

Sex and Gender Bias in Clinical Practice: Evidence from a Simulation-Based Quasi-Experimental Study in Spain

Cristina Pujol

Universitat de Barcelona

Lluís Medir

Universitat de Barcelona

Judit Vall Castelló

`judit.vall@ub.edu`

Universitat de Barcelona

Research Article

Keywords: bias, clinical decision-making, health disparities, medical education, clinical simulation

Posted Date: October 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11110582/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Sex and Gender Bias in Clinical Practice: Evidence from a Simulation-Based Quasi-Experimental Study in Spain

Cristina Pujol, Department of Political Science, Universitat de Barcelona

Lluís Medir, Department of Political Science, Universitat de Barcelona

Judit Vall Castelló, Department of Economics, Universitat de Barcelona & IEB.

Corresponding author: judit.vall@ub.edu.

Abstract

Sex-and-gender disparities in healthcare are often attributed to biological and social determinants, but implicit bias in clinical judgement may also contribute, though empirical evidence remains limited. Using mandatory clinical simulation exercises across four medical faculties in Spain, we examine sex- and gender-differences in clinical decision-making among medical students and resident physicians. Patients were allocated to male or female standardised cases through routine scheduling, without control over or prior knowledge of case characteristics, providing a quasi-experimental setting to compare performance across seven sequential stages of care. Female patients received significantly lower quality-of-care scores in symptom assessment (-14.79%), initial diagnosis (-18.50%), and clinician-patient communication (-15.18%), with no differences in later, more protocolised stages. Disparities persisted after controlling for training level and institutional fixed effects. Bias concentrates in the earliest, most interpretative stages of clinical management, with implications for quality of care and medical curriculum design.

Keywords: bias, clinical decision-making, health disparities, medical education, clinical simulation

Introduction

Sex and gender inequalities in health have long been explained through biological and social frameworks. Biological factors such as hormonal and physiological differences contribute to variation in disease susceptibility and progression. Social determinants, including education, socioeconomic status, labour market participation, caregiving responsibilities, and health-seeking behaviours, shape access to and quality of healthcare. Women, for instance, are more likely than men to seek medical advice or pursue specialised care [1, 2].

While foundational, these explanations overlook healthcare providers' role in perpetuating or mitigating gender and sex disparities. Importantly, differences in clinical management are not solely due to internalised stereotypes shaping clinician–patient interactions, but are also rooted in longstanding gaps in biomedical research that have historically limited sex- and gender-sensitive medical knowledge [3], thus constraining healthcare providers' ability to appropriately assess sex-and-gender differences in symptomatology, risk factors, and treatment responses.

Although the first evidence of gender biases in clinical management emerged over three decades ago [4–6], only recently has research provided large-scale evidence positioning these biases as a structural challenge for health systems [7] and documenting persistent deficits in patient-centred care for women [8–10].

In line with this call from scientific literature, a growing body of research has focused on the presence and persistence of gender biases in clinical settings. For example, women are more likely to receive substandard care and experience adverse safety events across several healthcare levels [11–13]. Even in well-studied conditions such as cardiovascular diseases (CVDs), where sex-specific presentations are well documented, women remain underdiagnosed and undertreated [14]. These patterns suggest that inequities also emerge within clinical reasoning and clinician–patient interactions. Implicit biases, defined as unconscious stereotypes and attitudes, can influence clinical judgment even among providers who consciously endorse egalitarian values [15, 16]. Cognitive biases, such as overconfidence, anchoring, and confirmation bias, further distort clinical reasoning and decision-making [17, 18].

Despite growing awareness, evidence on how and at what stage these biases affect clinical management remains limited. Our study addresses this gap by quantifying sex-and-gender disparities in core clinical management tasks performed by medical students and early-career healthcare providers. Specifically, we examine how these biases influence key clinical management tasks and whether they persist or attenuate across training stages, highlighting the role of experience. Understanding when bias emerges during the clinical reasoning and management

process and how it evolves across the training trajectory is key to design targeted, stage-appropriate interventions focused on bias awareness, before cognitive shortcuts and clinical heuristics become entrenched.

We employ standardised clinical simulations, a tool that replicates real-world scenarios in a controlled and rigorous setting. Simulations have proven effective in improving procedural skills, teamwork, and patient safety [19, 20]. By replicating real-world patient encounters in a controlled environment, simulations eliminate many of the confounding factors inherent in observational healthcare data: time pressure, resource constraints, patient non-cooperation, and systemic variation across care settings. This makes simulations particularly well-suited to detect systematic patterns in clinical reasoning, including those shaped by implicit bias, that would be difficult to observe or impossible to isolate in real clinical settings.

Our study draws on data from clinical simulations conducted in four medical faculties in Catalonia (Spain) over two academic years. Participants were assigned to male or female standardised patient cases through routine simulation scheduling processes in which they had no control over case allocation and were unaware of the patient characteristics in advance. This allocation mechanism provides a quasi-experimental setting to examine differences in clinical performance according to patient sex across seven sequential stages of care.

We find consistent disparities in early-stage clinical management: female patients receive lower quality-of-care scores in symptom assessment (10 percentage points lower, which represents a 14.79% decrease with respect to the mean score for this task), initial diagnosis (12 percentage points lower; an 18.50% decrease), and communication with the patient (10.7 percentage points lower; a 15.18% decrease). These differences persist across different levels of clinical experience. Disparities emerging in early stages of clinical management are particularly relevant for patient safety and health outcomes, since such deviations may cause important delays in diagnostic testing, definitive diagnosis and treatment, contributing to cumulative quality deficits.

