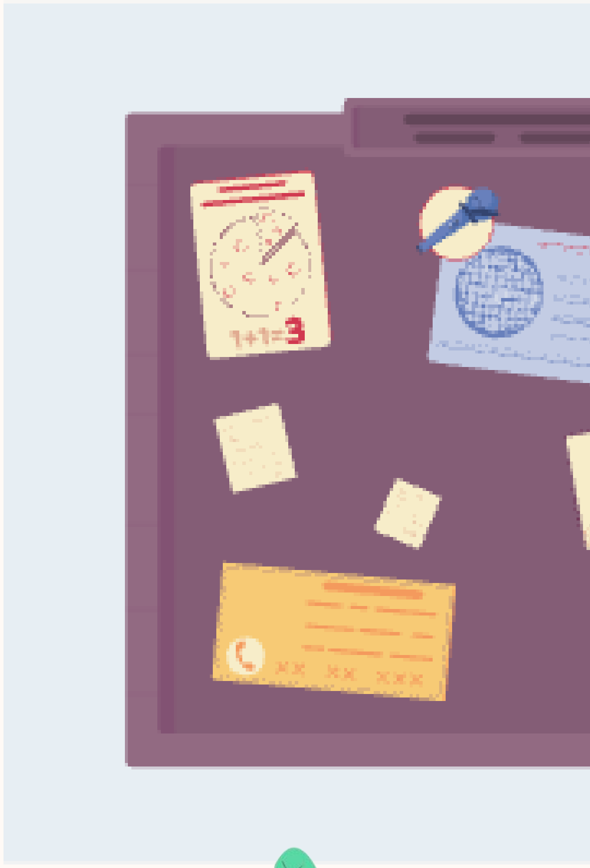
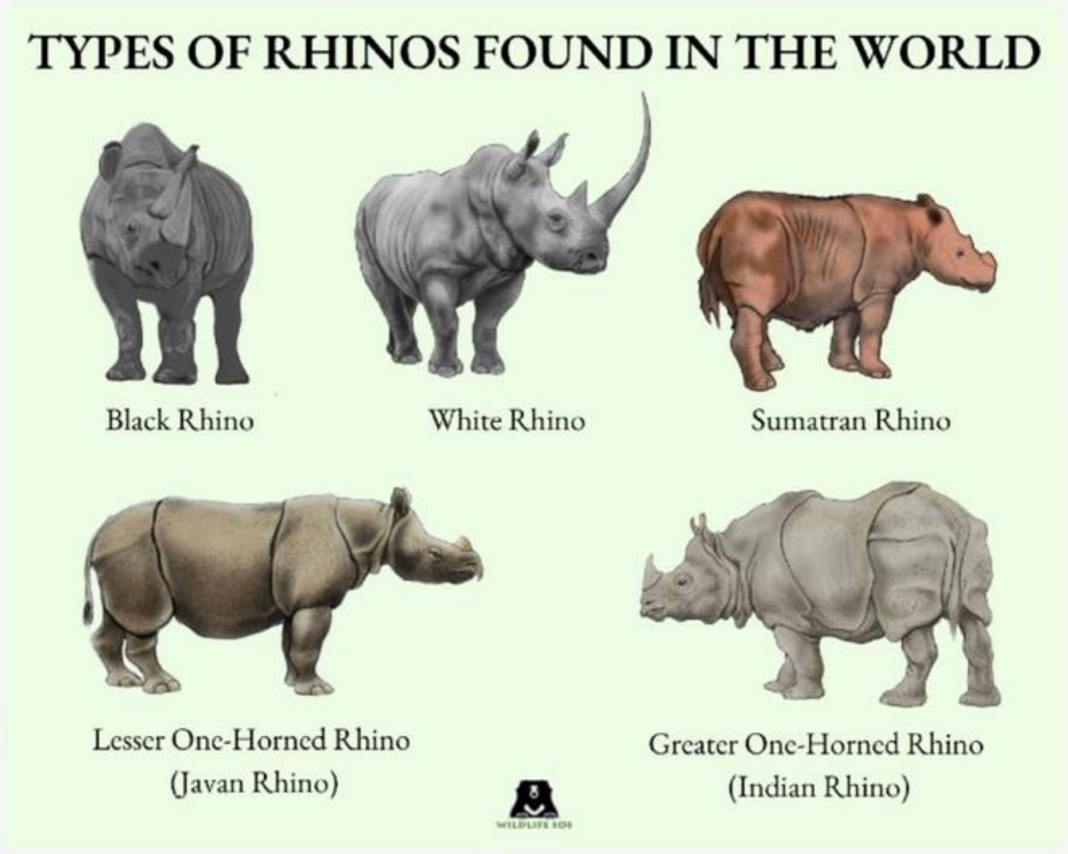


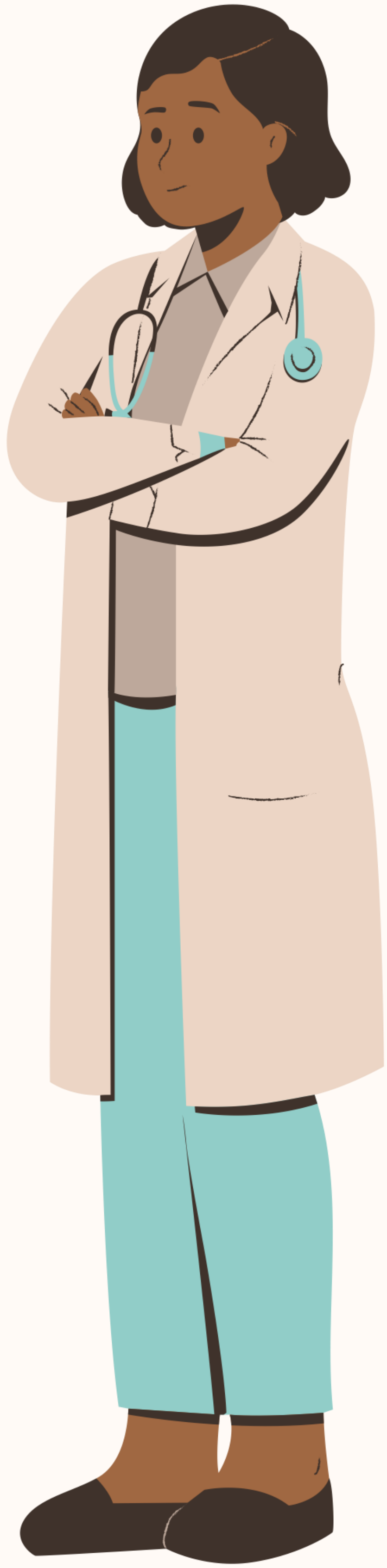
Vernon Rayford,MD, PharmD, FAAP, FACP

Primary Care IM: a rural update

Financial Disclosures: None



Meet the rhinos of the world. [Graphic (c) Wildlife SOS/Teesta Mukherjee]



Primary Care Update

Rural General Internal Medicine

Describe current challenges in Rural General Internal Medicine

Describe the General Internal Medicine Pipeline

Preview how artificial intelligence may impact General Internal Medicine

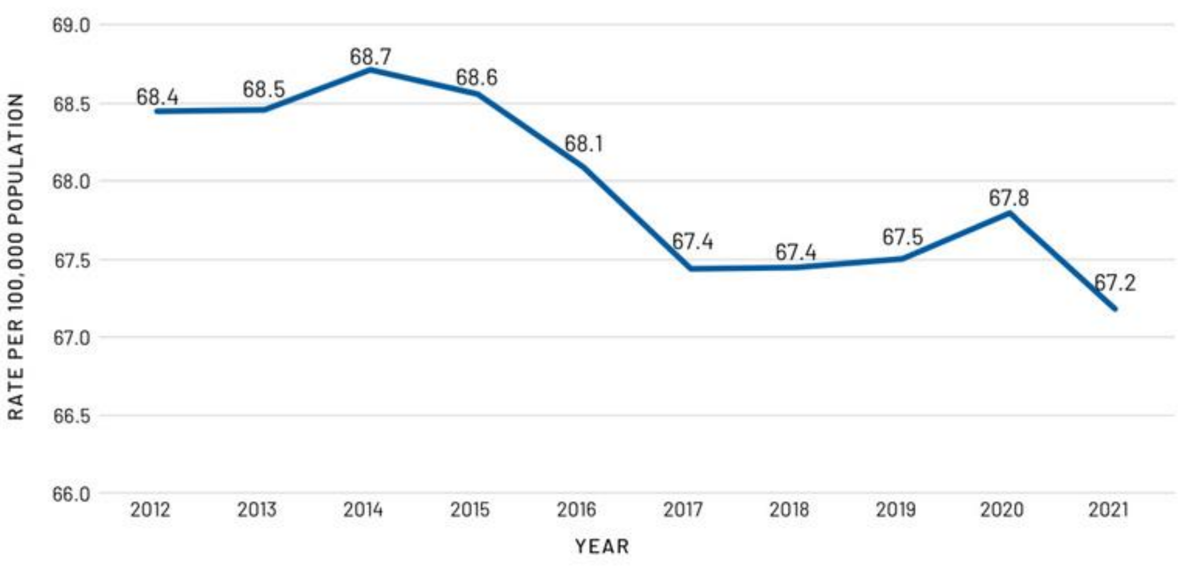
The Need

Table 2. Projected Primary Care Physician Need Under Various Conditions by Year

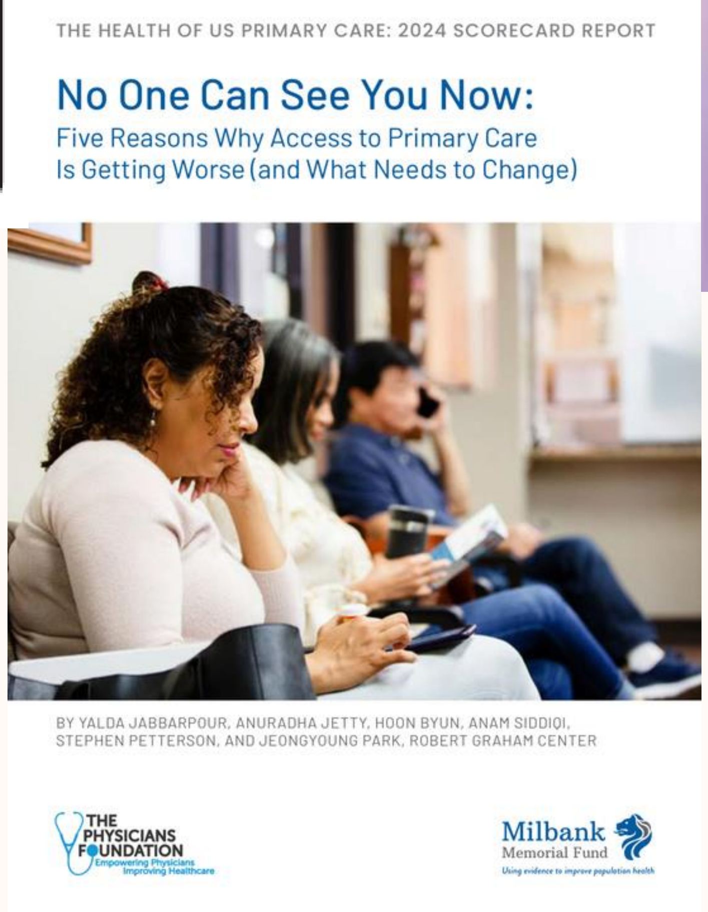
Condition	2010	2015	2020	2025
Baseline	209,662	209,662	209,662	209,662
Aging of population	–	2,693	6,264	9,894
Population growth	–	11,201	21,952	32,852
ACA coverage	–	7,104	8,097	8,279
Total	209,662	230,660	245,975	260,687

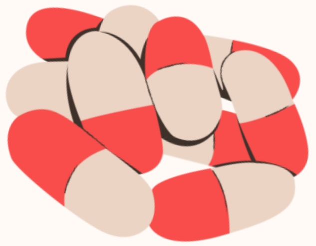
ACA = Affordable Care Act.

Figure 2. The Number of Primary Care Physicians per Capita Is Falling (2012–2021)



Data Source: Analyses of American Medical Association Masterfile (2012–2021), Centers for Medicare and Medicaid Services Physician and Other Practitioners data (2012–2021), and the American Community Survey Five-Year Summary Files (2012–2021).
Notes: Primary care specialties included family medicine, general practices, internal medicine, geriatrics, pediatrics, and osteopathy.





Stephen M. Petterson, Winston R. Liaw, Robert L. Phillips, David L. Rabin, David S. Meyers, Andrew W. Bazemore
The Annals of Family Medicine Nov 2012, 10 (6) 503-509; DOI: 10.1370/afm.1431

