admission. Secondary outcomes were the receipt of IV iron and blood transfusion at delivery.

RESULTS: There were 2,704 and 2,624 patients in the pre- and post-intervention groups of which 765 (28.3%) and 998 (38.0%) had antepartum anemia (P< 0.001). Of patients with anemia, 47/765 (6.1%) and 162/998 (16.2%) received IV Fe pre- and post-intervention (P=0.001) with no significant change in the rate of PO Fe (24.5% vs. 27.5%, P=0.17). Anemia on admission was lower in the post-intervention group (34.7% vs. 39.5%, P=0.04) and persisted at a threshold of Hct < 30% (4.5% vs. 6.7%, P=0.05). The rate of transfusion also decreased during the study period (2.5% vs. 4.6%, P=0.02).

CONCLUSION: The creation of an anemia protocol and obstetrics based IV iron center drastically increased utilization of IV Fe and decreased rates of anemia and blood transfusions at delivery at our institution. Given these findings, OB practices should strongly consider integrating similar interventions into their practices.

Table 1 - Rates of anemia and treatment before and after intervention

	Pre-intervention	Post-intervention	
	All patients		
	N=2,704	N=2,624	P-value
Hct <33%	765 (28.3)	998 (38.0)	<0.001
Hct <30%	163 (6.0)	205 (7.8)	0.01
Hct <27%	29 (1.1)	28 (1.1)	0.98
GA at anemia diagnosis (weeks)	27.8 (7.7)	27.0 (8.0)	0.041
	Patients with anemia		
	N=765	N=998	
IV iron (patients)	47 (6.1)	162 (16.2)	0.001
PO iron (patients)	210 (27.5)	245 (24.5)	0.17
Anemia on delivery admission			
Hct <33%	296 (39.5)	338 (34.7)	0.04
Hct <30%	50 (6.7)	44 (4.5)	0.05
Blood transfusion	35 (4.6)	25 (2.5)	0.02

Presented as mean (SD) or N (%) as appropriate

Abbreviations: Hct, hematocrit; GA, gestational age; IV, intravenous; PO, *per os* (oral)

244 An anemia protocol and obstetric clinic-based iron infusion center decreases disparities in IV iron access

Logan Mauney[1], Jonathan Y. siden[2], Kaitlyn E. James[1], Mark A. Clapp[1], Sarah N. Bernstein[1]

[1]Massachusetts General Hospital, Boston, MA, [2]Mass General Brigham, Boston, MA

OBJECTIVE: Racial disparities in antepartum anemia exist with Black pregnant people experiencing 2x greater prevalence of anemia than non-Hispanic Whites. The use of protocols for the diagnosis and management of anemia increases the use of intravenous (IV) iron and delivery hematocrit, but it is unknown if all racial and ethnic groups benefit. IV iron can be difficult to access because it is often administered in IV infusion centers or inpatient. To improve access, our institution established a standardized anemia protocol and infusion center housed in the obstetrics clinic.

STUDY DESIGN: This retrospective cohort study examined patients admitted for delivery at a large academic hospital in the 10 months pre- and post-intervention (3/2021-12/2021 and 8/2022-6/2023). Patients with Hct < 33% at any point in pregnancy were included. Those with hemoglobinopathy, renal disease, and bleeding in pregnancy were excluded. Self-reported race/ethnicity data was abstracted from the EMR. The primary outcome was anemia on admission and secondary outcomes were rates of IV iron and blood transfusion during the delivery encounter.

RESULTS: There were 2,704 and 2,624 patients in the pre- and post-intervention groups respectively, including 1,772 White (78%) and 206 Black (8%). Black patients had higher baseline rates of anemia than White patients (39% vs 27%). All racial and ethnic groups saw increases in IV iron administration after the intervention (p< 0.05), with Black patients experiencing an 11-point increase compared to a 10-point increase among White patients. The rates of anemia on admission for delivery and transfusion were not statistically different for any racial or ethnic groups pre- or post-intervention but were significant for the cohort as a whole.

CONCLUSION: Implementation of a standardized anemia protocol and an obstetrics clinic-based infusion center improved access to IV iron administration for all racial and ethnic groups, including for black patients. Given the urgent need to address disparities in maternity care, practices should consider implementation of similar interventions.

