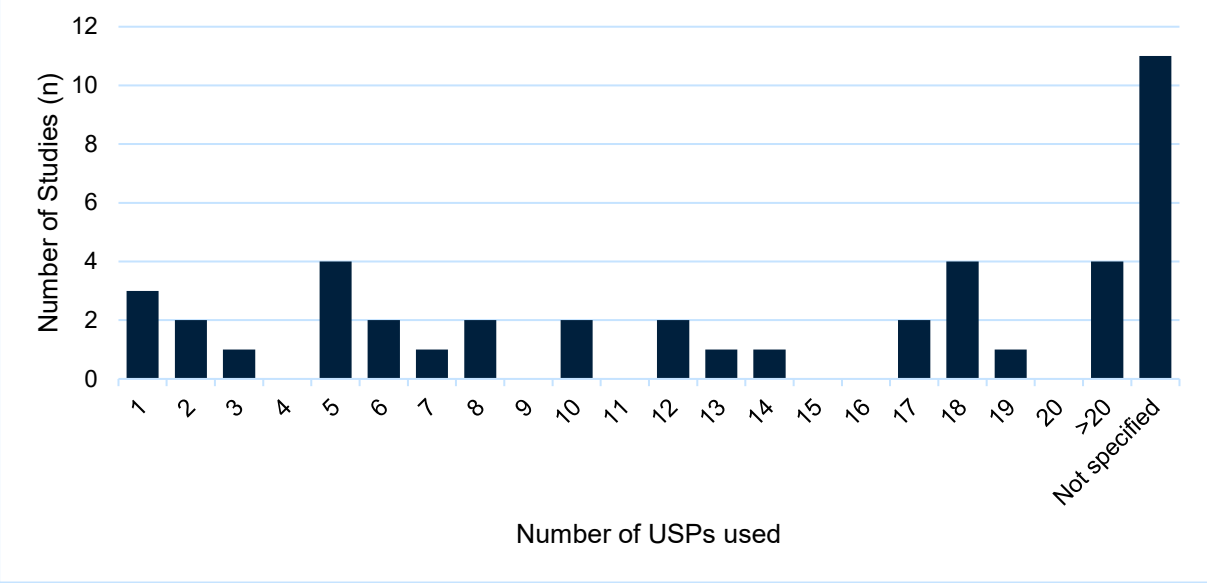


Figure 2. Distribution of studies by number of USPs used.

Training

In seven studies, USPs received 1–12 hours of training. An additional five studies reported training durations of 12–24 hours. In Su et al., a three-day training program was conducted [37]. In the studies by Bachmann et al. and Shirazi et al., USPs were trained over a period of five days [38,39]. Five studies reported training durations spanning several weeks. In total, 23 studies did not specify the length of the training.

Various training methods were used across studies. Fourteen studies included training focused on accuracy. In 25 studies, USPs were trained in the specific case scenarios. Seven studies trained USPs in completing checklists. Six studies provided training to avoid invasive examinations, while four studies included training on the use of recording equipment. Five studies trained USPs in performance evaluation. Sikombe et al. trained USPs to assess quality markers[40], and in Peabody et al.[41], training was done with establishing protocols that were already utilized, focusing on case training. In six studies, no details were provided regarding the training procedures.

Accuracy of USPs

Studies tried to standardize USP visits by evaluating the accuracy with which USPs can perform their cases. In 31 studies, no information was provided regarding the monitoring of USP accuracy. In three studies, feedback was given during the execution of USP visits[20,23,28]. Nine studies reported specific training focused on the accuracy of USP performance and case presentation.

The studies by Beaulieu et al., Zabar et al. and Zeng et al., also included USP evaluation and represent three of the five studies that assessed USP performance[20,23,28]. In Zabar et al. the project coordinator continuously gave feedback by debriefing after each visit and controlled the consultation with audio-tapes[23].

Credibility

A total of eighteen studies implemented measures to increase USP credibility. Some of these measures included physical adjustments (false pregnancy belly[42] or loose cloths for tuberculosis[25], use of fictitious personal documentation[22,23,26,28,35,41,43,44], practical adjustments [26,45–48] (for example: using private home calls[47] or self-payment of consultation(with no insurance involvement)[48], and consideration of local contextual factors[28,49,50] (for example: local addresses[28]). In one study[34], real patients (local recruited patients) were used to enhance the credibility of the USPs. Twenty-four studies did not mention efforts to enhance SP credibility.

Detection

Fourteen studies did not mention any suspicion of USPs by healthcare providers, nor measures taken to prevent detection.

A total of 23 studies reported on detection by GPs. Nine studies described self-reporting by the GP, seven used a survey to assess detection, six addressed the issue through other methods, and one study[51] mentioned exclusion of a GP in cases of suspected detection.

Preventative measures were reported in ten studies, most commonly focusing on reducing the detectability of USPs (n = 5). An other strategy included the use of a RCT design with blinded

intervention and control group, this had the advantage of not knowing when USP started respectively[52]. Additional strategies included training of USPs[24], local recruitment[49], adjusting for shift schedules of participants[45], and a study intervention aimed specifically at reducing detection by case design[51].

Data collection

In fifteen studies, data were collected using a checklist completed by the USP. Thirteen studies employed both audio recordings and checklists, while ten studies relied solely on audio recordings. Two studies[35,50] combined a checklist with a post-visit feedback conversation between the USP and a member of the research team. In Meuwissen et al.[53], an exit interview was conducted between the researcher and the USP, and Stevens et al.[34] utilized an online questionnaire.

Ethical approval

The vast majority of included studies (n = 37, 86%) reported having received ethical approval. Of these, 33 studies obtained approval from an institutional review board (IRB), while three LMIC studies were approved by a national ethics committee[25,39,50], and one Austrian study received clearance from a municipal ethics committee[44]. Comprehensive information regarding the type and issuing body of ethical approval for each study is available in **Appendix V**.

In three studies (7%) no formal ethical approval was obtained or required; the authors explicitly stated that ethical clearance was not needed due to the nature or their research. Lorenzo et al. (2015) clarified that the French local review board determined the project fell outside the scope of the French research regulation, as it was an evaluation of professional practices[54]. Similarly, Huibers et al. (2012) reported that ethical approval was not necessary under Dutch law[55], and Derkx et al. (2008) also stated that no approval was required for their nationwide telephone triage study[47]. Two studies (5%) provided no information on whether an ethical review was conducted or deemed necessary[27,41]. Finally Zabar et al. (2014) did not seek further IRB approval because data on resident physicians were already included in an existing IRB-approved research registry, while the system-level observations were classified as Quality Improvement and did not qualify as human subjects research[23].

Informed consent

Of the 43 included studies, the majority (n = 31, 72%) obtained informed consent from the participating PCPs. Three of these cases acquired consent verbally. In one of these studies, consent was requested via telephone[54], while the other two, both Chinese studies, used verbal consent in line with local cultural norms[46,49]. In three studies (n = 3, 7%), institutional consent procedures were applied. This means that the organization or healthcare institution, rather than the individual providers, formally agreed to participate in the study[47,53,56]. Six studies (n = 6, 14%) obtained a waiver of informed consent from a research ethics committee[20,33,36,50,51,57]. In one study, no consent was obtained, based on the rationale that the healthcare providers were not considered study participants and were not studied individually[34]. A Belgian study only offered participants the opportunity to opt out: participation occurred by default, with the option to decline[31]. In contrast, opting-out was explicitly not possible in another study[44]. For a complete overview of how each study approached consent and notification, including contextual details, see **Appendix VI**.

Regarding prior notification, PCPs were explicitly informed about the USP method in advance in 33 studies (n = 33, 77%). When informed, PCPs were told that they might receive an unannounced visit from a USP within a given time window which could range from several weeks to 18 months, without knowing exactly when the visit would take place. The group of studies in which healthcare providers were informed in advance overlapped substantially, but not entirely, with the group in which informed consent was obtained. There are four notable exceptions that account for the differences between the two sets. First, in Su et al. (2022), written informed consent was documented, but it was not clearly stated in the full text whether providers were informed in advance of potential visits[37]. Second, in Piribauer et al. (2012), physicians were informed that USPs could visit, but participation was mandatory, meaning that informed consent in the conventional sense was not possible[44]. Third, in Gillespie et al. (2023), medical residents were routinely assessed with USPs, and although they were aware of this, only institutional consent was obtained[56]. Lastly, in Peremans et al. (2010), GPs were informed of the study, but participation occurred by default unless they opted out, which does not constitute explicit informed consent[31]. This is clarified schematically in **Appendix VI**.

In three studies (n = 3, 7%), there was a partial disclosure: providers were informed about general data collection or quality evaluation activities but not that this would involve unannounced visits[46,53,57]. In six studies (n = 6, 14%), no prior information was given to the healthcare providers. This was due to methodological necessity except for one paper – as already mentioned above – where physicians were not informed because they were not the study subjects[34].

Harms and safeguards

Few studies reported harm to the PCPs as a result of their USP method. However most (n = 33) do mention safeguards to mitigate the possibly known harms. All safeguards taken to avoid harm to the PCPs are listed in **Appendix V**. The demanding schedules of PCPs were frequently recognized. To avoid taking up too much time, several studies limited the length of interactions, covered consultation fees, or scheduled visits outside peak hours[25,45]. Time burden as an explicit harm due to USP visits was mentioned by four studies[25,42,44,54].

Su et al. (2022) reported administrative burden, noting that follow-up questionnaires could disrupt workflow, especially in busy practices[37]. Schwartz et al. (2021) identified financial loss as a concern, as participating providers were not reimbursed[21]. In Bryant et al. (2009), PCPs raised privacy concerns and reported discomfort with being recorded[42].

Several studies in our review described clear efforts to minimize harm to USPs, listed in **Appendix V**. A frequently recurring strategy was to train USPs to avoid invasive procedures (n = 10, 23%). Meuwissen et al. (2006) for example trained their USPs to decline vaginal examinations[53]. Eight studies (19%) described training techniques to circumvent dangerous situations or detection. A specific example is given by Franks et al. (2006) where the USPs were taught to avoid referral to the emergency department or being administered medication[29]. Five studies (12%) selected their cases to ensure that no clinical examination or invasive procedures were required. De Castro et al. (2018) and Kwan et al. (2022) reported that the well-being of the USPs was actively monitored (n = 2, 5%)[50,51]. Peremans et al. (2010) provided access to a psychological support team should the USP encounter a negative experience[31].