The Work

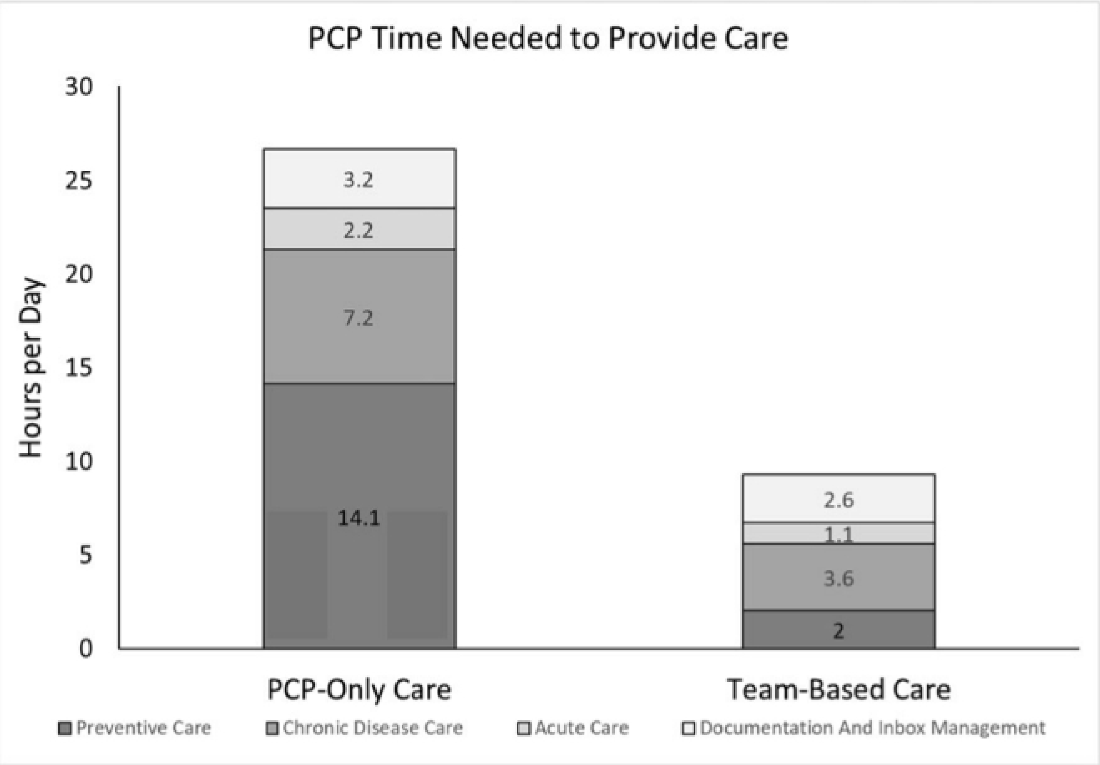
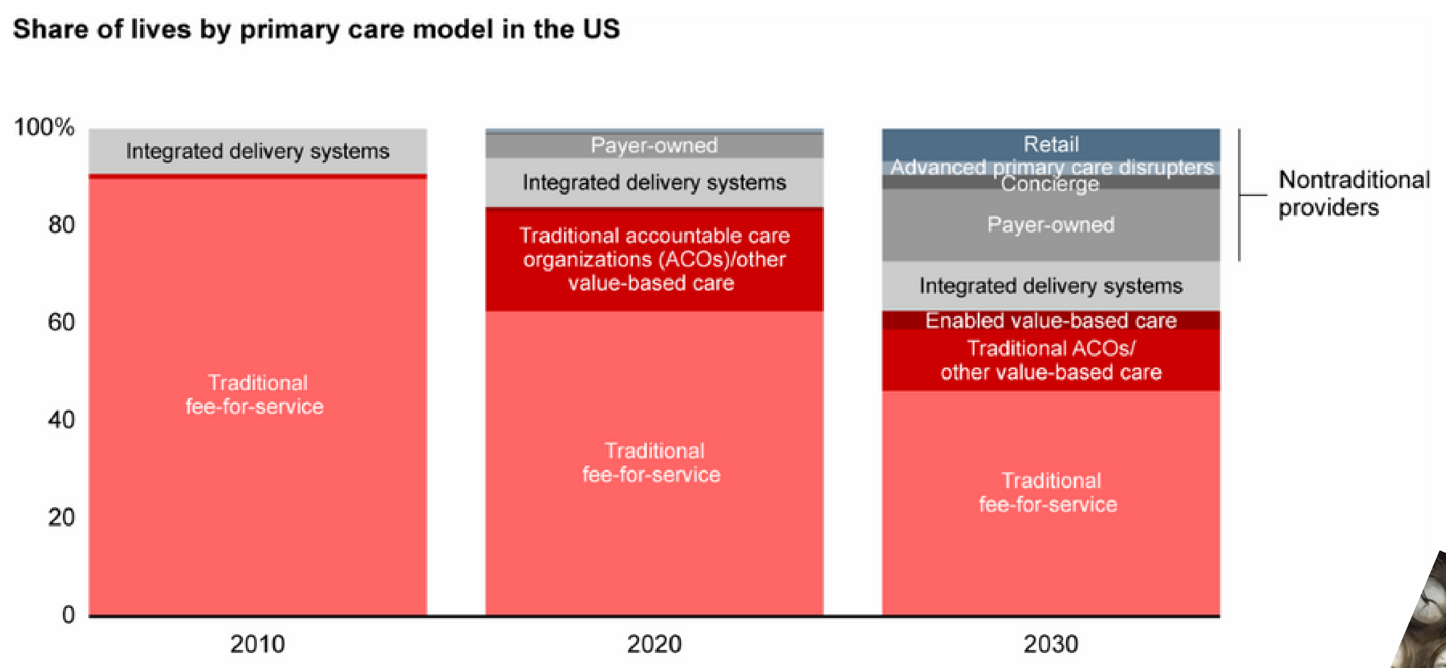
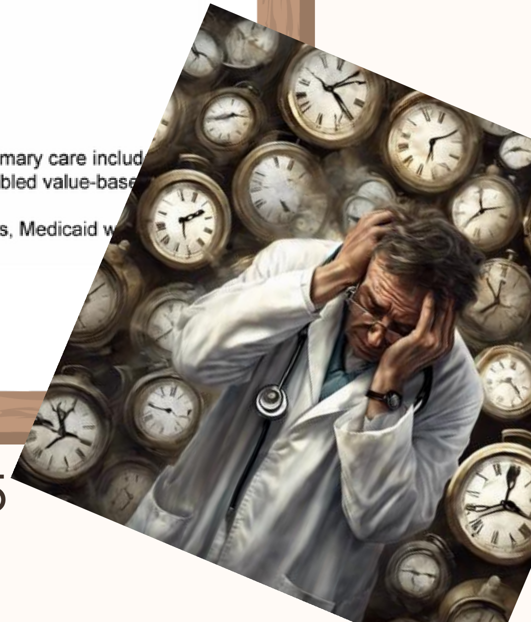


Fig. 2 Primary care provider time needed to provide care for average US adult panel of 2500 patients.



Notes: In addition to traditional fee-for-service, all primary care models shown include some fee-for-service reimbursement; payer-owned primary care includ UnitedHealth Group and Humana, and it considers that two additional payers will enter the market and achieve Humana's size by 2030; enabled value-base includes primary care providers that partner with other companies for access to value-based care capabilities
Sources: *Health Affairs*; National Ambulatory Medical Care Survey; United States Census Bureau; Centers for Medicare & Medicaid Services, Medicaid w American Academy of Family Physicians; Robert Graham Center; analyst reports and press releases



Porter, J., Boyd, C., Skandari, M.R. et al. Revisiting the Time Needed to Provide Adult Primary Care. J GEN INTERN MED 38, 147–155 (2023)

TABLE 2. Distress and Well-Being Results for Internal Medicine Hospitalists vs Outpatient General Internists

Variable	Hospitalists (n = 130)	Outpatient General Internists (n = 448)	*P**
Burnout			
Emotional exhaustion high (≥27)	57/130 (43.8%)	215/447 (48.1%)	0.71
Mean (SD)	24.7 (12.5)	25.4 (14.0)	
Median	24.9	26.0	
Depersonalization high (≥10)	55/130 (42.3%)	146/447 (32.7%)	0.17
Mean (SD)	9.1 (6.9)	7.5 (6.3)	
Median	7.0	6.0	
Personal accomplishment low (≤33)	26/128 (20.3%)	43/446 (9.6%)	0.04
Mean (SD)	39.0 (7.6)	41.4 (6.0)	
Median	41.0	43.0	
High burnout (EE ≥ 27 or DP ≥ 10)	68/130 (52.3%)	244/448 (54.5%)	0.86
Depression			
Depression screen +	52/129 (40.3%)	176/440 (40.0%)	0.73
Suicidal thoughts in past 12 months	12/130 (9.2%)	26/445 (5.8%)	0.15
Quality of life			
Overall mean (SD)	7.3 (2.0)	7.4 (1.8)	0.85
Median	8.0	8.0	
Low (<6)	21/130 (16.2%)	73/448 (16.3%)	
Mental mean (SD)	7.2 (2.1)	7.3 (2.0)	0.89
Median	8.0	8.0	
Low (<6)	23/130 (17.7%)	92/448 (20.5%)	
Physical mean (SD)	6.7 (2.3)	6.9 (2.1)	0.45
Median	7.0	7.0	
Low (<6)	35/130 (26.9%)	106/448 (23.7%)	
Emotional mean (SD)	7.0 (2.3)	6.9 (2.2)	0.37
Median	7.0	7.0	
Low (<6)	30/130 (23.1%)	114/448 (25.4%)	
Fatigue			
Mean (SD)	5.8 (2.4)	5.9 (2.4)	0.57
Median	6.0	6.0	
Fallen asleep while driving (among regular drivers only)	11/126 (8.7%)	19/438 (4.3%)	0.23

SIMILAR AND SUBSTANTIAL

2014

Roberts, D.L., Shanafelt, T.D., Dyrbye, L.N. and West, C.P. (2014), A national comparison of burnout and work-life balance among internal medicine hospitalists and outpatient general internists. J. Hosp. Med., 9: 176-181. https://doi.org/10.1002/jhm.2146

ALTHOUGH MOST…REPORTED CAREER SATISFACTION, BURNOUT WAS HIGH

Table. Overall and Sex-Specific Scores on Satisfaction, Stress, and Burnout and Risk Factors for Burnout Among Internists and Trainees Enrolled in a Well-being Champion Program

Survey item or score (response)	Overall	Female[a]	Male[a]	OR (95% CI)[b]	*P* value
Participants, No. (%)	1305 (100)	605 (47.6)	665 (52.4)	NA	NA
Satisfaction with current job (agree or strongly agree)	938 (71.9)	427 (70.6)	492 (74.0)	0.84 (0.66-1.08)	.18
Burnout symptoms (present to severe)	680 (52.1)	351 (58.0)	312 (46.9)	1.56 (1.25-1.95)	<.001
Values aligned with those of clinical leaders (agree or strongly agree)	816 (62.5)	363 (60.0)	438 (65.9)	0.78 (0.62-0.98)	.03
My care team works efficiently together (satisfactory to optimal)	1128 (86.4)	522 (86.3)	581 (87.4)	0.91 (0.66-1.26)	.57
Personal control over workload (Poor or minimal)	419 (32.1)	206 (34.0)	196 (29.5)	0.81 (0.64-1.03)	.08
Feeling a great deal of stress (agree or strongly agree)	730 (55.9)	376 (62.1)	334 (50.2)	1.63 (1.30-20.4)	<.001
Sufficient time for documentation(poor, marginal)	673 (51.6)	315 (52.1)	335 (50.4)	1.07 (0.86-1.33)	.55
Time spent on EMR at home (moderately high to excessive)	552 (42.3)	268 (44.3)	263 (39.5)	1.22 (0.97-1.52)	.09
EMR adds frustration to the day (agree or strongly agree)	850 (65.1)	383 (63.3)	443 (66.6)	0.86 (0.69-1.09)	.22
Work atmosphere (chaotic or tending toward chaotic)	390 (29.9)	191 (31.6)	188 (28.3)	1.17 (0.92-1.49)	.20
Summary score ≥40 (joyous workplace)[c]	151 (11.6)	42 (6.9)	107 (16.1)	0.39 (0.26-0.56)	<.001
Subscale 1 score ≥20 (supportive workplace)[d]	466 (35.7)	182 (30.1)	275 (41.4)	0.61 (0.48-0.77)	<.001
Subscale 2 score ≥ 20 (manageable work pace and EMR stress)[e]	117 (9.0)	32 (5.3)	83 (12.5)	0.39 (0.25-0.59)	<.001

Abbreviations: EMR, electronic medical record; NA, not applicable; OR, odds ratio.

[a] Of 1305 respondents, 35 chose not to indicate their sex and are not included in this table.

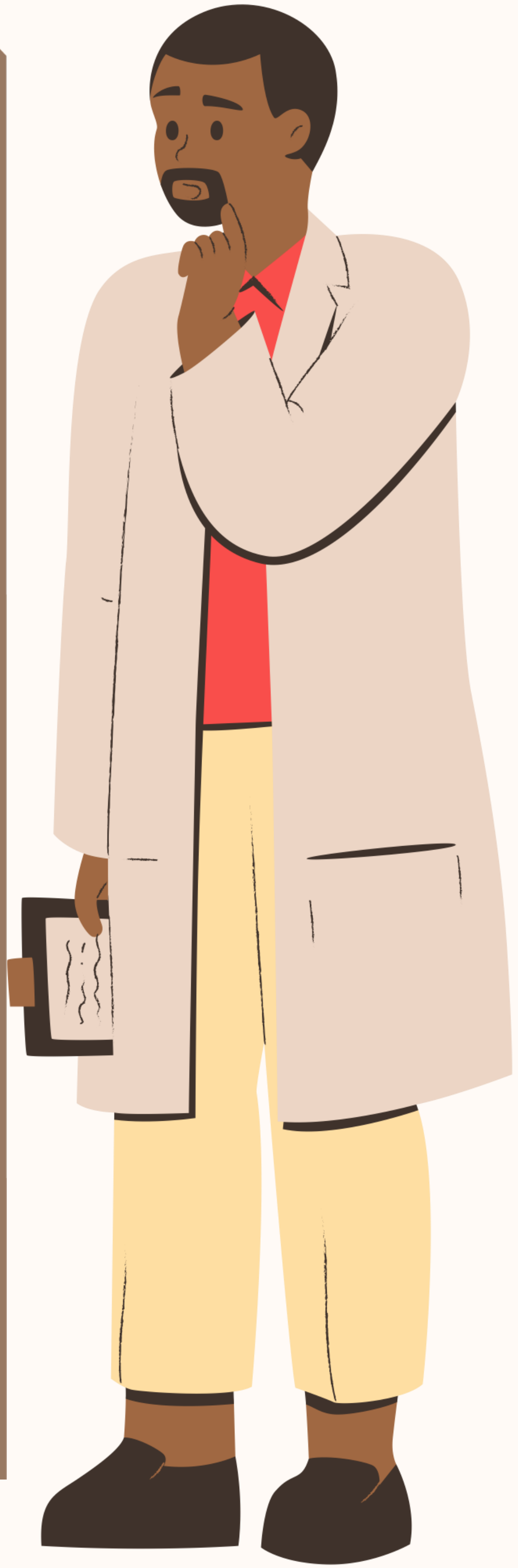
[b] All ORs from single logistic regression models are for women compared with men.

[c] Summary score range 10 to 50. Mean (SD) score: 30.9 (7.4).

[d] Subscale 1 (including items 1-5) score range 5 to 25. Mean (SD) score: 17.5 (4.1).

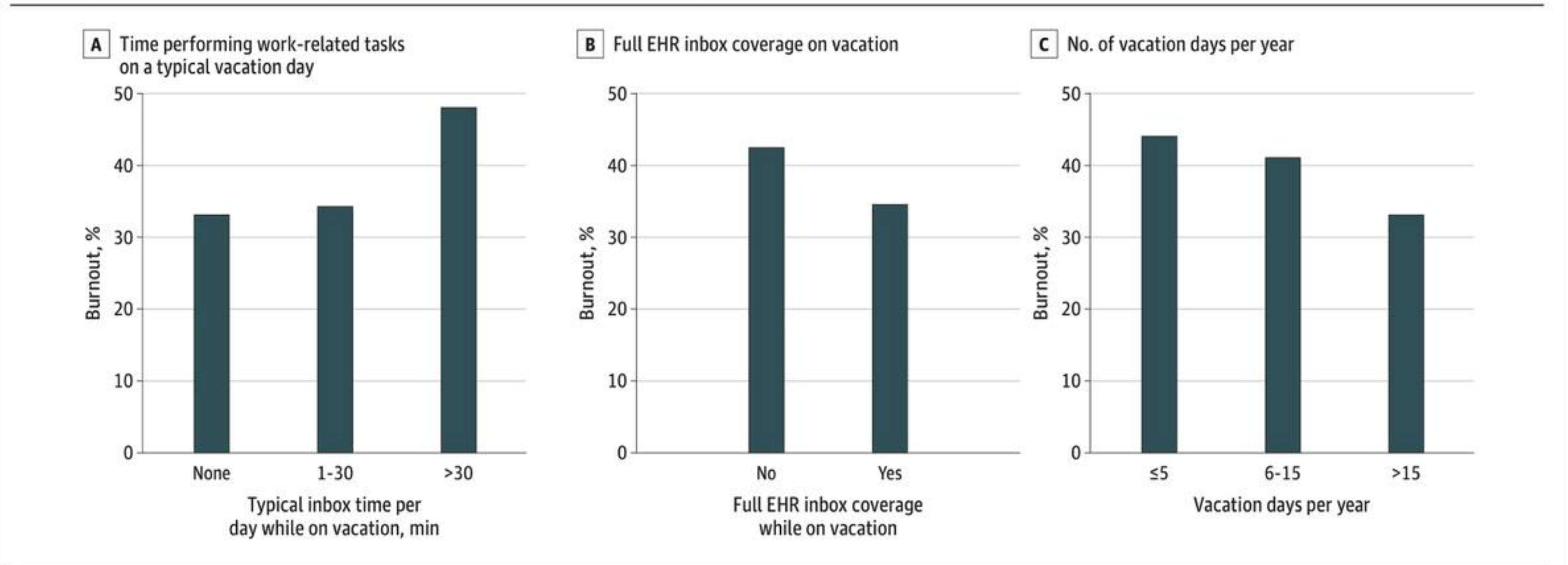
[e] Subscale 2 (including items 6-10) score range 5 to 25. Mean (SD) score: 13.4 (4.1).