Table 1 - Rates of anemia and treatment by race and ethnicity before and after intervention

	Pre-intervention (%)		Post-intervention (%)	
	All patients N=2704	With anemia N=765	All patients N= 2624	With anemia N=998
	Antepartum anemia	IV iron	Antepartum anemia	IV iron
Race				
All races	765 (28)	47 (6.1)	998 (38.0)	162 (16.2)*
White	478 (27)	26 (5.4)	592 (26)	86 (14.5)*
Black	80 (39)	10 (12.5)	113 (58)	27 (23.9)*
Asian	65 (22)	1 (1.5)	116 (35)	11 (9.5)*
Other	142 (34)	10 (7)	177 (26)	38 (21.5)*
Ethnicity				
All	739 (28)			
Non-Hispanic	576 (27)	34 (5.9)	147 (7)	113 (15.2)*
Hispanic	163 (32)	13 (8)	49 (9)	46 (21.7)*

Presented as mean (SD) or N (%) as appropriate

*=P<0.05 for pre- vs. post-intervention, Chi-square performed

Abbreviations: IV, intravenous

245 Accuracy of deep learning models in interpreting intrapartum fetal monitoring to predict fetal acidemia

Jennifer A. McCoy[1], Guangya Wan[2], Lisa D. Levine[3], Joseph Teel[3], John Holmes[3], William LaCava[2]

[1]University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, [2]Harvard University, Boston, MA, [3]University of Pennsylvania, Philadelphia, PA

OBJECTIVE: Up to 30% of cesareans in the US are performed due to false-positive interpretations of intrapartum electronic fetal monitoring (EFM). EFM interpretation is subjective and vulnerable to bias. Novel deep learning techniques can improve complex data processing and pattern recognition in medicine. We sought to apply deep learning approaches that could interpret EFM data to predict fetal acidemia.

STUDY DESIGN: The database was created using intrapartum EFM data from 2006-2020 at a large, multi-site academic health system. We included patients ≥34 weeks with a singleton, vertex fetus with EFM data available for ≥1 hour prior to delivery and an umbilical cord blood pH result available. We excluded those with >30% missingness in EFM data. Data pre-processing removed noise and artifact. Data was divided into training and testing sets with equal distribution of acidemic cases. Several different deep learning architectures were explored, including transformers, convolutional neural networks (CNNs), and long short-term memory (LSTM) networks. The primary outcome was low cord blood pH, investigated at four clinically meaningful thresholds: 7.2, 7.15, 7.1, and

7.05. Receiver operating characteristic (ROC) curves were generated with area under the curve (AUC) assessed to determine the performance of the models.

RESULTS: A total of 124,776 fetal monitoring files were available, 35,604 had a corresponding umbilical cord gas pH result, and the final sample size was 10,176. The prevalence of the outcome in the data was 20.9% with pH < 7.2, 9.1% < 7.15, 3.3% < 7.10, and 1.3% < 7.05. The median AUC values for each deep learning model at each different pH threshold are shown in Figure 1. The best performance was achieved with the CNN multiscale model and a pH threshold of 7.10, with an AUC of 0.82 95% CI [0.82-0.83].

CONCLUSION: A novel application of deep learning methods achieves excellent performance in predicting fetal acidemia on umbilical cord blood pH. This technology could improve the accuracy and consistency of EFM interpretation to prevent unnecessary cesarean deliveries and avoidable intrapartum fetal injury.

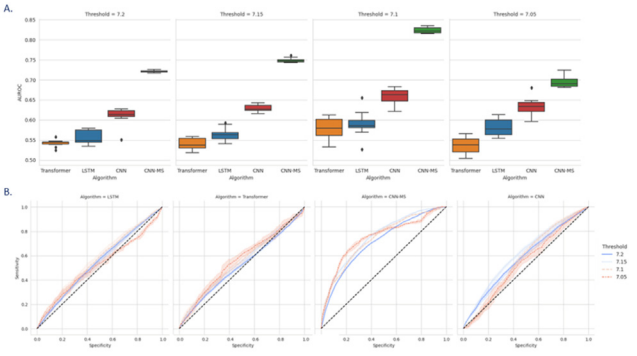

Figure 1. Panel A: box plots depicting area under the curve and 95% confidence intervals of each model type at each umbilical cord blood pH threshold. Panel B: ROC curves for each model type at each umbilical cord blood pH threshold. *LSTM: long short term memory; CNN: convolutional neural network; CNN-MS: convolutional neural network multiscale.*

246 The fragility index of randomized controlled trials in obstetrics

Jordan A. McKinney[1], Kelcey Day[1], Lifeng Lin[2], Luis Sanchez Ramos[1]

[1]University of Florida College of Medicine, Jacksonville, FL, [2]University of Arizona, Tucson, AZ

OBJECTIVE: To assess the fragility index (FI) and fragility quotient (FQ) of recently published obstetric randomized controlled trials (RCTs) in order to evaluate the robustness of study results beyond conventional statistical significance testing.