Compensation

Compensation of the research subjects (i.e., the PCPs) was not explicitly mentioned in the nearly half of studies (n = 20, 47%). When compensation was reported, the most common arrangement was that the USPs advanced a co-payment or full consultation fee, which was subsequently reimbursed by the study, or the study directly covered the physician's fee (n = 16, 37%). Usually the cost was borne by the research team, but in three cases the cost was borne by health insurance. Two of the three because of a collaboration[28,44], and in the last case because the USPs were real patients[34]. One of these studies also offered CME credits in addition to monetary compensation[58]. In three studies (7%), no financial compensation was provided, and participation was voluntary or incentivized through non-monetary means such as CME credits[21,38,52]. Finally, in one study, the USP personally paid for the consultation without reimbursement from the study; in this case, the USP was also the author[54].

Data protection

Because mystery shopping studies may involve sensitive or identifiable information, several studies reported specific measures to protect privacy. Data was either not identifiably collected (n = 6, 14%)[22,23,28,35,39,45], or de-identified after collection (n = 7, 16%)[21,36,44,51,53,56,57]. Five studies (12%) stated that data was not publicly available but could be accessed under specific conditions, such as data use agreements or reasonable requests[20,25,40,43,51]. Three studies (7%) indicated that data were stored in secure platforms where personal identifiers remained confidential and analyses were conducted at an aggregated level[33,35,36].

Debriefing

In this review we identified seven studies (n = 7, 16%) that explicitly implemented a form of debriefing after USP visits. Altshuler et al. (2023) shared USP-related findings with clinic leadership as part of a larger, ongoing educational evaluation of internal medicine residents[35]. Gillespie et al. (2023) provided audit and feedback reports summarizing screening and response rates to participating clinics in a statewide learning collaborative focused on social determinants of health[56]. Schwartz et al. (2021) delivered individualized feedback through performance reports, visual summaries, and coaching calls following USP visits[21]. Price-Haywood et al. (2014) offered structured verbal feedback at the end of visits, supplemented by access to a web-based platform and written summaries describing SP experiences[58]. Kwan et al. (2022) revisited participating clinics after data collection to explain the study and address concerns[51]. Franz et al. (2006) debriefed physicians after they completed a post-study questionnaire, although the format of the debriefing was not further specified[24]. In our review, we found no negative consequences of debriefing.

The remaining studies (n = 36, 84%) did not report debriefing procedures or recontacted participants only to assess whether they had recognized the USP. A notable variation was demonstrated by Culver et al. (2009) who included an additional step by first sending a detection questionnaire and subsequently followed up by sending a second mailing that revealed the identity of the USP[48].

DISCUSSION

Our review aimed to answer the following research questions: How were USPs implemented in Primary Care? What practical differences emerged between studies? How did researchers address the ethical challenges inherent to this method? And finally, how might such approaches be adapted to fit a healthcare system like that of Belgium?

Methodological Approaches and Implementation Strategies

The included studies mainly used cross-sectional designs, with some employing experimental approaches such as RCTs. Most research was conducted in primary care clinics and general practice, mainly in North America and Europe, settings broadly comparable to Belgium. Interventional studies offer opportunities for quality improvement, while observational designs reveal real-world clinical practices, both providing valuable insights for adapting USP methods to the Belgian healthcare system.

USPs were mainly used to evaluate guideline adherence, consultation quality, and intervention implementation, consistent with the findings of Wilhite et al.[59]. Rather than simply identifying gaps, USPs can actively support quality improvement in systems like Belgium's. For example, motivational interviewing on smoking cessation could be assessed and enhanced through tailored interventions. This approach could strengthen preventive care and generate meaningful public health gains.

The number and type of USP cases varied greatly across studies, largely shaped by practical constraints and study aims. For quality improvement, 1 or 2 standardized cases often sufficed, while broader assessments benefited from more diversity. Giesen et al. and Huibers et al. showed that using multiple case types improves telephone triage evaluations[27,55]. Simpler cases such as gastrointestinal symptoms and lower back pain were favored because they required minimal physical examination and could be portrayed consistently. Future research should balance case variety with feasibility to optimize study outcomes.

Each study adopted a unique approach to USP recruitment and qualification. While explicit justifications were often lacking, decisions were likely influenced by practical constraints such as funding and training resources. Some studies imposed specific requirements on USPs, criteria that are more challenging to meet with larger cohorts. These stricter demands were especially noted in research on sensitive topics such as racial bias[43] and female contraception interventions[31,50,53].

There was considerable discrepancy in the amount of training time reported for USPs. Reporting methods varied across studies, lacking standardization[12]. For instance, Amini et al. described 10 hours of training divided into five 2-hour sessions, though it remained unclear whether this occurred over multiple days. Training duration depended on various factors, including case-specific preparation, accuracy training, and checklist familiarization.

Ensuring fidelity and uniformity in case delivery was considered essential and should be emphasized in future USP training protocols[60]. This stands in contrast to studies employing real patient surveys accompanied by audio recordings, which can also yield reliable data. However,

such methods may at times be perceived by both GPs and patients as a form of 'surveillance', potentially influencing the dynamics of the doctor–patient relationship[16].

Quality assurance and validation of USP performance were addressed before, during, and after the intervention phases. The aim was to optimize accuracy and credibility while minimizing the detection rate to ensure consistency across study interventions. Despite this, not all studies reported implementing measures to enhance USP credibility. Moreover, while detection of USPs was mentioned more frequently, only a few studies took concrete steps to reduce the likelihood of detection. It is important to note that many studies did not provide detailed reporting on these aspects. Future research would benefit from more transparent and standardized reporting on USP detection and related influencing factors to better inform the design of low-detection, high-credibility USP interventions. It is important to address this issue, as highlighted in the study by Barrett et al.[61], since the presence of suspicion may influence the outcomes of an intervention. This phenomenon has been noted elsewhere and is reminiscent of the Hawthorne effect.

Data collection in the reviewed studies was predominantly conducted using USP-completed checklists immediately following the consultation, and audio recordings of the encounters. However, the use of audio recordings may raise ethical and legal concerns within the Belgian context, particularly in light of GDPR and the professional autonomy of GPs (as mentioned below). This is further complicated when post-intervention debriefings are considered. Nevertheless, as noted by Aujla et al., relying solely on checklists introduces a risk of recall bias, an issue that can be mitigated through the combined use of checklists and audio recordings[60].

Ethical Challenges and Considerations

Using USPs in primary care introduces potential risks for clinicians, actors, and the system. Most studies in our review acknowledged these risks and implemented measures to minimize them, aligning with ethical frameworks focused on transparency and truthfulness[11].

To protect USPs, studies often trained them to avoid invasive procedures and designed scenarios accordingly. Some also monitored well-being or offered psychological support when needed[31,50,51].

For PCPs, time and financial burden were common concerns, typically addressed through limited visit frequency, modest case scenarios, and reimbursement or CME credits[21,38,52]. While these strategies reduce objections, they increase the cost of implementation. Privacy concerns were handled by anonymizing or aggregating data, or by not making data publicly available. No study disclosed individual clinician identities, an important safeguard against reputational harm[14]. Sometimes, a small sample still poses a significant risk for re-identification. In one example, Fiscella et al. restricted data access by grouping ages into broad categories and removing key identifiers, requiring a data use agreement to ensure confidentiality[43]. Other differences in study acceptability may relate to setting, cultural expectations, and whether the purpose was regulatory or educational.

Informed consent remains a key ethical concern in USP study designs. While a previous review found near-universal consent[3], we found that 72% of studies obtained consent. In 14% of cases, participants were not informed beforehand, typically justified by low risk and high scientific value[14]. Guidelines vary on whether informed consent is required in deceptive research, reflecting ongoing debate in the literature[15]. Studies in high-income countries (HIC), especially in Europe and the

US[62], more often included consent or opt-outs[31], while LMICs relied more on ethics waivers[20,33,36,50,51,57]. This reflects patterns noted by Kingori and Douglas-Jones, who observed that deception-based methods are more readily approved in the Global South[63]. In Belgium, where trust is central to general practice, bypassing consent risks damaging relationships. GDPR further makes informed consent legally necessary. Yet, as shown in psychological research, even studies seeking consent sometimes fall short of GDPR standards[64]. Without consent, anonymization is mandatory, preventing feedback or linking findings to care. Wherever possible, informed consent should be prioritized to safeguard legal and ethical standards[65].

Debriefing, the structured act of informing participants after an incognito visit that they were part of a study and providing feedback, is considered an important ethical safeguard in USP studies[66]. Despite this, only 16% of the reviewed studies mentioned debriefing, consistent with earlier findings that also noted limited systematic feedback to physicians[3]. Guidance on when debriefing is required remains unclear and varies across ethical frameworks[15]. Some researchers emphasize its value for transparency and feedback, while others caution that debriefing could cause more harm than the original deception[11]. In educational contexts, however, debriefing is seen as essential to promote reflection and professional development[12].

None of the reviewed studies reported whether participants were given the option to withdraw their data after debriefing. Yet this is recommended by ethical guidelines[15]. This gap is relevant given the concerns raised by Hodson (2022), who warns that post-debriefing data withdrawal can introduce bias into study findings[67].

Since USPs have proven effective in identifying knowledge gaps that often go unnoticed by conventional methods[7], we suggest their use as a feasible educational tool in general practice. A model centered on voluntary participation, anonymity, and constructive feedback might be most feasible. For example, annual visits combined with individual coaching could offer a low-risk approach to support professional development. International evidence indicates that, with careful planning, this approach has little impact on daily operations and often goes unnoticed by clinicians[66].

Mystery Shopping in Belgian Primary Care

Mystery shopping has sparked debate in Belgium, particularly following media investigations into GP practices[68,69]. Initial political resistance[70] has softened, with acceptance growing in fields like aesthetic medicine[71]. However, this acceptance mainly reflects regulatory purposes and does not yet extend to educational quality improvement interventions. If the findings of our study were to be translated into recommendations for implementing USP methodology in a Belgian primary care setting, we would propose the following:

First, GPs should be informed in advance about the possibility of participating in a USP study and given the option to opt in. This approach maintains professional autonomy while safeguarding the scientific integrity of the research. To ensure feasibility and cost-effectiveness, the number of USPs should align with study goals, aiming for a limited number appropriate to the study's scope.