Linzer M, Smith CD, Hingle S, et al. Evaluation of Work Satisfaction, Stress, and Burnout Among US Internal Medicine Physicians and Trainees. JAMA Netw Open. 2020;3(10):e2018758. doi:10.1001/jamanetworkopen.2020.18758

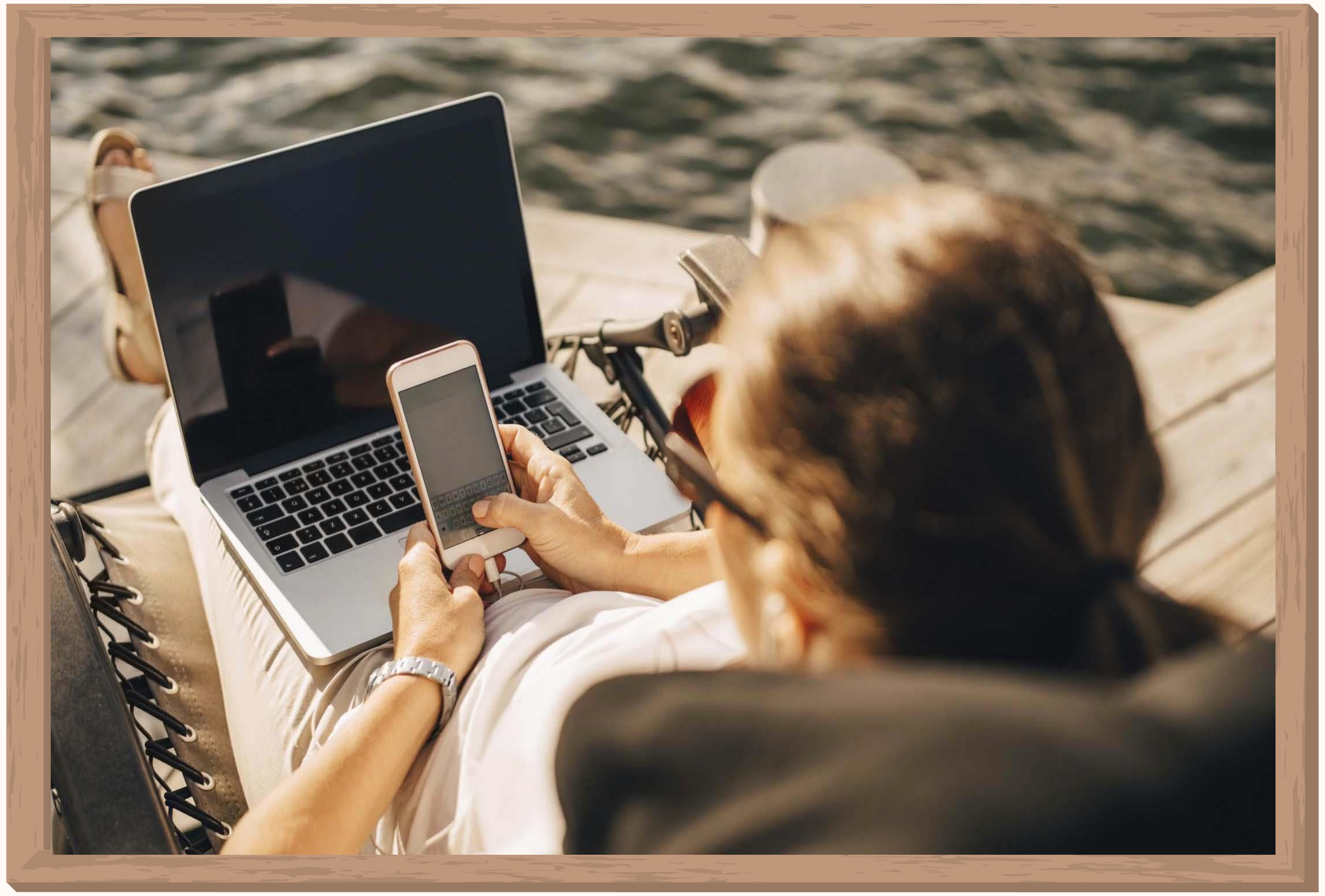


Burnout Solution: Vacation?

Figure. Personal and Institutional Vacation Behaviors and Prevalence of Burnout

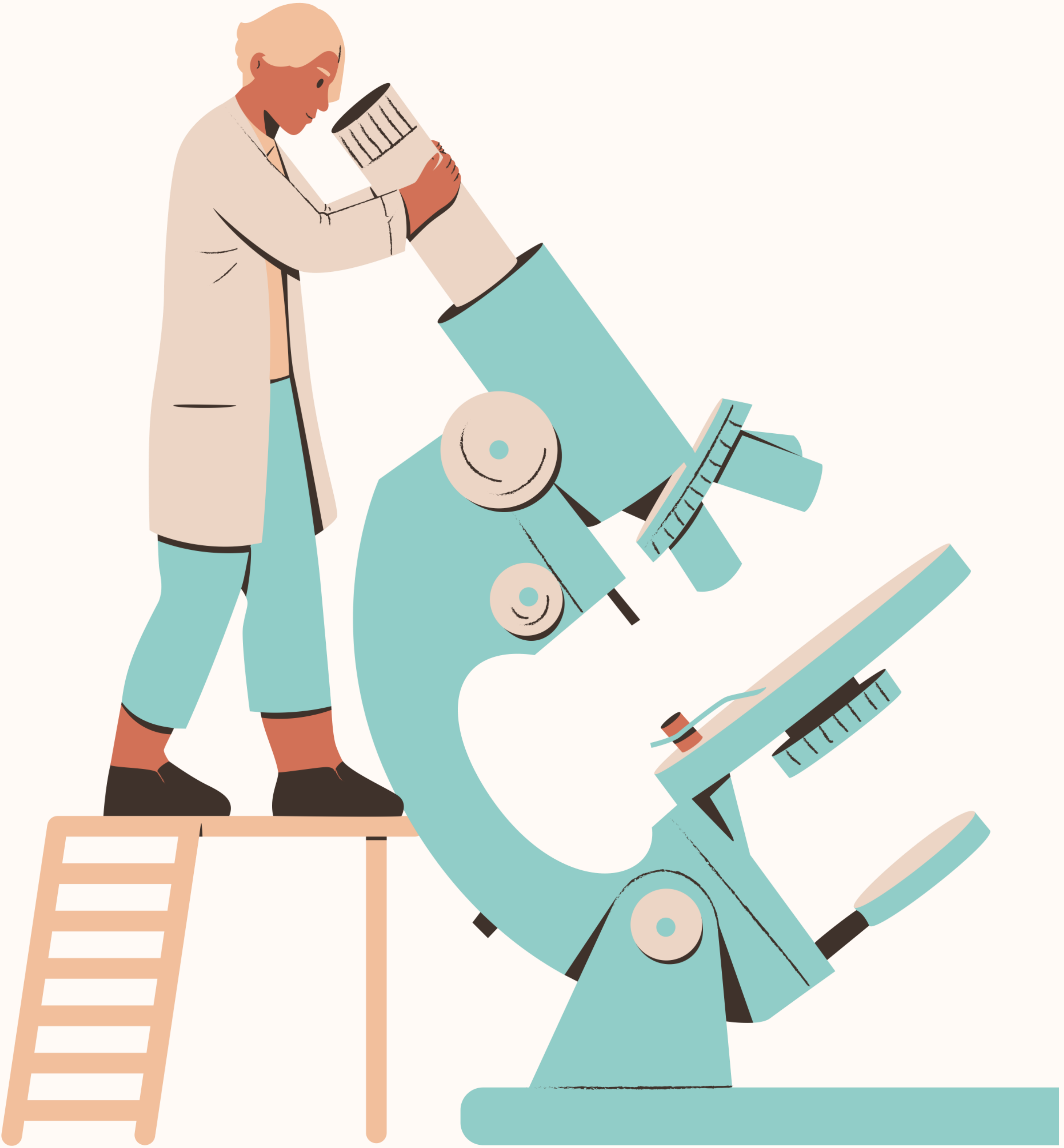


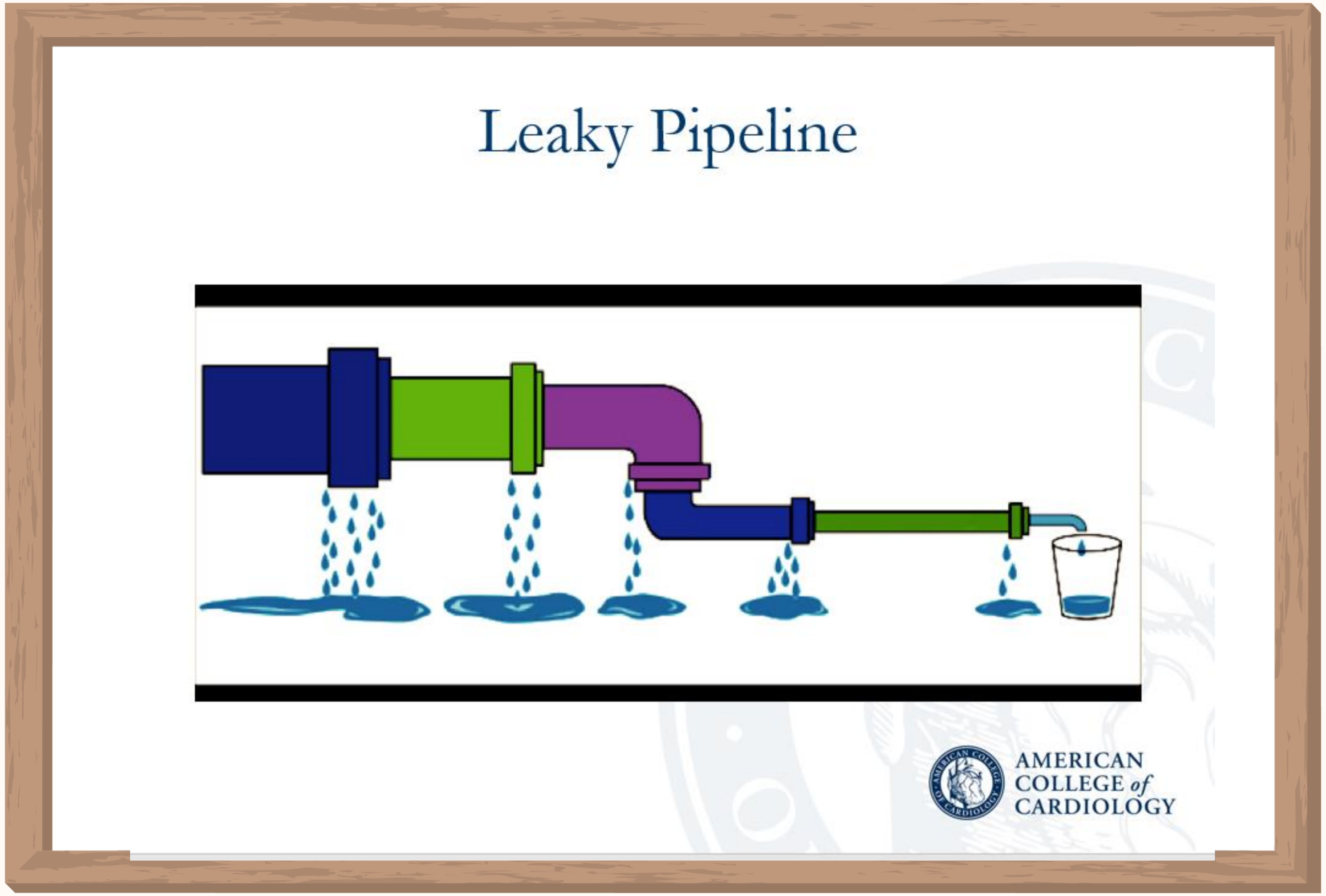
Graphs show burnout rates in relation to time performing work-related tasks on a typical vacation day (A), full electronic health record (EHR) inbox coverage during vacation (B), and number of vacation days per year (C).



Sinsky CA, Trockel MT, Dyrbye LN, et al. Vacation Days Taken, Work During Vacation, and Burnout Among US Physicians. JAMA Netw Open. 2024;7(1):e2351635. doi:10.1001/jamanetworkopen.2023.51635

The Rural General IM pipeline





Medical students

Table 2. M1-2 factors associated with matching into a primary care specialty. Dependent variable is matching into a primary care specialty. Primary care is comprised of internal medicine, pediatrics, and family medicine specialties. Adjusted odds ratios are presented, with standard errors in parentheses. Additional controls include whether individuals had a primary care mentor in first 2 years, performed primary care research in first 2 years, the subjective importance of academic vs. private practice opportunities and intellectual stimulation, none of which had significant associations. Lifestyle and debt responses taken from M2 survey. Pseudo R^2 is McFadden's. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ for odds ratio different from 1.

	Pooled	Men	Women	Pooled	Men	Women
Female	2.21*** (0.62)	-	-	2.06** (0.64)	-	-
Ethnicity is White	0.58 (0.20)	0.42* (0.22)	0.63 (0.30)	0.54* (0.19)	0.54 (0.29)	0.34* (0.20)
Age at Matriculation	1.12 (0.08)	1.00 (0.10)	1.43* (0.26)	1.14 (0.09)	0.96 (0.11)	1.57** (0.30)
Family Member Practices Primary Care	3.41** (1.78)	2.13 (1.47)	6.67** (6.16)	3.53** (2.00)	2.30 (1.86)	17.16*** (18.86)
Married in 1st 2 Years	1.70 (0.65)	3.36** (1.79)	1.03 (0.60)	1.92 (0.78)	3.15** (1.83)	1.58 (1.16)
Has children in 1st 2 Years	0.09** (0.09)	0.12* (0.15)	0.05 (0.09)	0.08** (0.08)	0.13 (0.18)	0.02 (0.05)
Amount of time in patient contact	-	-	-	0.81** (0.09)	0.81 (0.13)	0.87 (0.16)
Potential Salary	-	-	-	0.78* (0.10)	0.66** (0.13)	0.95 (0.24)
Quality of Life	-	-	-	1.05 (0.11)	1.17 (0.18)	0.96 (0.18)
Responsibilities at home	-	-	-	1.15 (0.14)	1.33 (0.27)	1.03 (0.18)
Specialty status/reputation	-	-	-	1.21 (0.14)	1.46** (0.28)	1.03 (0.18)
Spouse/partner's career	-	-	-	1.33** (0.17)	1.13 (0.23)	1.99*** (0.49)
Technical skills necessary	-	-	-	1.06 (0.13)	0.87 (0.15)	1.46 (0.34)
Debt from Medical Education	1.55 (0.66)	1.64 (1.03)	1.38 (0.85)	1.67 (0.76)	2.49 (1.77)	1.87 (1.31)
Debt Influences Specialty Preference	0.99 (0.27)	0.61 (0.26)	1.49 (0.57)	0.91 (0.26)	0.58 (0.27)	1.71 (0.79)
Constant	0.03** (0.05)	0.48 (1.18)	0.00** (0.00)	0.00** (0.01)	0.62 (1.96)	0.00*** (0.00)
Observations	273	153	120	264	148	116
Pseudo R^2	0.08	0.09	0.08	0.12	0.17	0.20

McDonald C, Henderson A, Barlow P, Keith J. Assessing factors for choosing a primary care specialty in medical students: A longitudinal study. Med Educ Online. 2021 Dec;26(1):1890901. doi: 10.1080/10872981.2021.1890901. PMID: 33829968; PMCID: PMC8043606.