Our findings therefore have implications for quality of care, medical education, and health system governance, highlighting the need to address bias in early-stage clinical decision-making processes. By identifying where in the clinical management process disparities emerge, this study provides policy-relevant evidence to guide targeted interventions in clinical reasoning and communication.

Throughout this article, references to bias and inequality explicitly encompass both sex and gender, which are distinct but interrelated phenomena [21]. This choice is deliberate and reflects both conceptual and empirical considerations. Sex-based biases stem from biological factors, as well as gaps in medical knowledge regarding symptomatology, disease presentation, genetic risk factors, and treatment

responses [22] with well-documented adverse consequences for women's health outcomes [3, 23]. Gender-based biases arise from socially constructed norms and stereotypes that influence clinical judgement, communication, and diagnostic reasoning [16].

However, these biases frequently operate simultaneously within doctor-patient interactions, making it empirically difficult to disentangle their individual effects [24], particularly in communicative and interpretative dimensions of care. Our study therefore refers to both sex and gender throughout, reflecting the multidimensional nature of the inequalities examined.

1. Materials and Methods

1.1. Study design

This study examines sex- and gender-related differences in healthcare delivery using data from clinical simulation exercises conducted between September 2023 and June 2025 across four medical faculties in Catalonia, Spain: Universitat de Barcelona (UB), Universitat Pompeu Fabra (UPF), Universitat de Lleida (UdL), and Universitat Rovira i Virgili (URV)¹. The clinical simulations were not specifically designed for the purposes of this study, nor were additional simulations exercises introduced into the participants' academic or clinical training programmes. Rather, they were part of the pre-existent mandatory curricula of the participating medical schools and training programmes. Data for the study was collected through structured observation of these routinely conducted simulation exercises.

Clinical scenarios were developed by the teaching staff at each participating university and represented specific clinical conditions. Scenarios were designed according to the learning objectives and expected competencies of participants at their respective levels of training. Participants were required to obtain the clinical information necessary to assess the patient and formulate an initial diagnosis through the history-taking, physical examination and diagnostic testing; conduct a differential diagnosis; establish a final diagnosis; initiate appropriate treatment; and communicate relevant information and decisions to the patient.

Participants were not formally randomised to clinical scenarios. Instead, case allocation was determined through routine educational scheduling procedures established independently of the study, and participants had no control over the case to which they were assigned. The content of the clinical scenario, including the patient's sex and clinical condition, was not disclosed to participants before the simulation. Thus, although allocation did not constitute formal randomisation,

¹ These four institutions host 56.98% of medicine students in public universities in Catalonia and 9.11% of those in Spain, according to 2024-2025 data from the Ministry of Science, Innovation and Universities

assignment to cases occurred independently of participants' choices and prior knowledge of scenario characteristics. This allocation mechanism provides a quasi-experimental setting for analysis, in which differences in clinical performance according to patient sex can be examined while accounting for relevant participant-, scenario-, and institutional-level characteristics.

Simulations covered 33 clinical conditions across thirteen ICD disease categories, spanning infectious diseases, neoplasms, genitourinary, respiratory, immune, mental health, musculoskeletal, traumatic injuries, and other conditions. All simulation exercises followed a standardised format: a pre-briefing; a simulated scenario with a standardised patient; and debriefing and feedback.

A total of 1,597 simulation exercises involving 1,947 participants were evaluated. Participants included third- through sixth-year medical students and early-career healthcare providers (medical residents in training, or MIRs in Spanish). First- and second-year students were excluded due to insufficient clinical training. All eligible participants taking part in the simulations during the study period were invited to participate. Opt-out rates were 2.22% in the UB-UPF-UdL group and 0% in URV. Close to 70% of the simulation exercises were performed by women, reflecting the underlying gender composition of medical students in Catalonia². Women also represented 66% of supervisors. Male standardised patients accounted for 35% of the simulated patient encounters (see *Table 1*).

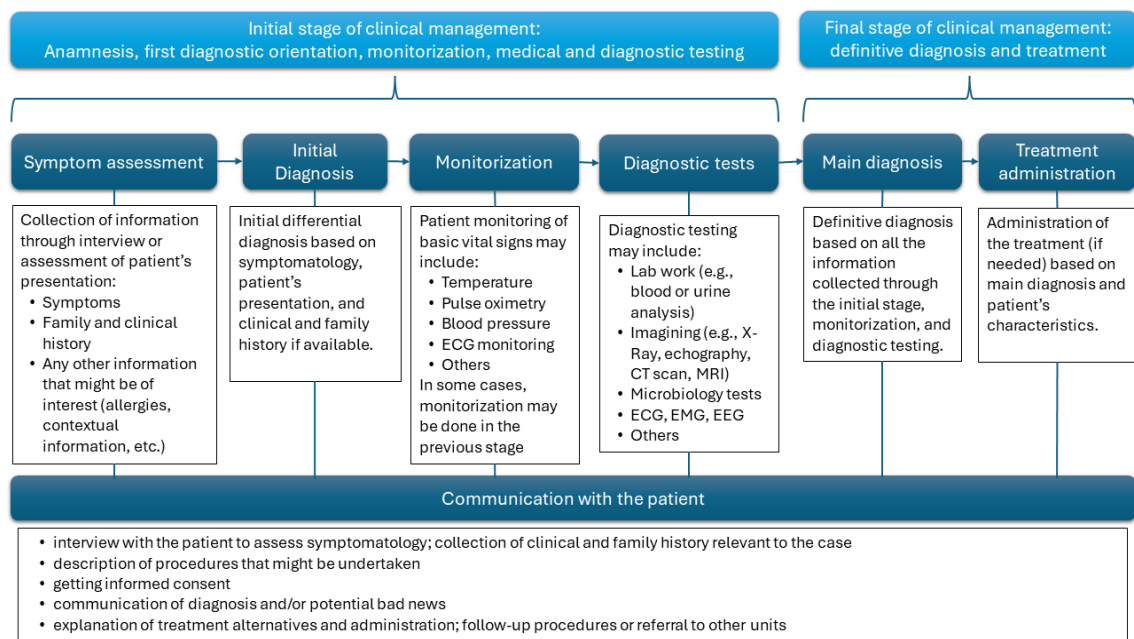
1.2. Data collection: simulation procedure and assessment of clinical management