Second, USP methodology should focus on evaluating current practices or assessing interventions aimed at quality improvement, with an emphasis on providing constructive feedback rather than

penalizing physicians for unfavorable results. Linking participation to incentives such as accreditation points could encourage engagement. USP encounters could specifically be used to assess physician communication skills, for example during smoking cessation counseling or cancer screening discussions.

Third, we recommend using audio recordings alongside standardized checklists to improve data reliability and reduce recall bias. Audio files should be transcribed, anonymized, and deleted after analysis to protect privacy. Alternatively, participants could be given the option to opt out of recording-based research.

Fourth, when reporting study results, all data should be anonymized and preferably aggregated to prevent any risk of individual identification. This protects both patient and physician privacy and ensures compliance with Belgian and European data protection standards. Aggregated reporting minimizes the possibility of re-identification, even in small samples or sensitive contexts.

Fifth, while GPs usually receive reimbursement for consultations via social security, direct billing for mystery shopping research is not appropriate under current Belgian regulations[72]. Instead, collaboration with the Belgian National Institute for Health and Disability Insurance (RIZIV) could ensure partial financial support for consultation costs, without burdening either the researchers or the participating GPs.

Finally, several practical challenges remain. Patient intake limits, variable consultation flow, and workforce shortages, particularly in suburban and rural areas, may hinder the implementation of USP studies. Moreover, the high value placed on clinical autonomy in Belgian primary care means that studies perceived as externally imposed could face resistance. Therefore, careful timing is essential, avoiding high-demand periods like the flu season.

Our findings provide practical direction for embedding mystery shopping in Belgian general practice while addressing its ethical complexity. By clarifying how studies handled consent, harm, and debriefing, we outline a pathway for ethically sound applications. In a system where professional trust and privacy protections are central, our work can guide the development of national frameworks that align with legal standards and practitioner expectations. If applied with care, mystery shopping could improve feedback culture, support lifelong learning, and contribute to transparent, data-informed quality improvement.

For future research, we recommend conducting a pilot study to assess feasibility and fine-tune procedures in the Belgian context. Furthermore, systematic reviews similar to ours could be conducted for other specialists, such as gynecologists and dermatologists, to explore the applicability of USP methodology in their fields. In those cases, reimbursement models would also need careful adaptation.

Strengths and limitations

This review provides a timely and structured update and a new perspective on the ethical and methodological challenges of using mystery shopping in international primary care. A major strength is the structured extraction of methodological and ethical practices using consistent

criteria, allowing us to identify recurring patterns and variations across 43 empirical studies. This enables concrete guidance for designing ethically sound USP protocols.

Nonetheless, our study has several limitations. First, despite a broad search strategy, we may have missed non-English studies or gray literature. Several relevant papers were found through snowballing, suggesting inconsistent indexing. Second, by limiting our scope to studies published after 2003, we may have excluded older yet still-relevant research. This threshold was chosen to ensure relevance to current ethical and methodological norms, and it also aligns with the endpoint of the previous systematic review by Rethans et al. (2007), thereby allowing our work to build on and update their findings[3]. Third, overlapping datasets appeared in multiple papers, sometimes inflating counts. We minimized this by focusing on unique descriptions of ethical practices, though some duplication remains. For example, parts of data overlapped between the studies of Franks et al. (2006), Franz et al. (2006), Srinivasan et al. (2006) and Fiscella et al. (2007)[24,26,29,73]; and also between Su et al. (2022) and Si et al. (2023)[37,74]. Finally, our synthesis relies on what was reported. Some procedures may not have been described so the number of studies that did not implement a particular method or ethical safeguard may have been overestimated.

Despite these limitations, this review offers a comprehensive synthesis of international practices and provides critical insights into ethical challenges, strengthening the foundation for future research and practical applications.

CONCLUSION

Mystery shopping provides a unique opportunity to observe real-world clinical behavior while minimizing bias, offering insights into guideline adherence, consultation quality, and intervention effectiveness.

In this review, we systematically analyzed 43 empirical studies across primary care settings to examine how USPs have been applied and how methodological and ethical challenges were addressed. Beyond updating an overview of common practices and gaps, our work introduces a new perspective by emphasizing the integration of ethical considerations and the structure of local healthcare systems into USP study designs.

Methodologically, USP applications were diverse, with wide variation in case types, actor training, and reporting standards. This diversity limits comparability across studies and highlights the need for standardized and transparent reporting practices.

Ethically, practices differed substantially, particularly regarding informed consent and debriefing. When appropriate safeguards were in place, USPs supported professional development and quality improvement. In trust-based systems like Belgian general practice, a careful ethical balance remains crucial.

We suggest future research explores physician experiences with mystery shopping, especially after debriefing. Similar reviews could also be performed in specialties such as gynecology or dermatology. A local pilot project in Belgian general practice, built on voluntary participation and educational feedback, could assess feasibility. While USP-based education requires time and

resources, its ability to identify behavioral gaps may outweigh these costs when other tools fall short.

When carefully designed and ethically implemented, mystery shopping holds significant promise as a constructive tool for improving care in general practice. It can facilitate critical self-reflection among clinicians, enhance a culture of structured and data-driven feedback, and strengthen core dimensions of quality such as accessibility, communication, and adherence to evidence-based practices.

What is already known on this topic?

- USP methods are increasingly used to assess healthcare quality in primary care settings across the world. These studies reveal critical insights into clinical practice gaps across diverse settings and reduce the biases typically found in direct observation.
- USPs are a well-established method for assessing both medical knowledge and communication styles, and are increasingly regarded as the 'gold standard' in this domain.
- Ethical concerns around informed consent and deception are frequently discussed in the literature but are often addressed inconsistently in practice.

What does this study add?

- We present a comprehensive update of 43 USP studies conducted over the past two decades in primary care, identifying a broad range of methodological designs, including differences in study objectives, training methods and data collection.
- Our review reveals significant contextual differences in use of waivers of consent, harm mitigation and debriefing practices, underscoring the need for clearer guidelines.
- These findings inform future research and policy efforts in quality improvement and support further exploration of USP methods as a potentially valuable educational tool in Belgian general practice.

ACKNOWLEDGMENTS

First, we would like to express our sincere gratitude to Prof. Dr. Pascal Borry, who guided us as promotor throughout this thesis project. His thoughtful input and clear feedback helped us to sharpen our research focus and frame our work within the broader ethical context. We are equally thankful to Dr. Kamiel Verbeke, our supervisor, for his continuous support and critical insights. His availability, careful reading, and pragmatic advice were invaluable, especially during the more challenging phases of the review process. Finally, we are grateful to each other for the smooth and constructive collaboration. Writing this thesis together has been both intellectually enriching and personally rewarding.

In line with the KU Leuven guidelines on the responsible use of generative AI, we used ChatGPT (OpenAI) solely as a language assistant to refine the clarity and coherence of texts we had written ourselves, and to support repetitive tasks in Excel. Generative AI was never used to create content or generate blocks of text included in this thesis.

All authors declare having no conflict of interest in the realization of this research article.

No funding has been received to support the development of this article.

REFERENCES

1. Kelley AT, Weiner SJ, Francis J. Directly Observed Care: Crossing the Chasm of Quality Measurement. J Gen Intern Med 2022, 38(1):203–7. doi: 10.1007/s11606-022-07781-1.
2. Collins JC, Chong WW, de Almeida Neto AC, Moles RJ, Schneider CR. The simulated patient method: Design and application in health services research. Research in Social and Administrative Pharmacy 2021, 17(12):2108–15. doi: 10.1016/j.sapharm.2021.04.021.
3. Rethans JJ, Gorter S, Bokken L, Morrison L. Unannounced standardised patients in real practice: A systematic literature review. Med Educ 2007, 41(6):537–49. doi: 10.1111/j.1365-2929.2006.02689.x.
4. King JJC, Das J, Kwan A, Daniels B, Powell-Jackson T, Makungu C et al. How to do (or not to do) ... using the standardized patient method to measure clinical quality of care in LMIC health facilities. Health Policy Plan. 2019, 34(8):625–34. doi: 10.1093/heapol/czz078.
5. Leonard K, Masatu MC. Outpatient process quality evaluation and the Hawthorne Effect. Soc Sci Med 2006, 63(9):2330–40. doi: 10.1016/j.socscimed.2006.06.003.
6. Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of Vignettes, Standardized Patients, and Chart Abstraction. JAMA 2000, 283(13):1715. doi: 10.1001/jama.283.13.1715.
7. Weiner SJ, Schwartz A. Directly observed care: Can unannounced standardized patients address a gap in performance measurement? J Gen Intern Med 2014, 29(8):1183–7. doi: 10.1007/s11606-014-2860-7.
8. Björnsdottir I, Granas AG, Bradley A, Norris P. A systematic review of the use of simulated patient methodology in pharmacy practice research from 2006 to 2016. International Journal of Pharmacy Practice 2020, 28(1):13–25. doi: 10.1111/ijpp.12570.
9. Beullens J, Rethans JJ, Goedhuys J, Buntinx F. The use of standardized patients in research in general practice. Fam Pract 1997, 14(1):58–62. doi: 10.1093/fampra/14.1.58.
10. Akuma I, Vaswani V. Exploring Partial Disclosure in Research: Challenges, Justifications, and Recommendations for Ethical Oversight. Asian Bioeth Rev 2024. doi: 10.1007/s41649-024-00311-7.
11. Verbeke K, Piasecki J, Baeyens D, Krawczyk T, Borry P. Truthfulness as the basis for ethical safeguards in deceptive research: An interview study with researchers. Account Res 2024. doi: 10.1080/08989621.2024.2362777.
12. Wilbur K, Elmubark A, Shabana S. Systematic review of standardized patient use in continuing medical education. Journal of Continuing Education in the Health Professions. 2018, 38(1):3–10. doi: 10.1097/CEH.0000000000000190.
13. Remmen R, Seuntjens L, Paulus D, Pestiaux D. Kwaliteitszorg in de huisartsgeneeskunde. Huisarts Nu 2008, 37(10):547–54.
14. Rhodes K V, Miller FG. Simulated Patient Studies: An Ethical Analysis. Milbank Quarterly 2012, 90(4):706–24. doi: 10.1111/j.1468-0009.2012.00680.x.
15. Verbeke K, Krawczyk T, Baeyens D, Piasecki J, Borry P. Informed Consent and Debriefing When Deceiving Participants: A Systematic Review of Research Ethics Guidelines. Journal of Empirical Research on Human Research Ethics 2023, 18(3):118–33. doi: 10.1177/15562646231173477.
16. Ball SL, Weiner SJ, Schwartz A, Altman L, Binns-Calvey A, Chan C et al. Implementation of a patient-collected audio recording audit & feedback quality improvement program to prevent contextual error: stakeholder perspective. BMC Health Serv Res 2021, 21(1):891. doi: 10.1186/s12913-021-06921-3.
17. Schwartz A, Peskin S, Spiro A, Weiner SJ. Direct observation of depression screening: Identifying diagnostic error and improving accuracy through unannounced standardized patients. Diagnosis 2020, 7(3):251–6. doi: 10.1515/dx-2019-0110.
18. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. The BMJ 2021, 372. doi: 10.1136/bmj.n160.