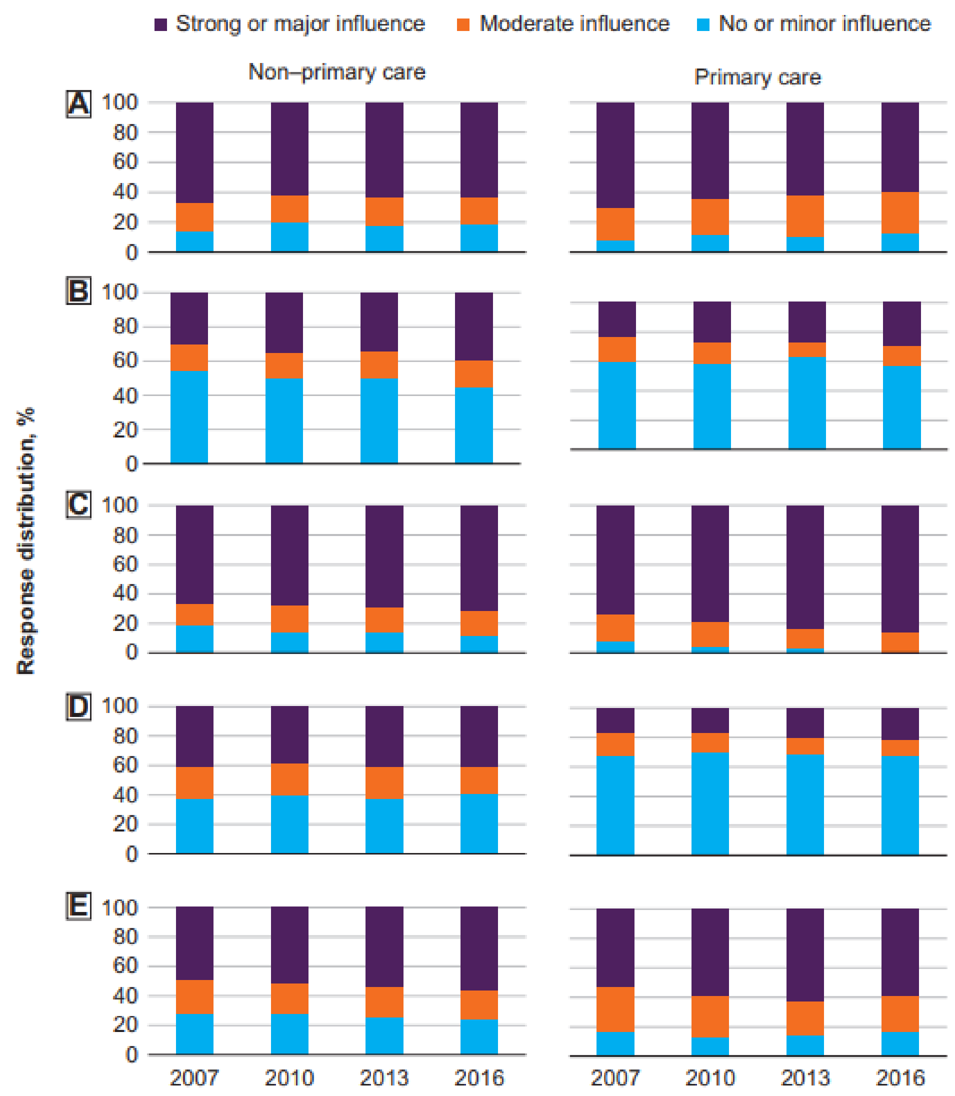


Figure 2.
Response distribution (%) of the influence of 5 key factors in an osteopathic medical student's choice to pursue primary care or a non–primary care specialty by year. (A) intellectual and technical content of the specialty; (B) debt level; (C) lifestyle; (D) prestige; (E) personal experience and abilities.

Stefani, Katherine M., Richards, Jesse R., Newman, Jessica, Poole, Kenneth G., Scott, Shannon C. and Scheckel, Caleb J.. "Choosing Primary Care: Factors Influencing Graduating Osteopathic Medical Students" Journal of Osteopathic Medicine, vol. 120, no. 6, 2020, pp. 380-387. https://doi.org/10.7556/jaoa.2020.060

Internal Medicine Residents Career Intentions

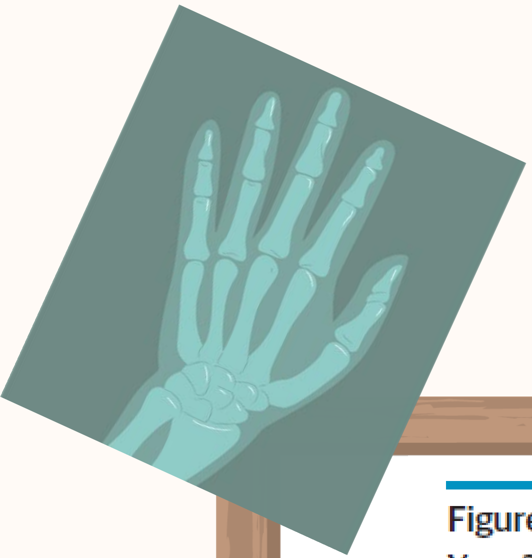
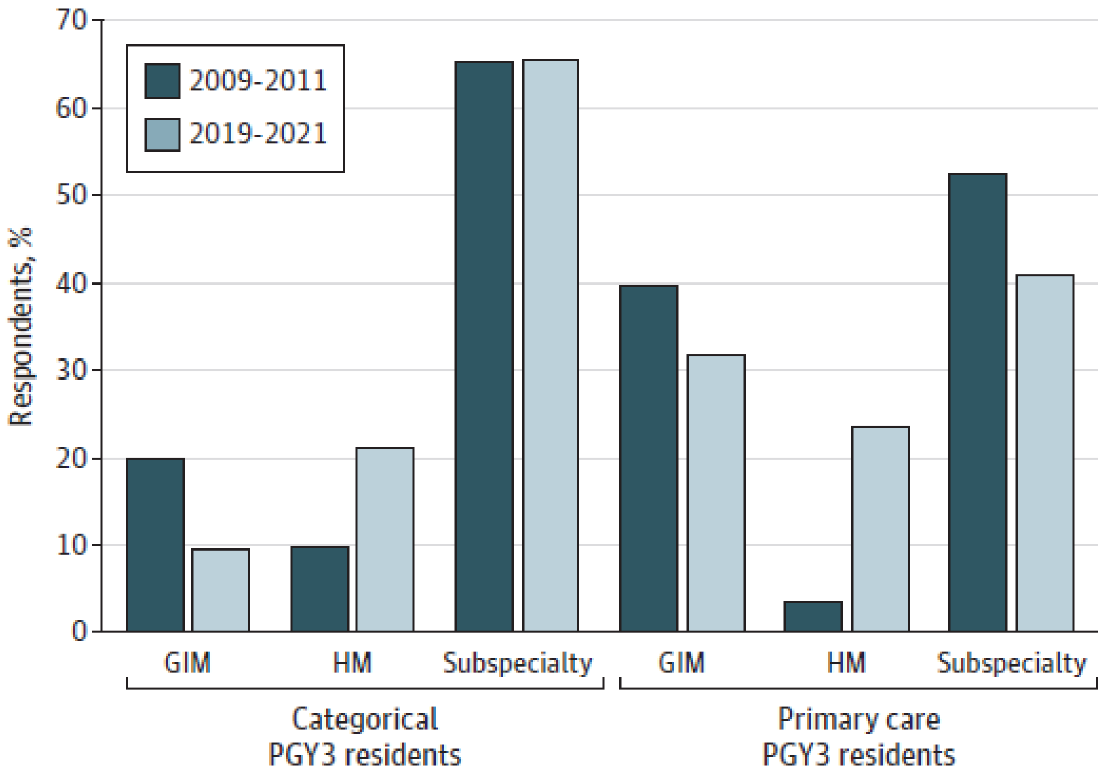


Figure. Career Plans for Categorical and Primary Care Postgraduate Year 3 (PGY3) Residents From 2009 to 2011 vs 2019 to 2021



GIM indicates general internal medicine; HM, hospital medicine.

Table 2 Interest in Primary Care Over Time

	Total *N*=172	Primary care *N*=94	Non-primary care *N*=78	*p* value
Interested in primary care prior to residency*	160 (94.1%)	88 (95.7%)	72 (92.3%)	.51
Interested in primary care at conclusion of residency	117 (68.0%)	88 (93.6%)	29 (37.2%)	<.001
Practicing primary care	94 (54.7%)	–		

**Percentages calculated based on the 170 participants who answered that they were interested in primary care prior to residency and 172 who answered they were interested in primary care at conclusion of residents and currently practicing*

Paralkar N, LaVine N, Ryan S, et al. Career Plans of Internal Medicine Residents From 2019 to 2021. JAMA Intern Med. 2023;183(10):1166–1167. doi:10.1001/jamainternmed.2023.2873

Kryzhanovskaya I, Cohen BE, Kohlwes RJ. Factors Associated with a Career in Primary Care Medicine: Continuity Clinic Experience Matters. J Gen Intern Med. 2021 Nov;36(11):3383-3387. doi: 10.1007/s11606-021-06625-8. Epub 2021 Feb 23. PMID: 33620629; PMCID: PMC8606375.

The Continuity Clinic

Table 3 Factors Influencing Towards a Career in Primary Care*

	Total	Primary care *N*=94	Non-primary care *N*=78	*p* value
Overall continuity clinic experience (*n*=172)	79 (45.9%)	57 (60.6%)	22 (28.2%)	<.001
Patient-physician relationship (*n*=165)	142 (86.1%)	86 (94.5%)	56 (75.7%)	.001
Access to role models (*n*=166)	114 (68.7%)	73 (80.2%)	41 (54.7%)	.001
Patient population (*n*=166)	106 (63.9%)	68 (74.7%)	38 (50.7%)	.002
Curriculum (*n*=164)	100 (61.0%)	59 (66.3%)	41 (54.7%)	.13

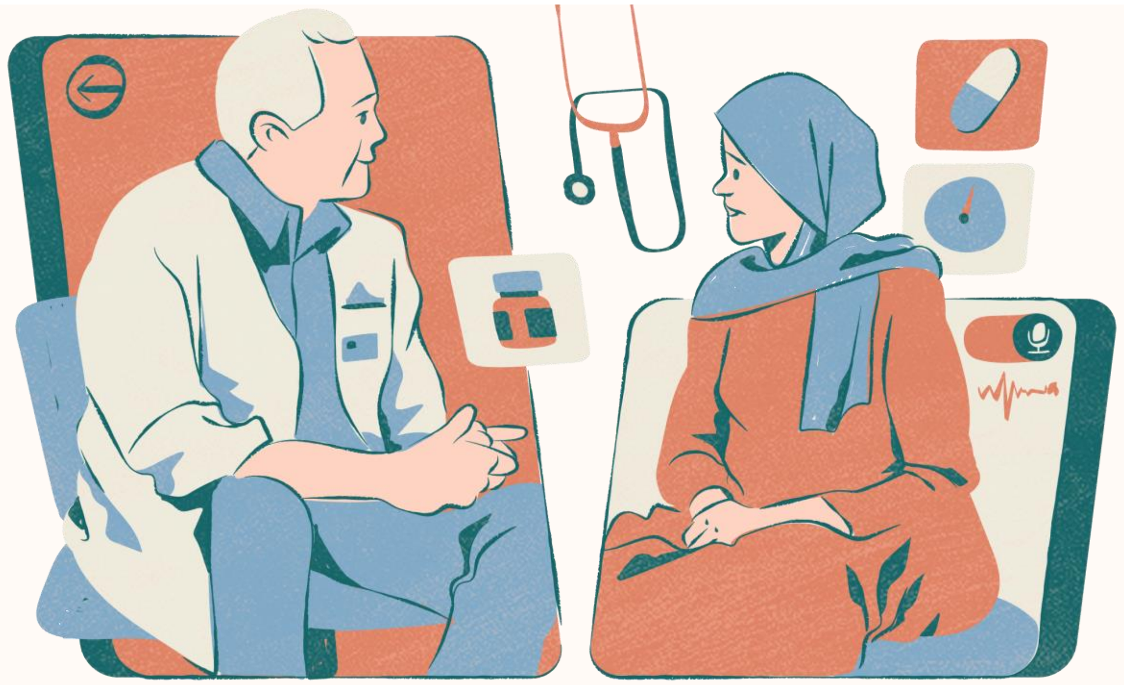


Table 4 Factors Influencing Away From a Career in Primary Care

	Total	Primary care *N*=94	Non-primary care *N*=78	*p* value
Support staff availability (*n*=165)	75 (45.5%)	41 (45.6%)	34 (45.3%)	1.0
Clerical duties (*n*=166)	99 (59.6%)	56 (61.5%)	43 (57.3%)	.64
Documentation (*n*=164)	74 (45.1%)	40 (44.9%)	34 (45.3%)	1.0
Time pressure (*n*=166)	108 (65.1%)	53 (58.2%)	55 (73.3%)	.05

Kryzhanovskaya I, Cohen BE, Kohlwes RJ. Factors Associated with a Career in Primary Care Medicine: Continuity Clinic Experience Matters. J Gen Intern Med. 2021 Nov;36(11):3383-3387. doi: 10.1007/s11606-021-06625-8. Epu
2021 Feb 23. PMID: 33620629; PMCID: PMC8606375.