Before each simulation, and following ethical standards, participants were invited to take part in a research study and informed that research participation was voluntary and not a condition for participating in the simulation exercise. Those who consented were provided with the study information, which was carefully framed: participants were told the study evaluated simulation as a tool for improving disease management. No reference was made to sex or gender bias to avoid influencing behaviour.

Clinical performance was assessed across seven domains or tasks representing sequential components of patient management: (a) symptom assessment; (b) initial diagnosis; (c) patient monitoring; (d) diagnostic testing; (e) main diagnosis; (f) treatment administration; and (g) overall communication with the patient (see *Figure 1*). Depending on the clinical scenario and the expected competencies associated with the participant's stage of training, some domains were not applicable and were therefore not assessed.

² Women represent 74% of total enrolment in Catalonia, according to data from the National Statistics Institute in Spain

Figure 1. Sequential nature of the clinical management process and steps evaluated in the clinical simulations



Note: This figure illustrates the seven core tasks evaluated during each clinical simulation, structured to mirror the patient clinical management process and the temporal logic of real-life patient care. Each task was evaluated independently according to standardised criteria. *Communication with the patient* was assessed throughout the entire simulation, as it is integral to all stages of clinical decision-making and therapeutic interaction. Source: authors' elaboration based on focus group interviews with simulation coordinators from each participating university in the study.

All evaluators were trained health professionals. They assessed performance using a structured checklist and according to the clinical complexity of the case and the competencies expected at the participant's level of training. The use of structured assessment criteria was intended to promote consistency across evaluators and participating institutions and to reduce subjectivity.

For each applicable clinical management domain, performance was classified using a standardised and predefined three-category rubric based on the proportion of expected actions successfully completed and the presence of clinically relevant errors or omissions, consistent with established checklist-based approaches to simulation assessment [26–29]. Each task was classified as correct when more than 70% of the required actions were completed without major errors; partially incorrect when only 50–70% were completed or clinically significant omissions were observed; and incorrect when fewer than 50% were completed or performance involved severe errors, major omissions, or failure to complete the task.

In addition to the participants' performance, patient characteristics (i.e., biological sex and health condition), participant characteristics (i.e., number, gender, academic year or residency year, team composition, peer observation), supervisor gender, institutional affiliation, and simulation date were also collected. Patient

sex, the primary exposure of interest, was recorded for all simulation exercises, with no missing observations.

The final dataset comprises task-level assessment of participants' clinical performance across the seven domains, together with patient-, participant-, supervisor-, simulation-, and institution-level characteristics.

1.3. Empirical strategy

We analyse performance across tasks according to our primary exposure variable, the biological sex of the simulated patient. Clinical performance was analysed separately for each of the seven domains described above. To facilitate analysis, the original three-category performance measure was dichotomised: a value of 1 was assigned for correct execution (>70% of the task completed without major errors), and 0 otherwise. This classification treats incomplete performance and clinically relevant errors or omissions as failure to meet the predefined standard of correct clinical management.

For each clinical management task, we estimate Ordinary Least Squares (OLS) regression models in which the performance indicator is regressed on patient sex. The coefficient on patient sex therefore represents the adjusted difference in the probability of correct performance between encounters involving female and male standardised patients, after accounting for the covariates included in the model. Models adjust for participants' level of training (academic year or residency stage), the gender composition of the participating medical team, and the gender of the simulation evaluator.

Analyses are conducted separately for two groups of participating institutions to account for an important difference in simulation design: at URV, the standardised patient also acted as the evaluator, limiting full separation between patient sex and evaluator characteristics. Hence, estimated differences may partially capture systematic differences in assessment associated with the standardised patient/evaluator. At UB, UP, and UdL, the roles were clearly separated. Therefore, estimates are reported separately for the URV and the UB-UPF-UdL samples rather than combining observations across these different simulation designs, ensuring transparency regarding these limitations.

Finally, we examine heterogeneity by participants' level of experience. Analyses are stratified, separating third- and fourth-year students from fifth-year students and MIRs, to assess whether disparities in clinical management attenuate or intensify with increasing clinical training experience.