19. Liao J, Chen Y, Cai Y, Zhan N, Sylvia S, Hanson K et al. Using smartphone-based virtual patients to assess the quality of primary healthcare in rural China: Protocol for a prospective multicentre study. BMJ Open 2018, 8(7). doi: 10.1136/bmjopen-2017-020943.

20. Zeng M, Cai Y, Cao J, He Q, Wang X, Lu Y et al. The Agreement Between Virtual Patient and Unannounced Standardized Patient Assessments in Evaluating Primary Health Care Quality: Multicenter, Cross-sectional Pilot Study in 7 Provinces of China. J Med Internet Res 2022, 24(12):1–13. doi: 10.2196/40082.

21. Schwartz A, Peskin S, Spiro A, Weiner SJ. Impact of Unannounced Standardized Patient Audit and Feedback on Care, Documentation, and Costs: an Experiment and Claims Analysis. J Gen Intern Med 2021, 36(1):27–34. doi: 10.1007/s11606-020-05965-1.

22. Hornung RL, Hansen LA, Sharp LK, Poorsattar SP, Lipsky MS. Skin cancer prevention in the primary care setting: Assessment using a standardized patient. Pediatr Dermatol 2007, 24(2):108–12. doi: 10.1111/j.1525-1470.2007.00353.x.

23. Zabar S, Hanley K, Stevens D, Murphy J, Burgess A, Kalet A et al. Unannounced standardized patients: A promising method of assessing patient-centered care in your health care system. BMC Health Serv Res 2014, 14:1–8. doi: 10.1186/1472-6963-14-157.

24. Franz CE, Epstein R, Miller KN, Brown A, Song J, Feldman M et al. Caught in the Act? Prevalence, Predictors, and Consequences of Physician Detection of Unannounced Standardized Patients. Health Serv Res 2006, 41(6):2290–302. doi: 10.1111/j.1475-6773.2006.00560.x.

25. Daniels B, Dolinger A, Bedoya G, Rogo K, Goicoechea A, Coarasa J et al. Use of standardised patients to assess quality of healthcare in Nairobi, Kenya: A pilot, cross-sectional study with international comparisons. BMJ Glob Health 2017, 2(2). doi: 10.1136/bmjgh-2017-000333.

26. Srinivasan M, Franks P, Meredith LS, Fiscella K, Epstein RM, Kravitz RL. Connoisseurs of Care? Unannounced Standardized Patients' Ratings of Physicians. 2006.

27. Giesen P, Ferwerda R, Tijssen R, Mokkink H, Drijver R, Van Den Bosch W et al. Safety of telephone triage in general practitioner cooperatives: Do triage nurses correctly estimate urgency? Qual Saf Health Care 2007, 16(3):181–4. doi: 10.1136/qshc.2006.018846.

28. Beaulieu MD, Rivard M, Hudon E, Saucier D, Remondin M, Favreau R. Using standardized patients to measure professional performance of physicians. International Journal for Quality in Health Care 2003, 15(3):251–9. doi: 10.1093/intqhc/mzg037.

29. Franks P, F Jerant A, Fiscella K, G Shields C, J Tancredi D, M Epstein R. Studying physician effects on patient outcomes: Physician interactional style and performance on quality of care indicators. Soc Sci Med 2006, 62(2):422–32. doi: 10.1016/j.socscimed.2005.05.027.

30. Feng B, Srinivasan M, Hoffman JR, Rainwater JA, Griffin E, Dragojevic M et al. Physician communication regarding prostate cancer screening: Analysis of unannounced standardized patient visits. Ann Fam Med 2013, 11(4):315–23. doi: 10.1370/afm.1509.

31. Peremans L, Rethans JJ, Verhoeven V, Coenen S, Debaene L, Meulemans H et al. Empowering patients or general practitioners? A randomised clinical trial to improve quality in reproductive health care in Belgium. European Journal of Contraception and Reproductive Health Care 2010, 15(4):280–9. doi: 10.3109/13625187.2010.492882.

32. van Onna M, Gorter S, Maiburg B, Waagenaar G, van Tubergen A. GPs' patterns of clinical assessment when faced with a patient suspected for spondyloarthritis: A prospective educational intervention study. BJGP Open 2017, 1(1). doi: 10.3399/bjgpopen17X100689.

33. Xie Y, He W, Wan Y, Luo H, Cai Y, Gong W et al. Validity of patients' online reviews at direct-to-consumer teleconsultation platforms: A protocol for a cross-sectional study using unannounced standardised patients. BMJ Open 2023, 13(5):1–11. doi: 10.1136/bmjopen-2023-071783.

34. Stevens SL, McManus RJ, Stevens RJ. Current practice of usual clinic blood pressure measurement in people with and without diabetes: A survey and prospective €mystery shopper' study in UK primary care. BMJ Open 2018, 8(4):1–8. doi: 10.1136/bmjopen-2017-020589.

35. Altshuler L, Fisher H, Wilhite J, Phillips Z, Holmes I, Greene RE et al. Comparison of Primary Care Patients' and Unannounced Standardized Patients' Perceptions of Care. J Patient Exp 2023, 10:1–10. doi: 10.1177/23743735231158940.

36. Xu D, Pan J, Dai X, Hu M, Cai Y, He H et al. Comparing quality of primary healthcare between public and private providers in China: study protocol of a cross-sectional study using unannounced standardised patients in seven provinces of China. BMJ Open 2021, 11(1):1–13. doi: 10.1136/bmjopen-2020-040792.

37. Su M, Zhou Z, Si Y, Fan X. The Association Between Patient-Centered Communication and Primary Care Quality in Urban China: Evidence From a Standardized Patient Study. Front Public Health 2022, 9(February). doi: 10.3389/fpubh.2021.779293.

38. Shirazi M, Lonka K, Parikh S V., Ristner G, Alaeddini F, Sadeghi M et al. A tailored educational intervention improves doctor's performance in managing depression: A randomized controlled trial. J Eval Clin Pract 2013, 19(1):16–24. doi: 10.1111/j.1365-2753.2011.01761.x.

39. Bachmann MO, Colvin MSE, Nsibande D, Connolly C, Curtis B. Quality of primary care for sexually transmitted diseases in Durban, South Africa: influences of patient, nurse, organizational and socioeconomic characteristics. Int J STD AIDS 2004, 15(6):388–94. doi: 10.1258/095646204774195245.

40. Sikombe K, Pry JM, Mody A, Rice B, Bukankala C, Eshun-Wilson I et al. Comparison of patient exit interviews with unannounced standardised patients for assessing HIV service delivery in Zambia: a study nested within a cluster randomised trial. BMJ Open 2023, 13(7). doi: 10.1136/bmjopen-2022-069086.

41. Peabody JW, Luck J, Glassman P, Jain S, Hansen J, Spell M et al. Measuring the Quality of Physician Practice by Using Clinical Vignettes: A Prospective Validation Study. Ann Intern Med 2004, 141(10):771–80. doi: 10.7326/0003-4819-141-10-200411160-00008.

42. Bryant KA, Wesley GC, Wood JA, Hines C, Marshall GS. Use of standardized patients to examine physicians' communication strategies when addressing vaccine refusal: A pilot study. Vaccine 2009, 27(27):3616–9. doi: 10.1016/j.vaccine.2009.03.048.

43. Fiscella K, Epstein RM, Griggs JJ, Marshall MM, Shields CG. Is physician implicit bias associated with differences in care by patient race for metastatic cancer-related pain? PLoS One 2021, 16(10 October):1–16. doi: 10.1371/journal.pone.0257794.

44. Piribauer F, Thaler K, Harris MF. Covert checks by standardised patients of general practitioners' delivery of new periodic health examinations: Clustered cross-sectional study from a consumer organisation. BMJ Open 2012, 2(4). doi: 10.1136/bmjopen-2011-000744.

45. Pasini A, Rigon G, Vaona A. A cross-sectional study of the quality of telephone triage in a primary care out-of-hours service. J Telemed Telecare 2015, 21(2):68–72. doi: 10.1177/1357633X14566573.

46. Liu H, Li H, Teuwen DE, Sylvia S, Shi H, Rozelle S et al. Irrational Use of Medicine in the Treatment of Presumptive Asthma Among Rural Primary Care Providers in Southwestern China. Front Pharmacol 2022, 13(February):1–11. doi: 10.3389/fphar.2022.767917.

47. Derkx HP, Rethans JJE, Muijtjens AM, Maiburg BH, Winkens R, Van Rooij HG et al. Quality of clinical aspects of call handling at Dutch out of hours centres: Cross sectional national study. BMJ 2008, 337(7671):676–9. doi: 10.1136/bmj.a1264.

48. Culver JO, Bowen DJ, Reynolds SE, Pinsky LE, Press N, Burke W. Breast cancer risk communication: Assessment of primary care physicians by standardized patients. Genetics in Medicine 2009, 11(10):735–41. doi: 10.1097/GIM.0b013e3181b2e5eb.

49. Wu Y, Zhou H, Ma X, Shi Y, Xue H, Zhou C et al. Using standardised patients to assess the quality of medical records: An application and evidence from rural China. BMJ Qual Saf 2020, 29(6):491–8. doi: 10.1136/bmjqs-2019-009890.