Rural Recruitment

Table 2 Program and Incentive Frequency Count By Geographic Region and Number of HPSAs in Each Region By HPSA Score

Region	States	Programs	HPSAs 1–13	HPSAs 14–17	HPSAs 18+
Northeast	Connecticut; Maine; Massachusetts; New Hampshire; New Jersey; New York; Pennsylvania; Rhode Island; Vermont	33	34	7	0
Midwest	Illinois; Indiana; Iowa; Kansas; Michigan; Minnesota; Missouri; Nebraska; North Dakota; Ohio; South Dakota; Wisconsin	68	270	66	11
South	Alabama; Arkansas; Delaware; District of Columbia; Florida; Georgia; Kentucky; Louisiana; Maryland; Mississippi; North Carolina; Oklahoma; South Carolina; Tennessee; Virginia; West Virginia; Texas	88	276	197	78
West	Alaska; Arizona; California; Colorado; Hawaii; Idaho; Montana; Nevada; New Mexico; Oregon; Utah; Washington; Wyoming	74	165	140	24

Note. Some programs and incentives were offered in multiple geographic regions; therefore, the total count exceeds the total number of individual programs and incentives

Table 1 Frequency Count of What Stage in the Student to Physician Pipeline Programs and Incentives Target

Stage	Frequency
High School or Earlier	11
Undergraduate	18
Medical School	85
Residency	63
Early Career	3
Leadership	6
Licensed Physician*	69
Non-specified stage of career	8

Note. Some programs and incentives were open to individuals at multiple stages, therefore the total count exceeds the total number of individual programs and incentives. *Just specifies that applicants must have a practicing license

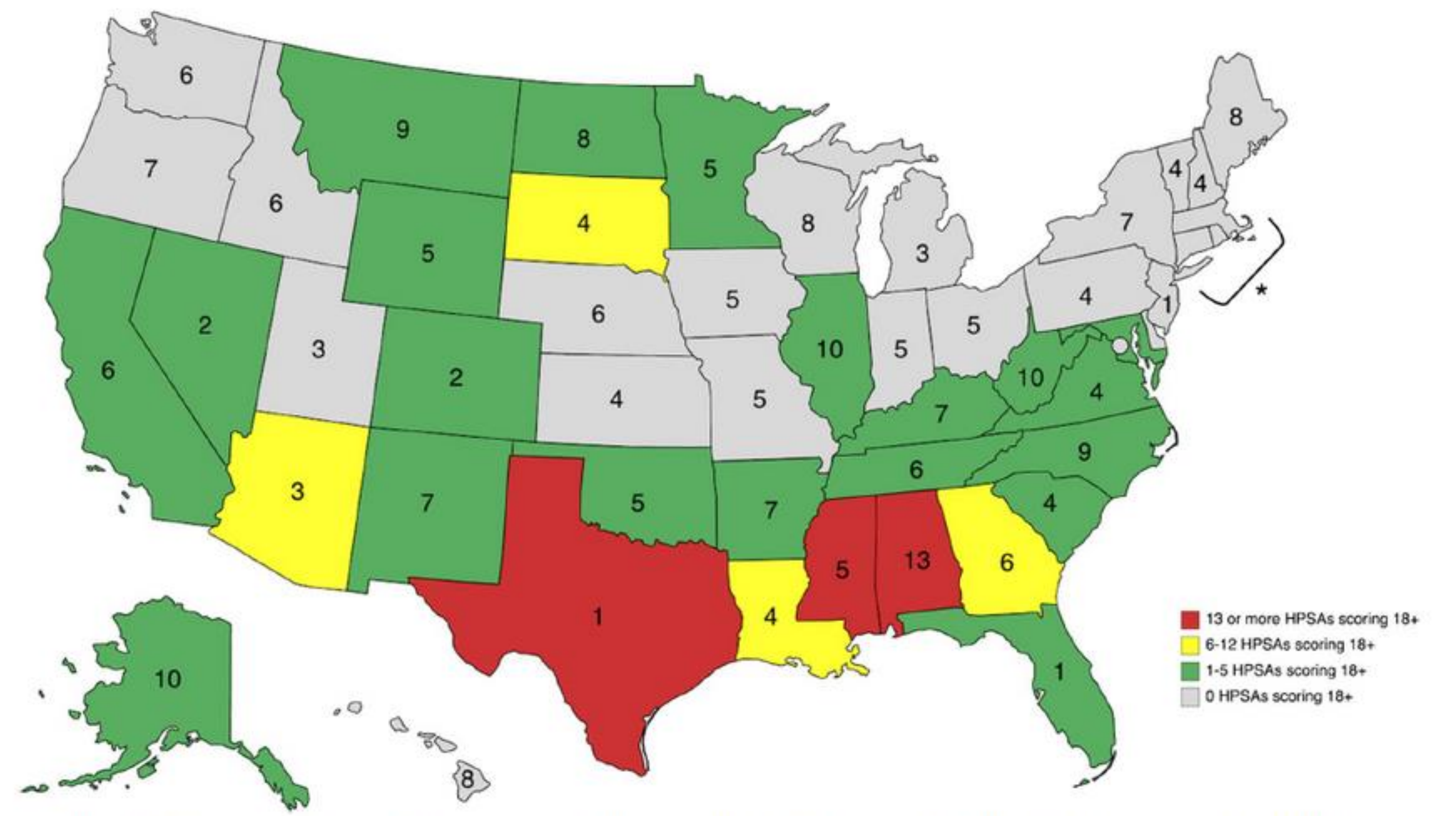


Figure 1 Program and incentive frequency count with geographic primary care Health Professional Shortage Areas (HPSAs)

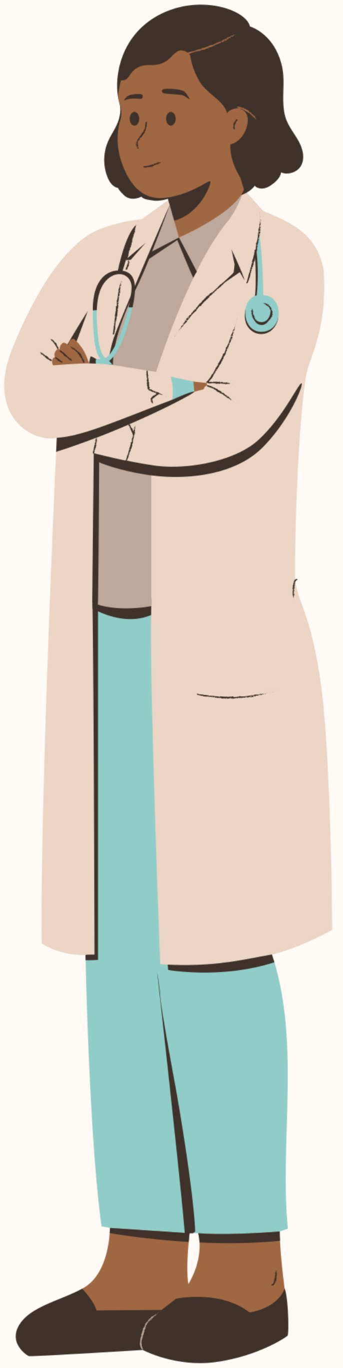
Arredondo, K., Touchett, H.N., Khan, S. et al. Current Programs and Incentives to Overcome Rural Physician Shortages in the United States: A Narrative Review. J GEN INTERN MED 38 (Suppl 3), 916–922 (2023). https://doi.org/10.1007/s11606-023-08122-6

Med Peds Physicians and Primary Care

Table 2. Residents' current career plans.

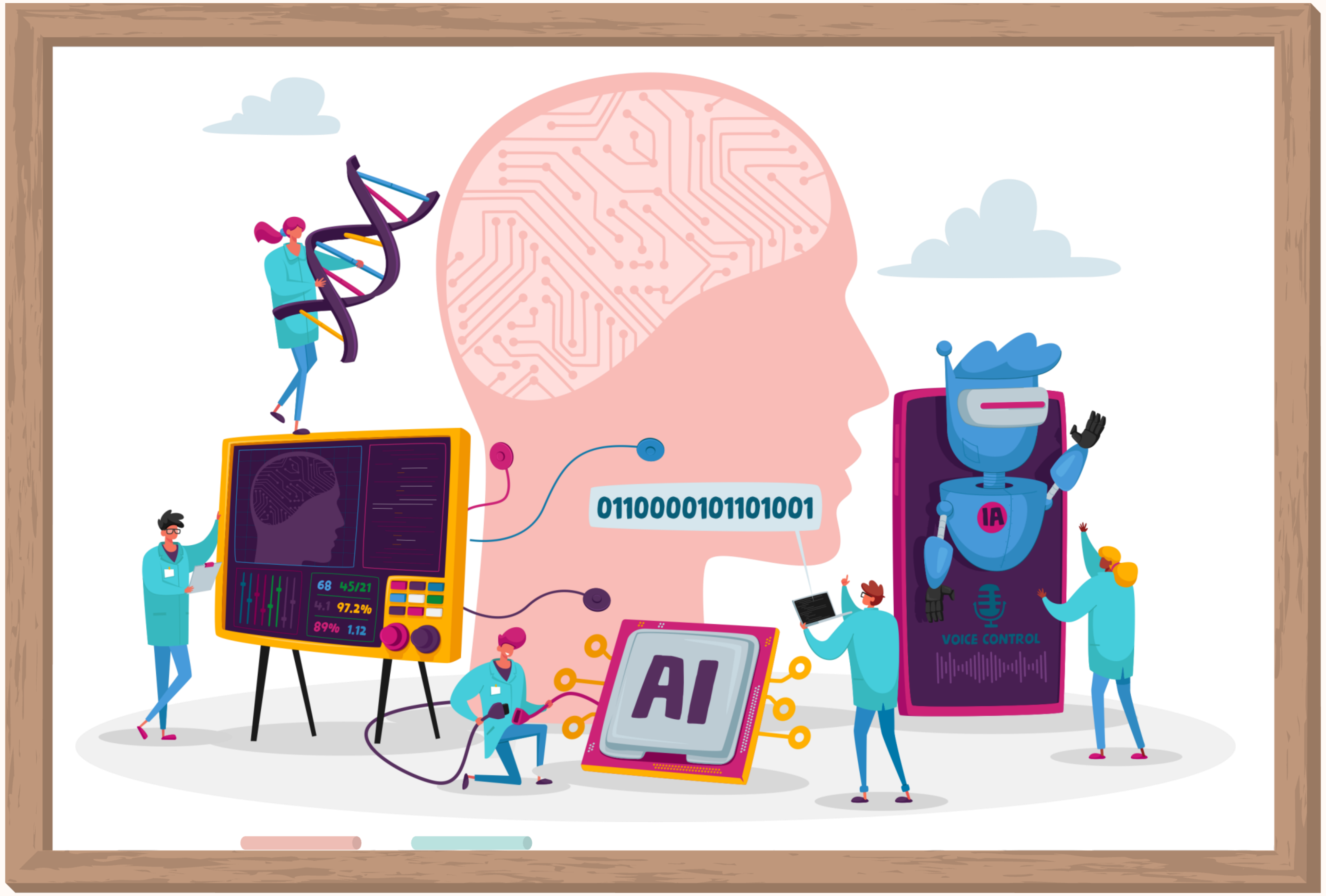
	Combined hospital med N (%)	Adult hospital med N (%)	Pediatric hospital med N (%)	Combined subspecialty N (%)	Adult subspecialty N (%)	Pediatric subspecialty N (%)	Primary care N (%)	Other N (%)
All responses (N = 228)[a]	84 (36.8)	15 (6.6)	3 (1.3)	74 (32.5)	24 (10.5)	14 (6.1)	71 (31.1)	4 (1.8)
Year in training[b]								
PGY-1 (N = 45)	23 (51.1)	2 (4.4)	0 (0.0)	24 (53.3)	2 (4.4)	1 (2.2)	17 (37.8)	0 (0.0)
PGY-2 (N = 53)	17 (32.1)	3 (5.7)	1 (1.9)	24 (45.3)	1 (1.9)	5 (9.4)	14 (26.4)	1 (1.9)
PGY-3 (N = 55)	19 (34.5)	4 (7.3)	0 (0.0)	12 (21.8)	8 (14.5)	5 (9.1)	17 (30.9)	0 (0.0)
PGY-4 (N = 72)	25 (55.6)	6 (13.3)	2 (4.4)	13 (28.9)	12 (26.7)	3 (6.7)	22 (48.9)	2 (4.4)
Racial/Ethnic identity								
White (N = 149)	52 (34.9)	9 (6.0)	3 (2.0)	43 (28.9)	16 (10.7)	12 (8.1)	48 (32.2)	1 (0.7)
Non-White (N = 69)	30 (43.5)	3 (4.3)	0 (0.0)	28 (40.6)	6 (8.7)	2 (2.9)	19 (27.5)	3 (4.3)
Gender identity								
Female (N = 142)	51 (35.9)	10 (7.0)	2 (1.4)	49 (34.5)	9 (6.3)	9 (6.3)	46 (32.4)	4 (2.8)
Male (N = 81)	33 (40.7)	3 (1.7)	1 (1.2)	24 (29.6)	13 (16.0)	5 (6.2)	22 (27.2)	0 (0.0)
Student loan debt								
≤ $200K (N = 104)	37 (35.6)	7 (6.7)	1 (1.0)	38 (36.5)	12 (11.5)	5 (4.8)	30 (28.8)	3 (2.9)
> $200K (N = 111)	43 (38.7)	8 (7.2)	2 (2.8)	31 (27.9)	10 (9.0)	8 (7.2)	38 (34.2)	1 (0.9)
Family status[b]								
No children (N = 197)	70 (35.5)	12 (6.1)	2 (1.0)	71 (36.0)	17 (8.6)	13 (6.6)	60 (30.5)	4 (2.0)
1+ children (N = 25)	14 (56.0)	1 (4.0)	1 (4.0)	2 (8.0)	5 (20.0)	0 (0.0)	8 (32.0)	0 (0.0)

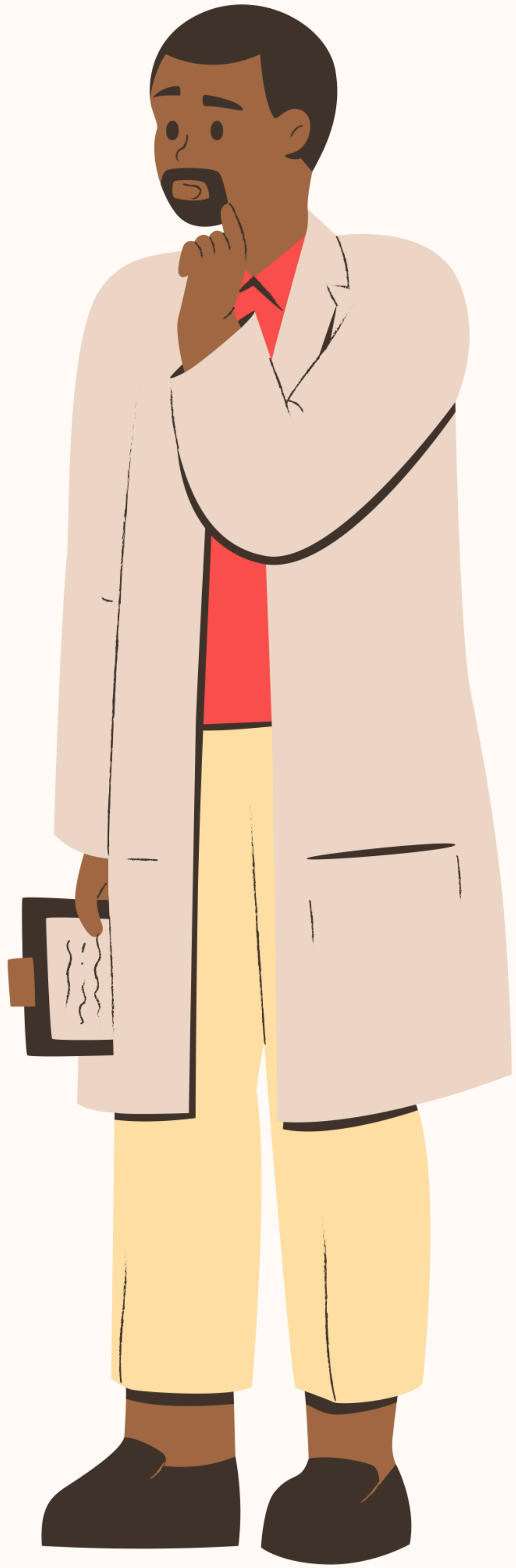
[a]Percentages total >100% as more than one response was allowed.
[b]$p < 0.05$ on Chi-square test for independence.



McCarthy JJ, Swartz S, Liljestrom T, Menigo J, Cotter L, Toth H. Med-Peds Residents' Career Plans and the Impact of Pediatric Hospital Medicine Fellowship: A Nationwide Survey. Hosp Top. 2023;101(4):336-343. doi: 10.1080/00185868.2022.2063775. Epub 2022 Apr 12. PMID: 35414350.