Table 1. Descriptive statistics

Characteristics	Total	URV	UB	UPF	UdL
Health providers					
Women	1351 (69,39%)	977 (69,50%)	204 (67,11%)	141 (71,94%)	29 (70,73%)
Men	594 (30,51%)	429 (30,50%)	99 (32,57%)	54 (27,55%)	12 (29,27)
Non-binary	2 (0,10%)	0 (0,0,%)	1 (0,33%)	1 (0,51%)	0 (0,00%)
Academic level					
3rd year	29 (1,82%)	0 (0,0,%)	2 (2,17%)	27 (31,03%)	0 (0,00%)
4th year	794 (49,72%)	707 (50,28%)	57 (61,96%)	26 (29,89%)	4 (33,33%)
5th year	346 (21,67%)	296 (21,05%)	33 (35,87%)	17 (19,54%)	0 (0,00%)
6th year	420 (26,30%)	403 (28,66%)	0 (0,00%)	17 (19,54%)	0 (0,00%)
MIR	8 (0,50%)	0 (0,0,%)	0 (0,00%)	0 (0,00%)	8 (66,67%)
Simulation information					
Female patient	1037 (64,93%)	925 (65,79%)	49 (53,26%)	60 (68,97%)	3 (25,00%)
Woman supervisor	1056 (66,12%)	925 (65,79%)	68 (73,91%)	60 (68,97%)	3 (25,00%)
Clinical case: disease group					
Certain infectious or parasitic diseases	107 (6,70%)	102 (7,25%)	0 (0,00%)	5 (5,75%)	0 (0,00%)
Developmental anomalies	6 (0,38%)	0 (0,00%)	0 (0,00%)	6 (6,90%)	0 (0,00%)
Disease of the nervous system	11 (0,69%)	0 (0,00%)	0 (0,00%)	11 (12,64%)	0 (0,00%)
Diseases of the blood or blood-forming organs	203 (12,71%)	203 (14,44%)	0 (0,00%)	0 (0,00%)	0 (0,00%)
Diseases of the circulatory system	147 (9,20%)	101 (7,18%)	38 (41,30%)	4 (4,60%)	4 (33,33%)
Diseases of the ear or mastoid process	4 (0,25%)	0 (0,00%)	0 (0,00%)	4 (4,60%)	0 (0,00%)
Diseases of the genitourinary system	111 (6,95%)	91 (6,47%)	16 (17,39%)	4 (4,60%)	0 (0,00%)
Diseases of the immune system	10 (0,63%)	0 (0,00%)	0 (0,00%)	10 (11,49%)	0 (0,00%)
Diseases of the musculoskeletal system or connective tissue	5 (0,31%)	0 (0,00%)	0 (0,00%)	5 (5,75%)	0 (0,00%)
Diseases of the nervous system	150 (9,39%)	150 (10,67%)	0 (0,00%)	0 (0,00%)	0 (0,00%)
Diseases of the respiratory system	7 (0,44%)	0 (0,00%)	0 (0,00%)	4 (4,60%)	3 (25,00%)
Endocrine, nutritional or metabolic diseases	103 (6,45%)	103 (7,33%)	0 (0,00%)	0 (0,00%)	0 (0,00%)
Injury, poisoning or certain other consequences of external causes	19 (1,19%)	0 (0,00%)	3 (3,26%)	13 (14,94%)	3 (25,00%)
Mental, behavioural or neurodevelopmental disorders	38 (2,38%)	0 (0,00%)	24 (26,09%)	14 (16,09%)	0 (0,00%)
Neoplasms	11 (0,69%)	0 (0,00%)	11 (11,96%)	0 (0,00%)	0 (0,00%)
Symptoms, signs or clinical findings of the nervous system	204 (12,77%)	204 (14,51%)	0 (0,00%)	0 (0,00%)	0 (0,00%)
Symptoms, signs or clinical findings, not elsewhere classified	461 (28,87%)	452 (32,15%)	0 (0,00%)	7 (8,05%)	2 (16,67%)

Note: Descriptive statistics for all study variables. Values are as counts (percentages) for all categorical variables. The table reports both the overall sample and the distribution across participating institutions (URV, UB, UPF, and UdL).