50. de Castro F, Barrientos-Gutierrez T, Braverman-Bronstein A, Santelli J, Place JM, Eternod-Arámburu M et al. Adolescent Access to Information on Contraceptives: A Mystery Client Study in Mexico. Journal of Adolescent Health 2018, 62(3):265–72. doi: 10.1016/j.jadohealth.2017.08.001.

51. Kwan A, Boone CE, Sulis G, Gertler PJ. Do private providers give patients what they demand, even if it is inappropriate? A randomised study using unannounced standardised patients in Kenya. BMJ Open 2022, 12(3). doi: 10.1136/bmjopen-2021-058746.

52. Houwink EJF, Muijtjens AMM, Van Teeffelen SR, Henneman L, Rethans JJ, Van Der Jagt LEJ et al. Effectiveness of oncogenetics training on general practitioners' consultation skills: A randomized controlled trial. Genetics in Medicine 2014, 16(1):45–52. doi: 10.1038/gim.2013.69.

53. Meuwissen LE, Gorter AC, Kester AD, Knottnerus JA. Does a competitive voucher program for adolescents improve the quality of reproductive health care? A simulated patient study in Nicaragua. BMC Public Health 2006, 6(1):204. doi: 10.1186/1471-2458-6-204.

54. Lorenzo A, Schildt P, Lorenzo M, Falcoff H, Noela F. Acute low back pain management in primary care: A simulated patient approach. Fam Pract 2015, 32(4):436–41. doi: 10.1093/fampra/cmv030.

55. Huibers L, Giesen P, Smits M, Mokkink H, Grol R, Wensing M. Nurse telephone triage in Dutch out-of-hours primary care: the relation between history taking and urgency estimation. Eur J Emerg Med 2012, 19(5):309–15. doi: 10.1097/MEJ.0b013e32834d3e67.

56. Gillespie C, Wilhite JA, Hanley K, Hardowar K, Altshuler L, Fisher H et al. Addressing social determinants of health in primary care: a quasi-experimental study using unannounced standardised patients to evaluate the impact of audit/feedback on physicians' rates of identifying and responding to social needs. BMJ Qual Saf 2023, 32(11):632–43. doi: 10.1136/bmjqs-2021-013904.

57. Amini H, Shakiba A, Sharifi V, Shirazi M, Sadeghi M, Abolhasani F et al. Evaluation of the performance of general practitioners in a collaborative care program by employing simulated patients. Soc Psychiatry Psychiatr Epidemiol 2016, 51(9):1311–9. doi: 10.1007/s00127-016-1226-3.

58. Price-Haywood EG, Harden-Barrios J, Cooper LA. Comparative effectiveness of audit-feedback versus additional physician communication training to improve cancer screening for patients with limited health literacy. J Gen Intern Med 2014, 29(8):1113–21. doi: 10.1007/s11606-014-2782-4.

59. Wilhite JA, Phillips Z, Altshuler L, Hernan G, Lambert R, Nicholson J et al. A systematic review of the use of unannounced standardized patients (USPs) in clinical settings: A call for more detailed quality and fidelity descriptions and expansion to new areas. Patient Educ Couns 2025, 130:108437. doi: 10.1016/J.PEC.2024.108437.

60. Aujla N, Chen YF, Samarakoon Y, Wilson A, Grolmusová N, Ayorinde A et al. Comparing the use of direct observation, standardized patients and exit interviews in low- And middle-income countries: A systematic review of methods of assessing quality of primary care. Health Policy Plan 2021, 36(3):341–56. doi: 10.1093/heapol/czaa152.

61. Barrett DW, Neuberg SL. Suspicion About Suspicion Probes: Ways Forward. Perspectives on Psychological Science 2023. doi: 10.1177/17456916231195855.

62. High-Income Countries 2025 [Internet]. 2025. Available from: https://worldpopulationreview.com/country-rankings/high-income-countries [Accessed April 2025].

63. Kingori P, Douglas Jones R. Revelation or confirmation? Med Anthropol Theory 2020, 7(2):214–29. doi: 10.17157/mat.7.2.776.

64. Hallinan D, Boehm F, Külpmann AI, Elson M. (Un)informed consent in Psychological Research: An empirical study on consent in psychological research and the GDPR [Internet]. Journal of Open Access to Law. 2023, 11:1–28. doi: 10.31234/osf.io/ugxam.

65. Laurijssen SJ, van der Graaf R, van Dijk WB, Schuit E, Groenwold RH, Grobbee DE et al. When is it impractical to ask informed consent? A systematic review. Clinical Trials 2022, 19(5):545–60. doi: 10.1177/17407745221103567.

66. Siminoff LA, Rogers HL, Waller AC, Harris-Haywood S, Esptein RM, Carrio FB et al. The advantages and challenges of unannounced standardized patient methodology to assess healthcare communication. Patient Educ Couns 2011, 82(3):318–24. doi: 10.1016/j.pec.2011.01.021.

67. Hodson G. Reconsidering reconsent: Threats to internal and external validity when participants reconsent after debriefing. British Journal of Psychology 2022, 113(3):853–71. doi: 10.1111/bjop.12561.

68. De Wolf L. Ontluisterende steekproef 'Factcheckers': 14 van 20 bezochte dokters schrijven onterecht ziektebriefje voor 1 dag [Internet]. VRT NWS. 2022. Available from: https://www.vrt.be/vrtnws/nl/2022/02/14/factcheckers-ziektebriefjes/ [Accessed April 2025].

69. Nuyts J. Wij deden undercover de test: geraak je nog aan Ozempic om af te vallen? [Internet]. Het Laatste Nieuws. 2023. Available from: https://www.hln.be/binnenland/wij-deden-undercover-de-test-geraak-je-nog-aan-ozempic-om-af-te-vallen~a1584cd4/?referrer=https%3A%2F%2Fwww.google.com%2F [Accessed April 2025].

70. Belgische Senaat. Antwoord op schriftelijke vraag nr. 5-7258 van Bert Anciaux aan de minister van Sociale Zaken en Volksgezondheid [Internet]. 2013. Available from: https://www.senate.be/www/?MIval=Vragen/SchriftelijkeVraag&LEG=5&NR=7258&LANG=nl [Accessed April 2025].

71. Rommers W. “Allemaal strafbare feiten”: overheid wil plastische chirurgen controleren met ‘mystery shoppers’ [Internet]. Gazet van Antwerpen. 2025. Available from: https://www.gva.be/binnenland/allemaal-strafbare-feiten-overheid-wil-plastische-chirurgen-controleren-met-mystery-shoppers/37907001.html [Accessed April 2025].

72. Rijksinstituut voor Ziekte- en Invaliditeitsverzekering (RIZIV). Richtlijnen I – Terugbetaling van prestaties verricht in het kader van wetenschappelijk onderzoek of klinische proeven [Internet]. Brussel; 2016. Available from: https://www.riziv.be/SiteCollectionDocuments/ib2016_1_richtlijnen_I_terugbetaling-van-prestaties.pdf [Accessed April 2025].

73. Fiscella K, Franks P, Srinivasan M, Kravitz RL, Epstein R. Ratings of physician communication by real and standardized patients. Ann Fam Med 2007, 5(2):151–8. doi: 10.1370/afm.643.

74. Si Y, Chen G, Su M, Zhou Z, Yip W, Chen X. The Impact of Physician-Patient Gender Match on Healthcare Quality : An Experiment in China. 2023, :1–11.

75. Geraghty EM, Franks P, Kravitz RL. Primary care visit length, quality, and satisfaction for standardized patients with depression. J Gen Intern Med 2007, 22(12):1641–7. doi: 10.1007/s11606-007-0371-5.

76. Wang Q, Adhikari SP, Wu Y, Sunil TS, Mao Y, Ye R et al. Consultation length, process quality and diagnosis quality of primary care in rural China: A cross-sectional standardized patient study. Patient Educ Couns 2022, 105(4):902–8. doi: 10.1016/j.pec.2021.08.006.

APPENDIX

APPENDIX I. DETAILED SEARCH STRATEGY

Pubmed search strategy

1. "simulated patient*"[tiab] OR "patient simulation"[MeSH] OR "patient simulation"[tiab]
2. "simulation training"[MeSH Terms] OR "simulation training"[tiab]
3. "standardised patient*"[tiab] OR "standardized patient*"[tiab]
4. "mystery shop*"[tiab] OR "mystery patient*"[tiab] OR "secret shop*"[tiab]
5. "pseudopatient*"[tiab] OR "pseudo customer*"[tiab] OR "pseudo patron*"[tiab]
6. "covert participant*"[tiab] OR "covert patient*"[tiab]
7. #1 OR #2 OR #3 OR #4 OR #5 OR #6
8. "General Practice"[MeSH] OR "General Practitioners"[MeSH] OR "general practi*"[tiab]
9. "Physicians, Family"[MeSH] OR "Family Practice"[MeSH] OR "family medicine"[tiab] OR "family practi*"[tiab]
10. "Physicians, Primary Care"[MeSH] OR "primary care"[tiab]
11. #8 OR #9 OR #10
12. "Ethics"[MeSH Terms] OR "ethic*"[All Fields]
13. "Quality Improvement"[MeSH] OR "Quality Indicators, Health Care"[MeSH] OR "Quality Assurance, Health Care"[MeSH] OR "quality in health care"[tiab] OR "quality health care"[tiab:~1] OR "quality in healthcare"[tiab] OR "quality healthcare"[tiab:~1]
14. #12 OR #13
15. #7 AND #11 AND #14
16. english[la] OR dutch[la]
17. 2003:2023[dp]
18. #16 AND #17
19. #15 AND #18

Web of Science search strategy

1. TS=("simulated patient*" OR "patient simulation")
2. TS=("simulation training")
3. TS=("standardised patient*" OR "standardized patient*")
4. TS=("mystery shop*" OR "mystery patient*" OR "secret shop*")
5. TS=("pseudopatient*" OR "pseudo customer*" OR "pseudo patron*")
6. TS=("covert participant*" OR "covert patient*")
7. #1 OR #2 OR #3 OR #4 OR #5 OR #6
8. TS=("general practi*")
9. TS=("family physician*" OR "family medicin*" OR "family practi*")
10. TS=("primary care") OR WC=(Primary Health Care)
11. #8 OR #9 OR #10
12. TS=("ethic*") OR WC=(Ethics) OR WC=(Medical Ethics)
13. TS=("quality improv*" OR "quality indicat*" OR "quality assur*" OR "quality in health care" OR "quality health care" OR "quality healthcare")
14. #12 OR #13
15. #7 AND #11 AND #14
16. LA=(English OR Dutch)
17. PY=(2003-2023)
18. #16 AND #17
19. #15 AND #18

APPENDIX II. RESEARCH QUESTIONS

1a. Country?