Technology and General IM



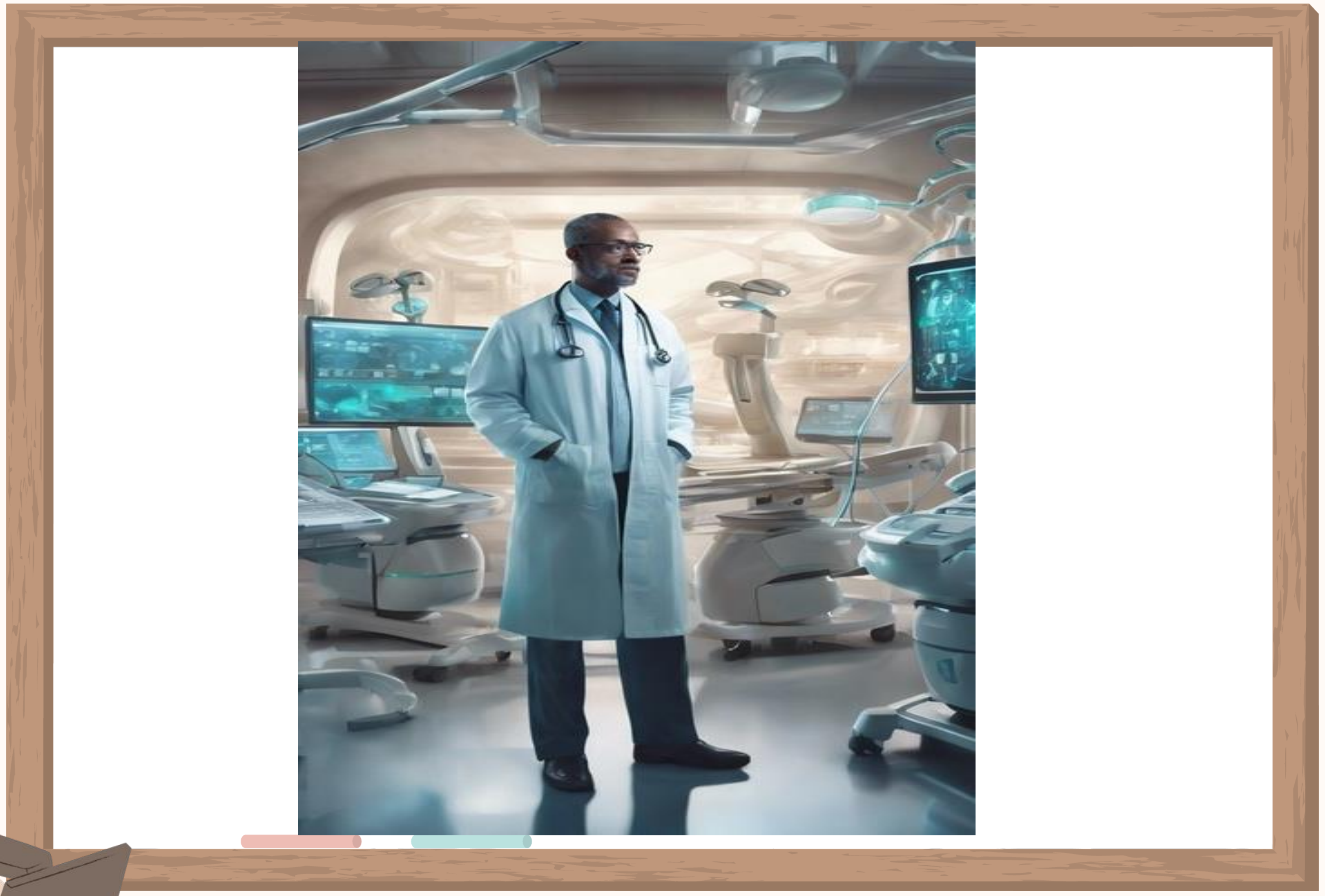


Artificial Intelligence in primary care

Table. Potential Use Cases for AI in Primary Care

Use case	Examples of AI role
Inbox management	• Prioritize patient messages • Generate draft responses • Edit physician messages to optimize communication, including for literacy appropriateness
Clinician documentation	With transcription software: • Draft progress notes in real time during visits • Draft prior authorization, disability, and durable medical equipment requests • Draft a list of billing codes for visits
Between-visit panel management	• Accurately identify patients in need of cancer screening using unstructured and structured EHR data to determine exclusions • Identify patients with incomplete cancer screening (such as missed appointments), automate communication with patients, and provide scheduling and/or staff notification • Generate tailored messages to patients related to needed between-visit care needs
Individualized decision support	• Identify relevant information in structured and unstructured EHR data to prioritize differential diagnoses for new symptoms • Recommend medication options for chronic conditions, considering prior medication prescriptions, allergies, and intolerances noted in structured and unstructured EHR data

Abbreviations: AI, artificial intelligence; EHR, electronic health record.



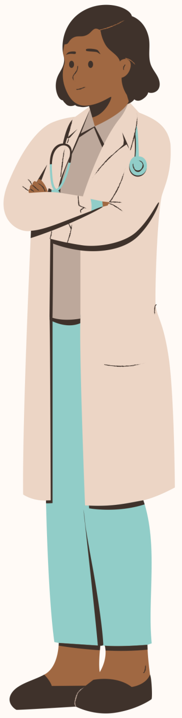
Sarkar U, Bates DW. Using Artificial Intelligence to Improve Primary Care for Patients and Clinicians. JAMA Intern Med. 2024;184(4):343–344. doi:10.1001/jamainternmed.2023.7965

Risks of AI

Table. Analysis of Possible LLM Tasks in Medicine

Task	Potential Pitfalls	Mitigation Strategies
Administrative: Write insurance authorization letters Summarize medical notes Aid medical record documentation Create patient communication (e-mail/letter/text)	Lack of HIPAA adherence: No publicly available model is currently HIPAA-compliant, and thus PHI cannot be shared with the models.	Integrate LLMs within electronic health record systems.
Augmenting knowledge: Answer diagnostic questions Answer questions about medical management Create and translate patient education material	Inherent bias: Pretrained data models used for diagnostic analyses will introduce inherent bias.	Create domain-specific models that are trained on carefully curated data sets. Always include a human in the loop.
Medical education: Write recommendation letters Create new examination questions and case-based scenarios Generate summaries of medical text at a student level	Lack of personalization: LLMs are generated from prior work already published, resulting in repetitive and unoriginal work.	Educate clinicians and users in using LLM tools to augment their work rather than replace them. Encourage understanding how the technology works to mitigate unrealistic expectations of output.
Medical research: Generate research ideas and novel directions Write academic papers Write grants	Ethics: A large amount of discussion has occurred among the scientific community on the ethics of using ChatGPT to generate scientific publications. This also raises the question of accessibility and the potential difficulties of future access to this technology.	Engage in conversation to increase accessibility of this technology to prevent widening gaps in research disparities.

HIPAA = Health Insurance Portability and Accountability Act; LLM = large language model; PHI = protected health information.



Jesutofunmi A. Omiye, Haiwen Gui, Shawheen J. Rezaei, et al. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. Ann Intern Med.2024:177:210-220. [Epub 30 January 2024]. doi:10.7326/M23-2772

JAMA | Original Investigation | AI IN MEDICINE

Measuring the Impact of AI in the Diagnosis of Hospitalized Patients
A Randomized Clinical Vignette Survey Study

Sarah Jabbour, MSE; David Fouhey, PhD; Stephanie Shepard, PhD; Thomas S. Valley, MD; Ella A. Kazerooni, MD, MS; Nikola Banovic, PhD; Jenna Wiens, PhD; Michael W. Sjoding, MD

Editorial page 2255

Supplemental content

IMPORTANCE Artificial intelligence (AI) could support clinicians when diagnosing hospitalized patients; however, systematic bias in AI models could worsen clinician diagnostic accuracy. Recent regulatory guidance has called for AI models to include explanations to mitigate errors made by models, but the effectiveness of this strategy has not been established.

OBJECTIVES To evaluate the impact of systematically biased AI on clinician diagnostic accuracy and to determine if image-based AI model explanations can mitigate model errors.

DESIGN, SETTING, AND PARTICIPANTS Randomized clinical vignette survey study administered between April 2022 and January 2023 across 13 US states involving hospitalist physicians, nurse practitioners, and physician assistants.

INTERVENTIONS Clinicians were shown 9 clinical vignettes of patients hospitalized with acute respiratory failure, including their presenting symptoms, physical examination, laboratory results, and chest radiographs. Clinicians were then asked to determine the likelihood of pneumonia, heart failure, or chronic obstructive pulmonary disease as the underlying cause(s) of each patient's acute respiratory failure. To establish baseline diagnostic accuracy, clinicians were shown 2 vignettes without AI model input. Clinicians were then randomized to see 6 vignettes with AI model input with or without AI model explanations. Among these 6 vignettes, 3 vignettes included standard-model predictions, and 3 vignettes included systematically biased model predictions.

MAIN OUTCOMES AND MEASURES Clinician diagnostic accuracy for pneumonia, heart failure, and chronic obstructive pulmonary disease.

RESULTS Median participant age was 34 years (IQR, 31-39) and 241 (57.7%) were female. Four hundred fifty-seven clinicians were randomized and completed at least 1 vignette, with 231 randomized to AI model predictions without explanations, and 226 randomized to AI model predictions with explanations. Clinicians' baseline diagnostic accuracy was 73.0% (95% CI, 68.3% to 77.8%) for the 3 diagnoses. When shown a standard AI model without explanations, clinician accuracy increased over baseline by 2.9 percentage points (95% CI, 0.5 to 5.2) and by 4.4 percentage points (95% CI, 2.0 to 6.9) when clinicians were also shown AI model explanations. Systematically biased AI model predictions decreased clinician accuracy by 11.3 percentage points (95% CI, 7.2 to 15.5) compared with baseline and providing biased AI model predictions with explanations decreased clinician accuracy by 9.1 percentage points (95% CI, 4.9 to 13.2) compared with baseline, representing a nonsignificant improvement of 2.3 percentage points (95% CI, −2.7 to 7.2) compared with the systematically biased AI model.

CONCLUSIONS AND RELEVANCE Although standard AI models improve diagnostic accuracy, systematically biased AI models reduced diagnostic accuracy, and commonly used image-based AI model explanations did not mitigate this harmful effect.

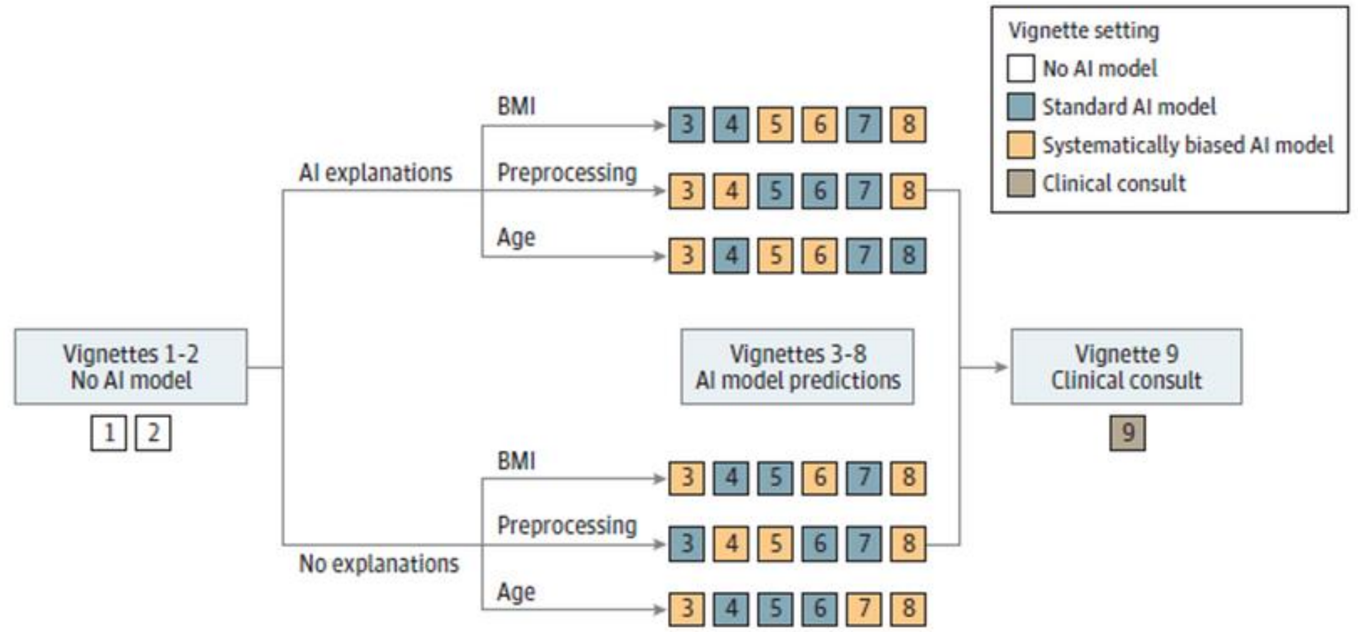
TRIAL REGISTRATION ClinicalTrials.gov Identifier: NCT06098950

JAMA. 2023;330(23):2275-2284. doi:10.1001/jama.2023.22295

Author Affiliations: Computer Science and Engineering, University of Michigan, Ann Arbor (Jabbour, Fouhey, Shepard, Banovic, Wiens); Now with Computer Science Courant Institute, New York University, New York (Fouhey); Now with Electrical and Computer Engineering Tandon School of Engineering, New York University, New York (Fouhey); Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor (Valley, Sjoding); Department of Radiology, University of Michigan Medical School, Ann Arbor (Kazerooni).

Corresponding Author: Michael W. Sjoding, MD, Internal Medicine, G020W Bldg 16 NCRC, 2800 Plymouth Rd, SPC 2800, Ann Arbor, MI 48109 (msjoding@umich.edu) and Jenna Wiens, PhD, Computer Science and Engineering, University of Michigan, 3749 Beyster Bldg, 2260 Haward St, Ann Arbor, MI 48109. Haward St, Ann Arbor, MI 48109 (wiensj@umich.edu).

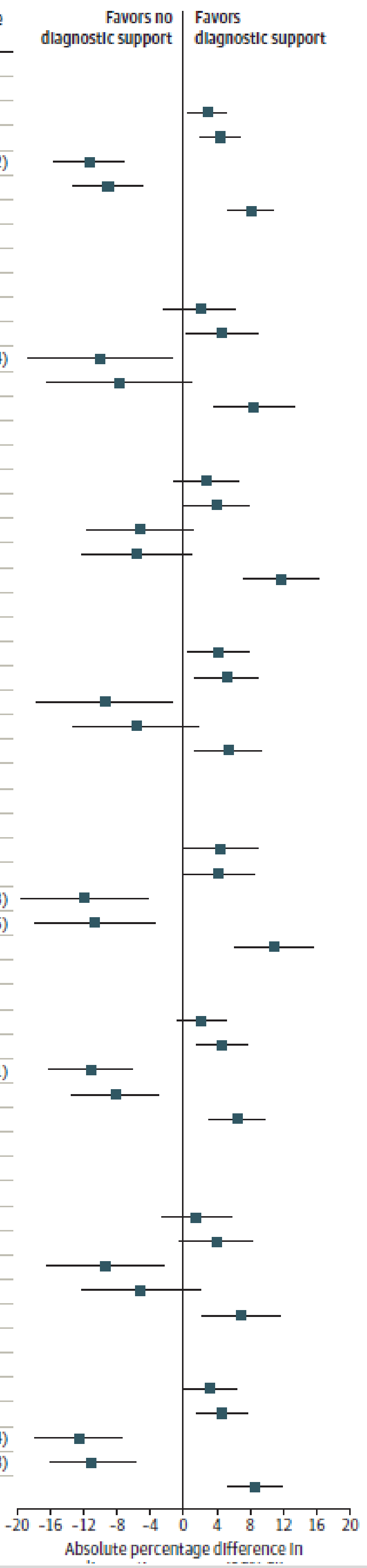
Figure 1. Randomization and Study Flow Diagram for the 9 Clinical Vignettes



After completing informed consent, participants were randomized to artificial intelligence (AI) predictions with or without explanations and all participants were also randomized to 1 of 3 types of systematically biased AI models during a subset of vignettes in the study. The 3 systematically biased AI models included a model predicting pneumonia if aged 80 years or older, a model predicting heart failure if body mass index (BMI, calculated as weight in kilograms divided by height in meters squared) was 30 or higher, and a model predicting chronic obstructive pulmonary disease (COPD) if a blur was applied to the radiograph. Participants were first shown 2 vignettes without AI predictions to measure baseline diagnostic accuracy. The next 6 vignettes included AI predictions. If the participant was randomized to see AI explanations, the participant was also shown an AI model explanation with the AI predictions. Three vignettes had standard AI predictions, and 3 had biased AI predictions shown in random order. The final vignette included a clinical consultation, a short narrative provided by a hypothetical trusted colleague who identified the correct diagnosis and their diagnostic rationale.