2. Results

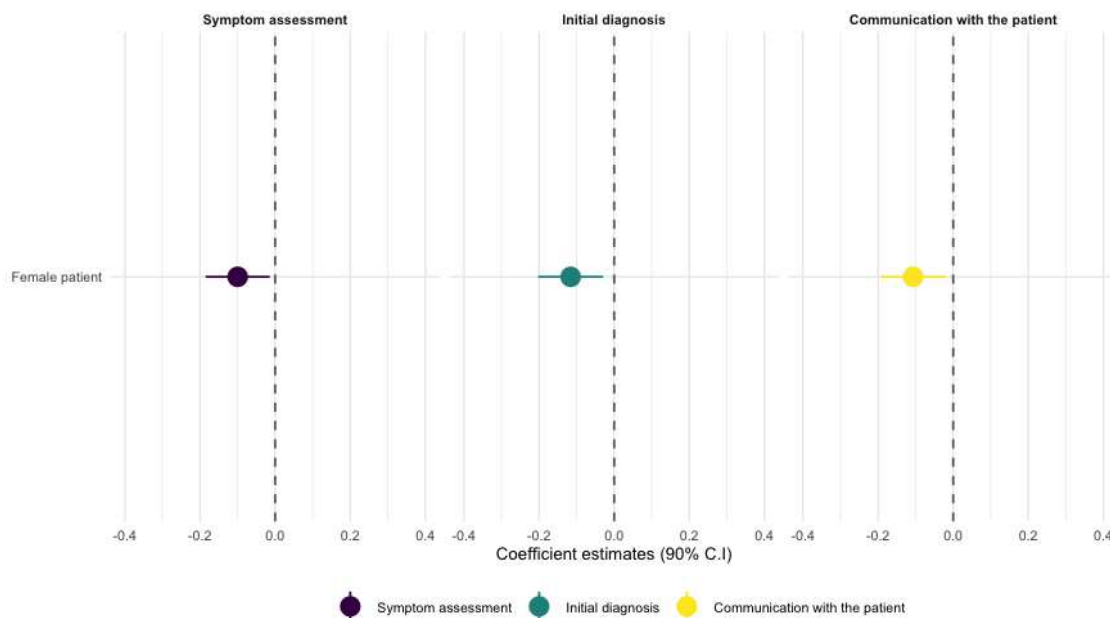
2.1. General findings

Data analysis examines healthcare provider performance across the sequential process of patient clinical management, which involves the seven tasks previously described (see Supplemental Material: Figure 1). The findings reveal significant sex disparities in early-stage clinical management, with lower quality-of-care scores for female patients in symptom assessment, initial diagnosis, and communication with the patient (see Figure 2). These differences remain robust after including institutional and experience-level fixed effects and controlling for the gender composition of the medical team and the supervisor gender (see Table 2).

The estimated coefficient for symptom assessment of female patients is -0.100, indicating a 10.0 percentage points reduction in the probability of receiving quality care when compared to male patients. This corresponds to a 14.79% decrease relative to the average performance level for this task. Similarly, being assigned a female patient is associated with a 0.12-point decrease in initial diagnosis (18.50% decrease). The probability of establishing a good communication with the patient is 10.7 percentage points lower in female versus male patients (15.18% decrease).

No statistically significant sex differences are found in later-stage tasks: patient monitoring, diagnostic testing, main diagnosis, or treatment administration (see Table 2).

Figure 2. Effect of patient sex on symptom assessment, initial diagnosis and communication with the patient



Note: This figure displays the estimated coefficients and 90% confidence intervals for the effect of *patient sex* on performance in three core clinical management tasks. Models include data from all four participant institutions: UB, UPF, UdL and URV. Each coloured point represents the coefficient estimate from an OLS regression, with fixed effects for the *participants' level of training and institution* and controls for the *Proportion of women in the medical team* and the *Supervisor gender*. Female patients are significantly associated with lower performance in all three tasks.

Table 2. Clinical management quality-of-care scores for symptoms, initial diagnosis, monitoring, diagnostic testing, main diagnosis, treatment administration and communication with the patient.

	Symptoms (1)	Initial dx (2)	Monitorization (3)	Diagnostic testing (4)	Main dx (5)	Treatment administration (6)	Communication with patient (7)
Female patient	-0.100* (0.052)	-0.116** (0.053)	0.002 (0.061)	0.016 (0.055)	0.019 (0.054)	0.095 (0.060)	-0.107** (0.053)
Women in the medical team	-0.012 (0.026)	0.106*** (0.026)	0.118** (0.060)	0.059** (0.026)	0.025 (0.042)	0.023 (0.037)	0.067*** (0.025)
Supervisor is a woman	-0.071 (0.052)	-0.063 (0.053)	0.035 (0.061)	-0.132** (0.056)	-0.141** (0.056)	-0.167*** (0.060)	0.044 (0.053)
Constant	0.752*** (0.113)	0.550*** (0.115)	0.125 (0.141)	0.390*** (0.121)	0.714*** (0.113)	0.388*** (0.129)	0.190 (0.116)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of dependent variable	0.676	0.627	0.639	0.564	0.818	0.55	0.705
N	1,596	1,595	366	1,575	495	889	1,585
R ²	0.063	0.100	0.166	0.154	0.083	0.091	0.035
Adjusted R ²	0.057	0.094	0.142	0.149	0.064	0.080	0.029
Residual Std. Error	0.455 (df = 1585)	0.460 (df = 1584)	0.445 (df = 355)	0.458 (df = 1565)	0.373 (df = 484)	0.477 (df = 878)	0.449 (df = 1574)
F Statistic	10.664*** (df = 10; 1585)	17.578*** (df = 10; 1584)	7.048*** (df = 10; 355)	31.594*** (df = 9; 1565)	4.404*** (df = 10; 484)	8.753*** (df = 10; 878)	5.764*** (df = 10; 1574)

Note: This table presents the OLS regression results for all clinical tasks evaluated. Estimates include data from all four institutions: UB, UPF, UdL and URV. Each model includes fixed effects for level of training and university. The primary independent variable is the patient sex (female vs. male, reference category). Coefficients represent the difference in the score obtained for each task relative to male patients. Thus, negative coefficients indicate lower scores when the patient is female, while positive coefficients indicate higher scores. Female patients are associated with significantly lower quality-of-care scores in symptom assessment, initial diagnosis, and communication with the patient. The proportion of women in the medical teams has a positive and statistically significant effect in the models for initial diagnosis and communication. The gender of the supervisor shows no effect. Standard errors are reported in parentheses. Significance levels: $p < 0.1^*$; $p < 0.05^{**}$; $p < 0.01^{***}$.

2.2. Heterogeneity by institution

Disaggregated analyses by institution are consistent with the overall findings: disparities in clinical management occur in its initial stage, particularly during symptom assessment and initial diagnosis. At this stage, doctor-patient interaction underpins early diagnostic reasoning through the collection clinical information. Consequently, differences in communication are also expected, given its central role in eliciting relevant clinical information prior to testing or further interventions.