1b. Region?

1c. Year study took place?

2a. What is the research purpose of the study?

2b. What is being studied (e.g., results)?

3a. Are participants informed of certain true aspects of the deceptive mystery shopping method (e.g., that mystery shoppers are coming)?

3b. Is there informed consent?

4a. Is there debriefing where the truth is told after the study? Are incurred negative study consequences identified and corrected?

4b. After debriefing, additional data collection on how participants experienced aspects of the mystery shopping study, ...?

5a. Possibility for participants to withdraw their data after participation?

5b. Is data protection in place?

6. Is there ethical committee approval?

7a. Who does the deception and how were they acquired?

7b. How were the standardized patients trained or educated?

7c. How were individuals (or physicians) chosen to be studied with (un)announced standardized patient?

8a. How do the standardized patients mislead?

8b. How was data collected from the SP? Checklists, audio recordings, video recordings, …?

9a. Is there an argument as to why a misleading method was or was not used or what are the advantages of using standardized patients?

9b. Is it argued why the burden or harm to subjects is acceptable? Or are measures taken to minimize harm?

10a. Are burden or harm to subjects mentioned, or negative consequences that persist after participation in the study? (burden to subject)

10b. Is burden or harm to the mystery shopper or standardized patient mentioned, or negative consequences that persist after participation in the study? Are measures taken to protect them from negative consequences?

11a. Did the deception include communication of false information, or information that was very far from reality? (What case study? What scenario?)

11b. Is there any mention of how the mystery shopping was made credible?

11c. Does it speak of participants having suspicions (“suspicions”) of what was really going on, and thus seeing through the deception?

12. Was the mystery shopping method standardized, or more flexible?

13. Educational component associated with this study? Do subjects or physicians learn anything?

APPENDIX III. OVERVIEW OF INCLUDED ARTICLES (N = 43)

Study ID	Design	Region, Country	Study period	Subjects; Setting	N° USP visits
Beaulieu 2003 [28]	RCT / Experimental	Montreal, Canada	1998-1999	GPs; General Practice	179
Bachmann 2004 [39]	Cross-sectional	Durban, South Africa	2004	Nurses; Primary care clinics	271
Peabody 2004 [41]	Quasi-Experimental	California, USA	2000	Internal medicine physicians; Primary care clinics	480
Franks 2006 [29]	Cross-sectional	New York City, USA	1999-2002	GPs and Internal medicine physicians; Primary care clinics	200
Franz 2006 [24]	RCT / Experimental	Sacramento, CA; San Francisco, CA; Rochester, NY, USA	2003-2004	GPs and Internal medicine physicians; Primary care clinics	298
Meuwissen 2006 [53]	Quasi-Experimental	Managua, Nicaragua	2000-2002	PCPs (unspecified); Primary care clinics	65
Srinivasan 2006 [26]	Quasi-Experimental	Rochester, NY; Sacramento, CA; San Francisco, CA, USA	2000-2001 2003-2004	GPs and Internal medicine physicians; Primary care clinics	484
Fiscella 2007 [73]	Observational	Rochester, NY, USA	2000-2001	GPs and Internal medicine physicians; Primary care clinics	193
Geraghty 2007 [75]	RCT / Experimental	Sacramento, CA; San Francisco, CA; New York, USA	2003-2004	GPs and Internal medicine physicians; Primary care clinics	294
Giesen 2007 [27]	Cross-sectional	The Netherlands	2003	Nurses; Telemedicine	370
Hornung 2007 [22]	Observational	Chicago, IL, USA	2007	GPs and Internal medicine physicians; Primary care clinics	15
Derkx 2008 [47]	Cross-sectional	12 provinces, The Netherlands	2004-2008	Nurses; Telemedicine	357
Bryant 2009 [42]	Observational	Louisville, KY, USA	2008	Mix of PCPs; Mixed: General Practice and Primary care clinics	9
Culver 2009 [48]	RCT / Experimental	Seattle–King County, WA, USA	<2009	GPs and Internal medicine physicians; Primary care clinics	86
Peremans 2010 [31]	RCT / Experimental	Flanders, Belgium	2005	GPs; General Practice	86
Huibers 2012 [55]	Cross-sectional	The Netherlands	2010	Nurses; Telemedicine	336
Piribauer 2012 [44]	Cross-sectional	Vienna, Austria	2008	GPs; General Practice	40
Feng 2013 [30]	RCT / Experimental	California, USA	2007-2008	GPs and Internal medicine physicians; Primary care clinics	118

Study ID	Design	Region, Country	Study period	Subjects; Setting	N° USP visits
Shirazi 2013 [38]	RCT / Experimental	Tehran, Iran	<2011	GPs; General Practice	354
Houwink 2014 [52]	RCT / Experimental	The Netherlands	2001-2012	GPs; General Practice	168
Price-Haywood 2014 [58]	RCT / Experimental	New Orleans, LA, USA	2008-2012	GPs and Internal medicine physicians; Primary care clinics	54
Zabar 2014 [23]	Observational	New York City, USA	2009	Internal medicine physicians; Primary care clinics	60
Lorenzo 2015 [54]	Observational	Paris, France	2013	GPs; General Practice	30
Pasini 2015 [45]	Cross-sectional	Verona, Italy	2001	GPs; Telemedicine	40
Amini 2016 [57]	Observational	Tehran, Iran	2013	GPs; General Practice	104
Daniels 2017 [25]	Cross-sectional	Nairobi, Kenya	2013-2014	Mix of PCPs; Primary care clinics	166
van Onna 2017 [32]	Quasi-Experimental	South Netherlands, Netherlands	<2016	GPs; General Practice	340
de Castro 2018 [50]	Cross-sectional	Mexico City, Mexico	2015	Mix of PCPs; Mixed: Primary care clinics and Pharmacies	327
Stevens 2018 [34]	Observational	United Kingdom	2015-2016	GPs and nurses; General Practice	n/a
Wu 2020 [49]	Cross-sectional	Rural areas, China	2015	PCPs (unspecified); Township Health Centers	610
Fiscella 2021 [43]	RCT / Experimental	New York; Indiana; Michigan, USA	2012-2016	Mix of primary and secondary care providers; Primary care clinics	183
Schwartz 2021 [21]	Quasi-Experimental	New Jersey, USA	2016-2019	PCPs (unspecified); Primary care clinics	217
Xu 2021 [36]	Cross-sectional	Gansu; Guizhou; Hunan; Inner Mongolia; Shaanxi; Sichuan; Guangdong, China	2020	Mix of primary and secondary care providers; Village clinics (tier 1) and Township Health Centers (tier 2)	2200
Kwan 2022 [51]	RCT / Experimental	35 counties, Kenya	2019	Mix of PCPs; Primary care clinics	200
Liu 2022 [46]	Quasi-Experimental	Rural village providers, Yunnan Province, China	2017-2018	PCPs (unspecified); Village clinics (tier1)	104
Su 2022 [37]	Cross-sectional	Xi'an, China	2017-2018	PCPs (unspecified); Community Health Centers	492
Wang 2022 [76]	Cross-sectional	Rural areas, Anhui; Shaanxi; Sichuan Provinces, China	2015	PCPs (unspecified); Township Health Centers	413
Zeng 2022 [20]	Cross-sectional	Gansu; Guizhou; Hunan; Inner Mongolia; Shaanxi; Sichuan; Guangdong, China	2019-2021	PCPs (unspecified); Community Health Centers	720

Study ID	Design	Region, Country	Study period	Subjects; Setting	N° USP visits
Altshuler 2023 [35]	Cross-sectional	New York City, USA	2017-2018	Internal medicine physicians; Primary care clinics	177
Gillespie 2023 [56]	Quasi-Experimental	New York City, USA	2018-2019	Internal medicine physicians; Mixed: Primary care clinics and Hospital	291
Si 2023 [74]	RCT / Experimental	Xi'an, China	2017-2018	PCPs (unspecified); Community Health Centers	492
Sikombe 2023 [40]	Cross-sectional	Lusaka, Zambia	2019-2021	Mix of PCPs; Primary care clinics	1111
Xie 2023 [33]	Cross-sectional	Online, China	2023	PCPs (unspecified); Telemedicine	506

GPs, general practitioners; PCPs, primary care providers; RCT, randomized controlled trial; USP, unannounced standardized patient

APPENDIX IV. SYSTEMATIC DATA EXTRACTION SHEET

See appended spreadsheet: StijnPeters_TiboVerplancke_appendixIV.xlsx

APPENDIX V. ETHICAL APPROVAL AND SAFEGUARDS

Study ID	Ethical approval	USP safeguards	PCP safeguards			
			Compensation	Data protection	Debriefing	Other safeguards
Beaulieu 2003	IRB approval		Yes, per visit	Data de-identified after collection		
Bachmann 2004	Ethics committee of the South African Medical Research Council			Data was not identifiably collected.		
Peabody 2004	Not mentioned.					
Franks 2006	IRB approval	Trained to avoid referral to emergency department; trained to avoid administration of medication	Yes, $100 per USP visit			Max 2 USPs per physician
Franz 2006	IRB approval				Physicians completed a Clinician Background Questionnaire (CBQ) at the end of the study and were then debriefed.	
Meuwissen 2006	IRB approval	Decline vaginal examinations	Yes, per visit	Data de-identified after collection		
Srinivasan 2006	IRB approval		Yes, $375 to $400 for participation			
Fiscella 2007	IRB approval	Reject invasive procedures	Yes, $100 per USP visit; Office staff $100 for logistics			Two visits per PCP; Max two physicians per practice
Geraghty 2007	IRB approval					
Giesen 2007	Not mentioned				No structured feedback, end of call USP revealed themselves	
Hornung 2007	IRB approval		Yes, cash per visit	Data was not identifiably collected.	No feedback given, called afterwards and asked if they could identify the USP	
Derkx 2008	Not needed				No debriefing, letter sent and asked if they had seen a USP beyond 12 months.	