STUDY DESIGN: A retrospective analysis was conducted to identify obstetric RCTs published between 2018 and 2022 in the top medical and obstetric journals, based on impact factor. Eligible studies included those with at least one statistically significant binary primary outcome. Two independent reviewers extracted data on sample size, event and control group sizes, loss to follow-up, blinding status, type of intervention, and other relevant factors. The FIs and FQs were calculated for each study.

RESULTS: A total of 245 RCTs were identified, and 31 met the eligibility criteria. The median and interquartile range FI for all studies was 6 (2-20), while the median FQ was 0.011 (0.004-0.034). Of the studies, 39% had an FI < 5. RCTs with placebo control groups demonstrated greater robustness compared to those with active control groups (5 [2-12]; P=0.028). Otherwise, no statistically significant relationships were observed between FI or FQ and other variables. Loss to follow-up exceeded the FI in 39% of the studies.

CONCLUSION: Incorporating the FI and FQ as complementary measures to traditional statistical significance testing is crucial for assessing the robustness of trial results in obstetrics. These metrics offer valuable insights into study reliability and can guide informed decisions regarding the safety and efficacy of interventions. By considering the FI and FQ, clinicians and researchers can navigate uncertainties associated with the evidence more effectively.

247 Exploring the limits of artificial intelligence for referencing scientific articles

Emily Graf[1], Jordan A. McKinney[1], Alexander Dye[1], Lifeng Lin[2], Luis Sanchez Ramos[1]

[1]University of Florida College of Medicine, Jacksonville, FL, [2]University of Arizona, Tucson, AZ

OBJECTIVE: To evaluate the reliability of three artificial intelligence (AI) chatbots (ChatGPT, Google Bard, and Chatsonic) in generating accurate references from existing obstetric literature.

STUDY DESIGN: Between mid-March and late April 2023, ChatGPT, Google Bard, and Chatsonic were prompted to provide references for specific obstetrical randomized controlled trials (RCTs) published in 2020, adhering to the AMA Manual of Style guidelines. RCTs were considered for inclusion if they were mentioned in a previous article that primarily evaluated RCTs published by the top medical and obstetrics and gynecology journals with the highest impact factors in 2020, as well as RCTs published in a new journal focused on publishing obstetric RCTs. The selection of the three AI models was based on their popularity, performance in natural language processing, and public availability. Data collection involved prompting the AI chatbots to provide references according to a standardized protocol. The primary evaluation metric was the accuracy of each AI model in correctly citing references, including authors, publication title, journal name, and DOI. Statistical analysis was performed using a permutation test to compare the performance of the AI models.

RESULTS: Among the 44 RCTs analyzed, Google Bard demonstrated the highest accuracy, correctly citing 13.6% of the requested RCTs, while ChatGPT and Chatsonic exhibited lower accuracy rates of 2.4% and 0%, respectively. Google Bard often substantially outperformed Chatsonic and ChatGPT in correctly citing the studied reference components. The majority of references from all AI models studied were noted to provide DOIs for unrelated studies or DOIs that do not exist.

CONCLUSION: To ensure the reliability of scientific information being disseminated, authors must exercise caution when utilizing AI for scientific writing and literature search. However, despite their limitations, collaborative partnerships between AI systems and researchers have the potential to drive synergistic advancements, leading to improved patient care and outcomes.

Summary of permutation analysis

	ChatGPT vs Google Bard			Chatsonic vs Google Bard			ChatGPT vs Chatsonic		
	ChatGPT	Google Bard	p-value	Chatsonic	Google Bard	p-value	ChatGPT	Chatsonic	p-value
Correct lead author (%)	6.8	31.8	.007	3.5	31.8	<.001	6.8	3.5	.626
Correct secondary authors (%)	4.5	22.7	.038	4.5	22.7	.022	4.5	4.5	1.00
Correct publication title (%)	18.2	68.2	<.001	15.9	68.2	<0.001	18.2	15.9	1.00
Correct journal name (%)	40.9	63.6	.053	29.5	63.6	.006	40.9	29.5	.358
Correct DOI (%)	2.3	38.6	<.001	0	38.6	<.001	2.3	0	1.00
Partially correct reference (%)*	52.3	84.1	.002	31.8	84.1	<.001	52.3	31.8	.063
Completely correct reference (%)	2.3	13.6	.122	0	13.6	.030	2.3	0	1.00

DOI: digital object identifier.
*If any component of the reference was correct