The absence of statistically significant differences in the UB-UPF-UdL sample for symptom assessment and communication likely reflects the limited sample size and reduced statistical power. Importantly, point estimates in the UB-UPF-UdL and URV samples are nearly identical in magnitude and direction, though standard errors are larger in the former. Thus, despite the lack of statistical significance in two of the tasks, the consistency in coefficients reinforces the robustness and validity of the findings (see *Figure 3*).

As for diagnostic testing, results are inconclusive: although the URV sample reveals statistically significant differences, the absence of such differences in the UB-UPF-UdL sample prevents firm conclusions (see *Supplemental Material: Figure 4*). For later-stage tasks (patient monitoring, main diagnosis, treatment), no significant sex differences are observed (see *Table 2; Supplemental Material: Figure 4, Table 3 & Table 4*).

a. Symptom assessment

Across all models, in the URV sample, being a female patient is significantly associated with a lower probability of accurate symptom assessment, even after controlling for team composition and supervisor gender. In the UB-UPF-UdL sample, the results point exactly to the same direction but are slightly smaller in magnitude and not statistically significant (see *Figure 3*).

Team composition and supervisor gender show no significant effect on performance (see *Supplemental Material: Table 3 & Table 4*).

b. Initial diagnosis

Being a female patient is significantly associated with lower scores in the assessment of initial diagnosis across all models. The magnitude of the effect remains stable when controlling for the team composition and supervisor gender, indicating a substantial and robust sex-related disparity in the first diagnostic orientation of patient management (see *Figure 3*).

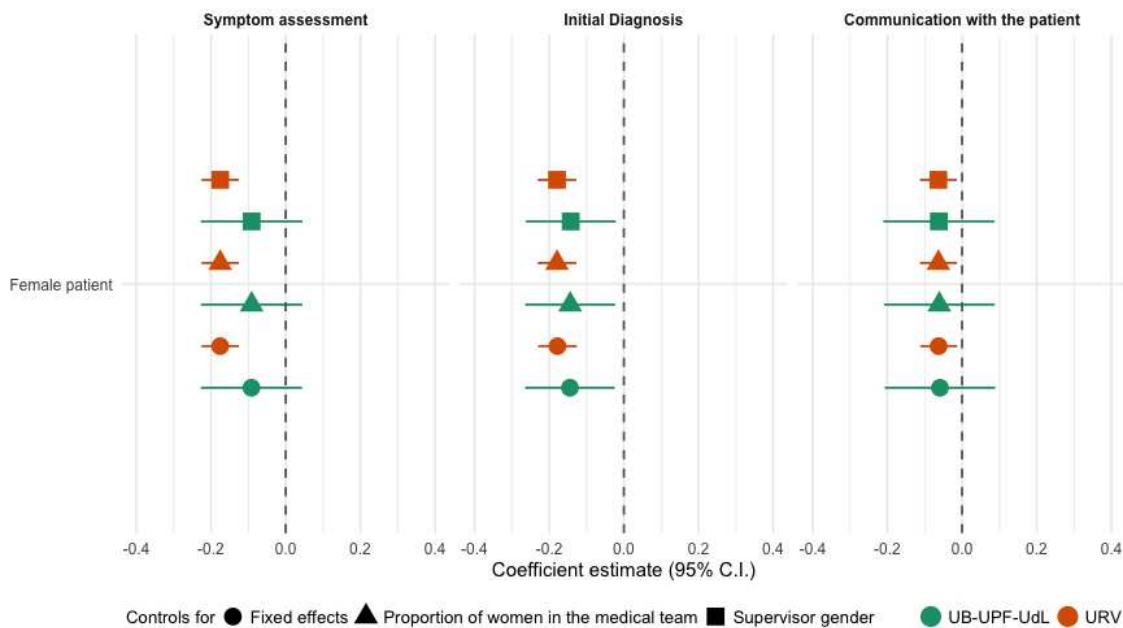
A higher proportion of women in the medical team is associated with better diagnostic accuracy (see *Table 2*). Supervisor gender shows no significant effect (see *Table 2 and Supplemental Material: Table 4*).

c. Communication with the patient

Being a female patient is consistently associated with a lower quality of communication with the patient. The effect is consistent in magnitude across all models and for both samples but is not statistically significant for the UB-UPF-UdL group due to its notable smaller size. This finding is highly relevant in this context because it not only indicates that interpersonal dynamics are affected by bias but might also partially explain why symptom assessment and initial diagnosis are also affected, since they are tasks that deeply depend on the use of communication and information exchange to be correctly performed (see *Figure 3*).

A higher proportion of women in the medical team is associated with better communication with the patient for both groups, though significance varies across samples (see *Table 2* and *Supplemental Material: Table 3 & 4*). This suggests that structural team dynamics related to gender may play a role in reducing bias when communicating with the patient. Supervisor gender shows no clear effect (see *Table 2* and *Supplemental Material: Table 4*).

Figure 3. Effect of patient sex on symptom assessment, initial diagnosis and communication with the patient



Note: The figure presents coefficient estimates and 95% confidence intervals for *patient sex* on *symptom assessment*, *initial diagnosis*, and *communication with the patient*. Estimates are shown separately for each and stratified by model specification. Each model progressively incorporates additional control variables. Round-shaped coefficients only include fixed effects for the *participants' level of training and institution* (only for the UB-UPF-UdL group). Triangle-shaped coefficients add a control for the *gender composition within clinical teams*. Squared-shaped coefficients further incorporate the *gender of the supervisor*.