Study ID	Ethical approval	USP safeguards	PCP safeguards			
			Compensation	Data protection	Debriefing	Other safeguards
Bryant 2009	IRB approval		Yes, usual rate		No structured feedback, post-study phone call was made to all participating physicians to detect whether they had identified the USP	Limiting to one participating physician per practice to avoid recognition and disruption; Avoid urgent care scenarios
Culver 2009	IRB approval	Trained to avoid detection	Yes, per visit		After the study, a mailing was sent to participating physicians announcing its completion and asking them to complete a detection questionnaire. A second mailing later revealed which patient was an SP.	Financial reimbursement
Peremans 2010	IRB approval	USPs had access to a support team if they experienced a negative consultation			No debriefing, detection form sent two months after last visit	
Huibers 2012	Not necessary according to Dutch law.				No structured feedback, end of call SP revealed themselves	
Piribauer 2012	Ethics committee the Statutory Public Health Ethics Committee of the City of Vienna		Yes, per visit	Data de-identified after collection		Healthcare providers were collectively notified (everyone equal before the law, no opt-out)
Feng 2013	IRB approval					
Shirazi 2013	IRB approval		CME, participation fee for CME waived			No administrative burden
Houwink 2014	IRB approval		CME and a book/book voucher		Informal comments after training	Physician receive genetic training
Price-Haywood 2014	IRB approval	Financial compensation; trained to avoid detection and harm.	Yes, $100 per USP visit, up to 20 CME credits from Tulane's CME Office		Feedback provided at end of visit to intervention physicians(structured verbal); access to WebSP to review their USP ratings, and written reports with narrative summaries of USP perceptions; all study physicians received two annual cancer screening status reports and aggregate baseline patient ratings of their communication.	
Zabar 2014	Not mentioned. Data was collected in an NYU IRB-approved research registry.			Data was not identifiably collected.		SPs did not disrupt normal clinical workflow

Study ID	Ethical approval	USP safeguards	PCP safeguards			
			Compensation	Data protection	Debriefing	Other safeguards
Lorenzo 2015	Not needed:		Yes, per visit, covered by author			
Pasini 2015	IRB approval	Avoided highly sensitive or potentially distressing cases.		Physician identities were not linked to performance data.	No structured feedback, post-study phone call was made to all participating physicians to detect whether they had identified the USP	Calls were made at appropriate times to avoid disruption (not late at night or during shift changes); Avoided cases requiring urgent or emergency interventions
Amini 2016	IRB approval			Data de-identified after collection; only aggregated result reported		
Daniels 2017	Ethical review board African Medical and Research Foundation & Government of Kenya	Reject invasive procedures; trained to avoid detection or dangerous situations.	Yes, whatever was charged	Data is not publicly available		Brief interactions; USPs let seriously ill patients go ahead in the queue
Van Onna 2017	IRB approval	Non-invasive protocol, reject probing or extended physical examinations				
De Castro 2018	Ethics committee of the National Institute of Public Health in Cuernavaca, Mexico	Monitored to ensure well-being; scripts were tested; parental consent was obtained for -18 year old.				
Stevens 2018	IRB approval		Yes, health insurance			
Wu 2020	IRB approval	Reject invasive procedures	Yes, per visit			
Fiscella 2021	IRB approval	Reject invasive procedures		Data is not publicly available; all users need to sign a data use agreement.		

Study ID	Ethical approval	USP safeguards	PCP safeguards			
			Compensation	Data protection	Debriefing	Other safeguards
Schwartz 2021	IRB approval		No, voluntarily	Data de-identified after collection	Feedback after two baseline visits and again after two follow-up visits; practice-wide feedback in the form of reports with visualized performance metrics, documentation fidelity analysis, and coaching calls; designed as a quality improvement tool rather than punitive assessment.	Physicians could receive MOC (Maintenance of Certification) credit from the American Board of Internal Medicine or American Board of Family Medicine.
Xu 2021	IRB approval	Trained to avoid detection and harm		Data was kept secure in a platform; clinician data de-identified.		Non-emergency settings only
Kwan 2022	IRB approval	Reject invasive procedures; monitored to ensure well-being; trained to avoid detection and harm	Yes, per visit	Data de-identified after collection; publicly available anonymized	The researchers revisited the clinic afterwards, explained study and addressed concerns.	After the study PCPs were given an opportunity to ask questions and raise concerns
Liu 2022	IRB approval	Reject invasive procedures; trained to avoid detection and harm	Yes, per visit			
Su 2022	IRB approval		Yes, all fees and medical expenses were paid		No feedback, but is praised as a way to improve	To avoid overburdening busy physicians, the questionnaire was left with them to complete
Wang 2022	IRB approval	Reject invasive procedures, case selection which do not require physical signs or invasive procedures				
Zeng 2022	IRB approval	Tuberculosis cases were excluded to protect USPs from unnecessary physical examinations and potential harm		Data is not publicly available, available upon request.		
Altshuler 2023	IRB approval		Yes, per visit	Data was not identifiably collected; data was kept secure in a platform.	Aggregate SP performance data was periodically reviewed with clinic staff and leadership for quality improvement.	

Study ID	Ethical approval	USP safeguards	PCP safeguards			
			Compensation	Data protection	Debriefing	Other safeguards
Gillespie 2023	IRB approval			Data de-identified after collection	Audit/feedback reports summarizing their screening and response rates; email, team meetings, and postings in staff conference rooms	Multiple modes of dissemination (email, meetings, posted materials) were used to reduce disruption, and feedback was integrated into existing team meetings rather than requiring additional sessions
Si 2023	IRB approval	Invasive tests were not necessary; trained to avoid detection and harm.	Yes, all fees and medical expenses were paid			
Sikombe 2023	IRB approval			Data is not publicly available; availaible on reasonable request.		The trained exit clients (TEC) method followed approaches that minimize harm to PCPs and USPs.
Xie 2023	IRB approval	No physical examinations or invasive procedures are required	Yes, all fees and medical expenses were paid	Data was kept secure in a platform; physicians identities remain confidential; data analyzed at an aggregated level.		

IRB, institutional review board; PCP, primary care provider; USP, unannounced standardized patient

APPENDIX VI. INFORMED CONSENT

Consent approach / Study ID	Informed in advance?	Way of informing
(Written) Informed consent[a]		
Beaulieu et al. 2003	Yes	Informed that they could be visited in the next 6 months
Peabody et al. 2004	Yes	Informed that they could be visited in the next 12 months
Bachmann et al. 2004	Yes	Signed consent was obtained from all professional nurses and clinic managers prior to the visits by simulated patients.
Franks et al. 2006	Yes	Informed that they could be visited in the next 12 months
Srinivasan et al. 2006	Yes	Informed about study method
Franz et al. 2006	Yes	Informed that they could be visited in 'several' months
Geraghty et al. 2007	Yes	Informed about study method
Giesen et al. 2007	Yes	The triage nurses were informed about the study's purpose and all consented to participate
Hornung et al. 2007	Yes	Informed that they could be visited in the next 12 months
Fiscella et al. 2007	Yes	Informed that they could be visited in the next 12 months
Culver et al. 2009	Yes	Physicians received mailed study info and gave consent for an unannounced SP visit with audio recording and reimbursement
Bryant et al. 2009	Yes	Informed that they could be visited in the next 6 months
Huibers et al. 2012	Yes	All triage nurses at the GPCs were informed and consented
Feng et al. 2013	Yes	Informed that they could be visited in the next 3 months
Shirazi et al. 2013	Yes	Informed about study method
Price-Haywood et al. 2014	Yes	Informed that they could be visited in the next 12 months
Houwink et al. 2014	Yes	Informed about study method, USP makes itself known prior the consultation
Zabar et al. 2014	Yes	Informed consent was both for their own training and for research purposes
Pasini et al. 2015	Yes	The doctors received an information letter and all consented to participate
Daniels et al. 2017	Yes	Informed that they could be visited in the next 6 months
Van et al. 2017	Yes	Informed about study method, USP makes itself known prior the consultation
Fiscella et al. 2021	Yes	They were informed they would receive two unannounced SP visits
Schwartz et al. 2021	Yes	Informed that they could be visited in the next 18 months
Wang et al. 2022	Yes	Informed that they could be visited in the next 6 months
Su et al. 2022	Yes/partial	Not specified whether physicians were informed they would be visited by a USP
Sikombe et al. 2023	Yes	Real patient exit client surveys; Trained Exit Clients (TECs) status was not disclosed to healthcare workers

Si et al. 2023	Yes	Physicians were informed three months in advance, provided written consent, but were not told which diseases were tested or when SP visits would occur.
Altshuler et al. 2023	Yes	Physicians consented during orientation to have their educational data included in a research registry
Verbal Informed consent		
Lorenzo et al. 2015	Yes	Informed via letter that they could be visited in the next 6 months, phone call to confirm participation
Wu et al. 2020	Yes	Informed that they could be visited in the next 6 months
Liu et al. 2022	Partial	Individual verbally consented to the overall study but they did not know they were being visited by a USP
Institutional consent		
Meuwissen et al. 2006	Partial	Clinics were informed that quality evaluations would take place after they signed a contract for quality measurement, but they were not told the methods used.
Derkx et al. 2008	Yes	The centers consented to be selected but were not informed if they were selected.
Gillespie et al. 2023	Yes	Medical residents are routinely checked for quality with USPs.
Opportunity to opt out		
Peremans et al. 2010	Yes	Informed but not told about selection; participating by default with option to opt-out
Impossible to opt out		
Piribauer et al. 2012	Yes	Physicians were collectively informed by the Vienna Medical Chamber that some GPs would be randomly tested by USPs in the coming weeks.
Waiver of consent		
Amini et al. 2016	Partial	Physicians were not individually informed or asked for consent; instead, a general announcement was made about the possibility of evaluation, with ethical approval obtained from the university's ethics committee.
De Castro et al. 2018	No	Not informed, methodologically necessary, approved by the ethics committee board
Xu et al. 2021	No	Waiver of informed consent
Kwan et al. 2022	No	Waiver of informed consent
Zeng et al. 2022	No	Not informed, methodologically necessary, approved by the ethics committee board
Xie et al. 2023	No	Waiver of informed consent
No consent needed		
Stevens et al. 2018	No	Real patient collects data incognito without doctor's knowledge; no consent was asked as the physicians are not enrolled as study participants

[a] Either explicitly stated that it is written consent or unspecified.