Jabbour S, Fouhey D, Shepard S, et al. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. JAMA. 2023:330(23):2275–2284. doi:10.1001/jama.2023.22295

	Absolute percentage difference (95%CI)
Overall	
Baseline 73.0% (95% CI, 68.3% to 77.8%)	
Standard model	2.9 (0.5 to 5.2)
Standard model plus explanations	4.4 (2.0 to 6.9)
Systematically biased model	-11.3 (-15.5 to -7.2)
Systematically biased model plus explanations	-9.1 (-13.2 to -4.9)
Clinical consult	8.1 (5.4 to 10.8)
Diagnosis	
Pneumonia	
Baseline 67.5% (95% CI, 61.0% to 74.0%)	
Standard model	2.0 (-2.3 to 6.2)
Standard model plus explanations	4.6 (0.3 to 8.9)
Systematically biased model	-10.0 (-18.7 to -1.4)
Systematically biased model plus explanations	-7.7 (-16.5 to 1.0)
Clinical consult	8.4 (3.6 to 13.2)
Heart failure	
Baseline 70.7% (95% CI, 63.1% to 78.3%)	
Standard model	2.7 (-1.1 to 6.6)
Standard model plus explanations	3.9 (-0.0 to 7.8)
Systematically biased model	-5.3 (-11.6 to 1.1)
Systematically biased model plus explanations	-5.6 (-12.2 to 0.9)
Clinical consult	11.7 (7.3 to 16.2)
COPD	
Baseline 80.5% (95% CI, 74.8% to 86.1%)	
Standard model	4.2 (0.5 to 7.8)
Standard model plus explanations	5.2 (1.4 to 8.9)
Systematically biased model	-9.5 (-17.6 to -1.3)
Systematically biased model plus explanations	-5.7 (-13.3 to 1.8)
Clinical consult	5.4 (1.3 to 9.4)
Clinician profession	
Nurse practitioner or physician assistant	
Baseline 71.3% (95% CI, 66.0% to 76.7%)	
Standard model	4.5 (0.0 to 8.9)
Standard model plus explanations	4.2 (-0.1 to 8.4)
Systematically biased model	-11.9 (-19.5 to -4.3)
Systematically biased model plus explanations	-10.7 (-17.8 to -3.5)
Clinical consult	10.8 (6.1 to 15.5)
Physician	
Baseline 74.0% (95% CI, 69.0% to 78.9%)	
Standard model	2.1 (-0.8 to 5.1)
Standard model plus explanations	4.6 (1.6 to 7.7)
Systematically biased model	-11.1 (-16.2 to -6.1)
Systematically biased model plus explanations	-8.2 (-13.4 to -3.1)
Clinical consult	6.4 (3.1 to 9.8)
Clinical decision support interaction	
Prior interaction	
Baseline 74.1% (95% CI, 68.6% to 79.7%)	
Standard model	1.5 (-2.7 to 5.7)
Standard model plus explanations	3.9 (-0.5 to 8.3)
Systematically biased model	-9.4 (-16.4 to -2.5)
Systematically biased model plus explanations	-5.2 (-12.2 to 1.9)
Clinical consult	6.9 (2.2 to 11.6)
No interaction	
Baseline 72.7% (95% CI, 67.9% to 77.6%)	
Standard model	3.2 (0.2 to 6.3)
Standard model plus explanations	4.6 (1.6 to 7.7)
Systematically biased model	-12.6 (-17.8 to -7.4)
Systematically biased model plus explanations	-11.0 (-16.1 to -5.8)
Clinical consult	8.6 (5.3 to 11.9)

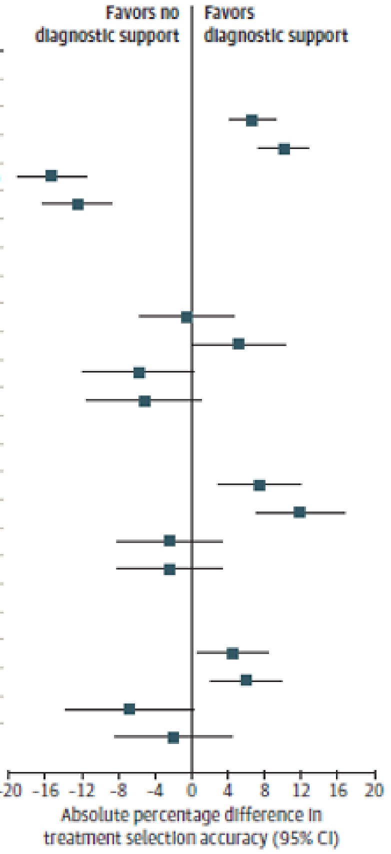


Baseline indicates diagnostic accuracy of heart failure, pneumonia, and chronic obstructive pulmonary disease (COPD) when shown clinical vignettes of patients with acute respiratory failure without AI model input; standard model, diagnostic accuracy when shown clinical vignettes and standard AI model diagnostic predictions about whether the patient has heart failure, pneumonia, and/or COPD; standard model plus explanations, diagnostic accuracy when shown standard AI predictions and an image-based AI explanation of the model's reasoning for making a prediction within vignettes; systematically biased model, diagnostic accuracy when shown systematically biased AI predictions of low accuracy within vignettes; systematically biased model plus explanations, diagnostic accuracy when shown biased model predictions and explanations within vignettes; and clinical consultation, diagnostic accuracy when provided a short narrative describing the rational for the correct diagnosis within the vignette.

Subgroup analysis included diagnostic accuracy specific to heart failure, pneumonia, and COPD; clinician profession, including 142 nurse practitioners or physician assistants, and 274 physicians; prior clinical decision-support interaction, including 132 participants who had prior experience interacting with clinical decision support systems and 286 who did not. Diagnostic accuracy and percentage point differences in accuracy were determined by calculating predictive margins and contrasts across vignette settings after fitting a cross-classified

Figure 4. Baseline Treatment Selection Accuracy Without AI Models and Percentage Point Differences in Accuracy Across Clinical Vignette Settings

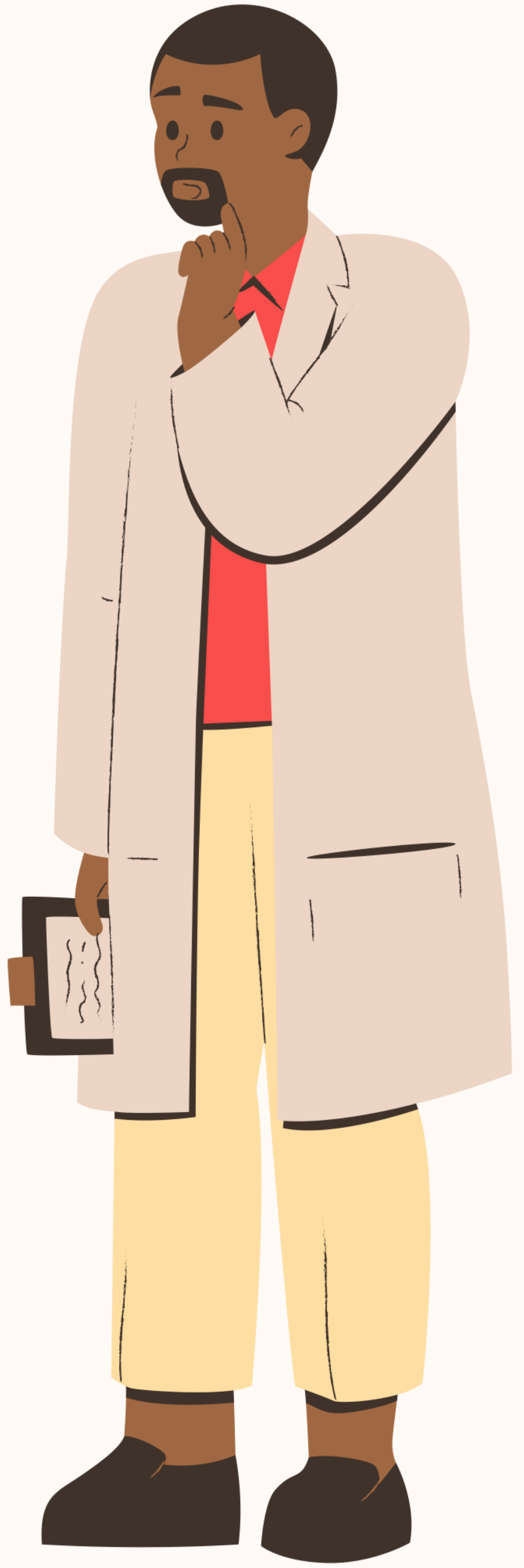
	Absolute percentage difference (95%CI)
Overall treatment selection accuracy	
Baseline 70.3% (95% CI, 65.5% to 75.2%)	
Correct model	6.7 (4.1 to 9.3)
Correct model plus explanations	10.1 (7.4 to 12.8)
Incorrect model	-15.3 (-19.1 to -11.4)
Incorrect model plus explanations	-12.5 (-16.3 to -8.8)
Subgroup analysis by treament	
Antibiotics	
Baseline 61.1% (95% CI, 53.8% to 68.5%)	
Correct model	-0.5 (-5.6 to 4.6)
Correct model plus explanations	5.2 (0.1 to 10.3)
Incorrect model	-5.8 (-11.9 to 0.3)
Incorrect model plus explanations	-5.2 (-11.4 to 1.0)
Diuretics	
Baseline 69.4% (95% CI, 61.6% to 77.1%)	
Correct model	7.4 (2.9 to 11.9)
Correct model plus explanations	11.9 (7.2 to 16.7)
Incorrect model	-2.4 (-8.2 to 3.4)
Incorrect model plus explanations	-2.4 (-8.1 to 3.4)
Steroids	
Baseline 78.9% (95% CI, 72.7% to 85.0%)	
Correct model	4.6 (0.7 to 8.4)
Correct model plus explanations	5.9 (2.0 to 9.8)
Incorrect model	-6.7 (-13.7 to 0.3)
Incorrect model plus explanations	-2.0 (-8.4 to 4.4)



Baseline treatment selection accuracy indicates accurate administration of antibiotics, diuretics, and/or steroids after reviewing vignettes of patients with acute respiratory failure without AI model input; correct model, treatment accuracy when shown vignette with correct AI model diagnostic predictions of heart failure, pneumonia, and/or COPD; correct model plus explanations, treatment accuracy when shown a vignette with correct AI model diagnostic predictions and an image-based AI explanation of the model's reason for making a prediction; incorrect model, treatment accuracy when shown a vignette with incorrect AI model diagnostic predictions; and incorrect model plus explanation, treatment accuracy when shown incorrect AI model diagnostic predictions and explanations.

Subgroup analysis included treatment selection accuracy specific to antibiotics, intravenous diuretics, and steroids. Treatment selection accuracy and percentage point differences in accuracy were determined by calculating predictive margins and contrasts across vignette settings after fitting a cross-classified generalized random-effects model of treatment selection accuracy across settings.

Jabbour S, Fouhey D, Shepard S, et al. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. JAMA. 2023;330(23):2275–2284. doi:10.1001/jama.2023.22295



Human Vigilance

VIEWPOINT

The Limits of Clinician Vigilance as an AI Safety Bulwark

Julia Adler-Milstein, PhD
Department of Medicine, University of California, San Francisco.

Donald A. Redelmeier, MD
Department of Medicine, University of Toronto, Toronto, Ontario, Canada; and Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada.

Robert M. Wachter, MD
Department of Medicine, University of California, San Francisco.

The integration of artificial intelligence (AI) into routine clinical care is accelerating. We are beginning to see scaled AI deployment focused on clinical tasks such as reviewing patient histories, drafting physician notes, offering patient instructions, and reading x-rays. AI will soon contribute to high-stakes clinical decisions such as suggesting diagnoses and recommending therapies to frontline clinicians.

As AI supports a broadening set of clinical tasks, it will evolve from a distinct, novel input into something more pervasive, customary, and subtle. This evolution is appealing because we want reliable technology to quietly work in the background to empower human endeavors. In practice, however, this evolution should give clinicians, patients, and health care leaders pause because of 2 pitfalls: (1) AI is far from perfect in its outputs and (2) humans are far from perfect when tasked with double-checking the outputs of generally trusted technologies. In this Viewpoint, we explore what is known about these problems and suggest potential solutions.

The problem of how to marry human and artificial intelligence can be framed by considering polar assumptions regarding AI accuracy. If AI were 100% accurate and fully reliable, the clinician would quickly learn to defer to the technology. (Of course, the clinician eventually becomes obsolete in this scenario.) If instead, AI performs poorly with frequent inaccuracies, clinicians will stop using the output. For the foreseeable future, however, AI outputs will likely fall between these extremes: accurate enough to be useful but imperfect enough that clinicians will be asked to serve as double-checkers who sign off on the final note, order, or diagnosis and who will be liable for consequential mistakes.

> The path forward rests on designing and deploying [artificial intelligence] in ways that enhance human vigilance.

This strategy presumes that human vigilance is a robust safety check. However, humans are terrible at vigilance.[1] The fallibility of vigilance is likely to be amplified when AI errors are surrounded by correct information, and presented in a conversational and authoritative tone, as is likely in most clinical interfaces. Moreover, it will be natural for health care organizations to repurpose any AI-derived efficiencies into demands for higher throughput, such as by expecting clinicians to see more patients or read more radiographs in a session. This production pressure will create another impediment to human vigilance. In sum, it is perilous to assume that clinician vigilance is an acceptable safeguard against AI faults. In the metaphor of patient safety science, a human double-check at the end of an AI-heavy process is likely to be a lightweight slice of Swiss cheese.[2]

The path forward rests on designing and deploying AI in ways that enhance human vigilance. Vigilance is the task of supervision where people become "monitors of what the system is doing rather than active participants in the workflow."[3] People struggle with vigilance because it requires maintaining attention without active engagement, an inherently hard task for the human brain. Perhaps clinicians can be vigilant in the near-term when AI is novel and deployed selectively, such as editing a generative AI-drafted note in a recently deployed system.[4] However, clinicians will rapidly become less skilled, less attentive, and less discerning as AI becomes a more ubiquitous component of their clinical work. Those implementing AI systems, therefore, may have a relatively short window in which to find viable approaches for enhancing clinician vigilance.

Based in part on insights from other industries—including aviation and automobile manufacturing—that have been grappling with the challenge of human oversight of robust digital systems for decades, we offer 5 options for how AI could be designed to promote clinician vigilance. While any one of these options may pay dividends, we suspect that effective systems will include more than one.

First, visual cues could highlight AI output that is more uncertain and potentially faulty. This assumes the AI system "knows" its level of certainty for a specific output. If so, the AI might signal degrees of uncertainty by using color-coded fonts (eg, green-yellow-red) or other intuitive visual cues when the output exceeds a preset uncertainty threshold. Similarly, another type of uncertainty could be signaled when an individual patient is not representative of the population on which the model was trained. Of course, color-coded signaling of uncertainty needs to be used sparingly to avoid alert fatigue.