2.3. Heterogeneity by level of training and experience

Clinical experience may also influence bias in patient management. At first glance, one might assume that greater experience and accumulated knowledge would be associated with fewer biases. However, this is not necessarily the case: increased self-confidence and repeated exposure to clinical cases may encourage reliance on cognitive shortcuts, potentially reinforcing unconscious biases. Analysing bias in relation to the level of clinical experience can help us understand how, and if, this phenomenon persists with training and educational progression, and how it may affect symptom assessment, initial diagnosis, and communication with the patient.

In the URV sample, female patients receive significantly lower scores in symptom assessment and initial diagnosis across both experience groups (see *Supplemental Material: Figure 5*). Although the magnitude of the effect decreases slightly amongst the more advanced participants, disparities remain persistent across training levels.

In the UB-UPF-UdL sample, scores for symptom assessment and initial diagnosis are worse for the more advanced group and statistically significant. However, the limited sample size, particularly in the more senior student subgroup, reduces statistical power and complicates definitive conclusions (see *Supplemental Material: Table 5 & Table 6*).

Communication with the patient displays a different pattern (see *Supplemental Material: Figure 5*). In the URV sample, disparities are not significant among 3rd-4th year students but become larger and statistically significant for more advanced participants, suggesting that communication-related disparities do not diminish with experience. In the UB-UPF-UdL sample, estimates are imprecise and show no consistent pattern (see *Supplemental Material: Table 7*).

Overall, the persistent and systematic disadvantage of female patients in symptom assessment, initial diagnosis and communication point to structural biases and patterns in clinical decision-making that remain present across levels of training, despite slight attenuation with experience.

3. Discussion

This study explores how the biological sex of simulated patients influences clinical decision-making among medical students and junior doctors. The analysis revealed consistent disparities in early diagnostic stages: female patients received lower quality-of-care scores in symptom assessment, initial diagnosis, and communication with the patient. These differences remained robust after controlling for participant experience, supervisor gender and gender composition of the medical team.

Notably, disparities emerged during the earliest stages of clinical management, where clinician-patient interaction and subjective interpretation are central.

Symptom assessment and initial diagnostic formulation rely heavily on narrative information, questioning strategies and interpretative judgement, precisely where gendered assumptions are most likely to interfere. Communication disparities further reinforce the role of interaction dynamics in shaping early clinical reasoning. In contrast, later stages of clinical management, such as monitoring, testing, main diagnosis and treatment administration, were less affected, likely because they rely more strongly on standardised protocols, objective clinical data and quantifiable indicators, which may constrain discretionary judgement and mitigate bias. This pattern highlights the potential protective role of structured, data-driven decision-making. However, in real-world settings, early-stage deviations may delay testing, misdirect diagnostic pathways or postpone treatment, ultimately affecting timeliness and overall quality of care, with consequences that extend well beyond the simulation context.

A key methodological strength of this study is the diversity of clinical cases included in the simulations. Rather than focusing on conditions already recognised as vulnerable to gender bias (e.g., CVDs, chronic pain or mental health disorders), our study employed 33 conditions across 13 ICD categories. This supports the interpretation that observed disparities reflect structural patterns in clinical reasoning, not disease-specific heuristics, strengthening the external validity of the findings.

Institutional heterogeneity further illuminates how simulation design shapes bias measurement. In the URV sample, where simulated patients also acted as evaluators, the penalisation of female patients was significant across multiple tasks and training levels. In contrast, in the UB-UPF-UdL sample, where roles were separated, significant differences appeared only in initial diagnosis. While partly explained by statistical power, this also underscores the importance of simulation design: overlapping patient-evaluator roles might heighten gendered interaction dynamics, influencing both behaviour and evaluation.

These findings contribute to the literature on sex-and-gender bias in clinical management, and carry direct implications for care delivery, medical education and health system governance. Patient sex influences core management tasks, particularly those dependent on communication and subjective interpretation, across conditions and training levels, suggesting implicit cognitive bias rather than disease-specific knowledge gaps.

The simulation-based design provides a replicable, scalable framework for identifying and evaluating disparities in clinical performance. Such approaches could support health systems and educational institutions in systematically assessing bias, tracking progress over time, and informing the design of targeted interventions aimed at improving equity in healthcare delivery.

Finally, bias awareness training should be mainstreamed across the entire medical curriculum, with particular attention to diagnostic reasoning and communication. More broadly, the concentration of disparities in the earliest stages of the clinical pathway, where diagnostic trajectories are shaped, positions sex-and-gender

inequity as a quality-of-care and health system governance challenge, not solely as a matter of individual behaviour or training issue. Targeted interventions in clinical reasoning and clinician-patient interaction are likely to yield greater equity than system-level reforms focused exclusively on protocols or resources.

Funding statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors

Competing interests

The author(s) declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Patient and Public Involvement statement

Patients and the public were not involved in the design, conduct, reporting, or dissemination plans of this study. This study examined clinical decision-making behaviour among medical students and intern resident trainees using standardised simulation exercises; no patient data were collected and no identifiable individuals were recruited as participants. Ethical approval was obtained from the corresponding Institutional Review Board, and written informed consent was obtained from all study participants (students and medical residents).

Data sharing statement

The data supporting the findings of this study are not publicly available due to institutional restrictions concerning the protection of personal data of participants. In accordance with institutional and ethical regulations, these data cannot be transferred or disclosed to third parties.