CENTER FOR BIOMEDICAL ETHICS AND LAW

Kapucijnenvoer 35 box 7001
3000 LEUVEN, België

pascal.borry@kuleuven.be
www.kuleuven.be

Ethical and Methodological aspects of Mystery Shopping Studies with Primary Care Physicians: a Systematic Review of Empirical Studies.

Dossiernr.:	MP028810
Masterproef:	Masterproef I HAIO, 2022-2025
Faculteit/opleiding:	Faculteit Geneeskunde - Master in de huisartsgeneeskunde
Studiefase:	eerste fase
Academiejaar:	2023-2024
Korte omschrijving / abstract:	Empirical studies with the use of mystery shopping in primary care settings. The domain of interest is general health and improvement of healthcare. We will look for studies involving general practice and mystery shopping or healthcare improvement with mystery shopping. Mystery shopping studies are often conducted in healthcare, and this study design has been leveraged to investigate a variety of topics such as patient-doctor communication. Studies that have reported on the impact of mystery shopping in primary healthcare will be included, irrespective of the originating country. The main outcome of this study is to evaluate the use of mystery shoppers in in general practice and the ethical and methodological questions that arise with their use. For instance, we will look at the role of debriefing or authorized deception procedures within these studies, as well as at the main methodological modalities through which mystery shopping is carried out. We will assess the empirical outcomes within mystery shopping studies to identify potential improvements achieved through mystery shopping. We will compare the diverse empirical outcomes utilized across various studies. When feasible, we will attempt to assess the impact of mystery shopping in primary care and its effectiveness as a quality improvement method.
Toegepaste technieken:	Systemtic Review
Trefwoorden:	Patient Simulation General practitioners Empirical Research Family Practice Quality of Health Care general practice Ethical Analysis Mystery Shopping
Gewenste taal van communicatie:	nl
Promotor:	Professor Pascal Borry (u0039446)
Begeleider/Contact:	Kamiel Verbeke (u0086860)
Organisatie:	KU Leuven
Dienst:	Interfacultair Centrum voor Biomedische Ethiek en Recht
Discipline:	Masterproef
Extra groepsleden:	Stijn Peters stijn.peters@student.kuleuven.be r0637565 Tibo Verplancke tibo.verplancke@student.kuleuven.be r0624501

De verstrekte informatie toont aan dat het onderzoek in het kader van de masterproef een zuivere literatuurstudie betreft en dat er op geen enkele wijze gebruik wordt gemaakt van proefdieren, proefpersonen of menselijk lichaamsmateriaal.

Dit type onderzoek vereist geen goedkeuring van een gemachtigde commissie voor medische ethiek.

Onder voorbehoud van de volledigheid en correctheid van de door u verstrekte gegevens, krijgt u hierbij het akkoord om het onderzoek in het kader van de masterproef te starten. Deze mail dient als bewijs van goedkeuring mocht u uw onderzoek wensen te publiceren.

Wij willen u erop attent maken dat u zelf verantwoordelijk blijft voor uw onderzoek. Bovendien doet elke wijziging aan de door u verstrekte gegevens omtrent de onderzoeksopzet deze goedkeuring vervallen. U dient in dat geval een amendement te maken aan uw huidig dossier.

NIHR | National Institute for Health and Care Research

PROSPERO
International prospective register of systematic reviews

Ethical and Methodological aspects of Mystery Shopping Studies with Primary Care Physicians: a Systematic Review of Empirical Studies.

Stijn Peters, Tibo Verplancke, Pascal Borry, Kamiel Verbeke

To enable PROSPERO to focus on COVID-19 submissions, this registration record has undergone basic automated checks for eligibility and is published exactly as submitted. PROSPERO has never provided peer review, and usual checking by the PROSPERO team does not endorse content. Therefore, automatically published records should be treated as any other PROSPERO registration. Further detail is provided here.

Citation
Stijn Peters, Tibo Verplancke, Pascal Borry, Kamiel Verbeke. Ethical and Methodological aspects of Mystery Shopping Studies with Primary Care Physicians: a Systematic Review of Empirical Studies.. PROSPERO 2024 Available from https://www.crd.york.ac.uk/PROSPERO/view/CRD42023482284

REVIEW TITLE AND BASIC DETAILS

Review title
Ethical and Methodological aspects of Mystery Shopping Studies with Primary Care Physicians: a Systematic Review of Empirical Studies.

Original language title
Idem.

Review objectives
1. What methodological approaches have researchers utilized in empirical studies that applied mystery shopping interventions within primary care settings?
2. What ethical challenges emerge during the implementation of empirical studies utilizing mystery shopping in primary care settings, and how have researchers addressed these challenges?

Keywords
Empirical Research, Ethical Analysis, Family, Family Practice, General Practice, General Practitioners, Health Care, Health Care, Health Care Evaluation Mechanisms, Medicine, Patient

Simulation, Physicians, Physicians, Physicians, Primary Care, Quality Assurance, Quality Improvement, Quality Indicators, Quality of Health Care, Simulation Training

SEARCHING AND SCREENING

Searches
We will perform a systematic search of the literature in the following databases: PubMed and Web of Science, utilizing the search terms as specified in the search strategy.
Search restrictions: We will exclude articles in languages not mastered by the review team. Additionally, studies published more than 20 years before the initiation of our research will be omitted, acknowledging the evolving ethical landscape and contextual changes during this timeframe.

Study design
We will focus on empirical mystery shopping studies in primary care settings, in which we will examine the methodological and ethical aspects these studies.

ELIGIBILITY CRITERIA

Condition or domain being studied
Empirical studies with the use of mystery shopping in primary care settings. We will look for studies describing current practices as well as those investigating ways to improve healthcare provision.

Population
Studies related to mystery shopping in primary healthcare. These studies include methodology and ethics sections, which will be the focus of our data extraction and subsequent analysis.

Intervention(s) or exposure(s)
The focus lies on empirical studies concerning mystery shopping in primary healthcare. As such, the mystery shopping methodology can be thought of as the methodological intervention our study is interested in. Mystery shopping studies may encompass various methodologies and ethical challenges. A systematic review will be conducted to extract and categorize these data accordingly.

Comparator(s) or control(s)
Not applicable.

Context
Mystery shopping studies are often conducted in healthcare, and this study design has been leveraged to investigate a variety of topics such as patient-doctor communication. Studies using mystery shopping in primary healthcare will be included, irrespective of the originating country.

OUTCOMES TO BE ANALYSED

Main outcomes
The main outcome of this study is to evaluate the ethical and methodological aspects of using mystery shoppers in primary care practice. For instance, we will look at the role of debriefing or

authorized deception procedures within these studies.

Additional outcomes
Not applicable.

DATA COLLECTION PROCESS

Data extraction (selection and coding)
Two review authors will independently review the titles and/or abstracts of studies identified through the search strategy to identify potentially eligible studies based on the aforementioned inclusion criteria. Full texts of potentially eligible studies will be obtained, and each review author will independently assess their eligibility for inclusion in the review.
During data extraction, the methodology sections will be carefully examined to identify the methodologies used, particularly focusing on the methodological and ethical considerations (e.g., consent protocols). Ethical aspects from the studies will be extracted, including debriefing, consent protocols, and privacy measures, facilitating a comprehensive analysis of ethical dimensions. Similarly, methodological aspects of the studies will be extracted to allow for a description of the potential variability of mystery shopping designs.

Risk of bias (quality) assessment
Any discrepancies regarding the eligibility of particular studies will be resolved through discussion between the two review authors. If consensus cannot be reached, a third reviewer will be consulted to provide a final decision.

PLANNED DATA SYNTHESIS

Strategy for data synthesis
Our strategy for data synthesis will systematically examine the extracted data from the included studies, focusing on mystery shopping methodology and ethical aspects. To facilitate comparison, we will create a comparative table of the methods used and ethical approaches, highlighting the concurrences and disparities across these findings. This approach aims to provide a comprehensive summary of the commonalities and variations in studies related to mystery shopping in primary healthcare, accentuating the diversities and limitations identified during data analysis.

Analysis of subgroups or subsets
Not applicable.

REVIEW AFFILIATION, FUNDING AND PEER REVIEW

Review team members

- Dr Stijn Peters, Interuniversitair Centrum voor Huisartsen Opleiding
- Dr Tibo Verplancke, Interuniversitair Centrum voor Huisartsen Opleiding
- Professor Pascal Borry, Centre for Biomedical Ethics and Law, Faculty of Medicine, KU Leuven
- Dr Kamiel Verbeke, Centre for Biomedical Ethics and Law, Faculty of Medicine, KU Leuven

Review affiliation
Centre for Biomedical Ethics and Law, Faculty of Medicine, KU Leuven

Funding source
None.

Named contact
Stijn Peters. Centre for Biomedical Ethics and Law Kapucijnenvoer 35 blok d - box 7001 3000 Leuven, Belgium
stijn-peters@hotmail.com

TIMELINE OF THE REVIEW

Review timeline
Start date: 01 September 2023. End date: 01 May 2025

Date of first submission to PROSPERO
12 November 2023

Date of registration in PROSPERO
23 November 2023

CURRENT REVIEW STAGE

Publication of review results
The intention is to publish the review once completed.The review will be published in English

Stage of the review at this submission

Review stage	Started	Completed
Pilot work	✓	
Formal searching/study identification	✓	
Screening search results against inclusion criteria		
Data extraction or receipt of IP		
Risk of bias/quality assessment		
Data synthesis		

Review status
The review is currently planned or ongoing.

ADDITIONAL INFORMATION

PROSPERO version history

- Version 1.1 published on 23 Nov 2023
- Version 1.0 published on 23 Nov 2023

Review conflict of interest

None known

Country

Belgium

Disclaimer

The content of this record displays the information provided by the review team. PROSPERO does not peer review registration records or endorse their content.

PROSPERO accepts and posts the information provided in good faith; responsibility for record content rests with the review team. The owner of this record has affirmed that the information provided is truthful and that they understand that deliberate provision of inaccurate information may be construed as scientific misconduct.

PROSPERO does not accept any liability for the content provided in this record or for its use. Readers use the information provided in this record at their own risk.

Any enquiries about the record should be referred to the named review contact