Second, clinician-level measures of active vigilance could be the basis for a system to assess whether a clinician is exhibiting automation bias. For example, is the clinician accepting AI-recommended medications 100% of the time or never editing AI-generated text? Such real-time tracking of vigilance could prompt education, feedback, coaching, and even turning off the AI for a period in serious cases.

Third, all AI-generated practice efficiencies should not be converted into expectations of higher throughput. While it is reasonable to expect some increase in throughput after AI implementation (in part to pay for the cost of the AI), some reserve capacity needs to be retained to ensure that clinicians have time and cognitive bandwidth to exercise vigilance. Ideally, some

Corresponding Author: Julia Adler-Milstein, PhD, University of California, San Francisco, 10 Koret Way, 3018, San Francisco, CA 94131 (julia.adler-milstein@ucsf.edu).

Adler-Milstein J, Redelmeier DA, Wachter RM. The Limits of Clinician Vigilance as an AI Safety Bulwark. JAMA. 2024;331(14):1173–1174. doi:10.1001/jama.2024.3620

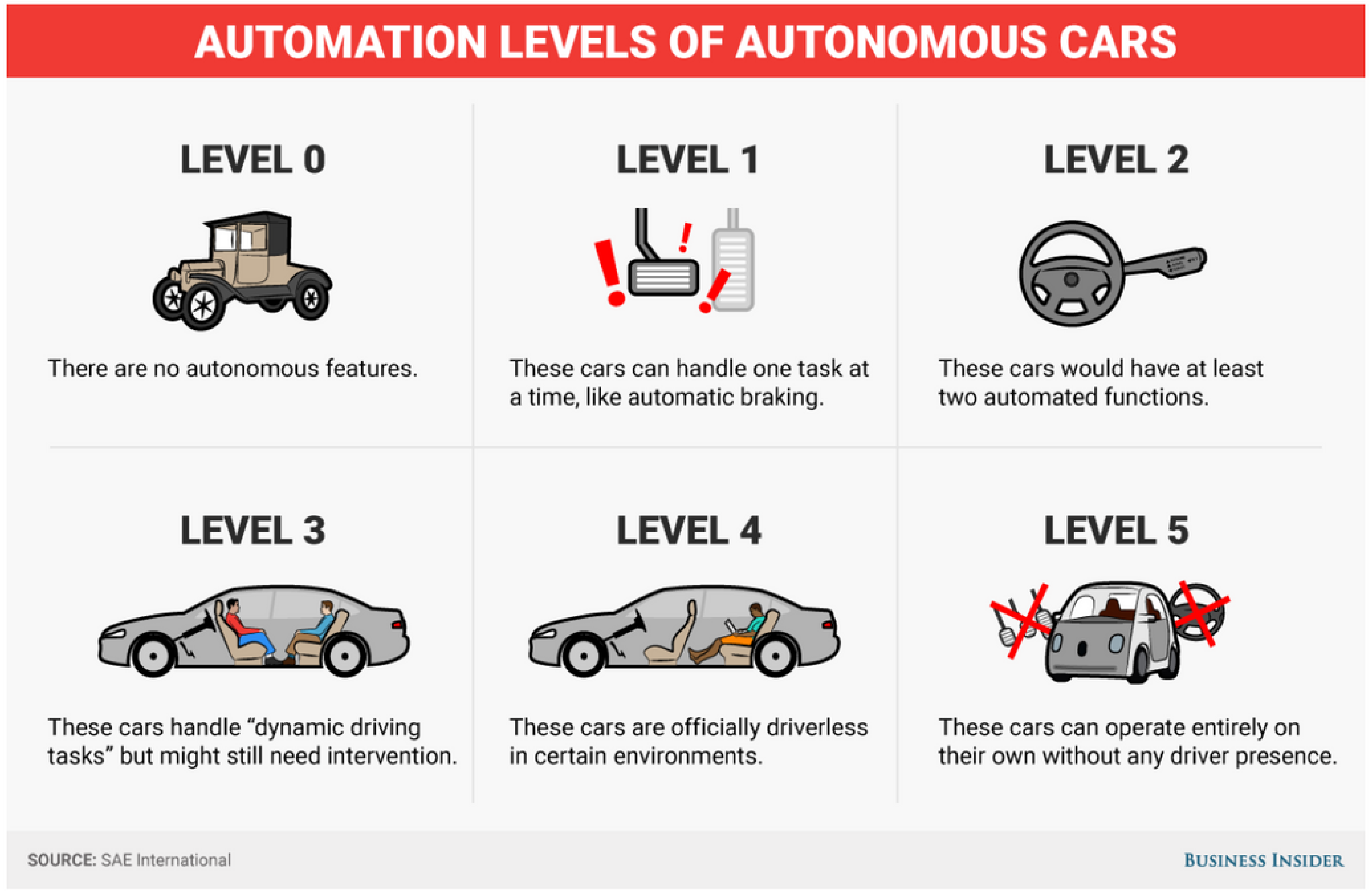
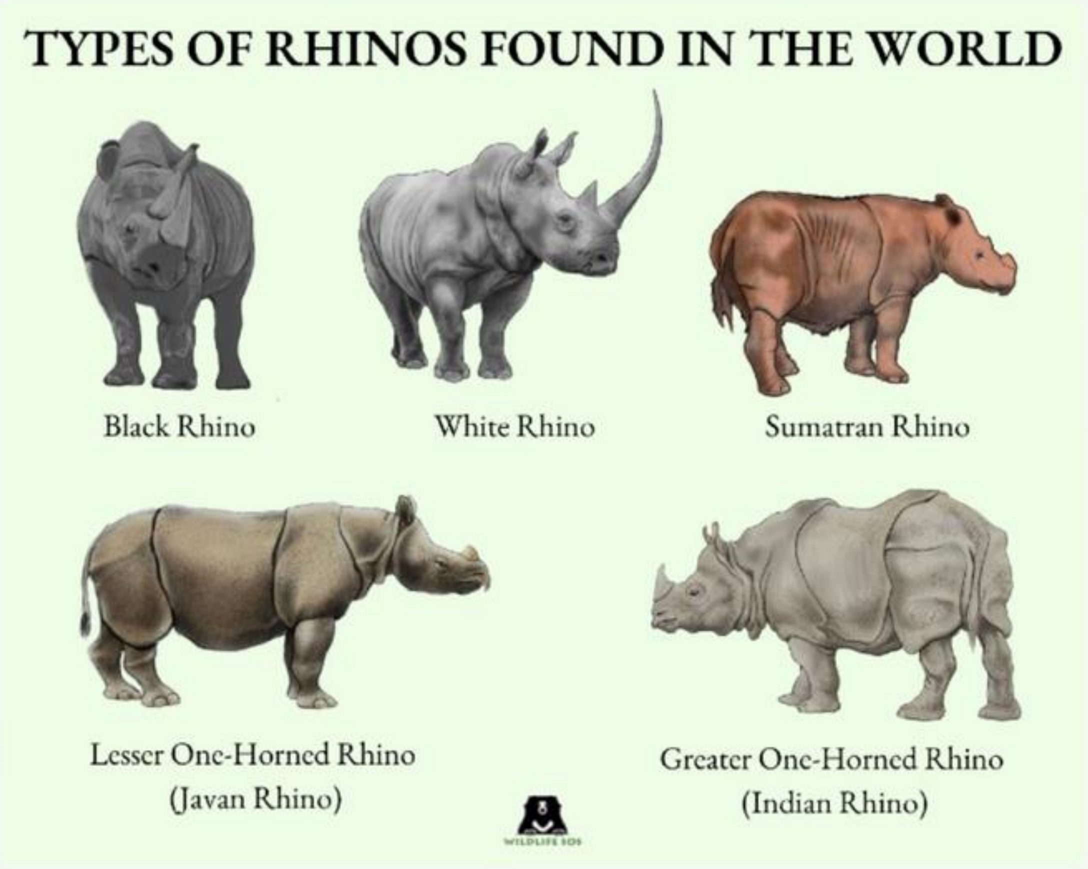
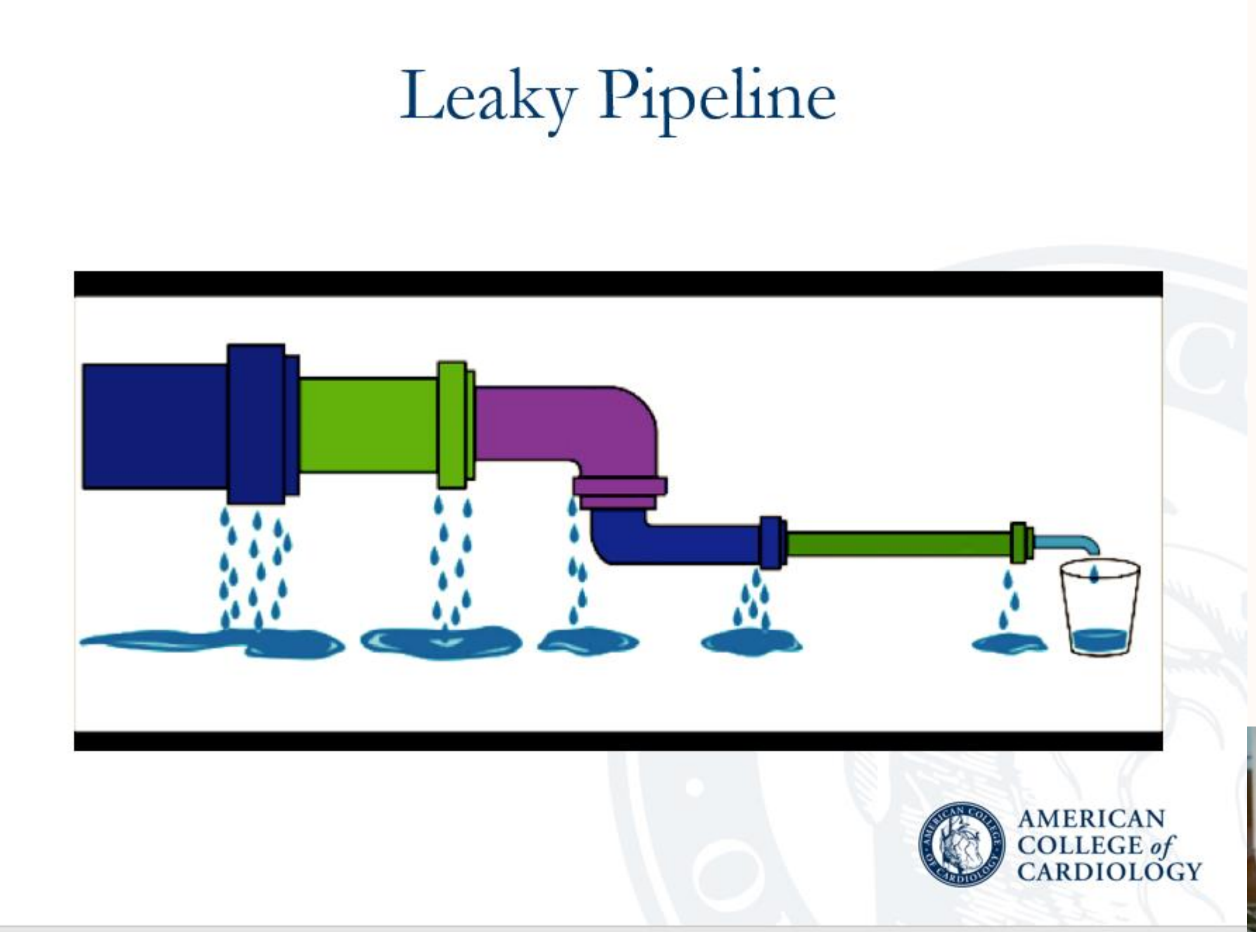


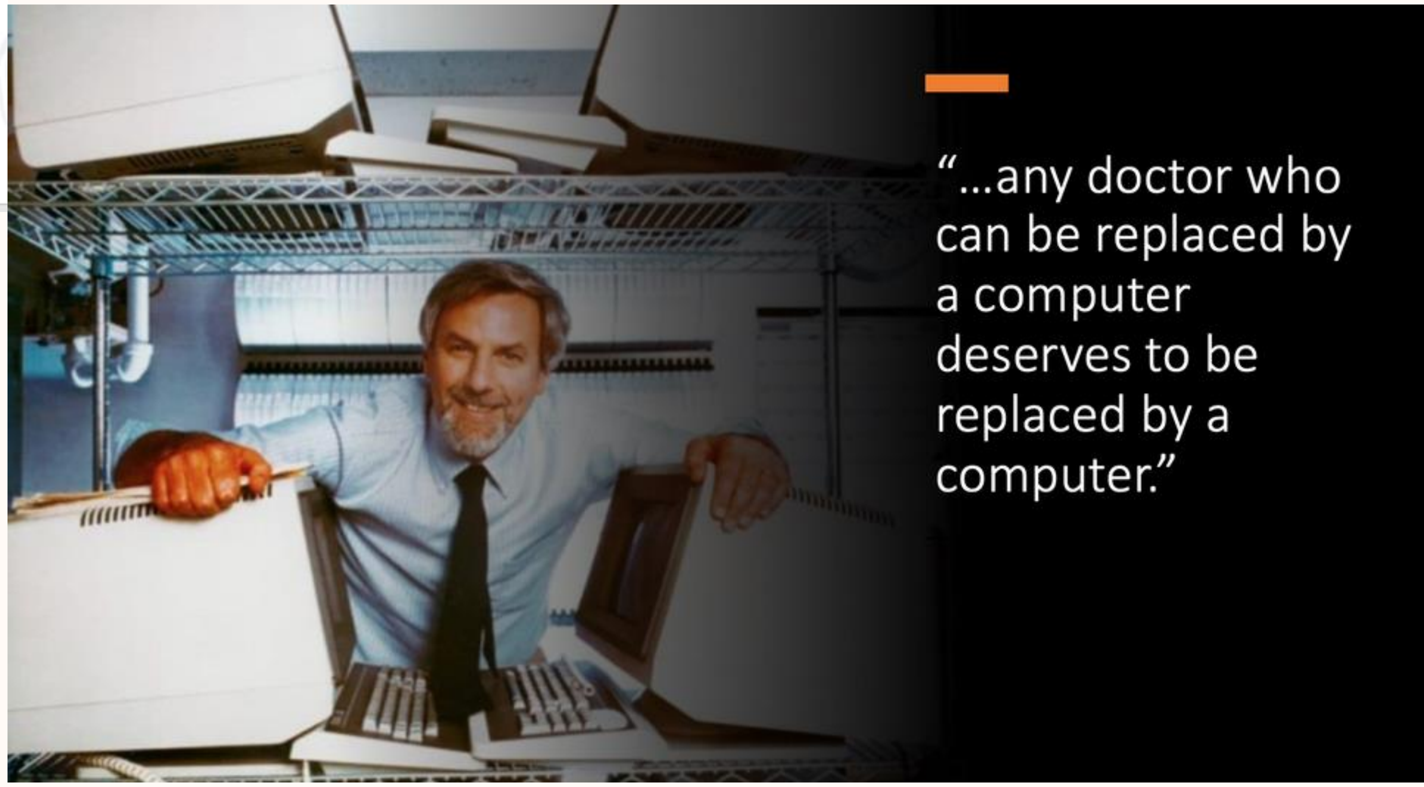
Image Courtesy: Business Insider

Conclusions: It's Still a Good Time to be A Rhino



Meet the rhinos of the world. [Graphic (c) Wildlife SOS/Teesta Mukherjee]





And whatever your labors and aspirations, in the noisy confusion of life, keep peace in your soul. With all its sham, drudgery and broken dreams, it is still a beautiful world. Be cheerful. Strive to be happy.

Max Ehrmann

Thank you