References

- [1] EIGE. *Gender in health*. European Institute for Gender Equality, https://eige.europa.eu/publications-resources/publications/gender-health?language_content_entity=en (26 January 2017).
- [2] Schmitz A, Lazarevič P. The gender health gap in Europe's ageing societies: universal findings across countries and age groups? *Eur J Ageing* 2020; 17: 509–520.
- [3] Merone L, Tsey K, Russell D, et al. Sex Inequalities in Medical Research: A Systematic Scoping Review of the Literature. *Womens Health Rep* 2022; 3: 49–59.
- [4] Ayanian JZ, Epstein AM. Differences in the use of procedures between women and men hospitalized for coronary heart disease. *N Engl J Med* 1991; 325: 221–225.
- [5] Healy B. The Yentl syndrome. *N Engl J Med* 1991; 325: 274–276.
- [6] Steingart RM, Packer M, Hamm P, et al. Sex differences in the management of coronary artery disease. Survival and Ventricular Enlargement Investigators. *N Engl J Med* 1991; 325: 226–230.
- [7] Westergaard D, Moseley P, Sørup FKH, et al. Population-wide analysis of differences in disease progression patterns in men and women. *Nat Commun* 2019; 10: 666.
- [8] Gagliardi AR, Green C, Dunn S, et al. How do and could clinical guidelines support patient-centred care for women: Content analysis of guidelines. *PloS One* 2019; 14: e0224507.
- [9] Gagliardi AR, Dunn S, Foster AM, et al. Is patient-centred care for women a priority for policy-makers? Content analysis of government policies. *Health Res Policy Syst* 2020; 18: 23.
- [10] Ramlakhan JU, Foster AM, Grace SL, et al. What constitutes patient-centred care for women: a theoretical rapid review. *Int J Equity Health* 2019; 18: 182.
- [11] Chen EH, Shofer FS, Dean AJ, et al. Gender Disparity in Analgesic Treatment of Emergency Department Patients with Acute Abdominal Pain. *Acad Emerg Med* 2008; 15: 414–418.
- [12] Piccardi C, Detollenaere J, Vanden Bussche P, et al. Social disparities in patient safety in primary care: a systematic review. *Int J Equity Health* 2018; 17: 114.
- [13] Conrad N, Judge A, Canoy D, et al. Diagnostic tests, drug prescriptions, and follow-up patterns after incident heart failure: A cohort study of 93,000 UK patients. *PLOS Med* 2019; 16: e1002805.

- [14] Vogel B, Acevedo M, Appelman Y, et al. The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030. *Lancet Lond Engl* 2021; 397: 2385–2438.
- [15] Blair IV, Steiner JF, Havranek EP. Unconscious (Implicit) Bias and Health Disparities: Where Do We Go from Here? *Perm J* 2011; 15: 71–78.
- [16] FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics* 2017; 18: 19.
- [17] Croskerry P. From Mindless to Mindful Practice — Cognitive Bias and Clinical Decision Making. *N Engl J Med* 2013; 368: 2445–2448.
- [18] Saposnik G, Redelmeier D, Ruff CC, et al. Cognitive biases associated with medical decisions: a systematic review. *BMC Med Inform Decis Mak* 2016; 16: 138.
- [19] Barsuk JH, Cohen ER, Williams MV, et al. Simulation-Based Mastery Learning for Thoracentesis Skills Improves Patient Outcomes: A Randomized Trial. *Acad Med J Assoc Am Med Coll* 2018; 93: 729–735.
- [20] Harrington DW SL. Designing a Simulation Scenario. In: *StatPearls [Internet]. Treasure Island (FL)*. StatPearls Publishing, <https://www.ncbi.nlm.nih.gov/books/NBK547670/> (2022, accessed 7 May 2023).
- [21] Gualtierotti R. Bridging the gap: Time to integrate sex and gender differences into research and clinical practice for improved health outcomes. *Eur J Intern Med* 2025; 134: 9–16.
- [22] Kaufman MR, Eschliman EL, Karver TS. Differentiating sex and gender in health research to achieve gender equity. *Bull World Health Organ* 2023; 101: 666–671.
- [23] Liu KA, Mager NAD. Women’s involvement in clinical trials: historical perspective and future implications. *Pharm Pract* 2016; 14: 708.
- [24] Maney DL, Duchesne A, Grossi G. Sex/gender entanglement: A problem of knots and buckets. *Biol Sex Differ* 2025; 16: 85.
- [25] von Elm E, Altman DG, Egger M, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007; 335: 806–808.
- [26] Ruiz-Manzanera J, Almela-Baeza J, Aliaga A, et al. Validation of Checklists and Evaluation of Clinical Skills in Cases of Abdominal Pain With Simulation in Formative, Objective, Structured Clinical Examination With Audiovisual Content in Third-Year Medical Students’ Surgical Clerkship. *J Surg Educ* 2024; 81: 1756–1763.

- [27] Ilgen JS, Ma IWY, Hatala R, et al. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ* 2015; 49: 161–173.
- [28] Adamson KA, Kardong-Edgren S, Willhaus J. An Updated Review of Published Simulation Evaluation Instruments. *Clin Simul Nurs* 2013; 9: e393–e400.
- [29] Cazzell M, Howe C. Using Objective Structured Clinical Evaluation for Simulation Evaluation: Checklist Considerations for Interrater Reliability. *Clin Simul Nurs* 2012; 8: e219–e225.